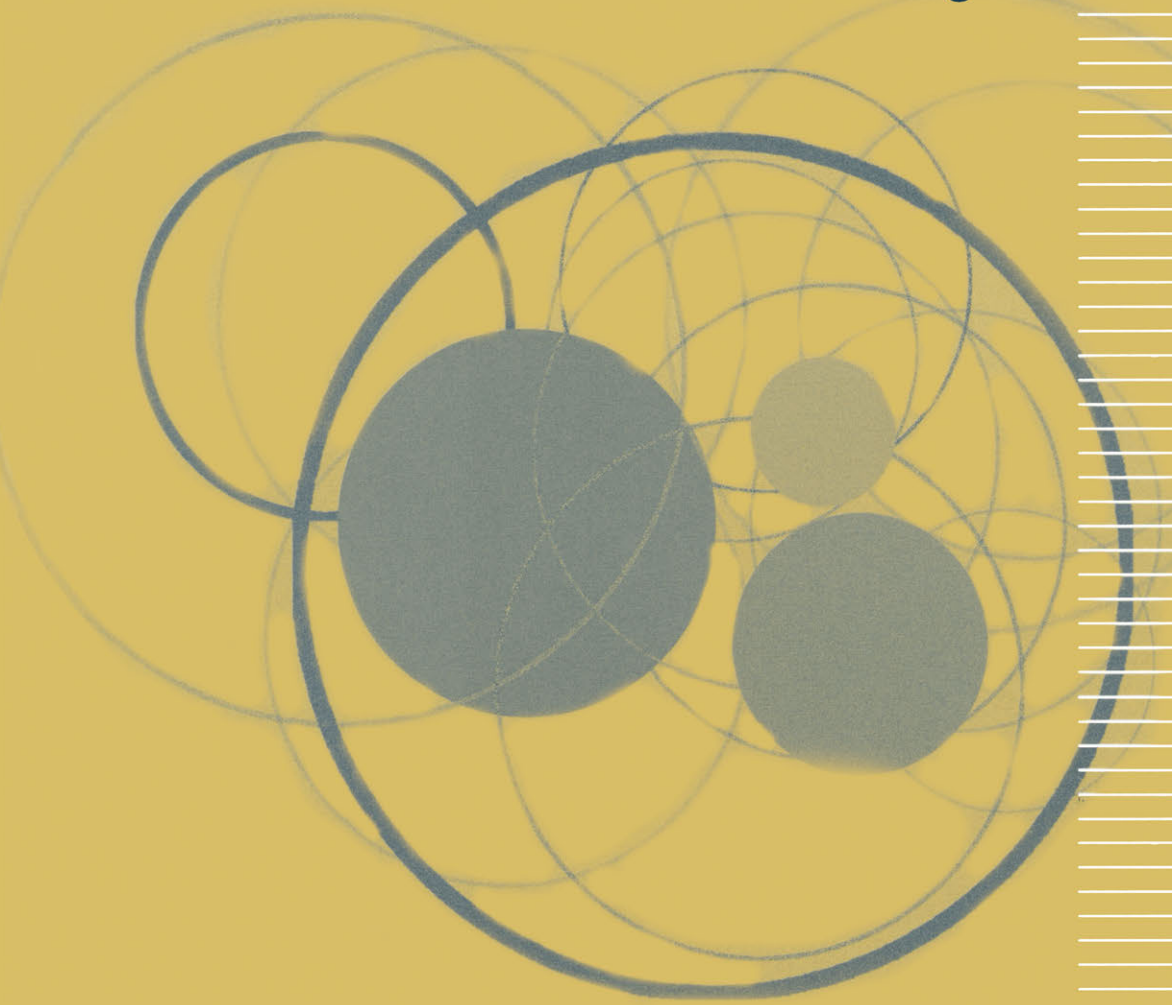
Statistical Consulting

Javier Cabrera
Andrew McDougall



Statistical Consulting

Springer Science+Business Media, LLC

Javier Cabrera
Andrew McDougall

Statistical Consulting

Springer

Javier Cabrera
Department of Statistics
Rutgers University, Busch Campus
Piscataway, NJ 08855-1179
USA
cabrera@rci.rutgers.edu

Andrew McDougall
CSAM-Statistics
Montclair State University
Upper Montclair, NJ 07043
USA
mcdougal@pegasus.montclair.edu

Cover design by James Ross.

Library of Congress Cataloging-in-Publication Data
Cabrera, Javier.
Statistical consulting / Javier Cabrera, Andrew McDougall.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4419-3177-1 ISBN 978-1-4757-3663-2 (eBook)
DOI 10.1007/978-1-4757-3663-2
1. Social sciences—Statistics. 2. Statistical consultants. I. McDougall, Andrew, 1961–
II. Title.
HA29.C2 2001
001.4'22—dc21 2001053286

Printed on acid-free paper.

Originally published by Springer-Verlag New York, Inc. in 2002
Softcover reprint of the hardcover 1st edition 2002

Production managed by Allan Abrams; manufacturing supervised by Joe Quatela.
Photocomposed copy prepared from the authors' LaTeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4419-3177-1 SPIN 10728650

Preface

The motivation for this book arose from the statistical consulting course that we have taught regularly for several years. In that course, we focus on the themes:

- Understanding the statistical consulting "process"
- Developing effective communication skills
- Obtaining experience through case studies.

In reality, there is no substitute for consulting directly with a client and for this interaction to be effective, good communication skills are essential. Unfortunately, this aspect of a statistician's training is often neglected and statistics graduates have little choice but to learn these skills on the job. One of the purposes of this book is to address this need.

Statistical consulting occurs in a diverse range of environments and for tackling real-life statistical problems, the statistician needs to have a strong interest in the scientific method. History itself provides the best examples for developing this interest and so we begin with a brief historical voyage in Chapter 1. There's no time like the present, of course, and in the remainder of this chapter we describe some of the environments in which statistical consulting plays a major role.

A detailed discussion on verbal and written communication skills that will be required in a consulting environment is presented in Chapter 2. Visualization is one of the most powerful tools available to the statistical consultant and the importance of quality graphics and effective presentations is also addressed in Chapter 2.

Statistical methodology is presented in Chapter 3. In describing the statistical methodology that a consultant can employ, the importance of engaging the client's understanding of the purpose and interpretation of a statistical procedure is emphasized. While it is assumed that the reader has the necessary technical skills to apply the statistical methods used in this book, our aim is to present the statistical methodology from the perspective of the client, rather than the statistician.

Putting all the components of the consulting process together is the next step, and in Chapter 4 an example of a statistical consulting project is presented in its entirety. This chapter concludes Part I of the book.

A wide range of case studies of varying complexity are presented in Part II of the book that will help the reader understand and appreciate the diversity of projects that can arise in statistical consulting. The case studies have been sorted into groups according to the level of technical difficulty associated with the statistical methods required for analysis. This provides the student and reader with some expectation of the level of statistical methodology and analytic complexity involved in a particular case study. Chapter 9 presents several case study exercises where the project and problem of interest are identified, but the analysis is left to the reader.

The Appendices provide information on resources (Appendix A), details on the SAS and S–PLUS software packages (Appendix B), and a collection of useful reference tables (Appendix C). This book would be a suitable text for a graduate course on statistical consulting and a course outline is provided for the instructor in Appendix A. All the datasets are available from the website address given in Appendix A.

Acknowledgments

We are indebted to many people who have contributed to this book. Special thanks go to our editor John Kimmel for all his efforts in making this book a reality and to the Springer-Verlag production team. We are especially grateful to the efforts of an anonymous reviewer whose *down* to earth *under*standing and insight led to considerable improvement of the initial manuscript. Personal thanks go to Dhammika Amaratunga, Zahur Islam, Makund Karwe, Dale Kopas, Brian Ludwig, Cristina Rameriz, Bonnie Ray, Patrice Roper, Robert Searson, Joan Skurnick, and Maria Drelich for their contributions. Finally, we would like to thank our respective families and friends for their enduring patience, support, and timely reminders to "Finish it!" Needless to say, we have.

Piscataway, New Jersey
Montclair, New Jersey

Javier Cabrera
Andrew McDougall
October 2001

Contents

Part I

The Methodology of Statistical Consulting

1
Introduction to Statistical Consulting

"What is statistics?" is a question often addressed at the beginning of a statistical text, which may partly explain why there is a need for statistical consultants: everybody skips to the next chapter only to be greeted by probability theory. So, what is statistical consulting then? Ultimately, it is about communication since the answer to the first question is that statistics is about any problem, in any field of study, that involves random variation. Of course, this does mean every field of study and problem necessarily falls into this category, but *c'est la vie* — we can accept that.

A statistical consultant is a problem solver and the purpose of this book is to address the skills needed to be an effective consultant. To fully appreciate these skills, the statistical consultant needs to have a strong interest in science and, in particular, the art of scientific discovery. In this chapter, we explain the importance of statistical consulting in the context of the scientific process. If this holds a special fascination for you, then this book is for you! We begin with an example.

Example 1.1 *Is the population symmetric?*

In a consulting meeting a question was posed about a machine filling bottles of pills. The nature of the discussion was whether a symmetric distribution could be theoretically justified for modeling the number of pills per bottle. It was suggested that the symmetry was justifiable on the basis that there was a theoretical line in the bottle neck and the bottle could be equally filled above or below the line.

The issue here is not that the theoretical argument is necessarily right or wrong, but simply that the "solution" remains as a "theory" when in fact what is needed is a direct observation of the facts. Why not check it out? Gather some data and draw a histogram.

Example 1.2 *The size of the moon.*

Is the moon larger when viewed on the horizon versus directly above?

Again, why not check it out? Of course, the temptation to elucidate some new theory that explains the perceived difference in size may be hard to resist here. For the scientifically inclined, perhaps the diffraction of light through a greater volume of atmosphere might appeal as an explanation (as in some kind of atmospheric lens effect), but, why not just check it out by direct observation? (Use a ruler.) The correct answer is that what we are seeing is a visual illusion and in both cases the moon is the same size.

The point of the above examples is that the best source of knowledge we have is data and it is this symbiosis between data and theory that provides us with scientific knowledge. The examples show that theory alone does not always lead us to valid conclusions and it is crucial to look at data.

This combination of observational study and theoretical knowledge is called the **scientific method** and it is the procedure followed in scientific discovery. It is worth noting that this paradigm was not introduced until the late seventeenth century. In the next section we consider some examples from the history of science where statistics and the scientific method have played an important role.

1.1 History of the Scientific Method

Imagine yourself in the Middle Ages, say fifteenth century Europe, about 200 years after the ancient Greek texts had been translated into Latin in Arab Spain. These translated texts had slowly filtered to the rest of Western Europe and represented the basic scientific and philosophical knowledge of the time. In this world fundamental knowledge was based on the traditions of ancient Greece. The role of the scientist or philosopher (synonymous terms at that time) was to study from the old Greek texts which constituted the complete body of knowledge — there was nothing to be added.

On the other hand all was not static. There were technological advances coming from the side of the artisans who were advancing practical scientific knowledge at a very slow pace. However, the applied knowledge acquired by artisans was not always a cumulative process as they followed an oral tradition and did not keep written records and publications. Thus, knowledge and skills were sometimes lost.

We should point out that Greek science was not the only basis of scientific knowledge. Indeed, writing systems and calendars were developed by many ancient cultures and some had also made technical advances such as bronze casting. In particular, Chinese philosophy and science had steadily developed since the time of Confucius (551–479 BC). It was challenges to traditional Greek knowledge, however, that led to the process of scientific discovery based on the scientific method.

The first well known challenge to Greek science came from Galileo (1564–1642) with his famous Tower of Pisa experiment that showed that the speed of a falling body did not depend on its weight, but the most serious and fundamental challenge was to Aristotle's cosmology by Kepler (1571–1630) and Galileo. Specifically, this cosmology described a universe as a celestial sphere with the Earth at its center, and the moon, sun, and five known planets orbiting around the Earth at distances proportional to the strings of the lyre. Outside the celestial sphere was fire that could be seen through little holes in the sphere. Today, these are more commonly referred to as stars.

Although challenges to the Aristotelian orthodoxy had arisen prior to Galileo and Kepler, scientific inquiry was closely tied to the metaphysical values and theological beliefs of the time. The traditional belief that the heavens and Earth were fundamentally different collapsed when Galileo observed the moons of Jupiter with the telescope[1] and found that they were **not** rotating around the Earth. Here we should perhaps note that knowledge can also be a dangerous thing — this revelation almost cost Galileo his life at the hands of the Inquisition! Around the same time, Kepler had made observations of the orbits of the planets that confirmed the general idea of the heliocentric model proposed by Copernicus (1473–1543): *The Sun is the center of the solar system, with the Earth and other planets orbiting around the Sun.*

These two overwhelming observational discoveries clearly established the fallacy of the Greek model and quickly led to an explosion of scientific studies. Below is a brief summary of scientific thinking and some common misconceptions that were challenged by experimental data during the sixteenth and seventeenth centuries.

---

[1]The telescope was developed for military and commercial purposes by Dutch merchants, although Galileo claims to have had an earlier prototype.

Scientific Thinking Before the Renaissance

- The ancient Greeks' scientific tradition.
- All knowledge was learned from books.
- Mathematical theory for understanding nature. It was disconnected from experimentation.

  Orbits of planets were circular because circles are mathematically perfect. Spheres are perfect.

  Bodies fall at speeds proportional to their weights.

  Heat from the sun is different to the heat of a fire.
- Applied knowledge acquired by craftsmen was lost because they did not keep organized records and publications.

The Development of the Scientific Method

Important discoveries such as print, gunpowder, and the magnetic compass began to illustrate the importance of modern science. Gilbert (1540–1603) did extensive work in magnetism and was the first to recognize that the knowledge in "books" was not enough to explain many physical phenomena:

> *"New tradition of men who look for knowledge not in books but in things themselves."*

Kepler, Galileo, and Gilbert were among the first to use the scientific method, applying it without a "formal" understanding of the paradigm. The first people to formalize and really understand the implications of the scientific method were Sir Francis Bacon (1561–1626) and Rene Descartes (1596–1650). The following quote from Bacon emphasizes the depth of his understanding of the scientific method:

> *"The union of theoretical interpretation and the practical control of nature will produce a line and race of inventions that may in some degree subdue and overcome the necessities and miseries of humanity."*

Descartes formulated an approach to solving problems based on the scientific method as

Analysis The way practical things are discovered.

Synthesis The theoretical way the same things can be deduced from first principles.

He also provided us with an important reminder:

> *"There is nothing more futile than to busy oneself with bare numbers and imaginary figures."*

For the purposes of this book we interpret this as

> **Data without context have little meaning.**

Scientific Discovery

Kepler's observations of the positions of the planets led him to conclude that the sun was the center of the solar system. To advance scientific knowledge, however, we also require an explanation of the process that gave rise to the observations. This led Kepler to formulate his famous three laws of planetary motion. Here it is worth noting that Kepler initially formulated his second law as, "The periods are proportional to the squares of the distance to the Sun." After further analysis of the existing data he reformulated his second law 10 years later as, "The *square* of the period is proportional to the *cube* of the distance to the Sun." This iteration between observation and model formulation underlies the scientific method:

Theory Proposed A theoretical model is postulated to explain a phenomenon of interest.

Experiment Based on Theory An experiment is designed to test the plausibility of the model.

Theory Modified by Experimental Data The theory and data are compared and modifications are made to account for discrepancies.

Experiment Based on Modified Theory The revised theory is tested by experiment.

This process continues, potentially indefinitely, or until a sufficient approximation is reached for the purposes of the study. This iteration of theory and data is what produces the advances in science and understanding. We list some of the main discoveries of science over the last three centuries. For more information, we refer the reader to Mason (1962).

18th Century

Newtonian mechanics. Universal gravitation. Calculus.

The problem of navigation. Astronomy and the chronometer.

Biology: Development of individual organisms. Circulation of the blood. Cell theory.

Chemistry: Lavoisier's new chemical theory. Composition of matter.

19th Century

Darwin and the Theory of Evolution.

Development of geology.

The atomic theory of matter.

The wave theory of light.

Electricity and magnetism.

Mathematics: Statistics and geometry.

Engineering and science.

20th Century — A Very Short List

Modern biology.

Relativity theory.

Quantum mechanics.

Data analysis and statistics.

Social sciences.

1.2 The Development of Statistics

Modern statistics has its roots in the scientific method as is clearly evident from the fundamental role that empirical data analysis has played in important scientific discoveries. On the other hand, there is also a long history of data collection arising from the study of population vital statistics such as the census of ancient Rome and statistics of state economics. Two interesting examples of the latter are the Tables of Mortality constructed by John Graunt in 1662, and the annual index of wheat prices sold in European markets, 1500–1869, constructed by Beveridge in 1921.[2] However, the basis of statistical thinking did not come until the late nineteenth and early twentieth centuries, long after the notions of probability theory had been well established.

It is worth noting that the emergence of interest in problems relating to probabilities arose primarily from the development of insurance.[3] What really sparked the interest in probability came from the requests of noblemen gambling in games of chance! Indeed, it was the Chevalier de Méré who approached Pascal (1623–1662) on the so-called *problème des points* — an early example of statistical consulting. Pascal then began a correspondence with Fermat (1601–1665), and both men established some of the foundations of probability.

---

[2]The Beveridge index is reproduced in Anderson (1971).

[3]Tables of annuities were constructed by Johan De Witt (1625–1672) in 1671.

While the mathematical theory of probability flourished, the move towards statistical thinking only really began to develop towards the late nineteenth century. Early versions of the Central Limit Theorem and Law of Large Numbers[4] had already appeared and Quetelet (1796–1874) introduced the concept of normal populations in 1835. Formal statistical thinking came with Galton (1822–1911) and Pearson (1857–1936) who developed the notions of regression and correlation with respect to analyzing relationships between measurements where an inherent random component (error) is present. Both were crucial in providing recognition by the scientific community of the importance of formal statistical analysis to establish the validity of scientific results as an important part of the scientific method. Until now, scientists had used data analysis in an informal way.

An example is illustrated by Boyle's discovery of the law, $PV = K$, relating air pressure to volume.[5] Boyle (1627–1691) performed numerous experiments with the air pump without a formal inference procedure and based on this empirical evidence, formulated the mathematical relationship of the law. (If only all models could be this easy!) Relationships of this type were discovered fairly quickly and a more sophisticated data analytic and inferential approach was required to investigate relationships involving random error.

Galton and Pearson were successful in introducing methods for analyzing these types of relationships, although Pearson (1900) originally introduced his chi-squared statistic for $I \times J$ contingency tables with $IJ - 1$ degrees of freedom. It was not until 1922 that this error was pointed out by Fisher (1890–1962), much to the indignation of Pearson who reacted angrily to Fisher's assertion that he had incorrectly stated the degrees of freedom. A brief account of this controversy is given in Agresti (1990; p. 69) and further details may be found in the biography of Fisher by Box (1978).

The discovery of the t-distributions by Gosset (1876–1937) gave rise to another important aspect of statistical inference: cost-effective experimentation. At the time Gosset was brewer-in-charge at the Guinness brewing company and used the t-test to identify the best variety of barley. In response, Guinness promptly bought up all the available seed. Gosset was permitted to publish his discoveries, but not under his own name. Hence he used the pen name "Student" and the sampling distribution of the t-statistic appeared in Student (1908).

Fisher was certainly a pioneer in the use of statistical methods for experimental design, working for many years on agricultural field experiments at Rothamsted in England. There he developed and first used the analysis of

---

[4] Although the Law of Large Numbers is commonly associated with Poisson (1781–1840), there exists some debate over the extent of Poisson's contribution. Stigler (1986; pp. 182–185) provides an interesting discussion on this issue.

[5] This was later modified to account for temperature to give the Boyle–Mariotte law, $PV/T = K$.

variance procedure for statistical analysis. This work led to his influential book *Statistical Methods for Research Workers* (Fisher 1925). Unintentionally, Fisher created the practice of using 5% as the conventional risk level for all known tests! However, his writings helped organize statistics as a distinct field of study and established the basis of statistical inference in the study of practical problems. In the U.S. statistics also played an important role in agricultural experiments and the work of the Statistical Laboratory at Iowa State University[6] is clearly evident in the book *Statistical Methods* by Snedecor (1937).

In the industrial sector, statistical methodology played a fundamental role in the development and philosophy of quality control in manufacturing and productivity. Following the introduction of the principles of scientific management that arose out of necessity during the industrial revolution, two distinct approaches to quality control were developed by scientists at AT&T Bell Laboratories. Shewhart (1931) introduced the *Control Chart* in 1924 to study process variation online for the purpose of improving the quality of product, while in 1928 *acceptance sampling* methods were introduced by Dodge and Romig (1959) as a system for assessing the overall quality of a product based on inspecting a small number of samples.

Acceptance sampling schemes dominated quality control procedures in the U.S. for many years, but they were primarily employed to simply "monitor" a production process. It wasn't until the overwhelming success of the Japanese automotive industry that statistical methods for quality improvement were universally adopted by manufacturers[7] throughout North America and Europe. The philosophy of Deming and the methods of Taguchi which inspired Japanese industry have had a major impact on the way quality is viewed from the manufacturing perspective. The more recent concepts of *total quality management* and *quality assurance* emphasize the need to provide a complete quality "package" as statistical methods alone will not solve quality problems.

From the point of view of the general public, perhaps the most important contribution of statistics has been in the health area. The early polio trials established statistical methodology as the most efficient tool for assessing the effectiveness of vaccination. The statistical approach has also

---

[6] It was then known as Iowa State College. We have mentioned this in part as there is an interesting anecdote concerning Fisher's one (and only) visit to Iowa State University. Unfortunately, his visit happened to coincide with a period when temperatures reached 100°F (37.8°C). Fisher found the situation so intolerable, that he apparently lay on the floor of his room for the entire visit and promised never to return to Ames, Iowa again.

[7] While the U.S. automotive industry suffered greatly from foreign competition, the U.S. chemical industry began adopting statistical process control methods in the 1950s and has remained one of the most competitive in the world.

been crucial in establishing the link between smoking and lung cancer.[8] The following table summarizes some of the developments and applications of statistics that we have discussed above.

- Early developments in data collection and processing in the seventeenth century: Tables of Mortality in England 1662
- Normal distribution. Quetelet, 1835
- Poisson, the law of large numbers
- Galton 1879. Introduced the concepts of correlation and regression. He used the Quetelet–Gaussian distribution
- Pearson. Established the foundations of modern statistics
- Fisher. The design of experiments, 95% confidence
- Shewhart. Quality control charts, 3-sigma limits
- The polio trials
- Connection between lung cancer and smoking

1.3 An Overview of Statistical Consulting

So, what exactly is statistical consulting? Why is it needed? While demographic and scientific studies are the two pillars of modern statistics, this is a subject that is somewhat unique in the sense that we are always analyzing somebody else's data. Statistics has certainly benefited from the rapid developments in computer technology and statistical software is now accessible to a wide audience, however, the complexity of the questions under study in most disciplines requires at least some form of expertise related to data analysis. The problem is that most researchers don't have time to acquire this specialized knowledge along with the practical experience to apply it appropriately. There is a need to involve someone who understands the scientific process and has the quantitative skills to fulfill this important role: the **statistical consultant**.

[8]The Liggett Group Inc. was the first tobacco company to admit that smoking is addictive and causes health problems, including lung cancer. A public statement was made on March 21, 1997.

Outline of Topics Covered in this Book

Chapter 1	The scientific method. Consulting environments
Chapter 2	The importance of verbal and written communication
Chapter 3	Statistical methodology and using resources
Chapter 4	Putting it all together. An example
Part II	Obtaining experience: the Case Studies

In this book, our emphasis is on providing the statistician with some of the skills that will be needed to be an effective consultant. The purpose of the first part of this chapter has been to explain the importance of understanding the scientific process. In the next section we consider a variety of environments where statistical consulting plays an important role. For example, we describe the procedure used by the pharmaceutical industry for creating a new drug: from the compound phase to postmarket studies. In describing these different environments, we also focus on the responsibilities of the statistician. In general, we can distinguish these types of consulting environments.

1. The statistician works as part of a research team on a major project and shares statistical knowledge with the team. The objectives are to pursue the research project as a "group." This means that the statistician has to learn the area of research in sufficient depth to be able to contribute effectively to the project. This would be typical of the telecommunication and pharmaceutical industries.

2. The statistician is a consultant for many projects, often of small scale, or where the statistician plays an important advisory role. This is typical of university environments or business consulting companies.

In either environment, the statistical consultant's role requires a combination of four types of knowledge.

Scientific Is able to quickly learn the basic facts of the subject area within which the project under study falls.

Statistical Needs to have the knowledge and experience to be able to apply appropriate procedures for design and analysis.

Computational Is computer literate. Has sufficient proficiency to perform the actual data analysis using software.

Communication Last and most important, the statistical consultant must develop good verbal and written communication skills.

In most cases, a statistician ends up communicating simple facts and ideas, therefore what really matters are good communication skills — both verbal and written. Remember that verbal communication includes the ability to *listen,* as well as the skill to articulate ideas orally. Writing skills are important as the "report" containing the results of the analysis must be understood (and believed) by the researcher. Chapter 2 is dedicated to a detailed discussion of the different aspects of written and verbal communication. A summary of the necessary statistical and computational background follows in Chapter 3.

Of course, the true test comes when the statistical consultant is actually required to interact with a client. To help the reader obtain a more realistic sense of what statistical consulting entails, we have reproduced the entire consultation process for a particular project in Chapter 4. Since many of the issues considered in that chapter would be generic to other consulting projects, we have chosen to present a range of projects of varying statistical difficulty in the form of case studies and case study exercises.

The purpose of these case studies is to expose the reader to a wide range of disciplines and methods of analysis that can arise in the practice of statistical consulting. Although the statistical analysis is emphasized in these case studies, we have tried to provide as much contextual information as possible so that the reader can appreciate the reality of the issues associated with each case study. The case studies constitute Part II of the book.

There are several other texts on statistical consulting that are worth mentioning. Hand and Everitt (1987) present a collection of examples that show the statistical consultant in action. Chatfield (1995) provides a statistical perspective of the role of a consultant as a problem solver. Derr (2000), and the classic text by Boen and Zahn (1982), consider the role of the statistical consultant as an effective communicator. Some relevant articles on the subject of statistical consulting are: Tweedie (1998), Williford et al. (1995), Hand (1994), Maindonald (1992), Joiner (1982), Greenfield (1979), Marquardt (1979), Lurie (1958), and Kimball (1957).

1.4 Statistical Consulting Environments

The practice of statistics takes place in diverse environments. In the remainder of this chapter we consider some specific environments and identify the role of the statistician. We begin with the pharmaceutical and telecommunications industries which constitute the first and second (legal[9]) revenue sectors of the U.S. economy. Statisticians play a big role in both of these environments. Other types of environments are business consulting companies and government agencies that provide monitoring and regulate various

[9] Illegal drugs are estimated to be the largest revenue sector.

functions of society. Finally, universities often have small or large statistical consulting operations related to biostatistics, agricultural and industrial applications, as well as research in general.

In discussing each of these environments we identify the role of the consulting statistician and the expertise that a statistician can bring to these environments. We also discuss what it means to be a team player by providing some details on interacting with other researchers and the type of basic knowledge that is necessary for a statistician to be an effective member of a research team.

1.4.1 Pharmaceutical

The Drug Development Process

Drug development is a complex and lengthy process that can take 7 to 15 years for a single drug at a cost that may reach tens of millions of U.S. dollars. Although the following description relates to drug development in the U.S., we should point out that this is a highly competitive industry on the international level. Indeed, one of the criticisms of the Federal Drug Administration (FDA, the agency that regulates the U.S. drug market) is the length of time needed to obtain approval for new drugs.

For the major international pharmaceutical companies, there are three main parts of the drug development process that occur sequentially. We refer the reader to Peace (1988) for a more detailed description of the drug development process and biopharmaceutical statistics.

- Discovery and decision
- Preclinical studies
- Clinical studies

Discovery and Decision

The process starts with the discovery of a new compound or of a new potential application of an existing compound. In the case of a new compound, initial in vitro and in vivo testing are used to assess the activity of the compound.

in vivo testing the compound in living organisms (cells).

in vitro tests performed in laboratory solutions (test tube).

Based on adequate results, the decision whether to develop the drug is then made. If the decision is positive, then the development process enters the stage of preclinical studies.

Preclinical Studies

In the preclinical stage, the initial toxicology of the compound is studied in animals. Initial formulation of the drug development and specific or comprehensive pharmacological studies in animals are also performed at this stage. At the end of this phase, the evidence of potential safety and effectiveness of the drug is assessed by the company. This is very important as the evidence must provide the company with reasonable confidence that the drug dosage will not be fatal and can be tolerated by humans. To proceed further, a U.S.-based company needs to file a Notice of Claimed Investigational New Drug Exemption. The intention of this notice is to allow the company to conduct studies on human subjects. This next stage involves several types of clinical trials.

Clinical Studies

At this point, there have been extensive tests done on animals and these studies have suggested that there is sufficient evidence that the drug will be of benefit to human subjects. Testing the drug in human subects is the next step and there are three types of studies which are performed in sequence.

Phase I The purpose of a Phase I clinical trial is to establish the initial safety information about the effect of the drug on humans, such as the range of acceptable dosages and the pharmacokinetics of the drug. These studies are normally conducted with healthy volunteers. This is to avoid as many confounding effects of the drug and disease as possible. The number of subjects typically varies between 4 to 20 per study, with up to 100 subjects in total used over the course of the Phase I trials.

Phase II The Phase II studies are directed towards patients who will potentially benefit from the new drug. Here, effective dose ranges and initial effects of the drug on these patients are assessed. Up to several hundred patients are usually selected without additional complications to avoid confounding factors.

Phase III Phase III studies provide assessment of safety, efficacy, and optimum dosage. These studies are designed with controls and treatment groups often consisting of hundreds, or even several thousand patients. Based on successful results obtained from these studies, the company can then submit a NDA: New Drug Application. The application contains the results from all three stages — from discovery to Phase III — and is reviewed by the FDA.

The FDA review panel of the NDA consists of reviewers in the following areas: medicine, pharmacology, biopharmaceutics, chemistry, and statistics. Each reviewer may request further information or tests to

be performed by the company related to the NDA before submitting a written review to the FDA.

Postmarket Activities

Once the NDA has been submitted to the FDA, promotional activities, outcomes research analysis, and followup studies to examine the longterm effects of the drug are initiated. These studies can take the form of a clinical trial and are referred to as Phase IV studies. The main purpose of these studies is to ensure that all claims made by the company about the new drug can be substantiated by so-called "clinical evidence." FDA approval is required before such claims can be used for promotional purposes. All reported adverse effects must also be investigated by the company and, in some cases, the drug may need to be withdrawn from the market.

Statistician's Responsibilities

The statistician clearly plays an important role throughout the drug development process and we list some of the responsibilities that would be expected of statisticians in the pharmaceutical industry:

- Participate in the development plan for studying a drug.
- Study design and protocol development. Randomization schemes.
- Data cleaning and database construction format.
- Analysis plan and program development for analysis.
- Report preparation. Produce tables and figures.
- Integrate clinical study results, safety and efficacy reports.
- Communication and NDA defense to the FDA review panel.
- Publication support and consultation with other company personnel.

1.4.2 Telecommunication

At the end of 1983, the American Telephone & Telegraph company (also affectionately referred to as "Ma Bell") was broken up. The breakup originally resulted in the "new" long-distance telecommunications giant, AT&T, along with the creation of seven regional companies that became the local-service providers within the U.S., the so-called "Baby Bells."

A lot has changed since then! Aided by the Telecommunications Act passed by the U.S. Congress in 1996, the grownup Baby Bells are now competing with each another and with AT&T and other long-distance providers. The Baby Bells have also branched out into the wireless and

Internet communication industries and some have invested heavily in international markets. Conversely, international telecommunications companies have also entered the U.S. market.

The original Bell Labs has a long history of statistical research and companies that arose from the breakup of Ma Bell, such as AT&T, Lucent, and Bellcore, have certainly continued with this tradition. These companies hire large numbers of statisticians in both the research and business divisions. The research division is typically divided between core and applied research, although the organizational divisions do not preclude individuals from different areas working together on joint projects. For those interested in the telecommunications industry, the articles by Kettenring (1995, 1997) are worth reading.

- Extremely large databases
- Research and business divisions
- Communications and information technologies

The unique feature of the telecommunications industry is the enormous amounts of data available to the statistician arising from many projects. Problems in information processing technologies, such as network testing, package transmission and delays, and software development and testing, generate gigabytes of data that need to be analyzed from many directions. The statistician has a lot of flexibility in terms of the way the data analysis is done as long as results are produced. There is also room for developing new methodology if necessary, and many of the jobs are research-oriented rather than production-oriented.

Typically the statisticians will be involved in one or more projects working on teams with engineers and computer scientists. Such a team would have a project manager that takes care of administrative work, a technical leader that directs the research, and a group of technical members. Most of the projects are generated by technology leaders, but may also come from other areas that request help. It is possible to initiate one's own project although this tends to happen less frequently. The statistician needs to become familiar with the ideas and terminology of communications technology and to understand what is (really) meant by networks, switches, packets, and the like. He or she will need to learn constantly the new innovations in the technology so it is important to have an interest in scientific issues.

There is a distinction between the business and research sides of the activities. The business side is dominated by product management, market research, and marketing whereas the research side is generally dominated by activities related to software applications which can be quite varied.

Salaries and benefits are commensurate with other research industries such as pharmaceuticals, and the bonuses can be very generous.

1.4.3 Business

Consulting Companies

Small companies provide the statistician with a consulting environment that encompasses a wide range of activities and often the consultant is involved in the full process, rather than just concentrating on a specific task or type of analysis. This may include direct interaction with clients from outside companies, report preparation and presentations, as well as budget proposals and other administrative tasks. Hahn and Hoerl (1998), Hoerl et al. (1993), and the text by Hamilton and Parker (1993) are worth reading.

Confidentiality of client information is a key component of successful business consulting and some typical areas in which small consulting companies and independent research contractors (IRC) specialize are:

- Market research
- Survey design and analysis
- Financial analysis
- Pharmaceutical and telecommunication projects
- Database management
- Expert witness for legal cases

Of course, this is only a portion of the areas covered by small businesses where statistical consulting services are involved and some companies may either subcontract, or actually specialize in subcontracted consulting projects. For example, a company may subcontract faculty or students from a university consulting program to assist with certain parts of a project.

This type of environment provides an opportunity for statisticians to learn all the details of how the company and the project process work, as well as giving them more exposure to outside interaction with clients. Thus, there is a wide range of activities in which the consultant can participate, as compared to a larger company where a consultant's role in a project may be more task-specific. Other advantages are that the consultant's work schedule can be more flexible, the type of work is more varied, and the remuneration can be more advantageous (e.g., the consultant's contract may include some form of profit sharing as a bonus).

On the other hand, the type of skills that are needed are more comprehensive. Good communication skills (verbal and written) are required in addition to statistical knowledge and business skills. The consultant may need to help on the initial contract proposal, prepare budgets, and advise or convince clients of the quality of the work. Presentations play an important role in making a pitch to the clients and keeping them satisfied. The consultant may also be entirely responsible for completing the statistical analysis — on time and within budget. Thus, there are certain constraints that take high precedence such as "deadlines" which *must* be kept; otherwise the company will lose its client. Some small consulting firms are known for having very hectic work schedules.

Example of a Market Research Survey Analysis

An outline of the stages and work items that would form the basis of a budget schedule for a market research survey project is presented in Table 1.1. The actual budget would allocate a certain cost for each work item (based on an hourly rate, for example), and also include certain overhead costs that are not considered here. Furthermore, this market research project was based on a survey design that had been used for the same client previously.

In this particular survey, the client was interested in assessing how the company and product was perceived by the consumer in terms of price, quality, and image. Although this type of survey had been conducted for the client previously, it is worth noting that a substantial number of work items in this budget concern database management, documentation, client interactions, and report preparation and presentation. We expand on these important aspects of statistical consulting in the next chapter.

Private Consultants

Prefer to keep your own hours? Boss yourself around for a change? Then become a full-time private consultant! But ... just before you quit your day job, here is a short list of some items you may want to review first:

- What computing hardware do you have? Can you afford to upgrade every few years?
- What software do you have? Can you afford to upgrade when new versions are released?
- Who will prepare your reports, accounts and invoices?
- What is your client base?

The last item is the most critical question to consider since a private consultant obviously needs clients to stay afloat financially. The main problem is that "small" projects are really **not** economically feasible because of the

TABLE 1.1. Budget Outline for a Market Survey Project

Stage and Work Items
Data Collection:
Pretest and manage data collection
Review response rates
Coordinate exchange of sample/data
Progress meetings with client
Survey Validation:
Clean, format, label data
Respondent profile (demographics)
Create tables for report
Profile of respondents' company/product usage
Examine brand and question order effects
Summary of responses by survey sections, companies
Standard errors for tables and graphs
Select results for graphical presentation in report
Meeting with project leader
Survey Analysis:
Scale analysis, decision trees for value/quality
Correlations and factor analysis: price, value, and questions
Correlations and factor analysis by section: client vs. others
Factor analysis for product questions: client vs. others
Logistic regression: client vs. competitor
Client meeting: discuss validation and analysis results
Modeling:
Set up data for modeling
Modeling to determine relationships between variables
Structural equation modeling
Client meeting: discuss modeling prior to write-up
Reporting Results:
Project meeting: discuss results and organize writing efforts
Creating/inserting tables and graphics in report document
Presentation to client: meeting scheduled for entire day

cost is too high. That is, it takes more time to meet with the client and prepare the necessary documentation (contract, invoices, and the report), than it does to analyze the data. The reader can find a detailed illustration of the overhead associated with a small project in Chapter 4.

Private consultants therefore need to be able to attract clients with "large" projects. Normally, this is achieved by establishing a network of clients long before the decision to go private is made by the consultant. Clients with large consulting projects clearly want to know to whom they are giving their precious data. We conclude with some additional observations from colleagues who have braved the murky waters of private consulting.

- Network, network, network.

  Some projects require a "team" of consultants with different specialties. A private consultant's network needs to extend beyond statistics.

- Find a niche or specialty that clients need. This may require the private consultant to possess or develop skills in areas outside statistics.

- Partnerships allow the workload to be shared and increase exposure. Larger projects can be solicited.

- Expert witness. Some private consultants specialize in legal cases where expert witness representation is required. We discuss some of the issues involved in being an expert witness next.

Expert Witness

At some point, you may become involved in a project where legal issues are important. In a discrimination case, for example, a plaintiff may be claiming they were unfairly treated in comparison to colleagues who were in a similar situation; the defendant, of course, would be trying to argue otherwise. Here, the role of a consulting statistician is to try to provide an objective assessment of the discrimination claim from a statistical perspective. That is, your specialized knowledge in the area of statistics makes you an "expert witness" in the case. See Kaye and Zeisel (1997), Finklestein and Levin (1990), and Fienberg (1989) for details concerning statistics and the legal profession.

To be an expert witness in a case does incur certain obligations. Hence, you must first decide whether to "take the case." This decision should be made as soon as you have fully apprised yourself of both the statistical and moral aspects of the case. If you have strong moral or ethical objections to the nature of the case, you should decline it; your effectiveness as an expert witness is likely to be of limited value. If you do agree, then you would be expected to provide testimony in a formal legal proceeding such as at a deposition or trial. Bear in mind that any testimony you do provide

will be interpreted entirely within the legal context. Maintaining statistical objectivity under these circumstances can be difficult. Some of the aspects of being an expert witness that you need to consider are:

- Duration of the case
- Answering questions
- Maintaining objectivity
- Working with the lawyer
- Efficient documentation.

Duration The outcome of a case can take a long time to decide. Hence, remembering the "exact" testimony you gave in deposition a year ago can be hard, if not impossible. This can be frustrating if used simply as a legal tactic, so be clear as to what your overall statistical conclusions are. This part of your testimony must be consistent with any previous testimony you provided.

Answers When you testify, your answers must be understandable and delivered in "layman terms." You cannot use statistical jargon (without explaining it first). Remember, this is not some practice session with your client's lawyer; whatever you say immediately goes on record. You should therefore answer any question with great care. Do **not** say more than you need to.

Objectivity Your purpose as an expert witness is to provide an objective assessment based on the statistical aspects of the case. This means you must not allow your testimony to be construed as implying causality. For example, a significant result from Fisher's exact test does not imply that discrimination occurred. The difficulty here is that the desire to assign causality can come from both sides of the case. Resist this temptation at all costs; your statistically flawed testimony will be used against you later.

Lawyers It is not your job to win the case. Just as you cannot expect your client's lawyer to become an instant expert in statistics, neither should you try to become a legal expert. If you think you have a great example to use in the case, talk to the lawyer about it **first**. Surprises and ad lib examples can easily go wrong in a legal context.

Documents Reports, computer printouts, and notes related to the case that you possess can all be requested. This can include email you sent or received from your client, as well as your verbal discussions (with

whom and about what) that were related to the case. Be discreet. This is not information to casually share around the watercooler. Similarly, try to keep your documentation efficient. Draft versions of your report do not need to be kept.

1.4.4 Government

The U.S. government and state agencies are some of the main sources of employment for statisticians. The article by Ross (1995) will be of interest for a statistician contemplating a career in government. Agencies that employ large numbers of statisticians are

1. Census Bureau
2. Food and Drug Administration (FDA)
3. Environmental Protection Agency (EPA).

The Census Bureau is widely known for performing the U.S. Census of Population. However, the Census Bureau mission is much broader. It conducts surveys in many areas such as retailing, housing starts and other housing topics, manufacturing, agricultural production, insurance claims, foreign trade, business, transportation, and many others. Statisticians play an important role in the gathering and processing of these data.

- Survey design.
- Data collection.
- Statistical data analysis and data mining.

Another aspect of the Census Bureau's mission is research and development of new methodology. Traditionally it has concentrated in survey sampling methodology, but more recently there has been interest in a broader range of methods such as data visualization, time series analysis, and data mining.

The FDA is a regulatory agency whose goal is the regulation of pharmaceuticals and food safety. Statisticians with an interest in medical science, biology, and food sciences may find a career at the FDA of interest. In drug regulation, there are three main areas where statisticians play an important role:

- Human drugs
- Animal drugs
- Medical devices.

Each of these areas deals with the regulating process of drugs and devices. The statistician plays an important role by helping to answer the basic questions of safety and efficacy of the drug or device. The statistician may participate in evaluation panels or write reports for the drug or device evaluation process.

The FDA in conjunction with other state and federal agencies is also responsible for monitoring of the food supply. Here the role of the statistician is in the analysis of the stream of data coming from the continuous testing of the food and bottled water supply according to established criteria. Two other areas of involvement of the FDA where statisticians play a role are the review and licensing of biological products and the regulation of cosmetics.

The EPA also involves statisticians in the regulatory process in projects such as the cleaning of toxic waste sites, and the enforcing of the clean air and clean water acts. The statisticians are involved with groups of chemists, biologists, or environmental scientists who conduct studies of the sites in question. Often these projects involve questions that are difficult to answer with standard methodology and the statistician may get involved with developing interesting and new approaches to the analysis.

There are other government areas that have hired a moderate number of statisticians in the past. Among them are the National Security Agency (NSA), the National Institute of Health (NIH), the U.S. military, NASA, and the CIA. While we have concentrated on agencies that are specific to the U.S., it can be seen that statistics is an important tool in many areas of government and this is certainly true of other countries as well. Finally, there are some well-known international agencies that employ statisticians such as the Organization for Economic Cooperation and Development (OECD) which collects and disseminates economic and environmental information.

1.4.5 University

University statistical consulting programs can provide an important resource for industry and small business consulting companies. The benefits of establishing an interaction with the private sector are clearly mutual: the overhead costs of contracting faculty members and students are relatively low. (No need to pay benefits, they have the statistical knowledge, and the university has the necessary computing equipment.) In particular, students get that all-important chance to gain real-world "experience." Information on established university consulting centers can be easily accessed via Internet connection to the university's website. Other references related to consulting in a university environment are Kirk (1991), Rustagi and Wolfe (1982), and the American Statistical Association (ASA) Section on Statistical Education.

University statistical consulting environments have actually existed for some time and were established as research centers to assist with improving agriculture.[10] While many of the U.S. land grant universities continue to maintain strong ties to the agricultural industry, there is a wide range of statistical consulting environments that exists at universities today. In some cases, statistical consulting may simply be provided on the basis of collegiality: a client contacts an appropriate faculty member to obtain statistical assistance. Below, we briefly consider the situation where there exists some type of formal structure within the university that clearly identifies a "center" for the purpose of statistical consulting.

Statistical Consulting Program (SCP)

Establishing a statistical consulting program at a university can provide invaluable internal support for researchers across a broad range of disciplines. Once established, the SCP will also provide exposure for the university through external consulting activities and outside grant solicitation. The three main objectives of the SCP are:

Consulting Center To serve the research needs of the university.

Consulting Course To expose students to the process of scientific thinking in the role of statistical consulting. This can be combined with a formal graduate course.

Industrial Advisory Board To establish a formal mechanism for interacting with the research community in industry. Students can be provided the opportunity of working in a nonacademic environment through internships, and industry benefits by having the chance to employ the best qualified students.

In the SCP environment a consultant will often work with one or more students on several projects at the same time. There is usually little time to learn more than the minimum science of the problem in order to resolve the statistical questions of the problem. However, as we show in the case studies presented in Part II, long-term research projects do arise.

[10] Indeed, much of the terminology associated with experimental design, such as *blocks, plots,* and *split-plots,* is directly related to the physical divisions of a parcel of land that were required by the design. Hence a "plot" was exactly that — a plot of land.

2
Communication

The ability of a consulting statistician to communicate effectively is very important. No surprise there, of course. Effective communication is certainly a desirable skill worth developing wherever there exists the need for interaction between two parties. During a statistician's training, however, considerable emphasis is often placed on developing the necessary technical skills, leaving communication as something that can be "picked up later." For the consulting statistician, "later" is no longer an option; good communication skills are required to be an effective consultant.

In this chapter, we specifically focus on the role of communication and explore some of the common elements and skills that are involved in effective communication. These are discussed in detail with a complete description of our approach and "how-to" guidelines are provided. Of course, developing good communication skills requires time and effort and our guidelines are not intended to be a substitute for this.

We begin by considering the following situation. When a statistician is involved in a project, a process takes place involving the transfer of information. Whether this is done with a group of collaborators or an individual client, there are certain common elements involved in the communication process:

1. Verbal interaction with the client(s) which continues until substantial progress has been made on a project.

2. Preparation of technical summaries, report writing, and presentation of results.

Certain skills will be required to perform these tasks effectively and in the next two sections we consider what is involved in interacting verbally with the client. Since the focus in this chapter is on communication, we proceed directly to the report writing stage of the project. While this presumes the analysis was able to be performed using the appropriate statistical methods, details concerning specific statistical techniques are covered in the next chapter and are not needed for our discussion on report writing.

We also discuss the role of oral presentations and provide details on how to make these presentations effective. The importance of quality graphics for presentation purposes is addressed in Section 2.6. A short introduction on the use of the PowerPoint software for enhancing presentations is given at the end of this section.

Before continuing, we must emphasize that any guidelines we provide for effective communication will not fit every consulting situation. However, it is hoped that the reader will benefit by having the opportunity to follow our approach in detail, and adapt it to specific situations as needed.

2.1 Verbal Interaction

There is a variety of situations where the statistician must interact directly with a client or group of collaborators. The main purpose of this interaction is to exchange information concerning a project of interest and to do this effectively the statistician needs to develop communication skills in the following areas.

1. Initiating the interaction process.

2. Understanding and defining the problem.

3. Evaluating the technical knowledge of the client/collaborator.

4. Assessing the overall issues and objectives of the project.

5. Identifying the statistician's specific contributions to the project.

For the purposes of this discussion we refer to the client(s) or collaborators as simply, the "client." While this word has certain overtones associated with it, the communication skills we need to develop are the same whether we are dealing with an individual client, or a group of collaborators. That is, these skills are not tied to a specific type of consulting environment where a certain nomenclature may be preferred.

Initiating the Interaction

So where do we start? How do we initiate the interaction? Deer (2000) and Bohn and Zahn (1968) both deal with this particular issue in detail and place considerable emphasis on the importance of creating positive first impressions. The physical setting of the meeting room and our initial nonverbal behavior towards the client will create their first impression of us — and we haven't even said "Hello" yet. We consider some of these nonverbal cues in the example presented in Chapter 4. Common courtesy and respect obviously go a long way towards creating a positive environment for our consultation meetings. Some simple things we can do to help make clients feel comfortable when we greet them are:

1. Stop what we are doing *immediately* and get up to greet the client. We may need to take the client's coat or indicate where they can put their briefcase.

2. Make eye contact and smile. This conveys the message that we are pleased to see the client and gives us an opportunity to assess the general demeanor of the client. If they appear rushed, give the client a little time to relax. Talk about the weather or other peripheral matters before asking about their project.

Once we are over the preliminary introductions, the client may want to show us some data, or mention a statistical procedure they want to use. However, it is necessary to start from the beginning. That is, we need to start with the context of the problem because without context, data have little meaning.

- How much do we need to know?
- Ask lots of questions.
- Always be prepared to take notes.

How much do we need to know? What we are really asking is how much information we need from the client in order to resolve the statistical aspects of the problem. Of course, the problem has yet to be well defined so some strategies are needed to elicit the appropriate information. The obvious approach is:

Ask lots of questions: At this early stage of the consultation it is often useful to have the client begin the session by describing the project in their own words. This gives us the opportunity to learn about the client's field and make appropriate interruptions whenever unfamiliar

or specialist terminology is introduced. Never heard of the "EQRT" scale? Then ask! We are not expected to be an expert in every field of scientific inquiry. Now read the last sentence again. Why? Because neither do we need to become an expert in the client's field.

When we do ask a question, we must also *listen* (carefully) to what the client says. Remember that clients come to us for statistical advice. They cannot be expected to know that certain terminology has quite specific meanings in statistics. Saying a factor was "significant" implies something quite different to us than if they had said it was important. The reverse could apply just as easily, of course. The client could have said "important" when, in fact, they meant significant (based on a previous study, for example). In our experience, clients often tend to do two things:

1. Use statistical terminology inappropriately. We should always double check what the client means.

2. Fail to mention important variables such as design factors that were employed in the experiment. That is, we also need to listen for what is *not* said.

Always be prepared to take notes: Naturally, we were ready to take notes during this question-and-answer session. ... Do **not** assume we will be able to remember all the details about the client's project later. Taking notes *during* the consultation session is an essential part of the documentation process (Section 2.3) and we emphasize the importance of adopting this practice.

Defining the Problem

Our initial task is to try to understand the context of a project from the client's perspective. This means we need to learn something about the client's field and its associated terminology before trying to define the problem. As we become more familiar with the "context" of the problem and begin to communicate with the client using a common basic terminology, the purpose of our questions can then be directed towards the following aspects of the project. This information will be helpful in defining the problem and identifying the statistical issues involved.

- Background of the project
- Status of the project
- Aims of the project
- What the client expects

Background Projects are often based on previous studies in which case there may be an established or accepted method of analysis. If so, obtaining a relevant reference from the client can help us ascertain whether the established method of analysis is reasonable and applicable to the client's problem.

Status What is the status of the project? If the study is in the pre-experiment or planning stage, our contribution can be important in ensuring the planned experiment will produce reliable data for the subsequent analysis. If the data have already been collected, we will need to direct our questions towards the collection process. How reliable are the data? Is the client aware of any outliers in the data? Was the experiment performed in accordance with the usual principles of statistical experimentation: control, randomization, and replication? Is there enough evidence (sample size and structure in the data[1]) to support the objectives of the project?

Aims What are the aims and hypotheses associated with the study? Are the client's objectives commensurate with the results that can be obtained from a statistical analysis? In some cases, certain hypotheses may need to be reformulated in order for the statistical analysis to provide valid conclusions. We should also make sure the client understands the distinction between causality and conclusions based on a statistical analysis.

Expectations What does the client expect from us? We are not magicians, nor are we directing the project. Our responsibility should always be to the statistical aspects of the problem; it is the client's responsibility to articulate the importance and motivation for the project.

[1] In one project, we were ready to start the analysis of a rather large dataset (100 MB) only to find out that a key variable had not been recorded in the original study. Hence the objectives of our project could not be met.

Technical Knowledge of the Client

Defining the statistical aspects of a client's problem for ourselves is the easy part. Now we need to explain them to the client. At the same time as we are defining the problem, it is useful (and sometimes revealing) to ascertain the client's knowledge and understanding of the statistical aspects of the project. For example, a client may ask, "How large a sample do I need?" How should we respond? In terms of the power or accuracy[2] of statistical tests, or with a more familiar notion like margin of error? This clearly depends on the client's technical knowledge. In our experience, margin of error often provides a useful starting point and avoids possible misinterpretation of terms such as accuracy or precision.

- How well does the client understand the project?
- How much statistical knowledge does the client possess?

As we indicated above, there may be a basic or established statistical methodology that is well accepted in the client's field. However, knowledge of a statistical procedure does not necessarily mean the client fully understands the concepts underlying the statistical procedure. Thus, part of our role can be an educational one. We consider some issues that may need to be addressed in this respect.

Educating the Client The client did not come to us for a statistics lecture! Our explanations should be given in the context of the client's project; provide the client with an interpretation of the outcome and purpose of a statistical procedure, not the mathematical details. For example, a P-value can be explained in terms of "risk" rather than a probability based on some type of distribution. Be patient, but avoid getting stuck on details that are not essential — *How much do we need to know?* also applies to the client.

Level of Sophistication The statistical methods employed for analysis need to be appropriate for the problem and this may require introducing the client to more sophisticated approaches. However, we should not try to make the statistical analysis more complicated than is really necessary. The client needs to be able to interpret the results of the analysis irrespective of the level of statistical sophistication.

[2]When asked, "How accurate do you want your test results to be?" a client promptly informed one of the authors they wanted the results to be 100% accurate. We quickly returned to the matter of determining an appropriate sample size.

Formalizing the Problem A more complicated issue is the potential need to formalize the ideas of the client. They may think about their work in a more intuitive way which needs to be carefully formalized before the statistical analysis of the problem can be performed. The time spent in formalizing the problem is well spent because it will help the client understand the research from a statistical perspective. In some cases, this may even lead to a better formulation of the research objectives.

Example 2.1 *There will be an interaction between* PST *and* PREF *by* GROUP.

In Section 9.1 (*Improving Teaching*), the hypothesis stated above was the client's best approximation to the "formal" hypothesis statement of the problem. What the client really wanted to know was whether the factors PREF and GROUP had an effect on the response PST. That is, was there was an interaction *effect* due to these factors.

Overall Issues and Objectives

At this stage we should have established a sufficiently good communication channel with the client and can now go into details concerning the overall statistical issues involved in the project. The following items should be able to be discussed in a language that we both understand.

- Aims and hypotheses of the project.
- Current or prior methodology, if any.
- Intended use of postexperiment results.

Objectives In some studies, the objectives may only be exploratory in nature and the appropriate hypotheses have yet to be formulated. On the other hand, if there are specific objectives of interest, we need to ensure that the experimental design will provide statistically valid results. If the experiment has not yet been conducted then we need to address issues related to the design such as sample size, randomization, and control, as well as implementation issues. If the data have already been collected, we need to consider whether the objectives of the study will be met by the current data. Additional data may be required.

Methodology To ensure that the statistical procedure is applied appropriately, specific issues will need to be addressed such as:

- What is the data type of each variable?
- Are there outliers or missing values present?
- Do certain constraints exist in the process?

This requires interacting closely with our client during a consultation session. (Avoid confusing the client: missing values generally tend to be, well . . . missing!) We present an example of this type of interaction in Section 3.5, where some regression-specific questions are posed. We should also compare the current or prior methodology that is being used in the project with established statistical procedures that may be more modern or more appropriate.

Postexperiment It can be worthwhile to consider the intended use of the postexperiment results. For example, does the outcome of the study depend critically on obtaining significance for a particular hypothesis. What are the consequences of getting a nonsignificant result? Make sure the client understands that we cannot simply change the result of an analysis because it doesn't support the client's initial objective.

Specific Contributions

The final stage of our verbal interaction with the client involves identifying our specific contributions. This is important to ensure that both we and the client understand clearly our respective roles in the project. The following items should be addressed as necessary.

Data management If we are responsible for performing the actual computations the client needs to provide us with the data in suitable format.

Data Analysis Error checking: The client needs to be aware that the initial stage of our analysis will involve checking the data. The client will need to provide us with any corrections.

Statistical Analysis Both we and the client have agreed on the method that will be used to analyze the data and the details of performing the computations.

Report writing Whether there is an expectation of presentation quality graphics and tables or any special requirements in the report that will be written.

Time frame There needs to be a realistic time frame to allow us to perform the analysis and complete the written report for the project.

2.2 Other Aspects of Verbal Interaction

Persuasive Communication

In practice, the different components of verbal interaction that we have just discussed may be performed simultaneously. The art of persuasive communication requires the creative combination of these components to make the interaction with our client more efficient. Handling a consultation session well relies on good organizational skills. We should be prepared to interrupt (politely) and redirect the client towards relevant issues as necessary. This will make our consultations sessions more productive and enable the analysis phase of the project to commence.

Initial Contact

Prior to our initial appointment with the client, we may have already established indirect communication and gained some information about the project. In this case, the focus and direction of the discussion during this initial meeting may be predetermined to some extent. However, it is important that all prior information be reiterated at the beginning of the consultation to ensure there is mutual agreement on the content of previous communications. Doing this also helps to develop the nature of the working relationship between ourselves and the client.

Decision Time

In certain cases we may need to refuse participation in the project due to various reasons. For example, we may have constraints which would prevent us from completing the project in the time frame required by the client. In legal cases, there may simply be a conflict of interest because of our work with a previous client. In these situations, the decision not to participate in the project would usually be able to be determined prior to the initial consultation session.

What if the need to make this decision arises during the consultation session? Informing a client of our decision not to participate in a project can be difficult, but needs to be done during the initial stages of the consultation process whenever possible.[3] It is crucial that the client is not led to false expectations of our intent to participate in the project. More important, the client has invested their time in discussing the project with us and now needs to look elsewhere. The key issue is:

Knowing when to walk away.

[3]Our decision to decline a consulting project certainly needs to be made before performing any actual analysis; we may be legally or contractually bound to complete the project analysis if it is started.

Of course, knowing when to walk away does **not** mean we just simply up-and-leave! Indeed, an important part of our decision will be informing the client about the problems in their project and providing advice or recommendations that are appropriate to overcoming the limitations of the study. We may even generate a "new" project; this time with the added bonus that we can participate in the design phase of the study. Some reasons why the best option may be to decline a consulting project are:

- The sample size is too small for any meaningful analysis or is simply too large for our current computing resources.
- The data is biased or poorly gathered and there is no opportunity for further planned data collection.
- The client may not really understand their project or their expectations of the analysis results are unrealistic.
- We may not understand the client's project or have limited expertise in the type of statistical analysis required for the project.
- We have moral or ethical objections to the project. This includes statistical ethics such as a client who "requests" a particular method of analysis which is clearly inappropriate.

Negative Outcomes

"Oh, I'm sorry. Here. You can put your gum in this (the waste basket)."

Since the context of the statement is missing, two possible scenarios in which this statement could have arisen are: we were a bit slow in picking up the client's nonverbal cues; after starting the consultation meeting, we noticed the client was chewing gum. In the first scenario, our statement was simply a response to a nonverbal request initiated by the client, prefaced by a genuine apology as a friendly gesture. In the second scenario, our statement clearly served a different purpose: we were attempting to address a "negative" situation that occured during a consultation session.

The above scenario is somewhat trivial, of course, but it does provide a useful example for illustrating some of the issues that the statistical consultant may need to consider when attempting to deal with negative situations. These are summarized below.

Perception What is the problem? Every client has mannerisms and idiosyncrasies that we may find "annoying," but this hardly counts as a negative "situation." Addressing unimportant problems will only make the situation worse. If a client chewing gum is not a big problem for you, let it slide.

Consequences What are the consequences of addressing a problem?

- By exposing the situation, we may convey a very negative impression to the client. (The client may be offended by the way we made the request: "What's with the insincere apology: 'Oh, I'm *sooo* sorry.' ")
- We may end up spending valuable time on issues unrelated to the client's project. (The client berates us for using the indirect approach. "You could've just asked me straightout." The situation is now worse and we need to spend more time trying to make the client feel less slighted.)
- It could backfire on us. (The gum is something the client needs for medical reasons. We assumed it ordinary gum and now face the two problems above.)

Timing When should we address the problem? It is usually better to address simple problems immediately, but in some cases it may be better to "sleep on it." That is, defer addressing the issue until the end of the consultation session. Perhaps even try to think of a creative solution to employ at the beginning of the next session which might circumvent the problem.

Win-Win What is the purpose of addressing a negative situation? The aim, of course, is to achieve a positive outcome. The consultant and client both benefit from understanding what the problem was and are able to move on to more important matters. Deer (2000) refers to this as a "Win-win outcome" and provides numerous examples of positive and negative situations that a statistical consultant may encounter.

Continuation ...

Developing good communication skills is an evolving process and while experience will certainly help, it is important for a statistical consultant to continuely reassess their performance. How might that difficult situation we got into last week, have been better resolved? Why did it take *two* meetings to clarify the objectives and work assignments for this project? Cultural differences may also have an impact on our interaction with a is client. Clearly, we have not addressed every aspect of verbal interaction in our presentation above, nor have we considered all the different types of consulting situations that a consultant can expect to encounter.

Fortunately, there are many other sources available that can help us improve our communication skills! The text by Deer (2000) and accompanying video are certainly worth looking at. Some articles relevant to this section are: Tweedie (1998), Finney (1982), and Lurie (1958). Finally, business orientated publications often provide useful advice that can be incorporated into the statistical consulting environment. See, for example, Hamilton and Parker (1993), and Yeatts and Hyten (1998).

2.3 How to Write Reports.

Our statistical analysis is now complete. The results supported most of the project's hypotheses, and we have an hour free to whip out that "report" the client requested. Great! Sounds simple enough so let's get started. What do we need? Some sort of introduction, the results, and our conclusions. The computer output? That can all go in an appendix at the end. Done! Not quite. In our contacts with leading researchers in industry, the most prevalent complaint we hear from these scientists is that our statistics graduates have great difficulty in writing reports. Indeed, there are cases where industry scientists involved in a project routinely reject the inclusion of a junior statistician who does not possess good written communication skills. While this may seem unfair, the "good job" we did on the analysis is of little value if nobody can understand our written report!

The objective of this section is to provide guidelines on writing reports. These guidelines do not guarantee "quality," of course — unintelligible content will remain so even when it is well formatted. Furthermore, reports produced in the workplace often need to satisfy specific constraints which may differ from some of the guidelines we suggest here. However there are some general principles that can assist with the process of writing a report.

- Our work needs to be well documented.
- We are subject to a finite time frame.
- Reports need to be concise but understandable.

Documentation We cannot emphasize enough the importance of making sure that all our work is well documented and that our findings reach the appropriate people in writing. When we write a series of reports on a particular project we are establishing a paper trail that documents our contribution to the project. Examples of the type of documentation involved in a project are provided in Chapter 4.

Time frame There is usually a finite time limit for which the documentation and reports are expected to be available in a presentable format. In order to accomplish our objectives efficiently, we should try to set up and follow realistic time frames for the completion of our reports.

Readability Every report we write is directed to a specific group of readers. The report needs to be understandable and should be written at the appropriate level for that group. The report should also be written in a reasonably concise manner. Information that is well-known to the reader may not need to be included in the text, or only

briefly stated as is necessary. The important points of the report, however, *must* be elaborated very carefully.

We do not pretend that our guidelines will teach the reader how to write in a general context, but they should provide some insight into the specifics involved in writing a report to our group, our superiors, or other readers of our work. In this spirit we introduce a very precise style of writing reports.

Project Outline

The project format that will be followed in this book has the structure shown below. In reality, we may need to make modifications to this structure but at least this provides the reader with the chance to learn a specific way of writing reports.

- Title page
- Introduction
- Results
- Conclusion
- References
- Appendices

Title Page

The title page is very important because what is written there reflects tremendously on the chances that our report will be read and understood. The information on the title page should give the reader a clear idea of the content and important points of our report. The structure of the title page is shown in the example in Figure 2.1. It contains:

- Project title
- Author(s)
- Date
- Executive summary

Project title The title, name(s), and date provide the citation information for the report. Try to choose an informative or "interesting"

title: *Final Project Report* tells the reader nothing! However, be careful when dealing with a subject where people's sensibilities may be offended. We comment further on this with regard to the PowerPoint title slide example shown in Figure 2.4.

Executive summary It must be very short and to the point. The executive summary contains a brief account of our conclusion. Do **not** describe the problem or discuss the type of methodology that was used; simply state the results and conclusion. For example,

> "The IBM stock price was at a higher level in October than in November. In October the mean price was ..."

The example of Figure 2.1 illustrates that the executive summary is short and to the point. Remember that the executive summary tell the reader what happened. It is **not** an "Abstract" which tells the reader what *will* happen. Abstracts precede articles published in journals. Consulting reports are specifically directed to our client.

Introduction

The purpose of the Introduction is to describe the project (insofar as we were involved), and to give the necessary background information. Try to be brief but do not leave out relevant information. Here, the basic descriptive statistics, graphs and summaries of the data can be included. If we have several graphs or large tables put them in an Appendix and refer to them in the text.

Results

State the points or hypotheses and prove them or disprove them. Go point by point showing how we performed the corresponding hypothesis test and how the results are to be interpreted. If we used statistical software that generates a sizable output file, extract the pertinent information and place it in the Appendix in table format. Do **not** simply dump entire output files into the Appendix. This practice is unprofessional on our part and it is not the reader's job to "find" which particular part of the output we are referring to in the text.

We should point out that certain report formats *do* require that entire output files be included. For example, the FDA requires that reports from the pharmaceutical industry also contain the source code used to generate the output file.

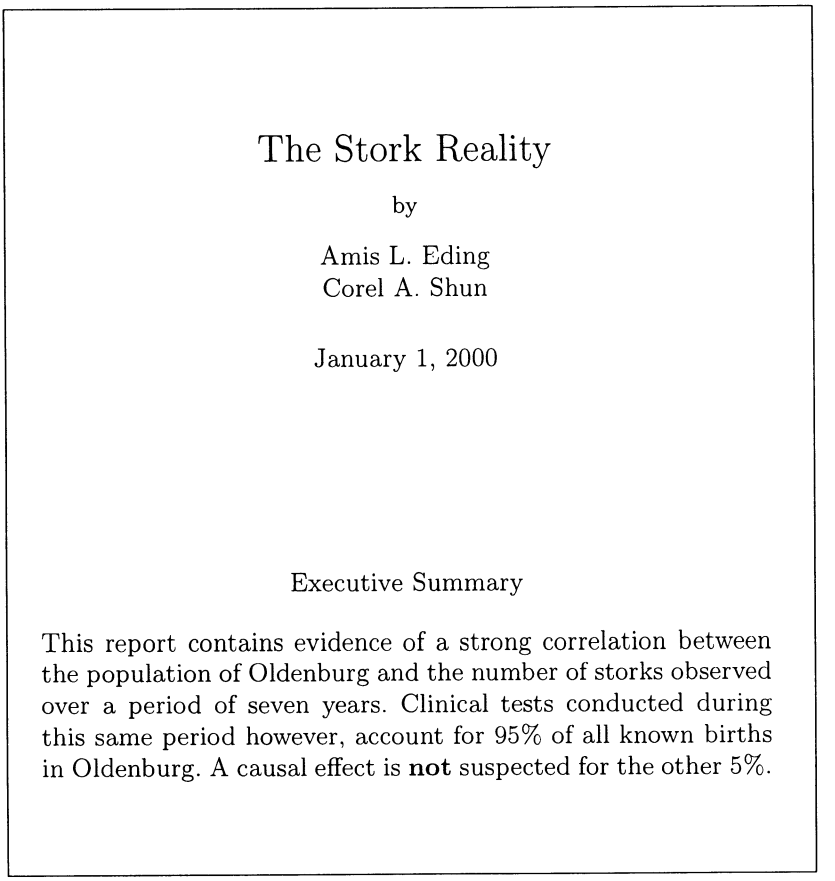


FIGURE 2.1. Title Page Example

Conclusion

The conclusion section should be an extended version of the executive summary. When writing the conclusion bear in mind that it should be in the form of a contextual interpretation of the results presented in the Results section. We should refrain from heavy reliance on excessive technical terminology. That is, the conclusions we make should be substantiated by the facts from the Results section and should explain what the results mean with respect to the goals and terminology of the project. In industry, for example, the language of real costs ($$$$) is extremely important in reports to senior management. Relating our conclusions to productivity and cost savings can provide a big boost to acceptance of the information.

References

The Reference section is the list of books, articles, software and other documents that were cited in the text. Do not include irrelevant references that are not specifically related to the project. Certain references such as legal documents or technical reports, would also be included in an Appendix if they have particular relevance to the project. If we do not have any references to cite, then omit the Reference section completely; we are not required to have it. The general rule for a citation is to write the author's names (last name, then initial of first name) followed by the year of publication, title of the work, and then the general publication information as appropriate: journal, volume, pages, publisher, etc. The bibliography in this book provides examples of many standard citation formats.

Appendix

The Appendix should be used for computer output, computer code, graphs, tables, or any other reference documentation that need not appear in the text. Again, do **not** use the Appendix as a dumping ground; it needs to be presented in an organized and appropriate format for the reader. Try to adopt a suitable indexing system within the Appendix that can be used as a reference for any place in the output we quoted in the text of the report. This could involve: underlining and numbering the relevant places in the output and referring to those numbers in the text, or employing designations such as Table 2B and Figure 3.4. Highlight the text or numbers using a light color marker (translucent yellow is a preferred color) to make them easy to find. If the amount of documentation is extensive, the Appendix can be broken into several sections. For example, graphs could be kept in one section, tables in another, letters from a relevant law case in another, and so on. If several analyses were performed, it may be useful to keep the results in separate appendices. Preceding the appendices with their own table of contents (in addition to the one for the entire report) is helpful.

2.4 Basic Guidelines for Writing

The purpose of the previous section was to provide a structure for writing reports. Now we need to fill in the structure of our report with actual details. A word of caution — the time needed to "fill in" the details can be deceptive. A report is not complete until *all* the parts have been written. While this would seem to be stating the obvious, remember that our submitted report can only be assessed on the basis of what *has* been written. The reader is not responsible for guessing what we meant to write, mistakenly omitted, or intended to "fix later." The style and quality of our writing also deserve attention and we consider some general guidelines that may be helpful in this regard. In reality, good writing takes effort and experience; guidelines are not a substitute for this learning process.

Of the numerous references on writing we mention Gopen and Swan (1990) which is posted on the the Jounal of Computational and Graphical Statistics (JCGS) section of the ASA Website: `www.amstat.org`

Some Basics

Spelling Typographical errors are distracting. Our favorite spell-checker can be helpful, but it is not infallible. A dictionary can help us check the appropriate usage of a particular word.

Fluency Our sentences must make sense. Reading the sentence aloud can often help identify problems with syntax and construction.

Paragraphs All the sentences in a paragraph should be related to a main point. Digressions tend to confuse the focus of the reader and this can result in our point being misunderstood.

Revisions Whether we continuously revise as we write, or wait until a complete draft is finished before revising, no final report should be submitted without performing a careful check. A good strategy is to have somebody else read our draft.

Example 2.2 *Outlier*

Statistical terms often have a strict technical meaning which will not be recognized by a spell-checker. Two common examples are "estimator" and "outlier" which certain spell-checkers may try to replace with estimate and *outliner*. Unfortunately, the latter term has appeared so often that it would seem that we have a new term for "Outlier" in statistics!

Example 2.3 *Their, there*

No, this is not an expression of sympathy! The first form, *their,* is used **only** as an adjective to indicate possession by a person or a group, "It

was their idea." We should note that *their* does not distinguish between a singular or plural identity: is this one person's idea or a group's idea? The writer needs to establish this. It follows that *there* should **not** be used to indicate possession. Incorrect usage of these two words will be very obvious to the reader.

Style

An important part of our preparation is knowing "who" will read our report. The client who initially brought us the project may well need to share our report with other people. Thus, our style of writing, should be commensurate with the overall level of statistical knowledge of the anticipated readership. In addition, the quality and style of our writing should be considered. It is not an impossible task to convey technical information across disciplines, it just requires more attention and care on our part as to what we write. The following example exaggerates a style of writing which we should **never** use!

> Based on the `Type III SS`, for the **LS** fitted **regression** model `MLR – IIb`, (Appendix II, **Part C**), which included the "X variables" **AGE88** AND **HGT-1**, `HGT-1` is **not** *significant* (NB: *P=0.2345*). This means **HGT-1**, **given AGE88**, *does NOT contribute in the* 'PREDICTION' of the **response**, `INC` (Y).

Emphasis Important words or phrases can be made to stand out by *changing* **fonts**, or using CAPITALS. However, if used too often (or inconsistently), the effectiveness is rapidly lost and our report becomes "tiring" to read. Similarly, excessive use of **acronyms** is unnecessary. The reader should not have to decode our report in order to read it!

Clarity The clarity of a sentence is more important than its length. Each sentence should be trying to convey a self-contained "packet" of information to the reader. Short sentences are fine, but always using short simple sentences will make our writing seem "choppy" and the reader feels like the flow of our narrative is being constantly interrupted. Similarly, long rambling sentences which keep adding divergent bits of information will lose the reader. They give up accumulating the information before the sentence is finished, make an interpretation at this point, then skip to the end.

Paragraphs A paragraph consists of a series of sentences which should all be related to a central theme. Although the structure of the paragraph will depend on the theme, there should be a natural flow of information between sentences. That is, we need to present the theme of the paragraph in an organized manner. In report writing, two types of paragraph structures are **Thesis** and **Linked** (see Example 2.4).

Technical Details Most reports will need to contain technical details concerning the statistical procedures we employed. If we are unsure of the readers' level of knowledge, keep the technical details of the statistics to a minimum. Citing an appropriate reference to a statistical procedure would be more appropriate than writing an entire textbook on the procedure!

Example 2.4 *Paragraph styles*

Thesis: Say the thesis $\Rightarrow$ Prove it $\Rightarrow$ Say it again.

This structure is a good way to write the Executive Summary. It is also useful for describing the results obtained from a particular type of statistical procedure, or for stating our conclusions based on the regression analysis we performed. Try not to add too much related information in a single paragraph. Rather than overwhelm the reader all at once, break the original theme into cohesive subthemes.

Linked: Linking sentences:

[Topic A] $\Rightarrow$ [New Info A]
[Topic B = New Info A] $\Rightarrow$ [New Info B] $\cdots$

By "$\Rightarrow$" we mean that the words in each sentence connect an introductory "topic" with a "new" piece of information. This new information becomes the introductory topic in the following sentence which then connects to the next "new" piece of information, and so on. This type of structure can be useful for the initial paragraphs of the Introduction section of our report where we need to describe the details of the project: aims, variables, sample size, etc. Thus, this structure is useful whenever we need to describe a *sequence* of related conditions, or the *progressive* steps of an investigation.

For example, the diagnostics associated with a regression fit could be linked as: Scatter plot $\rightarrow$ residuals. Residual-versus-fitted plot $\rightarrow$ normality. Normality tests $\rightarrow$ significance results. Nonsignificance $\rightarrow$ conclusions.

Example 2.5 *The apostrophe*

The apostrophe ' is used to indicate possession — *The client's data* $\cdots$, and for contractions — *won't, can't,* $\cdots$.

Although contractions can be avoided by using the expanded form (*Don't* $\longrightarrow$ *Do not*), this tends to make our writing cumbersome since the style appears unnecessarily formal. This formality imposes a certain emphasis which can detract from the point we are trying to convey to the reader. "Don't forget to plot the data" reminds the reader to "plot the data" whereas, "Do not forget to plot the data" admonishes the reader for "forgetting."

Example 2.6 *However, thus, hence*

Try to avoid writing the report in the style of a mathematical proof. Not *every* sentence needs to start with one of these qualifiers. Although we may be accustomed to their frequent use in technical expositions, this should not be necessary in the report. Eliminating the needless use of these qualifiers may take a conscious effort on our part, but simply stating a result is often sufficient.

English as a Second Language (ESL)

For writers whose native language is not English, performing the checks listed above may not be easy. Based on our experience, students in this situation can find the correct use of articles and punctuation to be particularly troublesome. If this situation applies to you, obtain a reference handbook on writing; preferably one that includes ESL advice. The handbook, *A Writer's Reference,* by Diana Hacker (1998) includes specific ESL advice and would be a useful reference in general. Having a native speaker of English read your drafts would also be helpful, provided you treat this as a learning experience and not as a way to avoid the problem. Of course, you still bear sole responsibility for quality and content of the final report. Finally, the benefits of enrolling in a writing course, whether ESL or not, should not be overlooked.

2.5 How to Make Effective Presentations

The art of communicating with an audience is not just a natural gift as some may pretend. There are people who are gifted with verbal communication, but for the rest of us it is a matter of preparation. We may not be able to emulate the great orators, but anybody who prepares properly will be able to deliver an effective presentation.

The type of presentation we prepare will obviously depend on our audience. If we are just presenting the results of our analysis to one or two people (our client), then a complex multimedia show would be completely inappropriate in most situations. (The case where the client happens to be the CEO of a large corporation is clearly not "most" situations.) While some of issues we present below do apply to one-on-one presentations, the emphasis in this section is on formal presentations. That is, the situation where we need to provide our client with a well prepared presentation of the "product" (the data analysis) that they paid for. Small business consulting companies need to do formal presentations all the time, and we should expect to do the same with any large consulting project.

In order to deliver an effective presentation we first need to consider the various aspects that are involved in nonverbal language and voice quality.

- Nonverbal language.
- Voice and speech.
- Preparing for the presentation.

Nonverbal Language

Nonverbal language has an important role in a presentation. Our body language can be very expressive, both from a positive and negative perspective. Some things to try to avoid are to appear too nervous, to move around too much, or to face away from the audience and look only at the screen. Obviously our attire and physical appearance should match the circumstances in which the presentation takes place. Equally important, our manner should invoke enthusiasm for the subject matter. Nonverbal language conveys a lot about ourselves and the quality of our work.

Voice Quality

Voice and speech intonation can enhance the delivery of a presentation. We have to learn to project our voice into and above the audience without shouting. Practice can help us with this. If our voice is too soft then we need to practise speaking louder and higher up in the air. Intonation is also very important; there is nothing more frustrating than trying to listen to a speaker droning on in the same tone. By varying the intonation of our voice we can stress specific components of our presentation so as to add emphasis to the important points. The text by Dunckel and Parnham (1993) may be useful here.

Preparing for the Presentation

Once we have finished the analysis and gather all the information that has or will appear in the report, we should prepare for a possible presentation. The first step is to try to establish the length of the presentation and to prepare the structure accordingly. It is common for inexperienced presenters to overextend themselves by trying to say too much or having far too many slides. Try to be reasonably conservative in terms of the material and the timing. Saying "less" is often much more effective than trying to include everything.

- Structure and timing.
- Multimedia, PowerPoint, overheads.
- Slide preparation.
- Objectivity.
- Handouts.
- Practice the presentation.
- Be prepared for possible questions.
- Get advice from a collegue.
- Time the presentation!

Structure

The structure of the presentation can be similar to the report structure. Start with a summary of what we are going to cover and then follow with a general introduction to the problem under consideration. Then continue with a more detail results description and finish with the conclusion. We do not have to include as much material as in the report and we also have a chance to be more creative. Sometimes it may be useful to make use of the introduction to motivate the audience if they are unfamiliar with certain aspects of our discourse.

Choice of Media

Another component to consider at this stage is the medium that we intend to use. This can vary from simple black and white transparencies to the most complex multimedia show. The preparation time is proportional to the complexity of the medium, but if it is done well a fancy multimedia presentation can be extremely effective. On the other hand the medium is not a substitute for content and it should not overshadow the substance of the presentation.

One way to enhance our presentation is to use color. A typical middle ground presentation requires the use of color and enough graphical sophistication that makes necessary the usage of a presentation software such as PowerPoint.

Slide Preparation

Next it is time to prepare our slides. This needs to done carefully. It is important to emphasize the content and not so much the form. Do not abuse

special effects and "junk-chart." The fact that those tools are available in the presentation software does not mean we have to use them *all* on every slide. Some other points to keep in mind when preparing the slides are:

1. Tables should be kept to a minimum. The use of charts and graphs is far more effective for conveying information.
2. Annotations should be kept to a minimum. These can be visually distractive.
3. Color choice is important. Never use yellow (it will be invisible to the audience) and avoid certain combinations like red on dark blue.
4. Lastly, make sure that we have the right number of slides; not too many, not too few.

If our presentation is very complex or requires careful comparisons of pieces of information (for example, tables or charts) we should consider preparing handouts as well. If we need to reinforce certain points then use slides and an additional projector.

Objectivity

Always remember that as a statistician and as a scientist, we are obliged to provide an objective and honest presentation of the information. Many of the tools that are available for presentation can be used to bias the evidence towards a certain conclusion which is not warranted. More important, we should **not** be afraid to say that we are unsure or uncertain about a result. Let the audience see the "problem." In our experience, we have often found the audience to be the best problem solvers!

Practice

Once the details of the slides have been completed we should practise the presentation in a realistic setting. Time the presentation! Presentations that end up going overtime lose audience interest rapidly. Be prepared to answer questions *during* the presentation. This time will need to be incorporated in the total time we plan to use for the presentation. Slide flexibility is a useful strategy here as we won't know how many, or what type of questions we may be asked until *after* the presentation. This just means that we should have certain slides that can be omitted entirely or discussed in varying degrees of detail depending on whether we need to "speed up" or "slow down" the presentation. This does take practice, but avoids the visual appearance of a "rushed" presentation:

- No questions are asked and we finish in half the time allotted to our presentation! The audience is left with the impression our presentation lacked content.

- Many questions are asked and we have to flip to the end of the presentation. Our timing is perceived to be poor and the audience is wondering what material they missed seeing.

Questions raised during the presentation are often short and quick, requiring simple clarification of a term, concept or entry on a graph or table. At the end of the presentation however, we should be prepared to answer more probing questions. We should make a list of questions that we can expect and prepare the answers beforehand. Our answers should be directed to the full audience (not just the person who asked the question), and need to be pitched at a level which everyone can understand. Finally, it is well worthwhile rehearsing the presentation in front of others who can help us correct obvious errors and polish the details of the presentation.

2.6 The Importance of Quality Graphics

Aesthetics:	Efficient use of the plot region.
Annotation:	Labels, legends, title and subtitles.
Contrasts:	Grey scale, color, lines and symbols.
Comprehension:	Conveying information graphically.

Many statistical procedures rely on graphical diagnostics and data visualization techniques for analysis purposes. Scatter plots, histograms, and bar charts[4] are examples of simple but very effective graphical displays that can be employed to communicate the results of our consulting efforts. In this section, we focus on some of the important principles of graphical design that are involved in creating *Quality* graphics for presentation purposes. These principles are indicated in the box above.

The first thing to note is that producing presentation quality graphics takes time and effort. Even with a good statistical software package such as S-PLUS, the "default" display will often need to be modified for presentation purposes. One advantage of S-PLUS is that it allows the user to exert precise control over the components which make up a graphical display.

The Plot Region

The term *figure* is used to describe a complete graphical display which is composed of a *plot region* surrounded by four *margins* (top, base, left

[4]Bar graph and bar plot are synonymous terms for bar graph.

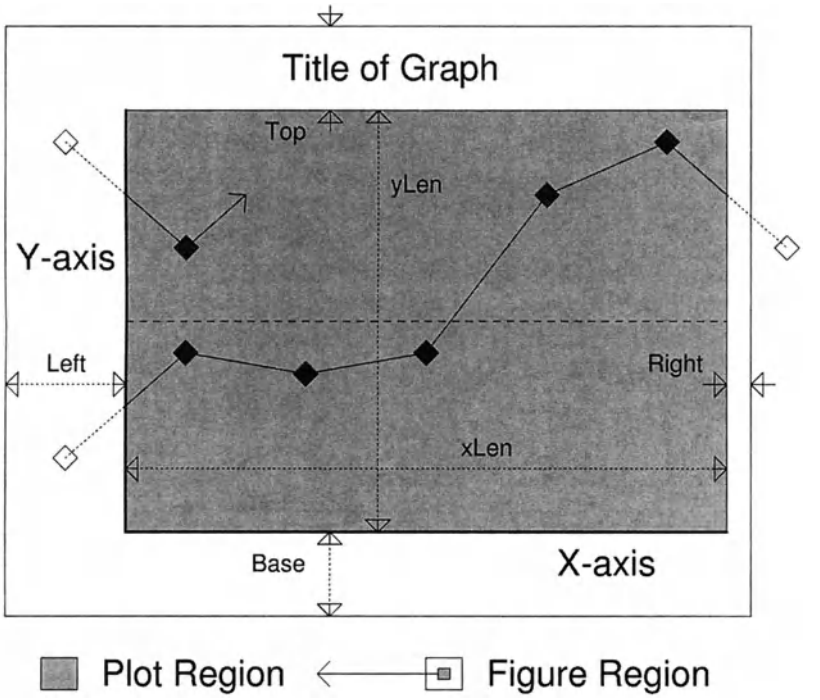


FIGURE 2.2. Dimension sketch of the Plot Region and Margins.

and right) as depicted in the time series schematic plot of Figure 2.2. The margins are used for titles, axis labels, tick marks and labels, and other annotations. The right margin usually contains no text and could be minimized to increase the plot region area. Where practical, legends can also be placed strategically inside the plot region (see Figure 2.3) to avoid reducing the area available for the graphical display.

Graphical Design

The example in Figure 2.3 illustrates the use of histograms to display the results of a two-sample t-test based on a study to compare the effect of incorporating learning style preferences (Experimental Group) versus traditional teaching methods.[5] We use this example to discuss the graphical design principles above.

Aesthetics The purpose of a figure is to convey information *visually*. The aim of enhancing the aesthetic quality of a figure — making it look "good" — is to focus attention on the displayed graphic. This

[5]Details of this study are presented in Chapter 4.

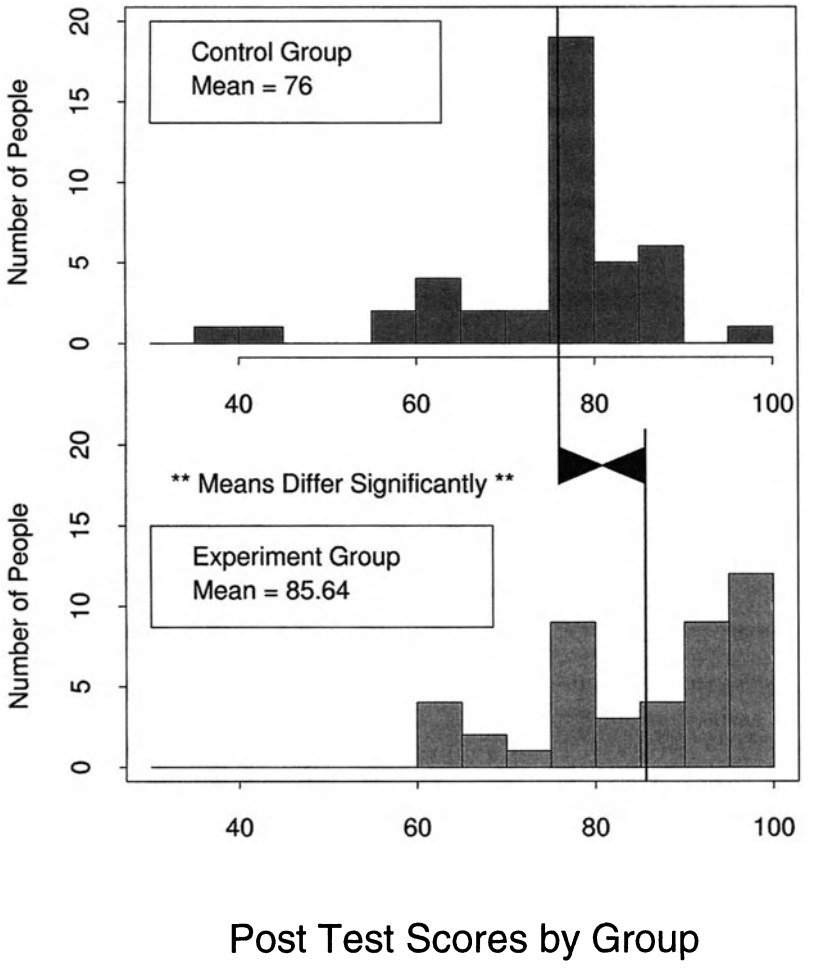


FIGURE 2.3. An example of comparative analysis using Histograms

involves making efficient use of the plot region and margin spaces which often requires a trial-and-error approach. Some general points to keep in mind are:

1. The displayed graphic should span the vertical and horizontal dimensions of the plot region.
2. Sufficient space must be provided in the margins for explanatory text in *larger* font sizes. The default font size employed for axis labels is often too small, making it impossible for the audience to read in a presentation setting.
3. Reducing the width of the right margin space (see Figure 2.3) has a side-effect of making the graph appear *off-center*. Compromises are a necessary part of improving a graph.
4. The title and axes labels are necessary text. Additional text will have the effect of adding visual clutter to the display and should therefore be used sparingly.
5. Color or grey scales, line types, point symbols, font sizes, and legend placement need to be experimented with. Visual contrasts are very important in helping the viewer distinguish groupings and other effects.
6. It is not necessary to change every default setting provided by the software package. If the default setting is acceptable, *don't* change it!

Annotation The margins are used for titles, axis labels, tick marks and labels, and other annotations. In Figure 2.3, the title also served as the X-axis label.

1. Keep axes labels simple (but informative), and use "nice" tick marks. For quantitative scales, the *unit* of measurement should also be included in the axis label.
2. Legends add to the visual clutter so choose the legend information carefully. In Figure 2.3, the histograms provided a visual proxy for the group spreads and sample sizes so only the mean of each group was included.
3. Other annotations such as the text indicating the significant result in Figure 2.3 may be helpful or necessary. For example, the baseline used to display relative percentages needs to be indicated in order for the viewer to make a meaningful assessment of the results.

Contrasts If we decide to use color to differentiate lines or points in the presentation **don't use yellow** — they may not be visible.[6]

1. Color provides the viewer with wider range of contrasts and can also add extra dimensionality to the graphical display. In Newton (1993) for example, color-encoding was employed as an aggregation technique for displaying very long time series. It also introduces more complexity in terms of interpretation and aesthetics. (We refer the interested reader to the extensive work by Carr (1998) on data visualization methods using color.)
2. The choice of colormap is important when translating color graphics to greyscale devices. Black-and-white photocopiers can often render a noninformative version of a color graph. Obtaining a reasonable greyscale for presentation graphics can be surprisingly difficult. In Figure 2.3, the top histogram "color" is still too dark.
3. The default size provided by many software packages for points in a scatter plot is usually inadequate for presentation purposes. These need to be replaced by larger symbols such as the opaque diamonds in Figure 2.2.
4. The number of contrasting lines that can be employed effectively in greyscale displays is quite limited. Simple dashed and dotted line types are better reserved for internal horizontal or vertical median lines; unbroken curved lines marked with different symbols are easier to differentiate than mixtures of dot-dash line types. The line width (thickness) is sometimes increased for the axes lines or for emphasis as in Figure 2.3, where the groups means are connected by a slighter thicker line.

Comprehension The most important component of graphical design is making sure the display conveys comprehensible information. For example, the time series schematic in Figure 2.2 is actually showing how the technique of "wrapping" a time series on itself is performed — the two points which connect across the right edge of the plot region are redrawn in the same position relative to the left edge of the plot region. This data visualization technique was employed by McDougall and Cook (1994a,b) in an interactive dynamic graphics environment. (Interactive graphics are discussed further in Chapter 3.) Although the concept of this real-time animation technique is straightforward, Figure 2.2 clearly required additional explanation in order for this transfer of information to be completed.

---

[6] One of the authors had the dubious distinction of providing their doctoral advisor with a transparency in which three time series plots were overlaid. Only two showed up at the presentation.

Conveying information through presentation quality graphics is not an easy task, but it is a skill that a statistical consultant needs to develop. Cleveland (1985, 1993) and the series by Tufte (1983, 1997, 1999) are two sets of references worth reading on the subject of statistical graphics and scientific graph construction in general. We close this section with some remarks based on our experiences:

1. Be aware of the audience — simple graphical displays, such as a histogram, can be just as effective as complex ones.
2. The ability of a statistical consultant to detect deceptive or misleading graphics is also important. Clients, unintentionally, sometimes produce graphical displays which are misleading.
3. Persuade the client to use a bar chart rather than a pie chart. It is more effective for comparison purposes which is the point of the display.
4. The points in a time series plot should always be connected. Trends, seasonality, outliers and other features of interest can be detected much more easily when the points are connected.
5. The plot region usually conforms to the rectangular shape of a sheet of paper in either portrait or landscape mode. As a result, the absolute scale of the X and Y axes in a scatter plot is distorted which can be misleading in some situations. For example, a circle would appear as an ellipse. A *square* plot region is a simple way to match absolute scales.
6. Complex graphical diagnostics require experience to interpret. Correlograms, Q–Q plots, and mosaic plots are examples of informative diagnostics that many clients would find difficult to interpret since they lack the prerequisite level of statistical knowledge and experience needed.

Using PowerPoint for Presentations

PowerPoint is an excellent tool for presentations and perhaps more importantly, provides "templates" which makes it easy to generate a good basic presentation. Of course, other presentation software can be used for this purpose and our choice of PowerPoint is mainly motivated by its ease-of-use at the default level and wide availability. PowerPoint will display a sequence of slides with the presenter advancing each slide in accordance with the pace of the presentation. We can also add special effects such as automatic timing, sound, and animation. Special effects can certainly en-

hance and help add vitality to the presentation,[7] but they can also be more of a nuisance factor and divert the attention from the important points. As in many situations, moderation usually wins out.

Preparing for our PowerPoint presentation follows the same steps as discussed in the previous section:

- Prepare the slides carefully — not too many, not too few.
- Avoid using too much "junk-chart" material.
- Provide handouts or use additional slides on an overhead projector if we need to reinforce certain points in the presentation.
- Above all, rehearse! Time the presentation under the type of conditions in which the presentation will take place.

The key difference is that this is an *electronic* presentation which means computer hardware, cables and software will be required. The most common setup is where our laptop computer is connected to an electronic display unit which sits on top of an overhead projector. Prior to the presentation, we need to do the following:

1. Find out what equipment we need to provide for the presentation.
2. Make sure the laptop is compatible with the electronic display unit and has the right cables for connecting.
3. Learn how to operate the display unit with the laptop. Certain control settings such as `Video Mirroring` may need to be turned on.
4. Know how to use the laptop! Spending several minutes trying to find PowerPoint or our presentation files would not be a good start.
5. Always, always, have a backup copy of the presentation. Computers do crash and can erase our carefully prepared PowerPoint slides.

The following instructions will enable us to produce a basic presentation. To enhance the presentation we can check out the many options available in PowerPoint and the Wizard tool. Examples of slides created by PowerPoint are shown in Figures 2.4 and 2.5. Before continuing, we again emphasize the following important point:

Presentation software is **not** *a substitute for content.*

---

[7] Presentations which are too technical may put people to sleep. Adding sound or animation would help maintain the interest of the audience even though the material may be quite technical.

Is the title, "A Cure for Cancer," used in Figure 2.4 appropriate? For people who have undergone treatment for cancer, this title is probably offensive and certainly disrespectful. Do **not** try to be too cute or provocative when dealing with sensitive issues. Our intention may have been innocent, but by then it is already too late: that slide has already reflected on our lack of respect for the sensibilities of the audience.

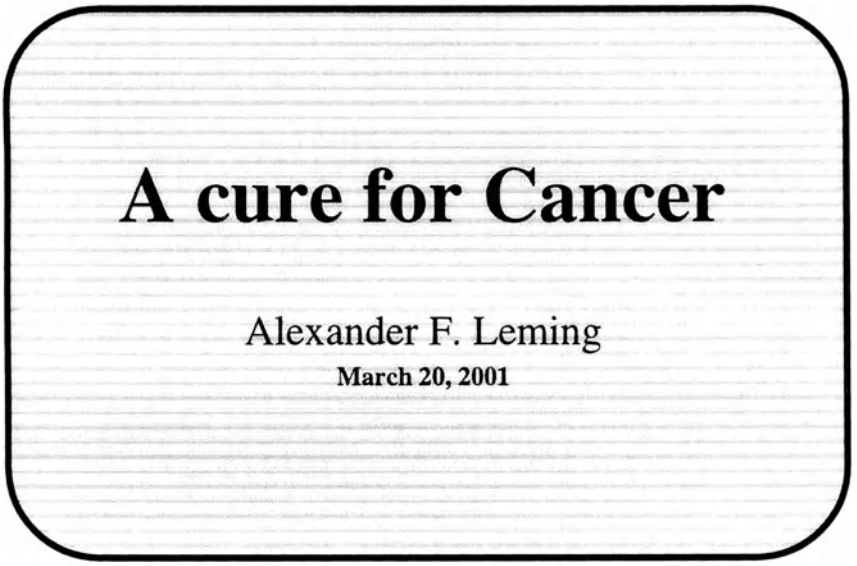


FIGURE 2.4. Example of a Title Slide using PowerPoint

`Click on PowerPoint` This will take you into a menu where you may choose to run the `Autocontent Wizard` or to select a `Template`. We recommend that you select a simple template. In the slide examples of Figures 2.4 and 2.5, we have used the `Blank Template`. We added a light background color from the color menu obtained by selecting the `Background` item from the `Format` menu.

`Title Slide` PowerPoint will bring the title slide to the screen and you can now fill in the presentation title, our name and date. Figure 2.4 shows a simple example.

`New Slide` Next you click in the `New Slide` of the `Common Tasks` window. This opens a window showing the slide `autolayouts`. Choose the one that best suits the intended purpose of the slide (e.g., text only, text with graphics, etc). If you want a different layout you can always

change it later. You can now enter the text of the second slide, adding graphics or pictures if applicable. Next you repeat the process until you finish all our slides.

`Pictures` You can insert a picture on a PowerPoint slide by using the `Picture + From file` item on the `Insert` menu. Figure 2.5 shows an example of this. PowerPoint will accept a variety of graphic format files including GIF, JPEG, EPS and PNG.

`Statistical Graphics` To incorporate a picture from S-PLUS-PC save it as a GIF file. In S-PLUS running under UNIX, use the PostScript graphics driver with the option `onefile=F`, so it will save each picture on a different file in the EPS format. In SAS you can select the export item on the file menu and select the GIF format. Any of these picture files formats can be read by PowerPoint.

`Reviewing` To play the presentation select `View Show` from the `Slide Show` menu. To go to the next slide press return.

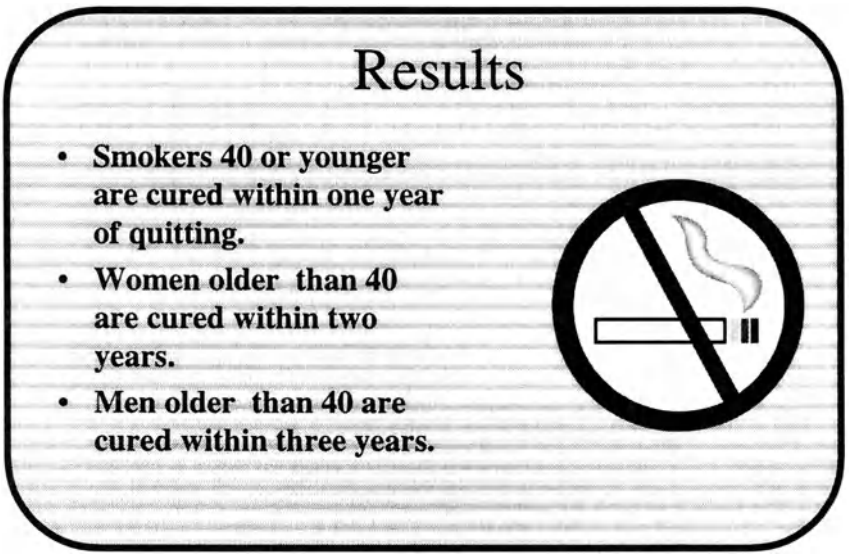


FIGURE 2.5. Example of a Results Slide using PowerPoint

An alternative method to generate slides for presentations is to create HTML files. An easy way to produce HTML file presentations is to use the Netscape Composer which is a Web page editor that is easy to use and

comes with later versions of Netscape. Netscape is available free of charge for most nonprofit organizations and many businesses have site licenses. Netscape Composer works more or less like an editor. We can change font type, face, size, and color and we can add pictures to the text in several forms. We can also add hyperlinks to other Web pages both on our local disk or on the Internet. Microsoft Office, which includes PowerPoint, Word, and Excel are predominant in many business environments and these applications also support HTML format for presentations.

Whether we use PowerPoint or a Web Composer, the presentation will require a reasonable amount of time to prepare and to rehearse. This is **not** a five-minute job.

3
Methodological Aspects

The aim of this chapter is to provide the reader with an overview of some standard techniques that are commonly used in statistical consulting. In keeping with spirit of this book the emphasis of our presentation is on the client's perspective rather than the technical details of the methods. Appendix C does provides some technical details and tables, but a more comprehensive exposition of these methods may be found elsewhere. We provide references to the methods we discuss and an overview of statistical software is provided at the end of this chapter.

Of course, before we can apply any statistical procedure we need some data. Hence, we begin by looking at methods for collecting data.

3.1 Data Collection

There are several methods that can be used to collect data, but in order to draw valid conclusions it is important to collect "good" data. Unfortunately, many clients seek the advice of a statistical consultant *after* the data have already been collected. This is largely our fault, of course. As statisticians, we are often guilty of focusing on technique and not emphasizing the important role that a statistical consultant can have during the design stage of a statistical investigation. Having said that, a client is clearly entitled to know what *are* the advantages of getting a consultant's advice during the planning phase of a project. Chatfield (1995) lists three compo-

nents of a planned study where the advice of a statistical consultant can be particularly useful:

- Identify the important variables.
- Clarify the objectives of the study.
- Formulate the statistical problem.

The objectives of a project need to be well formulated in order for the conclusions to have a meaningful interpretation. This requires specifying the "right" objectives for a client's project which is not always easy to do. The consultant needs to be careful not to end up giving the client the right answer to the wrong question. This is sometimes referred to as a "Type III Error" or an *error of the third kind* (Kimball 1957). We have addressed these issues in the previous chapter in terms of verbally interacting with the client during a consultation session and now turn our attention to the data collection process itself.

Data Collection Methods

The data collection process and objectives of a study may not always fall into the traditional distinction of a designed experiment versus an observational study. While it is certainly helpful to distinguish between the basic concepts associated with these data collection methods, there is a tendency for statisticians to be somewhat skeptical of conclusions based on observational studies. This is not without reason. Treatment effects in a well designed experiment have a clear interpretation whereas observational studies often involve subjective responses that can be difficult to quantify and interpret. Hence, interesting effects that emerge from an observational study may not necessarily be real.

We briefly consider the following types of data collection methods that a statistical consultant would commonly encounter.

- Observational Studies
- Sample Surveys
- Longitudinal Studies
- Clinical Trials
- Designed Experiments

Observational Studies

In an observational study, the investigator generally takes a passive role and simply observes a certain cohort of subjects and collects the responses. While the word "passive" may not be the first adjective that comes to mind when we have been confronted by yet another telemarketer, the solicitation process is effectively the result of an observational study: a profile analysis based on an existing database consisting of information related to our spending habits, demographics, Internet usage, and so on. The objective of the profile analysis is to identify a large source of potential respondents since even a small percentage of positive outcomes can be profitable. See Cochran (1983) for further details on the planning and analysis of observational studies.

Protocol

As the telemarketing example suggests, the data collection process and analysis involved in observational studies can be quite extensive. In larger studies, it is often necessary to have a written *protocol.* This is because most (if not all) of the data collection will be administered by people with no statistical training. In a sample survey, for example, the "interviewer" needs to know what to do when a designated respondent is unavailable — is the interviewer required to call back later? If so, when and how many times? The written protocol must therefore provide detailed instructions on every aspect of the data collection process. The purpose of a protocol is to try to control nonrandom errors.

Sample Surveys

Sample surveys and polls have become increasingly common and are used extensively in market research. Although the concept and importance of selecting a random sample is generally well understood, the response rate from many surveys can be very poor. In some cases, the results may also be of low quality. A good design format is therefore important to help maximize the response rate and improve the quality of the data collected. Well designed sample surveys provide a cost-effective method for estimating population characteristics accurately. Here, we summarize some of the main areas of concern in sample surveys, beginning with the following reminder:

Accuracy is **not** *guaranteed from an individual sample.*

That is, no matter how well designed and implemented a sample survey is, we are always subject to the profile represented by the (individual) sample we collect. Cochran (1977) is a classic text on the statistical theory associated with sampling methods. Dillman (1978), Thompson (1992), and Fink (1995) are other texts that deal with survey sampling methods.

Design Format The design format of a survey is very important since it can have a large impact on the quality of data collected. Questions need to be simple and clear to avoid interpretation problems. Make it easy for the respondent to answer a question, or not answer. Non-responses need to be distinguishable from a category such as "Not Applicable." Perhaps more important, do **not** try to ask too much.

Sampling Method The purpose of a sample is to infer something about a population and a random sample is necessary to ensure the statistical validity of the results. Simple random sampling (SRS) can be shown to be an optimal sampling strategy (Cochran 1977), but this is usually impractical to apply directly. The method of sampling needs to be consistent with the desired population profile, but must also satisfy practical overheads such as time and cost. Some standard sampling schemes are indicated below.

SRS Simple random sampling is equivalent to drawing names out of a hat without replacement. That is, every *unit* in the designated population has an equal chance of being selected. In most sample survey schemes, SRS is primarily used within subgroups identified in the population.

Stratification To avoid oversampling groups (strata) that dominate a population, stratified random sampling may be used based on the relative proportions of the strata present in the population. The population strata proportions need to be known in advance (or estimated from a pilot survey) and SRS would be used to select a sample within each stratum.

Cluster Randomly sampling across an entire population or stratum is often impractical. Instead, homogeneous clusters such as towns are first identified and a random sample (using SRS) of these clusters is selected to generate the final sample.

Multistage With large and diverse populations, several stages of clustering and stratification may be required to generate a cost-effective sample.

Multiphase Pilot studies, followups, and double sampling schemes can be used to optimize or adjust the overall sample size. Double sampling is often employed in situations where the quantity of interest (population parameter) need only be estimated to a certain precision.

Sample Size The purpose of a sample is to estimate some parameter of interest associated with the population. The precision of the sample estimate (statistic) depends on sample size through its sample variance. In practice, real costs are involved in collecting survey data

and these increase with sample size. A trade-off between cost and precision is usually required.

Implementation The key to obtaining a valid sample is random selection and to avoid selection bias, random numbers are used. Biased results arise when the sample profile differs from the desired population profile. Nonresponse and undercoverage are two common sources of bias that can be partially reduced using followup surveys. However, even the phrasing of a question can have a substantial influence on the response — leading questions are frequently used to help promote political positions and market products.

Longitudinal Studies

A more specialized type of observational study involves the collection of longitudinal data. This is where each unit in the sample cohort is measured on several occasions over a period of time. The texts by Plewis (1985) and Selvin (1996) deal with the statistical analysis of longitudinal data. The purpose of this type of study is to assess the effect of an explanatory variable X on a response variable Y under controlled conditions. A typical application would be:

Example 3.1 *What are the side-effects of a particular drug?*

The side-effects associated with a drug may only develop with long-term usage and may also depend on the dosage amount. In this case, a carefully selected group of individuals would be followed over a period of time and the effect of dosage (X) on various side-effects (Ys) would be observed.

If the X and Y records are obtained from historical data, then the longitudinal study is said to be *retrospective*. Retrospective studies clearly depend on the quality of the historical data, but have the advantage that the problem of interest can be immediately analyzed. When the Y is to be generated by the study, it is said to be *prospective*. Examples are presented below.

Retrospective The response variable Y is known from historical data and the effect of the explanatory variable X is of interest: 50 smokers and 50 nonsmokers are retrospectively classified by age at death.

Prospective The levels of the explanatory variable X are fixed at the start of the study and the future values of the response variable Y are of interest: 50 smokers and 50 nonsmokers are selected for a longitudinal study of the incidence of lung cancer.

Clinical Trials

Clinical trials were discussed in Chapter 1 in terms of the development and testing of drugs by the pharmaceutical industry. The full-scale testing conducted in Phase III studies is usually what is meant when a client or statistician refers to a clinical trial. Fleiss (1986), Pocock (1983), Meinert (1986), and Friedman et al. (1985) are texts that deal specifically with clinical trials. The key components of a clinical trial design are:

- Control group
- Written protocol
- Randomization

Control The control group consists of patients who receive the standard treatment or a *placebo*. A placebo is a "fake" version of the drug that is administered to patients in the treatment group. It contains purely inert substances and provides the investigator with a measure of the so-called "placebo effect" associated with the psychological influence of taking medicine. There may be several treatment groups employed to assess the efficacy of the real drug at different dosages.

Protocol Phase III clinical trials can typically involve a thousand or more patients spread out across many testing sites. In some studies, these sites may even be in different countries. The administration of clinical trials is therefore quite complex and, as we indicated above, the data collection process will be performed mostly by nurses, doctors, and other personnel who do **not** have any formal statistical training. In this situation, a written protocol is essential since deviations from the protocol are almost certain to occur. For example, a patient may need to be taken off the drug temporarily due to food poisoning or a patient may not follow the proper regimen for taking the drug. Similarly, deviations may also arise due to administrative personnel not following standard procedures correctly.

The written protocol for a clinical trial documents all information related to the study. This would include items such as

- The purpose of the study and the eligibility requirements for a patient to be included in the study.
- The sample size and method of assigning patients to the control and treatment groups.
- The treatment schedule and procedures for evaluating the patient's response. The training required to use certain evaluation instruments.

- Decision procedures to deal with protocol deviations.

The raw information from the study is usually collected on a form that is filled out by the person evaluating the patient's response. Some forms may contain several pages of information which all needs to be entered into an electronic database. A good design format is desirable to help ensure the quality of data obtained during the data collection process.

Randomization The randomization of the treatment levels to patients is particularly important from a statistical perspective since this is the only objective method to counter potential sources of bias. However, randomization by itself is not enough to remove nonrandom sources of bias such as the doctor who "knows" what treatment a patient is receiving. The doctor may provide excellent care to all the patients in the study, but clearly won't be looking for any treatment effect from a patient receiving the placebo. A *double-blind* implementation where both the patient and doctor do not know which treatment the patient is receiving can be employed to overcome this problem.

Randomized drug trials have certainly proved to be very effective for assessing new treatments, but they also pose certain ethical issues. Is it really fair to allow patients in the control group to go untreated? This question is made considerably more difficult when the patient has a disease or virus that would be fatal if left untreated. The statistical answer is that a control group is needed for comparative purposes. The ethical question clearly does *not* have an easy answer.

Designed Experiments

The last data collection method we consider concerns studies based on experimental designs. In these investigations, the effect of one or more treatments on a response variable is of interest. In Section 3.7 we revisit the statistical design of experiments and consider some specific designs that are commonly used in practice. The practical application of experimental design is emphasized in the text by Box et al. (1978). (This should be on every statistical consultant's bookshelf.) The basic principles of statistical design of experiments are listed below.

- Control
- Randomization
- Replication

Control The importance of the control group and randomized assignment of treatments to patients in clinical trials was discussed above. Comparing several treatments is the simplest application of the *principle* of control. More generally, it refers to the problem of eliminating sources of systematic error (bias) and controlling for the effects of other variables on the response. *Blocking* factors, for example, exploit natural groupings of the observations. By controlling for the between-block variation in an experimental design, a more precise comparison of the treatment means is possible.

Randomization The main advantage of well designed experiments over observational studies is that experiments can provide good evidence for causation. By controlling for other factors and eliminating bias, differences in the observed response can be attributed to the effect of the treatment. By removing the subjective judgment of the investigator and making the assignment of subjects to treatment groups independent of any characteristic of the experimental units, randomization relies on chance alone to create comparative groups. A completely randomized design is not necessarily practical or even desirable (e.g., blocking). Some form of restricted randomization is needed in this situation.

Replication This is not an issue in clinical trials, but not every branch of science has the budget required of pharmaceutical companies. Indeed, the purpose of many experimental designs is to provide the investigator with an efficient method to analyze the effect of several treatments simultaneously. That is, the total number of treatment combinations or runs that need to be performed can be kept to a minimum. (We defer a more formal description of these designs until later in this chapter.) However, replication of at least some runs is required in order to assess experimental error. As always, there is a trade-off between sample size (number of runs) and precision (sensitivity of experiment to detect treatment differences) which the investigator must face. The role of a statistical consultant is to provide guidance on selecting a suitable compromise.

Occasionally, the consultant may encounter the problem of *pseudoreplication* (Hurlbert 1984). To study regrowth after a forest fire, the number of new seedlings per square meter is counted at a particular location. If the counts from 10 square meter blocks are obtained, do we have 10 replicates? Chatfield (1995) provides a similar example involving a decomposition study of several bags of leaves in a pond.

Designed Experiments (?)

No, this is not a pseudoreplication of a previous heading! In practice, many "designed" experiments fail to provide useful results because they do not

fulfill the objectives of the study. The subjects, treatments, or implementation of the experiment may not realistically duplicate the conditions that the investigator really wants to study. Lack of realism can even be due to the experiment itself — if the subjects know "it's just an experiment," are their responses representative of the general population profile?

The issues involved in actually designing an experiment often receive very little attention during a statistician's training and far greater emphasis is placed on the technical details associated with balanced and orthogonal designs. For the statistical consultant, experimental results are more likely to come from unbalanced and nonorthogonal designs. This usually does not present a problem, at least computationally, since good statistical software packages exist.

It is worth noting that our primary concern with data collected from a designed experiment is with the quality of the design. There are numerous examples of botched experiments and misleading conclusions based on the analysis of the wrong design. (Andersen (1990) is an entire book on flawed experiments.) Unless we have collected the data ourselves, the first task of a statistical consultant is to determine precisely how the data were collected. Although this obviously applies to any of the data collection methods we have discussed in this section, it is especially important in designed experiments.

3.2 Data Processing

Now that the database has been collected and transferred to electronic format it will need to be read into the statistical software program. Here we assume that the data transfer to our system has been successfully completed and the database for the project now resides on our system in appropriate format.

Always make a **backup copy** *of the original database.*

(Details concerning these issues are addressed later in this chapter.) The first stage of our statistical analysis begins with **Data Processing**. Thus, our first task is to read the data into our statistical software program and perform an initial assessment of the data quality.

Data Quality

The three main steps for assessing initial data quality are:

- Backup — Make a backup copy of the original data.
- Format — Codes and data types for reading the data.

- Values — The data entries associated with the variables.

Backup Keeping a copy of the original data serves several purposes. It provides the standard precautionary measure in case our *working* data file gets corrupted or deleted — this can happen at any time, often without warning, and we may be completely powerless to stop it (e.g., a bad drive wipes the disk, we mistakenly type `rm *` on a UNIX machine, etc.). It allows us to communicate problems with data in the format that the client knows. Line 53 in our (sorted) working data file may be quite different from that of the original. Similarly, it allows us to compare specific entries in our working dataset with those of the original. This can be particularly helpful when trying to resolve input errors.

Format Statistical software packages such as SAS and S-PLUS require format statements in order to read a data file correctly. For example, character variables need to be declared properly in SAS and data tables are assumed to be rectangular in S-PLUS. Problems often arise when the original data contain missing values that are represented by blank spaces (such as an empty cell in an Excel spreadsheet). Since the default is to ignore blank spaces, the remaining data entries become assigned to the wrong variable. In some cases it may be easier to edit certain aspects of the data file (our working copy) rather than try to employ complex format statements. Alternatively, it may be possible to read in the data file with a proxy format (as character variables, for example), then output the data in the desired format. Of course, providing clear instructions to our client as to the type of format we expect their data file to appear in, can certainly help reduce the amount of time we need to spend on this activity.

Values Once the data appear to have been read correctly, the next step is to check the statistical quality of the data. This includes detecting obvious or potential errors in the data values themselves. Summary statistics and frequency tables provide a useful way to isolate and assess potential errors. Some examples are presented below.

We should emphasize that an assessment of data quality also applies in the case where the clients have assumed responsibility for performing the statistical analysis using their own software. In this situation we should carefully examine the summary statistics associated with the client's data processing, checking for the type of potential errors or problems we discuss below. This needs to be done before we provide the client with assistance on the interpretation of the results — remember that inferences based on invalid data are rather meaningless. A more succinct expression for this scenario is "garbage in, garbage out."

Potential Error	Possible Cause
Some categorical classes contain very few observations	Data entry error Artifact of the database
A categorical variable contains more levels than it should	Changes in coding used Column shift during read-in
Variables that contain a high proportion of missing values	Wrong format used in read-in Column shift during read-in Programming error
Values associated with user-defined variables do not make sense	Data entry error Programming error Illegal operation performed
Values that lie outside the expected or allowable range of a variable	Outlier: data entry error Read-in or programming error Poor quality data

1. Classes of a categorical variable that contain very few observations.

   This may simply be a typing error that results in a nonexistent class being created for a categorical variable. If not, then it may be necessary to combine classes or exclude that class from the analysis.

2. Changes in the coding used for the classes of a categorical variable.

   Categorical variables whose classes have been numerically encoded tend to be more susceptible to miscoding errors. Although numerical values are easier to enter for categorical variables, it can be difficult and time consuming to diagnose systematic errors. In cases where a variable has more levels than it should, a careful check is needed to ensure that the data are being read correctly (e.g., 30 .0 will result in 0 being assigned, incorrectly, to the next variable).

3. Data values that lie outside the expected or allowable range.

   A quick check of the summary statistics (min, max, mean, standard deviation) can indicate potential errors for continuous variables. Obvious errors are usually revealed as gross outliers. A large standard deviation should also be regarded as suspicious, particularly when it is of similar magnitude to the mean. ID and count variables can be checked via frequency tabulations. (There should be a frequency count of 1 for each ID label.)

4. Variables that contain a high proportion of missing values.

   This can occur if a column-shift has occurred due to a blank space since character values are set to missing if the next variable is declared as continuous.

5. Data values associated with user-defined variables.

If the client has preprocessed the data and provided us with variables computed from other quantities, we should use our software to compute the same quantity and check for mismatches.

Missing Values and Errors

There are several important issues that arise when the database contains missing values or errors. In some cases, the impact on a statistical analysis can be substantial and simply ignoring missing values in textbook fashion is **not** recommended! Similarly, decisions need to be made regarding errors that remain after the data have been processed. Again, simply assigning a missing value to any identified error may not be the best approach. The following provides a list of some situations where errors and missing values typically occur.

Data Collection

- Misreading a measurement
- Misrecording a measurement (20 instead of 2.0)
- Estimating a measurement
- Faulty apparatus used in study
- Different instruments or personnel used in the study
- Truncation or biased rounding of measurements

Data Entry

- Mistyping values during transcription to electronic format
- Misrepresenting data values
- Duplication of values in columns or rows
- Misalignment of values in column or rows
- Incorrectly assigning values as missing
- Assigning a value when observation was missing

The main issue that arises with errors and missing values is that there is a loss of information. In statistical terms, this corresponds to a loss of degrees of freedom and hence, less reliable inference. The first step is therefore to work with the client and try to recover these data where practical. Having the client define precisely how certain variables were recorded and entered is often a useful place to start the recovery.

We may be surprised how our "explicit" instructions concerning data entry, coding, and formats were not quite as clear as we thought. Do **not** blame the client. They would have done their

best to do what we asked — this is our turn to learn from experience.[1]

Once the recovery from the original data source has been effectively exhausted, decisions need to be made with regard to the remaining errors and missing values. At this point we should emphasize that only "potential" errors actually remain since the correct value cannot be recovered from the original source. The difficulty here is that there is a tendency for the analyst to make ad hoc decisions such as setting all remaining "potential" errors to be missing, or dropping an entire multivariate observation even if only one component is missing. On the other hand, when a value is clearly an error (e.g., it lies outside the allowable range), modifying an error or missing data value is clearly a subjective decision and introduces bias (or worse) into the analysis. This includes the often-used practice of setting a missing value to the mean or median of the data associated with the variable. A more objective approach to imputing missing values is described in Little and Rubin (1987).

3.3 Statistical Issues

Having completed the data processing phase of the project, the investigator is now ready to perform a "statistical analysis" of the data. Even in relatively simple projects, this would typically consist of:

1. Computing numerical summaries and graphical displays

2. Performing a statistical "test" on the data

3. Interpreting the results and making conclusions.

The last item usually presents the greatest difficulty for clients since this involves the concept of statistical inference. The theoretical foundations underlying statistical methodology are extensive and reside within the domain of expertise of the statistical consultant — keep them there! The client came to us for assistance, not a course on probability and inference theory. Of course, the issue of statistical inference is important and needs to be addressed. We consider some of the issues with respect to the following areas of inference.

---

[1] One way to help avoid these types of misunderstandings is to actually sit with the client and enter some example cases. The client can then see what we *really* mean about binary coding and distinguishing missing values.

- Estimation
- Hypothesis tests
- Sample size and power

Estimation

A summary statistic such as the sample mean $\overline{X}$ is an example of a **point estimate** that can be used to infer something about the unknown mean (μ) of the population from which the sample was drawn. While this statement may seem relatively innocuous, it rests on top of a rather considerable amount of estimation theory. The "inference" about the population parameter μ is based on the properties of the sample mean **estimator**: the theoretical random variable quantity defined *prior* to the experiment ... and about now, we probably just lost the client! What the client really needs to know is what type of estimation method is appropriate for the project. The following examples are used to illustrate this point.

Example 3.2 *The average AGE of a patient*

Databases provide the date of birth of a patient, but not necessarily a corresponding `AGE` variable. The difference between the current date and the patient's date of birth can be used to compute a truncated version of `AGE` in whole years.

The problem with using a truncated version of `AGE` is that when the *average* `AGE` is computed, how should a value such as 68.75 years be interpreted? While this "looks" like the average age is 3 months shy of 69 years, the association of "months" with the fractional year part (0.75) is misleading since it was derived from a truncated `AGE` measure that does **not** reflect a patient's age in terms of months. That is, it clearly underestimates the average age of the patient cohort since the fractional year component (months + days) of an individual is not included.

Example 3.3 *Normal blood pressure*

The "normal" blood pressure for a person is often quoted as a range.

In this situation, an **interval estimate** is a more appropriate measure of the normal blood pressure for a person. A single point estimate such as the mean value clearly does not convey the information a person needs if their blood pressure happens to be higher than this "normal" mean value.

Remarks

1. The two simple examples above are not the only types of estimation problems that the consultant will encounter. An interesting example is given in Thisted (1988) where the E–M algorithm (Dempster et al. 1977) was employed to estimate the proportion associated with a mixed distribution. This is known as the "Widows" dataset where the mixing proportion of the Poisson and binomial distributions needs to be estimated for the "zero" counts: widows with no dependent children.

2. In some cases, the asymptotic normal theory that is often employed to derive standard errors may not hold. In this case, resampling methods (discussed later) can be employed to generate standard errors and confidence intervals. The client may need some convincing that this is "allowed" and perhaps more important, the client needs to understand that resampling methods do **not** make the sample "larger" than it is.

Hypothesis Tests

The aim of a statistical inference procedure is to enable the analyst to make conclusions about the problem under investigation based on the experimental evidence. In traditional applications of statistical inference, the analysis of a problem involves several components. First, the problem needs to be described in mathematical terms, which serves to identify the nature of the randomness. Some simplification of the precise nature of the randomness is often necessary in order to make the analysis tractable.

Having translated the problem into a "random process," the next step is to select a suitable statistical procedure for evaluating the experimental evidence. Here, the choice of procedure is dependent on both the context of the problem and the type of evidence to be evaluated. Matching this information with an appropriate probability model provides the analyst with a procedure that can be used to measure the likelihood of any outcome from the experiment. However, in order to compute the correct probability of an actual outcome, the "true" state of the random process needs to be represented by the model. Clearly, if this were known, statistical inference would not be needed. Thus, the following strategy can be employed.

1. The analyst makes a "hypothesis" about the true state of the process and then evaluates the probability of obtaining the outcome reported from the experiment, under this assumption.

2. Should this computed probability seem too unlikely, it is reasonable to infer that the stated hypothesis was incorrect.

3. In this case, the analyst would conclude that the evidence does not appear to support the hypothesis that was assumed.

This "conclusion" and associated hypothesis are dependent on the context of the problem. In summary, the stages involved in this type of statistical inference procedure are:

Assumptions Statistical procedures require certain assumptions to be made about the "random process" under study. These assumptions enable a probability model to be derived that mathematically describes the random process. The experimental evidence is then analyzed on the basis of this model.

Hypothesis In the hypothesis testing approach to statistical inference, the evidence is assumed to have occurred when the experiment was performed under a particular condition. This is referred to as the null hypothesis (H_o). Under H_o, the probability of obtaining a result at least as "unlikely" as the observed outcome from the experiment is computed from the probability model. The computed probability is called the P-value and will depend on the form of the alternative hypothesis (H_1) of the hypothesis test. However, if the P-value is less than 5%, then H_o may be rejected in favor of H_1. In this case, the result of the hypothesis test is said to be statistically *significant*.

Conclusions It is important to note that a statistically significant result from a hypothesis test does **not** imply causation. Indeed, the P-value provides a measure of the probability of making a Type I error: rejecting H_o when it is true. We would need to repeat an experiment many times in order to establish causation. In practice, we usually do not have the opportunity to completely reproduce an experiment and need to make conclusions based on the results at hand. Considerable care is necessary when making conclusions based on a statistical inference procedure to ensure that unsubstantiated claims are not implicitly associated with the statistical nature of the analysis.

Inference Aside

Statistical inference is perhaps better described as a philosophy since there is considerable debate concerning the theoretical framework on which the above *frequentist* approach to inference is based. Indeed, there are philosophical problems involved in the very notion that we can assess probabilities and make inferences. This book is clearly not the place to address this issue, but it is important to note that there are other approaches to statistical inference. (We refer the reader to Barnett (1982) for a comparative account of the different approaches to statistical inference.) In particular, *Bayesian* inference combines "prior" information that we may have about

the data (usually expressed in terms of a probability distribution) with the sample data which enables the "posterior" information to be evaluated.

In this book, we take the pragmatic view that for most practical purposes, the significance testing approach to statistical inference does work. (Provided, of course, it is performed correctly.) However, there are many situations where the Bayesian approach is intuitively more appealing. In market research, for example, our prior belief about the specified population for a new product should be incorporated in planning our test market survey. Statistical consultants therefore need to be flexible in their approach to statistical inference.

Remarks

1. The statistical emphasis that we have (deliberately) employed in summarizing the stages of a hypothesis test is more likely to convey an impression of arrogance on our part, than to further the client's understanding of statistical inference. Although clients may be quite familiar with "performing" a hypothesis test, most would find this type of summary to be intractable.

2. So what approach should we take? If the client is already familiar with this type of statistical test procedure, then we can begin by emphasizing the importance of using the evidence to check whether the assumptions of the test are satisfied prior to drawing conclusions based on them. This would be followed by examining the *contextual* meaning of the stated hypothesis and, finally, making sure the client understands the limitations of a "statistical" conclusion.

3. For clients who are not familiar with statistical tests, the remark above still applies, but we have found it useful to precede this discussion with a simple example such as the following.

   > A stranger claims to have obtained 10 heads in 10 tosses of coin. If the coin is assumed to be fair, then the likelihood of obtaining this outcome by chance alone is less than 1 in 1000. Since this is less than the minimum standard of 5%, we would conclude that the result is due to something other than chance. Accordingly, we might accuse the stranger of lying.
   >
   > While this "conclusion" may seem quite reasonable, suppose we are given the opportunity to reproduce the experiment by tossing the coin in question ourselves. If we were to obtain 9 heads in 10 tosses, then our conclusion would agree with the statistical notion that the result is due to something other than chance. That is, we would reject the assumption that the coin is fair. However, the implication that the stranger is lying is clearly invalid.

4. Clients often state their objectives in terms of *directional* hypotheses. When translated into statistical notation, a directional hypothesis represents H_1. The problem is that the P-value provides a measure of the likelihood of the experimental outcome with respect to the null hypothesis H_o. The key point is that a decision can be made about the client's directional hypothesis based on the P-value associated with the statistical test. Here, it is usually the case that the client needs help with the conclusion.

5. It is often instructive to consider the conclusion that would be associated with a *nonsignificant* test result. In some fields of study, the 5% cutoff is set in stone. (Even a result such as 0.052 would be ignored!) This is unfortunate, but we should at least try to convince clients about the fallacy of adopting a strict cutoff — perhaps they can begin to change the attitude of their peers. One way to do this is to consider the *power* of the test. This may need to be done indirectly at first (starting with the relationship between sample size and the "reliability" of a test), but providing the client with a new type of "P-value" (power) and a statement of what the number means, can be useful. The issue of sample size and power is discussed next.

Sample Size and Power

Determining an appropriate sample size is an important concern in observational studies and planned experiments. It is directly related to the cost of implementing the study and affects the quality of the statistical inference associated with the results. Methods for sample size determination address the problem of optimizing the amount of information that needs to be collected with respect to a particular application. Some examples are:

Longitudinal Studies How long should the study run and how often should the data be collected?

Sample Surveys In complex surveys, cost functions need to be optimized with respect to quality to ensure maximal return of information.

Clinical Trials The effect of a new drug needs to satisfy certain government regulations. How many patients are required in each treatment level and for the control group?

Quality Assurance Inspection schedules and sampling schemes need to satisfy quality protocols.

Experimental Designs How many items per cell are required in a particular design to detect a specified difference in the effects?

In general, an appropriate sample size can be obtained by specifying a level of accuracy or by maximizing a specified objective function. Explicit methods exist for many standard problems and comprehensive tables based on power analysis are available: Kraemer and Thiemann (1987), Cohen (1988), and Desu and Raghavarao (1990) are three key monographs. For a review of sample size determination methods, including Bayesian approaches, see Adcock (1997). Power analysis is also incorporated into a number of statistical software packages.

In practice, however, determining an appropriate sample for the client may be quite difficult. Consider, for example, the following definition of power $= 1 - \beta$, where $\beta = P[$ Type II error $]$.

> The **power** is the probability of obtaining a significant result. It is a function of the sample size n, the effect-size δ, the standard deviation σ, and the significance level α. The power tells you how likely it is that your experiment will detect a specified difference δ at a given significance level α.

Determining the sample size n therefore involves specifying each of the other quantities. Conventional values for the level of significance and power are 5% and 80%, respectively. In many standard tests, σ can often be absorbed to give standardized effect-sizes. Otherwise modifications in the sample size computations are necessary to account for the variability associated with estimating σ. The main problem occurs in trying to specify a reasonable effect-size since this represents the degree to which an investigator believes the null hypothesis (H_o) to be false. This is a subjective decision and is further complicated by the different scales of measurement involved since effect-size depends on the type of test used.

In the behavioral sciences, tables (see, for example, Cohen (1988, 1992)) to determine sample size are often based on classifying effect-sizes as *small, medium*, and *large* for certain tests. In engineering and quality assurance applications, sensible sample sizes are often well established and the operating characteristic (OC) curve is used to assess β versus standardized effect-sizes. The OC curve is simply a plot of β versus δ for a given sample size n. An OC chart consists of overlaid curves at several different sample sizes. For simple experiments and designs, these tables and OC charts are usually sufficient for the purposes of sample size determination. We conclude with some additional comments.

1. Explaining to a client the concepts involved in power analysis can be very difficult. Adopting a confidence interval approach may be more effective. For example, the conservative margin of error calculation $2\sqrt{p(1-p)/n} \leq 1/\sqrt{n}$ that is used in opinion polls is an easy way to familiarize a client with the issue of sample size versus quality.

2. Power calculations are performed prior to the experiment. They do not explain why a postexperiment result is not significant.

3. A specified sample size does not guarantee that the experiment will produce a valid result. There are usually assumptions underlying the test procedure that may not be satisfied by the experiment. It is important to emphasize this to the client.

4. For large sample sizes, significant results are expected. Whether these are meaningful is not always clear. Larger sample sizes are also more prone to data processing errors and implementation problems, and may contain poor quality information. Larger is *not* always better.

5. For planned experiments, sample sizes are determined assuming balanced designs: equal group or cell sizes. Missing values can badly affect the power associated with an unbalanced design. The client needs to be made aware of the consequences of allowing a designed experiment to become too unbalanced.

6. In multivariate experiments, the power and sample size will also depend on the number of response variates to be measured.

3.4 Statistical Methods Used in Consulting

In the following two sections we review some of the statistical methods commonly used in consulting environments. In the next section, we concentrate on "standard" methods. These include descriptive methods that would be employed during the exploratory phase of the analysis as well as some basic statistical procedures (as listed in the table below). For the more specialized procedures, our presentation is necessarily terse since a general exposition on these methods can be found elsewhere. Perhaps more important, details concerning the application of some of these methods are provided in the context of the case studies presented in Part II. References to other texts and the relevant case study are provided. Our shortlist of statistical methods is summarized in the following table.

Standard Methods

- Exploratory Data Analysis
- Contingency Tables
- The t-Test
- Analysis of Variance
- Regression

General Methods

- General Linear Models
- Multivariate Analysis
- Time Series Analysis
- Categorical Data Analysis
- Specialized Procedures

3.5 Standard Methods

In describing the statistical methods and techniques included in this section, we have tried to emphasize the importance of engaging the client's understanding of the purpose, details, and interpretation of each method. That is, these are the methods about which we may need to educate the client. The statistician has no problem understanding these elementary methods, but clients certainly do! That's where the statistical consultant comes in.

Exploratory Data Analysis (EDA)

has become a rather amorphous terminology in the sense that there is an enormous range of statistical methods that can be associated with a so-called exploratory data analysis. For our purposes EDA includes any type of graphical display, numerical summary, or statistical procedure that is used to investigate the distributional and structural properties of a dataset during the initial stages of analysis. Thus, EDA represents an "informal" analysis of the data in an attempt to reveal unusual or interesting features.

Descriptive Statistics

The calculation of numerical summary statistics and construction of simple graphs is referred to as *descriptive* statistics. For report purposes, a

descriptive analysis provides an overall summary of individual variables in the database. Additional summaries within important subgroups may also be useful. From the consultant's perspective, these descriptive diagnostics will already have played a fundamental role during the data processing phase of the project. Pertinent aspects such as data quality of certain variables can be helpful in guiding the direction of the project's analysis.

Shape:	Distributional features of the data
Location:	The position of the data (center, extrema)
Spread:	Measures of variability in the data

For the three descriptive properties listed above, the histogram, sample mean, and standard deviation are certainly the most widely used to summarize individual variables. However, it is important to recognize when these statistics may be misleading.

Graphical Displays

The importance of quality graphics was discussed in Section 2.6. Here we focus on the statistical aspects associated with a selection of graphical displays that are useful in EDA. For example, although the pie chart is often employed to display distributional information, a bar chart is a more effective data visualization tool.

Distribution:	Histogram, Bar Chart, Stemplot, Boxplot
Y versus X:	Scatter Plot, Time Plot, Mosaic Plot
Multivariate:	Contour Plot, Trellis Plot, Dynamic Graphics
Diagnostic:	Q–Q Plot, Periodogram, Residual-vs.-Fitted

Distribution Histogram, Bar Chart, Stemplot, Boxplot

The histogram and bar chart are widely used to display the frequency distribution of quantitative and categorical data. For quantitative variables, features such as symmetry, skewness, multimodality, groupings, and outliers are of interest. Two other effective techniques for investigating these "shape" features are the stemplot and boxplot displays. A histogram was shown in Figure 2.3 of the previous chapter (Section 2.6). Examples of the other display types are shown in Figures 3.1 through 3.3.

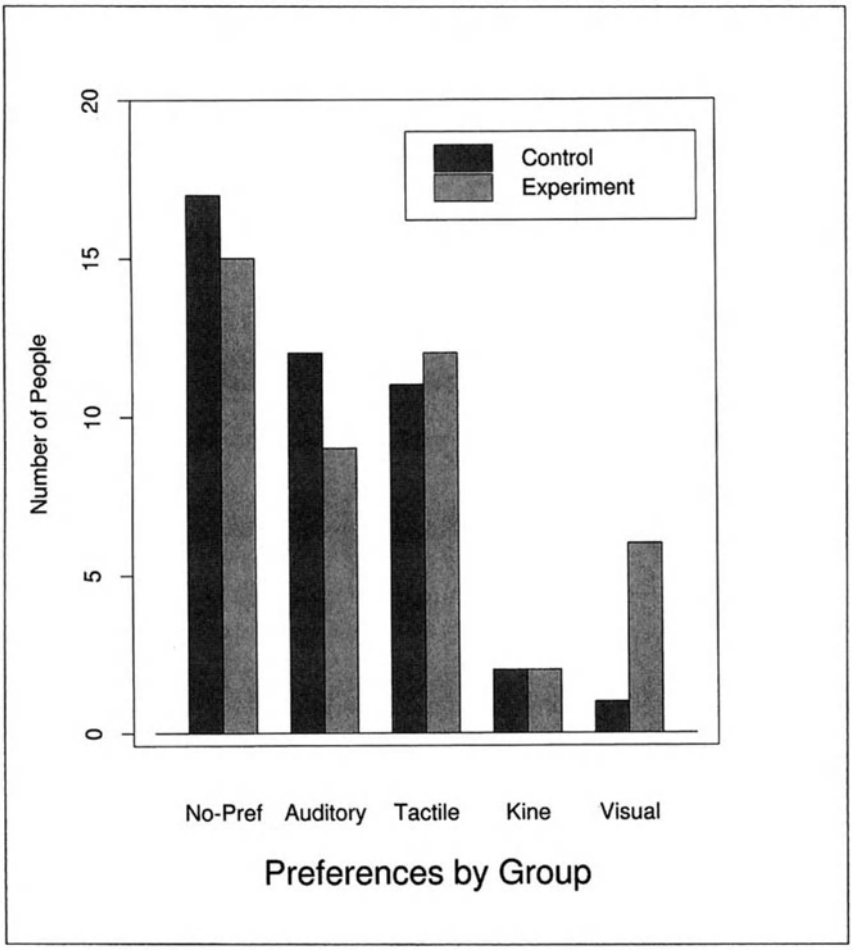


FIGURE 3.1. Split Bar Chart

The purpose of the histogram or stemplot[2] is to provide a representative display of the underlying distribution that generated the sample data. Thus, an important issue is the choice of the interval width to use in a histogram display, and number of lines to use in a stemplot display. This represents a compromise between resolution and sample size (n):

> Too few intervals and important distributional features will be missed; too many and artificial features of distribution will be introduced by the display itself.

For a sample size of n, the following criteria have proved useful.

Gaussian Process $h_n = 3.49sn^{-1/3}$ was proposed by Scott (1979) as a data-based choice for the interval width to use in a histogram display for data assumed to be normally distributed. Here, s is an estimate of σ (such as the sample standard deviation). A rule proposed by Freedman and Diaconis (1981) is $h_n = 2(IQR)n^{-1/3}$, where IQR is the interquartile range of the sample.

Stemplot Let $L_{\max}$ denote the **maximum** number of lines to use in a stemplot display. In Hoaglin et al. (1983), $L_{\max} = [10 \log_{10} n]$, where $[x]$ denotes the integer part of x, is proposed as a reasonable upper limit for the number of lines to use in a stemplot display. This result was attributed to Dixon and Kronmal (1965) who used it for histograms. It works well for small to moderate sample sizes ($n \leq 300$). For larger sample sizes, the benefit of retaining the actual data values becomes less useful.

Remarks

1. Make sure clients understand "how" to interpret a histogram. The concept of symmetry may need to be explained.
2. The construction of bar charts can sometimes be misleading, particularly if several variables are represented on a single display. For example, relative frequencies need to be employed when the group counts are based on different sample sizes.
3. A disadvantage of the stemplot and boxplot displays is that they may be unfamiliar to some clients, or to the area in which the clients would normally present their results. If the construction

[2]This was originally called a "stem-and-leaf" display (Tukey 1970); the name deriving from splitting the most significant digits of a data value into a "stem" component (leading digits) and the "leaves" consisting of the trailing digits from all data values with a common stem.

```
> stem(lottery.payoff)

N = 254   Median = 270.25
Quartiles = 194, 365

Decimal point is 2 places to the right of the colon

   0 : 8
   1 : 00001112223333333333344444
   1 : 55555566666677777778888888899999999999
   2 : 000000011111111111122222223333333344444444
   2 : 55555666666666677777888899999999999999999
   3 : 00000000111111222233333333444
   3 : 55555555666667777777888888899999999
   4 : 0122234
   4 : 55555678888889
   5 : 111111134
   5 : 555667
   6 : 44
   6 : 7

High: 756.0 869.5
```

FIGURE 3.2. Stemplot: Lottery Data (Source: S-PLUS)

of these displays requires more explanation than their interpretation, stick with a histogram.

4. The key difference between the stemplot and histogram displays is that effects due to the data themselves, such as rounding or multiplicities, can be detected using the stemplot. The histogram is sensitive to these types of perturbations: a small shift in the class boundaries (breakpoints) can produce large changes in the frequency distribution.

5. The boxplot provides a very compact visual display of the data based on the quartiles and extremes: $\{\min, Q_1, M, Q_3, \max\}$, where M is the median (see **Location** below). It is particularly useful for comparing the distributions of several related variables or for examining the distribution of an individual variable across subgroups.

Y versus X Scatter Plot, Time Plot, Mosaic Plot

The term "Scatter Plot" refers to the situation where the relationship between two quantitative variables, labeled X and Y for convenience, is displayed as points on the Cartesian plane. The strength and nature of the association (linear or nonlinear) is of interest, along with influential points in the two-dimensional plane. In the regression context Y is regarded as the response variable and the predictive quality of the input or explanatory variable X is of interest. Examples are shown in Figure 3.4.

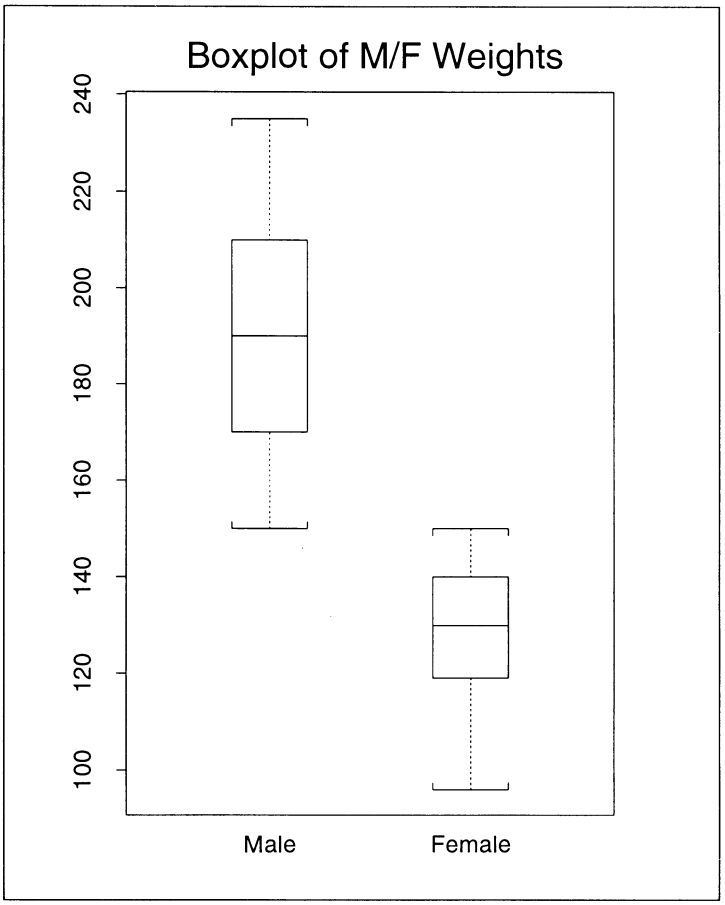


FIGURE 3.3. Parallel Boxplots

A special case is when X denotes a time index. In this situation, the Y values are usually connected to form a continuous "time series" plot. Periodicities, trends, and outliers are of interest in time series analysis. (See Figure 7.4, Case Study 7.4.)

Remarks

1. While the definition of a scatter plot may have seemed somewhat redundant, it served to point out a fundamental limitation in our ability to "visualize" data: we are essentially restricted to a two-dimensional view. For this reason, it is important to regard a scatter plot with caution. The absence of an observed relationship between X and Y does **not** necessarily imply there is no association between these variables.

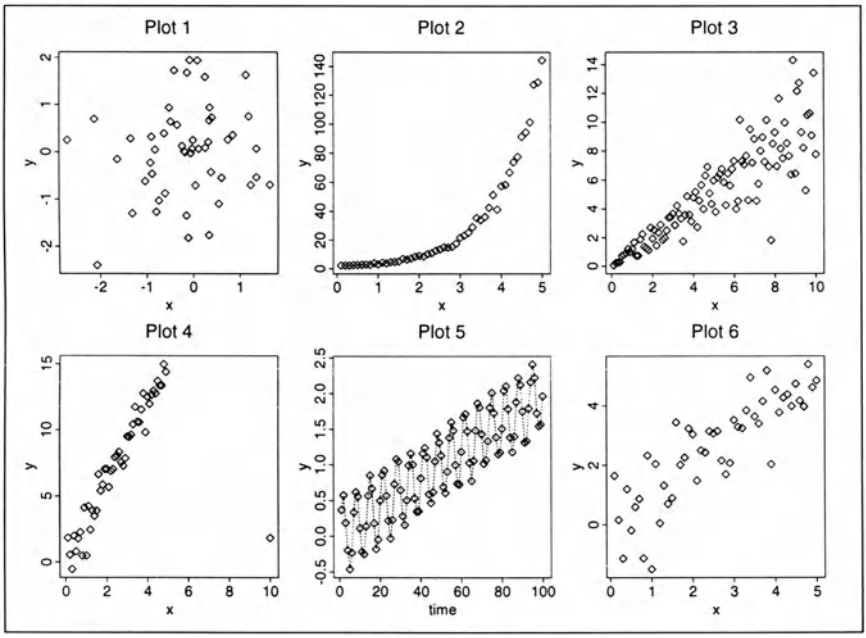


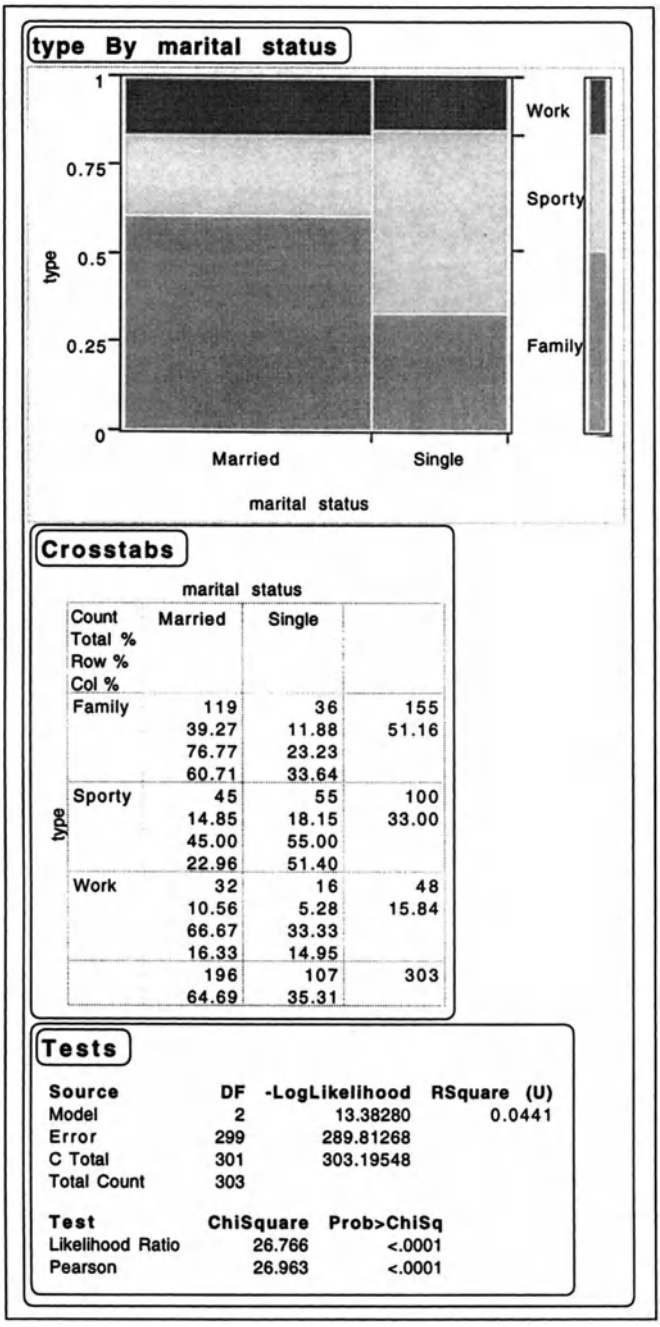
FIGURE 3.4. Scatter Plots

2. The presence of seasonality, cycles, and trends can usually be detected from a time plot. If these structural effects are removed (e.g., by differencing or fitting a time series model), interest then centers on whether serial correlation exists in the "residual" series, $r(t)$ say. A scatter plot of $Y = r(t)$ versus $X = r(t-1)$ provides a simple check for autocorrelation.
3. When X and Y are both categorical variables, a mosaic plot provides a graphical display of the relative frequencies in the two-way frequency table. This is discussed in more detail in the section on **Tabulation** below.

Multivariate Contour Plot, Trellis Plot, Dynamic Graphics

As indicated above, our data visualization techniques are restricted to 2D. Of course, this has not prevented the investigation of multidimensional relationships and contour plots provide an effective technique for examining three quantitative variables simultaneously. The contour plot is analogous to the altitude contours on a topographical map. Perspective plots, 3D plots, and 3D histograms can also be employed, but obtaining a "good" view of the data can be difficult due to the opaque drawing methods employed with these techniques.

Trellis plots (Mathsoft 1997) attempt to provide an "automatic" graphical system for displaying multivariate data in 2D. The initial



Crosstabs

	marital status		
Count / Total % / Row % / Col %	Married	Single	
Family	119	36	155
	39.27	11.88	51.16
	76.77	23.23	
	60.71	33.64	
Sporty	45	55	100
	14.85	18.15	33.00
	45.00	55.00	
	22.96	51.40	
Work	32	16	48
	10.56	5.28	15.84
	66.67	33.33	
	16.33	14.95	
	196	107	303
	64.69	35.31	

type

Tests

Source	DF	-LogLikelihood	RSquare (U)
Model	2	13.38280	0.0441
Error	299	289.81268	
C Total	301	303.19548	
Total Count	303		

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	26.766	<.0001
Pearson	26.963	<.0001

FIGURE 3.5. Mosaic Plot: Car Poll Data (Source: JMP)

"plot" can be any of the displays discussed above — bar chart, histogram, boxplot, scatter plot, contour plot, 3D plot — which is then reproduced conditional on the levels associated with all combinations of the remaining variables. An example of a trellis plot may be found in Case Study 7.2, Figure 7.1.

Another multivariate graphical technique is to "animate" the display, referred to as dynamic graphics. Spinning plots are often incorporated into statistical software that allows the user to perceive a 3D view of the data. XGobi (Swayne et al. 1991) provides the user with the ability to perform a guided EDA "tour" of multivariate data.

Remarks

1. Parallel boxplots and pairwise scatter plots are two examples of simple graphical displays for multivariate data. Although there have been many creative techniques developed for the simultaneous display of $k \geq 3$ variables, it is worth remembering that any 2D view of multivariate data will never be entirely complete.
2. How important are multivariate graphical displays to the consultant? To a certain extent, diagnostic displays such as those presented below will often suffice. Perhaps a better question is: "Can we interpret the results from more sophisticated multivariate displays?" In our experience, the ability to "look" at data will always outweigh the potential to be misled by numerical diagnostics.

Diagnostic Q–Q Plot, Residual-vs.-Fitted Plot, Periodogram

The Q–Q plot is often used to assess normality of residuals associated with a residual-versus-fitted plot. The periodogram is employed in time series analysis to assess the presence of a significant frequency or cycle in the data. All three plots are common examples of diagnostic plots that are used extensively in statistical analyses.

Remarks

1. The Q–Q plot consists of quantiles from a probability distribution $f(x)$ plotted against the ordered data. If the data were indeed generated by $f(x)$ then the scatter plot should exhibit straight-line behavior. In practice, the distribution of interest is usually the normal distribution and $n > 30$ observations are needed to provide a reliable diagnostic. See Figure 3.6 for examples of Q–Q plots.
2. The residual-versus-fitted plot represents the results from fitting a model to the data. If the model is correct, then the residuals

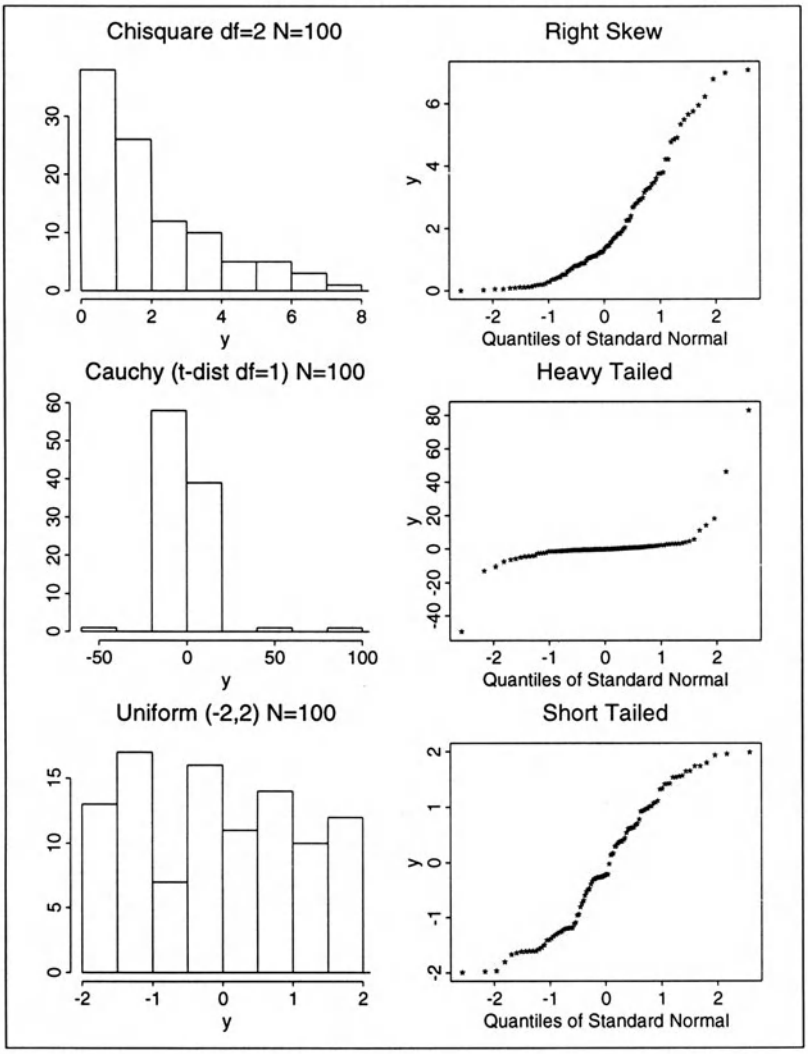


FIGURE 3.6. Q–Q Plots

should **not** exhibit any nonrandom patterns. It is used extensively in many statistical methods and the examples presented in Figure 3.7 are relevant to the regression methodology discussed later.

3. The Periodogram is essentially a one-way ANOVA partition of the total variation into frequency components defined by the Fourier decomposition of a time series $y(t)$. If a particular frequency, ω_j is dominant, then a large "spike" $I(\omega_j)$ will appear in the periodogram plot of $Y = I(\omega_j)$ versus $X = \omega_j$.

Numerical Diagnostics

Tabulation:	Frequency and summary tables
Location:	Position of the data: center, quantiles
Spread:	Measures of variability
Correlation:	Measures of association between X and Y

Tabulation Frequency and Summary Tables

A one-way frequency table is simply a list of the number of observations — the frequency or *count* — associated with the levels of a categorical variable. Quantitative data usually need to be grouped into nonoverlapping intervals (or otherwise categorized) to be useful in frequency table analysis. A **contingency table** is a two-way frequency table that lists the "cell" count associated with each level combination of the two variables. For EDA purposes it is useful to include the relative frequency or percentage for each level or cell in the frequency table. In a contingency table, cell percentages can also be calculated conditional on the "marginal" row and column counts.

The contingency table example presented in Figure 3.5 was generated by the statistical software JMP (SAS 2000)[3] and includes the mosaic plot which is a graphical display of the relative frequencies associated with the cells of the contingency table. The area of the rectangular "tiles" within the unit square correspond to the relative frequencies of the cells. The use of color (greyscale) to differentiate the levels of a variable gives the display a mosaic appearance — hence the name of this plot.

[3]The plot shown was produced using Version 3. The citation refers to Version 4.

Multiway frequency tables can be presented as a sequence of two-way tables, conditional on the level combinations of the remaining variables. In practice, **summary tables** are more effective when several variables are involved. The basic format of a summary table is a two-way table generated by the levels of row and column factors, A and B say. The levels of any additional factor are then "nested" within the levels of A and B, with the row factor nested more often. This is partly due to the dimensions of a sheet of paper. It is also easier to read a table when the rows have most of the nesting. The entries in the body of the summary table can be counts, means, or any other descriptive statistic desired by the client.

Remarks

1. A table should be informative! Some examples of where this is **not** the case are:
   - Only the percentages are provided. The sample size used to calculate these percentages should *always* be given.
   - Percentages are given to insufficient precision.
   - One level accounts for a large proportion (90% say) of the data. If the other levels are of interest they should be presented in a separate table.
   - The sample size is too small to support a valid frequency analysis.
2. Sparse tables can occur when there are too many levels involved. A large proportion of cells have zero or single counts. For analysis purposes it will be necessary to recategorize a variable by combining certain levels. These decisions should be made with the client.
3. Frequency tables are sometimes constructed using nonexclusive categories. That is, an item can belong to more than one level. This clearly affects the interpretation and statistical analysis of the table and it is important to check whether this situation exists whenever a client provides us with the frequency table.
4. The number of tables generated in a frequency analysis can be considerable. For example, a survey consisting of 100 questions (variables) would generate $\binom{100}{2} = 4950$ tables if all two-variable interactions were considered. Similarly, the number of conditional two-way tables generated in a multifactor frequency analysis increases geometrically. Judicious selection of the variables to use in cross-tabulation and multiway frequency analysis is important.

The main aim is to tabulate summary information in a concise but clear fashion. Although the basic arrangement of a table is a two-way layout, employing subgroupings within rows and/or columns allows more complex relationships to be presented. However, the format of the table is important and usually requires judicious selection of the row and column variables and subgroupings. Some points to consider when constructing a summary table are:

Statistics The cells of a table normally consist of simple summary statistics such as frequencies, percentages, averages, sums, and quantiles.

Standard Errors These are often combined with the corresponding estimate in the form of ±SE or (SE). It is important to make clear whether a standard *deviation* or standard *error* is being used.

***P*-Values** Only significant P-values are usually of interest and often it suffices to list certain thresholds (e.g., 0.05, 0.01) in a footnote to the table.

Table Size A summary table should fit on a single page. Since this is rectangular in shape, there is less physical space available for columns than rows (in portrait mode). Subgrouping will usually need to be done within the row variable first.

Number of Cells The number of cells in a table is determined by the product of all the grouping levels used. When necessary, a table can be split into several smaller tables, each "conditional" on the levels of a particular variable.

Landscape Tables presented in landscape mode (sideways) allow more columns to be included. Since this requires turning the report to view the table, then back again to read the associated text, landscape mode should be avoided where possible.

Location Central Tendency, Quantiles, Resistant Measures

The sample mean $\overline{X}$ and median M are the most commonly used statistics to describe the "average" value of a process under study. For exploratory purposes, *order statistics* provide a convenient way to investigate the distributional properties of a process. For example, the *quartiles* of a dataset, Q_1 and Q_3, are location measures of the 25 and 75% percentiles, respectively. (That is, 25% of the data values lie below Q_1.) Unlike the median, there is no "standard" calculation for Q_1 and Q_3. Thus, minor differences do exist between statistical software packages.

Remarks

1. The difference between M and $\overline{X}$ is easily demonstrated by considering the following dataset, {27, 29, 31, 33, 180}, for which $\overline{X} = 60$ and $M = 31$. If these numbers represented salaries in a small company looking to hire new personnel, an "average" salary of \$60K would appear very attractive! Clearly, a location measure such as $\overline{X}$ by itself, can be quite misleading — a measure of the variability should always be provided.
2. **5-Number Summary** The boxplot is a graphical display of the 5-number summary which is the collection of the following five location statistics: $\{\min, Q_1, M, Q_3, \max\}$. An important component in the construction of a boxplot is the determination of potential outliers which is based on the following criteria.

   Outliers Any $X_i < C_L$ or $X_i > C_U$ is considered to be a potential outlier where[4]

   $$C_L = Q_1 - 1.5 * (Q_3 - Q_1) \text{ and}$$
   $$C_U = Q_3 + 1.5 * (Q_3 - Q_1).$$
3. As shown in the example above, the sample mean is not *resistant* to outliers. Although resistance to gross errors is clearly desirable from a practical perspective, the statistical efficiency[5] of an estimator is also important. This has led to the study of **robust** estimators and a simple example is the trimmed mean:

   Trimmed Means A 5% trimmed mean discards the smallest 5% of the data values and the largest 5% of the data values. That is, a total of 10% of the dataset is discarded. The sample mean is then computed based on the remaining 90% of data. In Hoaglin et al. (1983), it is shown that a 25% trimmed mean retains reasonable efficiency.
4. From a client's perspective, the endless array of different measures for the same location can seem somewhat disturbing. It is important to emphasize the exploratory purpose of these different measures. In reality, the question is not, "Which one should we use?" but "Do the data support the assumptions required for the intended analysis?" If not, at least we can show why.

Spread Standard Deviation, Standard Error, Range, IQR

A location estimate simply provides a reference point for the dataset; it gives no information about the "variability" of the data values.

---

[4]The constant 1.5 is a common choice; 2 is also used.

[5]This is actually a relative measure that compares the precision of two estimators. For example, the median M requires approximately 64% more data to be able to measure the "center" of a normal distribution (μ) with the same precision as $\overline{X}$.

Two common measures of spread are the standard deviation S and range R. The interquartile range $IQR = Q_3 - Q_1$ can be used to provide a resistant measure of spread.

Remarks

1. **Standard Deviation** Although it is as commonly used as $\overline{X}$, it is also one of the most "unintuitive" summary statistics. The definition of S via the sample variance S^2 certainly doesn't help matters in this regard. The main issue for some clients is trying to understand what exactly S measures. "Is my value of S good or bad?" may seem like a silly question, but in our experience this is not what the client is really asking. Often, the real issue is about the validity of the analysis which supposedly had something to do with S. Introducing the concept of standard error in terms of margin of error can be helpful here.

2. **Standard Error** The statistics defined above all provide a measure of the "raw" data variability. For analysis purposes, we require a measure of the sampling variability associated with an *estimate* (such as $\overline{X}$) that was computed from the dataset. The problem here is that we only have "one" estimate! This measure of spread is called the *standard error* of the estimate and is usually based on the theoretical variance of the estimator.

   For example, the standard error of the sample mean is $\sqrt{\sigma^2/n}$, where σ^2 is the unknown process variance. Thus, $S/\sqrt{n}$ provides an estimate of the standard error for $\overline{X}$. Again, this description of a standard error may not mean much to a client (e.g., they happen to be looking at regression output). In this case, we may need to simply state the standard error and explain what it represents in terms of repeating the client's experiment many times.

3. **Confidence Interval (CI)** This combines a location and spread estimate to give an *interval* $[a, b]$ which represents a range of possible values for the (unknown) process parameter of interest. Often the limits of the CI are chosen so that the interval has a 95% chance of containing the process parameter.

   For example, the conservative 95% CI given by $\hat{p} \pm n^{-1/2}$ is often used to report poll results. Here, $\hat{p}$ is the proportion of yes responses in a yes/no survey poll of size n, and $n^{-1/2}$ is the margin of error which is usually stated as, for example, ± 3 percentage points.

4. **Resampling** methods such as the jackknife and bootstrap are computational techniques that can be used to estimate standard errors. The jackknife method allows n different estimates

of, for example, $\overline{X}$ to be constructed from the same dataset by leaving out one observation each time. Thus, an estimate of the standard error of $\overline{X}$ can be calculated based on the usual standard deviation formula applied to the n different values obtained for the sample mean. The bootstrap method resamples from the dataset, *with replacement,* and constructs the estimate each time. It is more computationally intensive, often employing several thousand resamplings.

5. **Skewness** and **Kurtosis** coefficients are two measures that can be interpreted in terms of properties associated with the *shape* of the distribution. However, they are very sensitive to outliers and can give quite misleading information.

   A much more reliable method of determining skewness and "tail" weight (kurtosis) is to examine a graphical display of the data such as a histogram or Q–Q plot.

Correlation Measures of Association Between X and Y

The correlation coefficient r is a dimensionless measure of the **linear** association between two quantitative variables. The linearity distinction is important since a value of r near zero implies that no significant correlation exists between X and Y — it does not necessarily mean there is no *association* between these variables since the association could be nonlinear. Furthermore, r only measures pairwise dependency; it cannot detect dependencies that may exist between multiple variables.

With the warning label now in place, Pearson's (product moment) correlation coefficient is defined by

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} .$$

It can be shown that $-1 \leq r \leq 1$ with perfect linear association occurring when $r = \pm 1$. The concept of correlation can also be applied to variables consisting of *ordinal* data such as those obtained from survey data. In this situation, the measure of linear association is between the "ranks" of the two variables and Spearman's rank correlation coefficient is often used. The simple version is defined by

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} ,$$

where $d_i = \text{rank}(X_i) - \text{rank}(Y_i)$. When there are "ties" in the ranks of either group, then Pearson's product formula can be used with

rank(X_i), rank(Y_i) replacing X_i, Y_i. Alternatively, the simple version can be adjusted[6] to account for the ties.

Remarks

1. Pearson's (product moment) correlation coefficient provides a measure of the direction and strength of the *linear* association between two quantitative variables. When the variables are ordinal, Spearman's correlation should be used, since this is based on the ranks of each variable. This avoids the problem of assuming the actual interpoint distances of an ordinal scale are quantitatively meaningful.

2. Note that clients may sometimes use the term "correlation" in a more colloquial sense implying causation. This can lead to misconceptions about what the results of a (statistical) correlation analysis really represent.

3. In practice, observational studies can yield low correlation values, $|r| < 0.3$. The statistical significance of r can be assessed under the assumption that no (linear) association exists between X and Y. Under this assumption, the standard error of r is approximately $n^{-1/2}$ so $|r| > 2n^{-1/2}$ would suggest a significant correlation does exist.

4. If n is large then it is possible to obtain a "significant" correlation with a low value of r. The regression interpretation of r^2 is useful in this situation: *The variability in Y explained by X is equal to r^2*. Hence $|r| < 0.3$ implies that X explains less than 10% of the variability in Y.

5. One particular measure that is often requested with regard to survey test data is Cronbach's Alpha statistic (Cronbach 1951). This is used to estimate the reliability of the survey instrument with higher values of this coefficient supposedly indicating good reliability.[7]

---

[6] See Press et al. (1992; p.640).

[7] Survey instruments often contain two versions of the same question (separated by other questions). Thus, if many respondents "agree" with the question, *"Do you feel confident about $< x >$?"* but also "agree" with *"Does $< x >$ cause you to lack confidence?"* then the reliability of the survey instrument would be questionable.

Transformations

- Symmetry
- Spread-vs.-Level
- Residual Analysis

The dataset that is actually used for analysis is often several stages removed from the original or raw data values. The process of data *cleaning* "transforms" the original dataset by performing tasks such as removing errors, adjusting outliers, estimating missing values, encoding categorical variables, and standardizing variables. Simple operations such as the sum, difference, or ratio may also be applied to create new variables and in some applications, variables are automatically transformed by an "accepted" procedure. For example, stock *returns* are defined as $R_t = \log Y_t - \log Y_{t-1}$, where Y_t is the stock price. After the data processing phase has been completed, further transformation may be appropriate for statistical purposes. Some common reasons for transforming are:

- The distribution of a variable is strongly skewed.
- There is a spread-level effect across batches.
- The residuals from a fitted model exhibit a systematic pattern.
- The data do not satisfy the assumptions of a statistical procedure.

The main difficulty that arises in these situations is the presence of *nonlinearity* which can substantially increase the complexity of the statistical analysis. By applying a nonlinear transformation to the data, we may be able to alleviate these problems and produce a structure that supports an uncomplicated analysis, allowing simpler interpretation of the results.

Power Transformations The class of power transformations is the simplest and most widely used technique for the *reexpression* of a dataset that consists of positive values. These transformations have the form: $T_p(x) = \text{sign}(p)x^p$ for $p \neq 0$ and $T_p(x) = \log x$ for $p = 0$, where $\text{sign}(p) = \pm 1$ is used to preserve the order of the data in the dataset. A slightly more specialized version is the **Box–Cox** class of power transformations defined below. This incorporates the log_e transformation as a special case[8] which is useful for comparing the properties of different transformations.

[8] It can be shown that $(x^p - 1)/p \rightarrow \log_e x$ as $p \rightarrow 0$.

$$Y_i(p) \equiv T_p^*(X_i) = \begin{cases} \frac{1}{p}(X_i^p - 1) & , p \neq 0 \\ \log_e(X_i) & , p = 0 \ . \end{cases}$$

In practice, a "nice" power is used to avoid interpretational problems. Transformations that can be given a contextual rationale will have much more appeal to a client than a power such as $p = 0.43$. Table 3.1 lists some commonly used powers for transforming a dataset. The purpose of the transformation is to try to induce symmetry in the distribution of the transformed data.

TABLE 3.1. Transformations to Symmetry

		Original Dataset	
Transformation	p	Shape	Application
Square (X)	2	Left skew	Upper threshold
Identity	1	Symmetric	No change
Square Root	0.5	Right skew	Area measure
Cube Root	$\frac{1}{3}$	Right Skew	Volume measure
Logarithm	0	Right skew	Population growth
Reciprocal Root	-0.5	Right skew	Inverse area
Reciprocal (X)	-1	Right skew	Extreme outliers

Symmetry An important application of power transformations is to promote *symmetry* in the distribution of a variable. This has the desirable property of providing a natural "center" to the dataset which makes estimates of location conceptually easier to understand. There are also theoretical advantages associated with working with location estimates based on symmetric distributions. The Box–Cox transformation derives from their procedure for choosing p under the assumption that the transformed data are normally distributed (Box and Cox 1964). In the spirit of EDA, we describe a method due to Hinkley (1977) which is simpler to implement and can be made resistant.

Hinkley's Transformation to Symmetry Hinkley (1977) developed a method for selecting a suitable transformation to symmetry of a data sample based on the simple premise that random samples obtained from a "symmetric" distribution will tend to reflect the identity: **mean = median**. Hence, a relative measure of the asymmetry in a sample is:

$$D_p = \frac{\text{mean}\{Y_i(p)\} - \text{median}\{Y_i(p)\}}{\text{spread}\{Y_i(p)\}} \ ,$$

where $Y_i(p)$ is the transformed value of X_i and $D_p = 0$ implies symmetry. Two choices for the "spread" statistic are S and the IQR. A trimmed mean could also be employed to obtain resistance.

D_p Plot The diagnostic consists of plotting D_p versus p, where p goes on the x-axis and solving $D_p = 0$ by interpolation.

Spread-vs.-Level For K batches of related data, a spread-versus-level plot provides a method for *stabilizing* spread across batches via a common power transformation. It is closely related to the classical problem of variance stabilization wherein $\sigma_X^2 = g(\mu_X)$ and a transformation $Y = \psi(X)$ is sought such that σ_Y^2 is approximately independent of μ_Y. In the following diagnostic, introduced by Tukey (1977), the dependency between the IQR and median is assumed to have the form $IQR_X = kM_X^b$, where k, b are constants. Hoaglin et al. (1983) show that the function $\psi(X) = X^{1-b}$ stabilizes the IQR and that b corresponds to the slope of the following plot.

Spread-vs.-Level Plot Plot $\log(IQR_j)$ versus $\log(M_j)$ for batches $j = 1, 2, \ldots, K$, where $\log(M_j)$ goes on the x-axis. Then a suitable power can be obtained by taking $p = 1 - b$, where b is the slope of a straight line fitted to these points.

Remarks

1. Applying a transformation for analysis purposes is sometimes met with strong resistance.

   - The dataset has already been "set up" for the analysis — why the need to further complicate the analysis?
   - Transforming will make the interpretation more difficult.

   It can be quite difficult to persuade a skeptical or resistant client of the need to transform the data and in some cases they may still remain unconvinced even when the benefits of transforming are demonstrated. In our experience, explaining the results based on the transformed data can be easier than trying to advocate the reasons for a transformation.

2. Increasing spread with increasing level in batch data is often combined with right skewness. In this situation, the serendipitous effect of applying a transformation is that both sources of nonlinearity are alleviated. However, there are situations where transformations will not always work:

   - A data structure that exhibits features such as multimodality, or both left and right skewness, will not benefit from a power transformation.

- A power transformation cannot be applied (directly) to a dataset that contains negative values. In some cases, adding a constant or applying separate transformations to the negative and positive values "may" be appropriate.

3. Many statistical procedures assume that an additive model:

   DATA = MODEL + ERROR

   can be used to analyze a data structure. Transforming the DATA can sometimes help promote an additive structure by removing interaction effects between the MODEL and ERROR and stabilizing the ERROR variance. For example:

   - Y versus X scatter plot — applying a power transformation to one or both variables can allow a straight line model to be fitted which is easier to analyze and employ for predictive purposes.
   - ANOVA — The assumption of equal variances in an *analysis of variance* model corresponds to an assessment of the spread-versus-level plot.

Contingency Tables

The methodology of contingency tables is directed at the analysis of the association between the levels of two categorical variables. The standard method of analysis is to perform a test of *independence* by computing Pearson's chi-square statistic or, in the case of small samples, Fisher's exact test. Both are discussed in detail below. The **Crosstabs** panel in Figure 3.5 is an example of a contingency table with the **Tests** panel containing the results of Pearson's chi-square test. Case Study 6.1 and the Case Study Exercise 9.5 both provide examples of the applications involving frequency tabulations between two categorical variables.

Contingency tables can be generated from different types of studies and the nature of the particular study needs to be considered. For example, in longitudinal studies (Section 3.1), the interpretation of the association between two categorical variables X (rows) and Y (columns) will depend on whether the study is retrospective or prospective. There is also another variation called *cross-sectional* studies. The differences between these three types of studies are indicated below.

Retrospective The column totals for Y are fixed and the distribution of the X levels is of interest. For example, 50 smokers and 50 nonsmokers are retrospectively classified by gender.

Prospective The row totals for X are fixed and the distribution of the Y levels is of interest. For example, 50 smokers and 50 nonsmokers are

selected for a longitudinal study of the incidence (yes or no) of lung cancer.

Cross-Sectional The total sample size is fixed and the joint distribution of X and Y is of interest. It is assumed that X and Y are "meaningfully" related.

Chi-Square Test

Pearson's chi-square statistic can be found in most introductory texts (see, for example, Moore and McCabe (1998)) and provides a test for the null hypothesis of no association (independence) between the two variables. However, the chi-square test should only be used when the sample size is sufficiently large. In practice, the usual criterion is that no more than 20% of the cells should have expected counts less than 5. Although the following details will be familiar to most readers, for completeness we illustrate the inference process with respect to this particular test statistic.

Details of the Chi-Square Test

For notational purposes, let A, B denote two categorical factors with n_a, n_b levels (classes), respectively. The contingency table associated with a sample of observations that are crossclassified by the levels of these factors consists of $n_a \times n_b$ "cells" with n_{ij} observations in each cell. Pearson's chi-square statistic is computed as

$$W = \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} ,$$

where $O_{ij} = n_{ij}$ is the observed cell count (frequency) and E_{ij} is the *expected* cell count.

Under the assumption that A, B are independent, $E_{ij} = n_{i\bullet} n_{\bullet j}/n_{ij}$, where $n_{i\bullet}$, $n_{\bullet j}$ are the marginal totals for the A_i row and B_j column. The assumption of independence is equivalent to the probability statement $P[A_i \cap B_j] = P[A_i]P[B_j]$ which can be expressed as the hypothesis:

H_o : *there is no association between* A *and* B.

For large samples, it can be shown that under H_o (the hypothesis is correct), W has an approximate chi-square distribution with $(n_a - 1)(n_b - 1)$ degrees of freedom. This is often abbreviated as: $W \sim \chi^2_{(n_a-1)(n_b-1)}$.

Since any difference between O_{ij} and E_{ij} will contribute positively to the test statistic, a large value of W would suggest evidence *against* H_o. Thus, a significant result (P-value $= P[\chi^2_{(n_a-1)(n_b-1)} > W] < 0.05$) would support the conclusion that an association does exist between the factors A and B. As mentioned above, the chi-square test should only be used when the sample size is sufficiently large. Combining levels may be necessary to avoid having too many $E_{ij} < 5$.

Fisher's Exact Test

Fisher's exact test was originally applied in the 2×2 case, where the exact sampling distribution under the assumption of independence, conditional on fixed marginal totals, can be shown to be the hypergeometric distribution (see Appendix C, Table C.1). Thus, the advantage of Fisher's exact test is that it can be applied even when the sample size is small.

For 2×2 tables, the P-value for Fisher's exact test is determined from the hypergeometric probability associated with the (1,1) cell count. Since the marginal totals are fixed, variations in the cell counts from the original 2×2 table can be ordered by the *odds ratio* given by

$$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Thus, one- and two-sided P-values can be computed that correspond to the alternatives associated with the equivalent hypothesis of independence: H_o' : $\theta = 1$. The two-sided P-value can be defined as the sum of hypergeometric probabilities of all the tables (variations) that are no more likely to occur than the observed table.

For small samples, however, the P-values will be highly discretized and testing at a fixed level of significance is usually not possible.[9] Intuitively, of course, there is always the question of whether a statistical analysis of a small sample is necessarily meaningful. Usually power calculations can provide a measure of the quality of a test, but this is somewhat more complicated with Fisher's exact test (see, for example, Gail and Gart (1973)). Instead, confidence intervals for the odds ratio θ can be calculated. The software package, StatXact, provides exact confidence intervals for θ.

Fisher's exact test for independence can be generalized to any $n_a \times n_b$ contingency table, albeit at a rather substantial computational cost in some cases. This is **not** the sort of test to run on several arbitrary sized contingency tables just for exploratory purposes. Fisher's exact test is discussed in further detail in Agresti (1990).

Other Tests

There are many other tests and measures of association that can be employed for contingency tables. For a more detailed summary, the reader is referred to the discussion presented in the `PROC FREQ` section of the SAS manual (SAS 1990).

---

[9] By "highly discretized" we mean that only few possible P-values can actually occur. Hence comparing the P-value to 0.05 is somewhat misleading since that boundary may not be achievable for the given sample size. Randomized tests could be employed, but then we need to justify (to the client) why we tossed a (biased) coin to decide whether a result was significant!

Interpretation of Significance

Another important issue in the analysis of contingency tables is the interpretation of the results when the calculated P-values from any of the above tests are significant. In these cases we search the table for clear patterns of association between the variables that explain the finding, but it is quite possible that such patterns are unclear or that there may be competing interpretations. For 2×2 tables the interpretation is easy since the only interesting patterns are to have higher than expected frequencies in the diagonal or in the off-diagonal cells. But for larger tables the interpretation of the relationship may be difficult in some cases, even when the test shows a significant relationship. We return to this point in Case Study 6.1 with regard to the analysis of Tables 6.1 and 6.3.

Remarks

1. We have presented the analysis of contingency tables in more detail than was perhaps necessary primarily because this type of analysis is frequently (!) employed in discrimination cases. As is apparent from Case Study 6.1, trying to explain the application and interpretation of Pearson's chi-square test or Fisher's exact test to an opposing lawyer can be a rather frustrating experience.

2. Other clients will be much more receptive than the opposing lawyer above, but they will still struggle with the same details. The `row %` and `col %` entries shown in the **Crosstab** panel of Figure 3.5 provide an easy way to illustrate the assumption of independence (they will be approximately the same if H_o is accepted; otherwise there will be obvious differences).

3. If there is evidence of a significant association, the client may need to be "walked" through the `row %`, `col %`, and `total %` entries. Being able to determine significance via the P-value doesn't automatically mean the client will be able to interpret the association. This is particularly important when a variable contains a category such as "Other." In this case, the significant association may not have a meaningful interpretation.

The t-Test

In addition to the one-sample and two-sample t-tests, the t-test is used (and abused) extensively in many other statistical procedures such as regression and analysis of variance. There is also a *paired* t-test which can be employed in *before-and-after* experiments as illustrated by the following example.

Example 3.4 *Double-blind experiment*

In a placebo versus active drug study, the response of each patient is measured for the two drugs at separate times. The order of the drugs is randomized and it is assumed that similar conditions[10] exist at both times. To avoid "interview" bias, neither the patient nor the person who administers the drug to the patient should be aware of what drug is given. This is referred to as a Double-Blind implementation of a *before-and-after* experiment.

Paired *t*-Test A paired t-test may used to analyze the differences between the two responses from each patient. Since this paired design accounts for a patient's baseline (natural response to the condition of interest), the average difference ($\overline{D} = \overline{X}_{\text{after}} - \overline{X}_{\text{before}}$) will be less sensitive to the large within-group variability that often exists in these types of experiments.

Two-Sample *t*-Test A simpler implementation would be to use two independent groups. With randomization, each patient receives only one of the drugs and a two-sample t-test can be used to assess the significance of the difference of the two response means. However, if the responses within each group are highly variable (as is common), this will dominate the experimental error and mask a meaningful difference between the group means.

In the next chapter, we follow a detailed (consultation session) presentation of the results from a two-sample t-test to a client. The two-sample t-test is also employed for the preliminary analysis in Case Study 6.3 and Appendix C contains tables of some standard test procedures, including the one- and two-sample t-test. Thus, we focus on some of the statistical properties of the two-sample t-test. We first consider the two-sided and one-sided forms of the *alternative* hypothesis for this test.

Two-Sided Test The statistical translation of a client's two-sided t-test of the difference between the means of two independent groups, A and B say, is:

$$H_o : \mu_A = \mu_B \text{ versus } H_1 : \mu_A \neq \mu_B \ .$$

This is usually a test of the status quo: nothing has changed versus something has happened. The P-value of this test is $P[\,T > |t_o|\,]$, where $t_o = (\overline{x}_A - \overline{x}_B)/se(\overline{x}_A - \overline{x}_B)$ is the observed value of the test statistic. We discuss the standard error term $se(\overline{x}_A - \overline{x}_B)$ shortly.

One-Sided Test A one-sided test is often stated as:

---

[10]The assumption of "similar conditions" implies that there is no residual or carry-over effect from the first drug.

$$H_o : \mu_A = \mu_B \text{ versus } \begin{cases} H_1 : & \mu_A > \mu_B \quad \text{(right-tailed)} \\ H_1 : & \mu_A < \mu_B \quad \text{(left-tailed)} \end{cases}.$$

The key issue is whether testing in only one direction is warranted. The direction **cannot** be decided postexperiment. The form of the H_o represents the point at which the Type I error of the test will be the largest. The test statistic is the same as t_o above, but the P-value is $P[T > t_o]$ for the right-tailed test; $P[T < t_o]$ for the left-tailed test.

Standard Error of the Two-Sample t-Test

This is where a client is most likely to get confused. Why are there *two* t-tests? Is one of them testing something else? How can you have "fractional" degrees of freedom? Here, the consultant needs to be careful not to get too technical since a significant result from either test has the same interpretation in terms of providing evidence that the process means appear to be different.

One version assumes the two samples came from independent (normal) processes that have *equal* variances ($\sigma_A^2 = \sigma_B^2$). In this case, the pooled sample variance estimate is used (see Appendix C, Table C.6) and $T \sim t_{n_A+n_B-2}$ under H_o. In the second version, the variances of the normal processes can be *unequal*, but the test statistic does **not** have a t-distribution under H_o. The Smith–Satterthwaite test approximates the null distribution by t_ν, where the degrees of freedom parameter is computed from a weighted average of the sample variances (see Table C.6).

The key issue is whether it is reasonable to assume the process variances are equal. The sample standard deviations provide empirical evidence, but a contextual rationale would be preferable. (For example, measuring male and female performance on a science test: there may be a gender difference, but to assume that *variability* in performance would be the same for both groups seems reasonable.) This issue is discussed again in the next chapter.

Remarks

1. Prior to interpreting or performing a t-test, of course, we need to check that the conditions of the test are supported by the data. Do *not* forget to check simple things such as: do the data constitute a random sample? Are the two samples independent?

2. The two-sample t-test is robust to departures from distributional assumptions. Moore and McCabe (1993) provide an excellent illustration of this robustness property. Their simulations show that the two-sample t-test will provide reliable inference even when the distributions are far from normal.

3. There are situations where performing a t-test is "silly." For example, the sample size is so large ($n > 1000$ say) that significant results are

to be expected. The "significance" of these results, however, may not be meaningful. Another example is where one of the samples is very small. In this case, the power of the two-sample t-test may be too poor to provide a reliable result.

Analysis of Variance (ANOVA)

To compare several group means, the analysis of variance (ANOVA) procedure provides a method for assessing significant differences between (some of) the means. It is a general technique pioneered by Fisher for partitioning the overall variability of a response variable into components associated with one or more *factors* and with the random error. The literature on the ANOVA technique is extensive and our presentation is only able to touch on some of the issues associated this methodology. Montgomery (1997) and Box et al. (1978) provide good coverage of this methodology with relevant examples.

One-Way ANOVA

In the simplest case of a single factor or *treatment* with two or more levels, the one-way ANOVA procedure generalizes the two-sample t-test. That is, instead of comparing $I > 2$ treatment level means in "pairs," the ANOVA procedure tests the null hypothesis of no treatment effect $H_o : \mu_1 = \mu_2 = \ldots = \mu_I$, directly. The (simultaneous) comparison of the treatment level (group) means is performed by comparing the variability **between** the group means with the (error) variability **within** the groups. Thus, a significant result from a one-way ANOVA simply provides evidence of a difference between the group means; it does not indicate "how" the group means differ.

The results from applying the ANOVA procedure can be presented in the form of an "ANOVA table." For a balanced one-way design, with I treatment levels and K observations per level, the ANOVA table summarizes the treatment effect, A say, in terms of the sum of squares decomposition of the response (Y) variation: SST = SSA + SSE, where SST = $(n-1)s_y^2$ and $n = IK$. Thus, the terms SSA and SSE correspond to the decomposition of SST into the "between group" and "within group" variation, respectively.

Source	DF	SS	MS = SS/DF	F-Value
Treatment A	$I-1$	SSA	MSA	$F_o = \dfrac{\text{MSA}}{\text{MSE}}$
Error	$I(K-1)$	SSE	MSE	
Total	$n-1$	SST		

The F-value defined by F_o provides a test statistic for assessing the treatment effect. To obtain a meaningful comparison, F_o is contructed using mean sums of squares (MS = SS/DF), where DF denotes the degrees of freedom parameter. MSA and MSE therefore represent the "average" variability associated with the treatment and error components, respectively. If normality and constant variance across the treatment levels can be assumed then $F_o \sim \mathcal{F}_{I-1,I(K-1)}$ under the null hypothesis H_o : *no treatment effect.* Any difference between the observed treatment level means will contribute positively to SSA (and hence subtract from SSE since SST is fixed), so a large value of F_o provides evidence against H_o.

The P-value for the ANOVA test of H_o is $P[\mathcal{F}_{I-1,I(K-1)} > F_o]$ which is included in the ANOVA table output by most statistical software packages. Thus, if the one-way ANOVA model is adequate for the data and H_o is rejected, a posthoc analysis can be performed to investigate the relationship between the treatment level means. These issues are dealt with in the two-way ANOVA procedure discussed below.

Two-Way ANOVA

Two-way ANOVA is employed when the purpose of the study is to investigate the effect of two factors, A and B say, *simultaneously.* For the statistician, this makes the analysis more interesting since there may be an **interaction** effect (denoted by AB) between the two factors. That is, the effect of factor A on the response may not be the same at all levels of factor B. If this occurs, then comparisons between the treatment level means of one factor (e.g., A) can be obscured by the AB interaction, suggesting (incorrectly) that factor A has no effect on the response. Thus, the variability associated with the AB interaction needs to be examined first.

While these issues may open up the interest of a statistician, for some clients, two-way ANOVA may seem more like opening up Pandora's Box! Repeated measures, blocking factors, crossed and nested designs, interaction effects, Type III hypotheses, a *mixed* effects analysis. ... Too late now. That client just made the statistician's day. In our experience, introducing clients to the methodology and concepts underlying two-way ANOVA can sometimes be a difficult process. Be patient. Spending the time and effort to clarify these difficult concepts for the client is very worthwhile. And the benefit to our clients? ... Well, they did open it, so let's take a look inside.

Pandora's Box

The transition to two-way ANOVA is often predicated by studies where one of the factors has only two levels (e.g., control versus treatment). The client may simply want to know whether there is a "statistically significant difference" between the two levels with respect to the response variable. ... And, "Oh yes. We also need to know if this is true for both males and females." (Oops! The client just opened Pandora's box.)

Multiple *t*-Tests The problem with performing multiple t-tests is that the experimentwise (overall) Type I error is not controlled. That is, the chance of an individual Type I error (*falsely* detecting a significant difference between two levels of a treatment factor) for any comparisonwise t-test is α (typically 0.05). But the chance of making *at least one* comparisonwise Type I error with multiple t-tests can be much larger. For k comparisions, an upper bound for the experimentwise Type I error is $1 - (1 - \alpha)^k$. Since this converges to 1, it follows that the *simultaneous* use of multiple t-tests should **not** be used to assess the significance of ANOVA factors. Instead, the ANOVA hypothesis H_o : *no treatment effect*, should always be tested before any comparisonwise (posthoc) analysis is performed.

In our Pandora's box scenario, this essentially means that there could be (up to) a 25% chance[11] of *falsely* detecting a significant difference between some combination of the male/female and treatment/control levels. That is, these t-tests do **not** provide us with reliable evidence of a significant effect due to gender, or the control/treatment factor. Hence we need to carry out the two-way ANOVA analysis.

One Factor at a Time This approach does have a certain intuitive appeal: to investigate the effect of one factor on the response, just hold the other factor constant. However, the problem with the one factor at a time approach is that this method is not only less efficent than the corresponding two-factor factorial design, the results may be quite misleading when *interaction* is present. This is demonstrated in the following example.

Example 3.5 *One factor at a time versus the* 2×2 *factorial design.*

One Factor at a Time

	B_1		B_2	
A_1	8	12	38	42
A_2	47	53		

2×2 Factorial

	B_1	B_2
A_1	10	40
A_2	50	5

As illustrated by the tables in Example 3.5, the advantage of the 2×2 design is that the estimates of the main effects for A and B are just as *precise* as those obtained from the one factor at a time design, but only four observations are required. More importantly, the conclusion that the unobserved A_2B_2 combination in the one factor at a time

[11] Actually, it is $0.2649 = 1 - 0.95^6$, but the client probably wouldn't have considered performing the two interaction t-tests. The main point is that it helps us emphasize the problem that arises when the experimentwise error is not controlled.

design might produce an even larger response, may be completely wrong if interaction is present.

Randomization The above example may seem somewhat contrived as the A_2B_2 combination would probably have been run anyway. (If for no other reason than to see what the responses were.) Hence, this experiment would also allow the interaction effect to be estimated since we now have *two* replicates of a 2×2 factorial.... Or do we? This depends on whether *randomization* was employed correctly in the experiment. For the statistical analysis of a two-factor factorial experiment to be considered valid, *all* of the treatment-level combinations need to have been run in random order. Determining precisely "how" the observations were collected is therefore important.

To better illustrate the problem, suppose factor B has $k > 2$ levels. In the one factor at a time approach, it is quite possible that the two-level factor A is held constant (at A_1 say), and the response observed at each level of factor B. The same procedure is then used with the A_2 level held constant. Even if the levels of B were run in random order, the results do **not** conform to a factorial design since the randomization was restricted to "treatments *within* blocks." That is, A has effectively become a blocking factor and this affects how the statistical analysis can proceed.

Fixed Effects Design

The ANOVA table for a balanced two-way **fixed effects** factorial design, with I treatment levels for factor A, J treatment levels for factor B, and K replicates, is presented below. The assumption the both treatment factors are "fixed" means that inferences drawn from this analysis are applicable only to particular levels of A and B chosen by the investigator. In similar fashion to the one-way case, the ANOVA table summarizes the **main** effects of A and B, and the interaction effect AB, in terms of the sum of squares decomposition of the response variable variation.

Source	DF	SS	MS	F-Value
Treatment A	$I-1$	SSA	MSA	$F_A = \dfrac{\text{MSA}}{\text{MSE}}$
Treatment B	$J-1$	SSB	MSB	$F_B = \dfrac{\text{MSB}}{\text{MSE}}$
Interaction AB	$(I-1)(J-1)$	SSAB	MSAB	$F_{AB} = \dfrac{\text{MSAB}}{\text{MSE}}$
Error	$IJ(K-1)$	SSE	MSE	
Total	$n-1$	SST		

The two-way ANOVA model can be written in the form

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk} \ , \tag{3.1}$$

where μ_{ij} represents the combined effect of the i,jth level of A and B on the response y_{ijk}, and $\epsilon_{ijk} \sim N(0, \sigma^2)$ independently. The test statistic for an individual effect can be constructed using the reparameterization: $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, where μ is the overall mean response, α_i is the effect of the ith level of factor A, β_j is the effect of the jth level of B, and γ_{ij} is the effect of the interaction between α_i and β_j. If the assumptions underlying this model are satisfied by the data, then the null distribution of the corresponding F-value for each effect is $\mathcal{F}_{\text{DF(effect)},\text{DF(error)}}$.

As in the one-way case, a large F-value provides evidence against the null hypothesis for the corresponding effect. However, the AB interaction test should *always* be examined first. The reason for this is that there is little point in testing H_A or H_B if H_{AB} : *no interaction effect* is rejected, since the difference between any two levels of a main effect also includes an "average" interaction effect. In other words, regardless of the test results for H_A and H_B, "both" treatment factors are important if there is a significant interaction effect. (How could the interaction exist otherwise?)

Diagnostics

Before adopting any conclusions based on the ANOVA table, we need to check model adequacy. EDA methods can be employed to examine the response data directly, but we really need the "residuals" from the fitted ANOVA model in order to perform diagnostic checks of model adequacy. From (3.1), it follows that the residuals may be calculated as $e_{ijk} = y_{ijk} - \overline{y}_{ij}$, where $\overline{y}_{ij}$ is the average over the K responses in the i,jth cell.

The analysis of residuals plays a fundamental role in checking model adequacy. If the ANOVA model is adequate for the data, the residuals should not exhibit any obvious structure. For inference purposes, the errors are assumed to be normally distributed with constant variance. The following diagnostics are used to check whether these features are reflected in the sample of residuals.

Normality The Q–Q plot can be used to graphically assess departures from normality. The Shapiro–Wilk statistic provides a formal statistical test for normality. For small to moderate size experiments (i.e., total number of runs $n < 30$), these diagnostics should be interpreted cautiously with regard to evidence of nonnormality.

Variance The residual-versus-fitted plot can be used to graphically assess whether the within-sample variability appears to be constant. Plotting the residuals with respect to the levels of a factor can be used to assess whether the variability appears homogeneous across groups. Note that standardizing the residuals before plotting makes it easier

to compare their magnitude to ± 2 (see Figure 3.7). There are also formal "homogeneity of variances in ANOVA" tests that can be employed such as Bartlett's test. We refer the reader to Anderson and McLean (1974) for a review of equality of variance tests.

Posthoc Analysis

The F-value associated with the interaction effect AB needs to be examined first. If the interaction effect is significant, comparisons between the levels of a main effect may be obscured by the AB interaction. An *interaction plot* is a simple graphical display of the response mean plotted against the levels of one factor, A say, at a given level of factor B. The means are joined by lines to create a "profile" of the average response across the levels of A. This is repeated for each level of factor B and the profiles are overlaid on one plot. If the profiles are markedly different, this suggests interaction is present. If all the profiles have similar form, no interaction is suggested.

When the interaction effect is not significant, differences between the levels of a (significant) main effect can be investigated using various multiple comparison methods:

- Fisher's least significant difference (LSD) corresponds to the use of pairwise t-tests. As the number of levels increases (more pairwise comparisons need to be made), the Type I error of the experiment can become large using the LSD method.

- Duncan's multiple range test is one of the more popular multiple comparison methods since it has good power. That is, it is very effective at detecting differences between means when real differences actually exist. This was developed by Duncan (1955).

- Tukey (1953)[12] proposed the studentized range test (HSD) which specifies one critical value for all pairwise comparisons (the highest significant difference). It is more conservative than Duncan's method.

Parameter Estimation

In general, the main focus of the discussion between a client and the statistical consultant would be on the correct interpretation of the two-way ANOVA table and conclusions based on the posthoc analysis. Furthermore, direct reference to the reparameterized version of the ANOVA model (3.1) can be avoided using pseudo-code: $Y = \mu + A + B + AB + error$. However, it is instructive to consider the problem of estimating the parameters of a two-way ANOVA model since this leads to the issue of "estimable functions" and the different types of sums of squares that can be generated from an ANOVA analysis.

---

[12]This citation refers to unpublished notes on *"The problem of multiple comparisons."* However, the studentized range test is described in detail by Scheffé (1959).

As we indicated above, the two-way ANOVA model (3.1) can be reparameterized as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \ , \qquad \epsilon_{ijk} \sim N(0, \sigma^2) \ , \tag{3.2}$$

where, as before, μ is the overall mean response, α_i is the effect of the ith level of factor A, β_j is the effect of the jth level of B, γ_{ij} is the effect of the interaction between α_i and β_j, and ϵ_{ijk} is the random error component. Given sufficient data (i.e., $n = IJK > 1 + I + J + IJ$), it follows that we "should" be able to estimate the model parameters that appear in (3.2). In addition to constructing the hypothesis tests that appear in the ANOVA table, the parameter estimates should also allow us to construct posthoc hypothesis tests of interest. For example, α_1 = `CONTROL` versus *all other treatment levels* $(\alpha_i, i > 1)$.

The problem is that the model is inherently overparameterized and regardless of how much data is available, the parameters in (3.2) are linearly dependent. That is, the usual (normal) equations associated with estimating these parameters do not permit a unique solution unless certain identifiability constraints are imposed. The standard approach is to impose the *sum to zero* constraint which requires that the parameter estimates satisfy

$$\sum_i \hat{\alpha}_i = \sum_j \hat{\beta}_j = 0 \ , \quad \text{and} \quad \sum_i \hat{\gamma}_{ij} = \sum_j \hat{\gamma}_{ij} = 0 \ . \tag{3.3}$$

Under the constraints in (3.3), the resulting parameter estimates obtained from the normal equations are

$$\hat{\mu} = \overline{y} \ , \quad \hat{\alpha}_i = \overline{y}_i - \overline{y} \ , \quad \hat{\beta}_j = \overline{y}_j - \overline{y} \ , \quad \text{and} \quad \hat{\gamma}_{ij} = \overline{y}_{ij} - \overline{y}_i - \overline{y}_j - \overline{y} \ ,$$

where $\overline{y}_\bullet$ denotes the average of the y_{ijk} over the missing subscript(s).

While this solution to the normal equations is intuitively appealing, most statistical software packages do **not** present the parameter estimates in this fashion. Instead, transformations of the original parameters are employed and there are differences in the method of transformation used by each software package. To simplify the discussion we consider the situation where factor A has three levels. The following table indicates the different transformations employed by JMP, S-PLUS, and SAS.

		SAS			JMP		S-PLUS	
Coded Parameter		A1	A2	A3	A1	A2	A1	A2
Original	α_1	1	0	0	1	0	−1	−1
Parameter	α_2	0	1	0	0	1	1	−1
	α_3	0	0	1	−1	−1	0	2

Sum to Zero JMP sets the parameter for the last level equal to the negative sum of the other two levels. This satisfies the sum to zero constraint for the A effect. However, the parameter estimates (and

tests) contained in "JMP Parameter Report" are **not** $\hat{\alpha}_1$ and $\hat{\alpha}_2$. In fact, the correct interpretation of the parameter labeled as `"A[1]"`[13] is $\alpha_1 - (\alpha_1 + \alpha_2 + \alpha_3)/3$.

Helmert Contrasts The default transformations used by S-PLUS are *Helmert contrasts*. These contrasts are equivalent to comparing the ith level of A with the "average" of the preceding $(i-1)$ levels. S-PLUS provides the sum to zero transformation as an option in its analysis of variance function `"aov()"` and also allows user-defined constrasts to be employed in this function.

Estimable Functions SAS does not transform the parameters directly. Instead, SAS searches for *estimable functions* using generalized inverses.[14] In the above example, the solution chosen happens to be such that $A3$ is set to zero. This method computationally more expensive, but allows SAS to generate four types of estimable functions which can be used to construct four types of sums of squares for testing the ANOVA hypothesis.

Types of SSs

The *sums of squares* of the effects in the SS column of the ANOVA table above can be computed in different ways. In SAS, four types of sums of squares are available which are imaginatively called Type I, Type II, Type III, and Type IV SSs. In S-PLUS (Version 6 for UNIX), only Type I and III are available. The underlying difference among these sums of squares is in the construction of "estimable functions" for testing the null hypothesis H_o : *no effect*. We briefly illustrate the differences for the two-way ANOVA model. We refer the reader to Chapter 9 of the SAS/STAT manual (SAS 1990) for further details.

Type I The SS for each effect is computed **sequentially**. That is, the "order" in which the effects are added to the model (3.2) is important. Since factor A was added first, SSA represents the change in sums of squares due to factor A alone: the null model $y_{ijk} = \mu + \epsilon_{ijk}$ versus $y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$. SSB, however, represents the change *after* the effect of factor A has already been accounted for. That is, $y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$ versus $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$. SSAB is the last change due to the effect of AB which compares the previous model with (3.2).

Type I hypotheses therefore test the "incremental" effect of each new term in the model, conditional on all earlier terms being in the model.

---

[13] Version 4. In JMP 3.x, this parameter was even more confusingly labeled as `"A[1-3]"` making it seem like the corresponding test was of α_1 versus α_3.

[14] We refer the reader to SAS (1990; Chapter 4) for specific details.

This would be appropriate for purely nested models and polynomial regression models.

Type II For balanced designs, main effects models, and pure regression models, Type II hypotheses are appropriate. Types II, III, and IV SSs are referred to as **partial** sums of squares since they provide a measure of the contribution of an effect, factor A say, after accounting for all the other effects that do not contain A. Hence, the Type II SS for factor A compares the models: $y_{ijk} = \mu + \beta_j + \epsilon_{ijk}$ versus $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$. (The interaction term AB is not included since this contains the effect of A.) Except for the models listed above, Type III SSs are preferred over Type II SSs.

Type III, IV Type III SSs are constructed in similar fashion to Type II SSs, except that the hypotheses are based on estimable functions that are orthogonal. That is, the Type III hypotheses for A is based on estimable functions that are orthogonal to those of AB. This means the Type III hypothesis are still appropriate when the design is unbalanced or contains missing cells (no observations for certain factor level combinations). Type IV hypotheses correspond to Type II hypotheses, but Type IV can be applied when there are missing cells. The difference between Type III and IV SS is in the way estimable functions are selected. Type IV tries to make these functions "balanced" whereas Type III makes them orthogonal.

Random Effects

The F-tests shown in the two-way ANOVA table above are based on the assumption that both treatment factors are **fixed**. That is, inferences drawn from this analysis will be specific to the particular levels of A and B chosen by the investigator. When the levels of a factor are randomly selected from a range of possible values, however, inferences drawn about the effect of that factor will be applicable to *all* levels. Since the effect of this factor will depend on which levels are selected, it is referred to as a **random** effect.

Although the basic construction of the ANOVA table remains unchanged, the F-tests need to be modified to account for the additional variability introduced by a random effect. The expected mean squares are used to construct the appropriate test statistics in analysis of variance problems. For the two-way ANOVA model, the appropriate F-tests are:

Both Random MSAB is used instead of MSE in the denominator of both F_A and F_B.

Mixed Models If A is fixed and B random, then the interaction effect AB is also random. However, different constraints can be imposed on the γ_{ij} in (3.2) which affects how F_A and F_B are modified. This is discussed in more detail in Case Study 8.1 where the appropriate F-tests are presented.

Remarks

1. Phew! ... Do we really expect the client to understand every detail we covered about ANOVA?

   No, not "every detail" of course, but it is important that we are able explain why we are using Type III hypotheses, a random effects model, or Duncan's multiple range test. Clearly, the client does not need to know the technical details of these diagnostics, but they do need to know why they are being used.

   The main point that we have tried to make here is that two-way ANOVA is a "nontrivial" method of analysis, both statistically and conceptually. However, it is employed extensively in many different fields of study (even if the client is unaware of this), and the consultant's role is to guide the client through the design and analysis. As we stated earlier, this may not be an easy process. Be patient. Our clients are not incapable of making the transition to two-way ANOVA; they just need some guidance.

2. Our description of the ANOVA procedure is clearly contingent on the experimental design and the factors involved in the study. For example, one of the factors in the two-way ANOVA may be a "blocking" factor. In Section 3.7, we examine some specific block designs in more detail. As we have already mentioned (Section 3.1), one of the critical issues when dealing with a client's postexperiment study, is finding out precisely "how" the data were collected.

3. In practice, experiments often result in "unbalanced" data: that is, the number of observations per cell (levels of A in a one-way ANOVA, or AB in a two-way ANOVA). Unless the design is far from balanced, this should not present a problem in terms of the analysis.

4. The ANOVA method can obviously be extended to examine the effect of more than two factors. However, the number of factor level combinations can grow rapidly and specialized experimental designs should be employed instead. In some experiments, the levels of one factor may be **nested** within each level of the other factor. Again, we consider more specialized experimental designs in Section 3.7.

5. In the above presentation, we avoided explicit definitions of the sums of squares terms. While these formulas are certainly not that hard to explain to a client, there is the obvious question of relevancy: the result displayed in the ANOVA table will be a number, not a formula! A more productive exercise would be to explain the "concept" of the ANOVA procedure in terms of, for example, parallel boxplots. This leads the client towards the important notion of checking model adequacy *before* adopting conclusions based on the results of the ANOVA table.

6. The degrees of freedom (DF) parameter has already appeared in our descriptions of contingency tables and t-tests. Apart from the (approximate) two-sample t-test where the variances are not assumed to be equal, clients who require assistance with these methods are often indifferent to the DF issue. However, this is not always the case with ANOVA and, occasionally, clients do get "hung up" over the DF parameter. In this situation, introducing the concept of observations as pieces of "information" can be helpful. That is, we need to explain what it an "effective" sample size means with regard to modeling data.

7. There are many other aspects of two-way ANOVA that we have not addressed (such as the situation when a factor has ordinal levels: use orthogonal polynomials). However, there is a limit to how much information we should regurgitate and our emphasis in this chapter is to present information on statistical methodology from the perspective of a statistical consultant attempting to explain, for example, two-way ANOVA to a client. Thus, we conclude our description of the ANOVA procedure with the following:

> Respect the client's study. Our problems depend on their research efforts.

Regression

Regression analysis is concerned with the problem of fitting a linear model to describe the relationship between a quantitative response variable Y and one or more explanatory[15] variables. The existence of such a relationship could be guided by background knowledge of the variables involved or motivated by empirical evidence from previous studies. In practice, regression analysis is often used on observational data where such knowledge may be tenuous at best and the relationship between the response and explanatory variables (if any) is unlikely to be exactly linear. There is also a tendency for clients to want to include as many explanatory variables as possible in the regression model — "just in case they might be important." Despite its potential for misuse, regression analysis is a powerful tool and can provide useful insights about the process and variables under study.

Many texts on regression also deal with ANOVA. Seber (1977) covers the standard theory of regression and the text by Draper and Smith (1981) is almost synonymous with the term regression. Regression diagnostics are considered in detail in Besley et al. (1980).

---

[15]Synonymous terms are: regressor, predictor, input, and independent variables.

Multiple Linear Regression

The multiple linear regression model has the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \ ,$$

where ϵ is the random error component. For inference purposes, ϵ is usually assumed to be independently normally distributed with zero mean and constant variance σ^2. A single "observation" consists of the k explanatory values together with the observed response value:

$$(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i) \ .$$

The linearity imposed by this model is less restrictive than it might first appear. Simple transformations of the response variable can be employed and the explanatory variables can be defined as functions of other variables. This includes, for example, indicator or dummy variables for categorical predictors, polynomial terms ($X_j = X^j$) in curvilinear models, and interaction terms (e.g., $X_3 = X_1 * X_2$) in response surface models. In the simplest case when only one explanatory variable is involved ($k = 1$), this model is referred to as simple linear regression (SLR).

Preliminary Analysis

Prior to the actual fitting process, there are several important issues that first need to be addressed to ensure the regression methodology is applied appropriately. This requires interacting closely with the client during a consulting session when the overall issues and objectives of the project are considered. The following regression-specific questions are illustrative of the type of verbal communication skills outlined in Section 2.1. Before considering any of these questions, however, there is one fundamental rule in regression analysis:

Plot the data.

- What is the purpose of the regression model?
- Are all the explanatory variables really necessary?
- Is there a problem with missing values?
- How should outliers be treated?
- Are any of the variables ordinal or qualitative?
- Do certain constraints exist in this process?
- Is there a potential time series effect?

Plots Always, always look at plots of the data. Scatter plots of Y versus each of the X_j can provide invaluable information about the strength of each explanatory variable as a predictor and can reveal important features such as outliers, nonlinearity, and clustering. Pairwise plots (X_j versus X_k) can also be useful for revealing potential redundancies in the form of high correlation between two explanatory variables. Of course, collapsing multivariate data onto two dimensions will not provide a "complete" picture of the regression sample, and can sometimes be misleading. Thus, the purpose of these scatter plots is exploratory — they should not be used as a means for deciding whether to eliminate explanatory variables.

Purpose If the model is to be used for predictive purposes then it needs to be of sufficient quality in order to provide reliable predictions. A sufficiently large database and background knowledge about the process is clearly desirable in this case.

Number of Predictors Overfitting[16] a regression model has the adverse effect of inflating the variance associated with the parameter estimates. Multicollinearity problems may also arise due to linear dependencies between the explanatory variables. Justifying the need for each explanatory variable at the outset can be a worthwhile exercise.

Missing Values Statistical software procedures routinely eliminate all observations containing missing values before starting the requested computation. Explanatory variables containing a large proportion of missing values should therefore be dropped from the regression model (or recoded, provided this makes sense).

Outliers Influential points such as outliers and high leverage points can have a substantial impact on the fitted model sometimes giving results that are nonsensical in the context of the process. Preliminary detection of potential outliers from individual summary statistics and pairwise scatter plots may be possible, but other influential points may also be detected during the analysis. Where appropriate, decisions regarding these points should be made on a case-by-case basis with the client. That is, the client should be made aware that further interaction may be necessary before the final analysis can be completed.

Categorical Variables Qualitative data (e.g., `Gender`) are sometimes excluded from consideration in a regression analysis, or inappropri-

[16] For some clients this may seem counterintuitive: didn't R^2 increase? Demonstrating the problem by adding a random vector or using a saturated model ($p = n$, where p is the number of explanatory variables) in a simple example can be helpful.

ately converted to an ordinal scale. Similarly, an apparent quantitative variable such as **Age** may actually represent a grouping variable since only certain values were recorded. These situations affect how the regression analysis proceeds and particular attention needs to be given to explanatory variables involving ordinal or qualitative data types.

Of course, if the *response* variable is categorical, a **generalized** linear model would need to be employed. This model is briefly discussed in the Specialized Methods section below.

Constraints The regression effect associated with an explanatory variable may be different over certain domains. For example, the effect of temperature on residential electricity consumption is negatively correlated over low temperature ranges, but positively correlated over high temperature ranges. Change points, thresholds, and other physical constraints need to be incorporated in the regression analysis.

Serial Correlation Observations that are collected over time may be subject to serial correlation. This can badly inflate the error variance leading to inefficient regression estimates. It is worthwhile probing the client about the data collection process so that a potential time series effect does not arise unexpectedly. Modifications to the regression procedure such as two-stage estimation based on autoregressive models could be used in this situation. This is discussed in Seber and Wild (1989). Alternatively, time series analysis may be required using lagged response and explanatory variables. This is referred to as transfer function modeling which is discussed in Brockwell and Davis (1991).

Regression Analysis

Fitting a regression model is typically an iterative process, alternating between assessing the quality of the fitted model and making modifications to the model. However, it is worth emphasizing the following:

Don't overdo the fitting procedure.

There are an unlimited number of regression models that could be tried and almost as many diagnostic checks that can be performed. In practice, the best model is often the one that has the simplest interpretation.

The most common method for fitting regression models is least squares which can be computed very efficiently. From the least squares parameter estimates $\hat{\beta}_j$, $j = 0, 1, \ldots, k$ we obtain the fitted values: $\hat{y}_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ji}$ (where $x_{0i} \equiv 1$) and the residuals: $e_i = y_i - \hat{y}_i$ for $i = 1, \ldots, n$. These quantities enable the usual ANOVA and parameter estimates table to be constructed as shown below.

Analysis of Variance

Source	DF	SS	MS = SS/DF	F-Value
Regression	k	SSR	MSR	$F_o = \dfrac{\text{MSR}}{\text{MSE}}$
Error	$n-k-1$	SSE	MSE	
Total	$n-1$	SST		

Parameter Estimates

Variable	Estimate	Std. Error	T-Value
Intercept	$\hat{\beta}_0$	$\text{se}(\hat{\beta}_0)$	$t_o = \hat{\beta}_0/\text{se}(\hat{\beta}_0)$
X_1	$\hat{\beta}_1$	$\text{se}(\hat{\beta}_1)$	$t_o = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	$\hat{\beta}_k$	$\text{se}(\hat{\beta}_k)$	$t_o = \hat{\beta}_k/\text{se}(\hat{\beta}_k)$

Here, the F-value provides a test statistic for assessing whether there is an overall regression "effect." If normality and constant variance can be assumed then $F_o \sim \mathcal{F}_{k,n-k-1}$ under the null hypothesis H_o : *no regression effect from* **any** *of the explanatory variables.* The T-value is the t-test statistic[17] for the null hypothesis $H_o : \beta_j = 0$ meaning *no regression effect due to the explanatory variable* X_j.

Residual Diagnostics

The next stage is to perform diagnostic checks on the fitted regression model. If the regression model and assumptions are appropriate for the process under investigation, then the residuals $e_i = y_i - \hat{y}_i$ should reflect the properties imposed on the ϵ_i. Residual analysis therefore plays an important role in regression diagnostics. The main areas of interest are:

- Model adequacy: residual-versus-fitted value plot.
- Assessing normality: the Q–Q plot, Shapiro–Wilk test.
- Detecting influential points: deletion diagnostics.

[17]The `Std Error` entries correspond to the square roots of the diagonal elements of $\text{MSE} * (X'X)^{-1}$, where X is the $n \times (k+1)$ regressor matrix.

Residual vs. Fitted Plot This plot should **not** exhibit any obvious nonrandom pattern. The e_i are often standardized as $r_i = e_i/s_e$, where $s_e^2 = \text{MSE}$. Large r_i can be compared with ± 2.

Q–Q Plot For moderate sample sizes ($n > 30$ say), a Q–Q plot of the residuals e_i can be used to assess the assumption of normal errors: $\epsilon \sim N(0, \sigma^2)$. The ordered residuals should exhibit a straight-line pattern on a Q–Q plot. As we mentioned in the residual diagnostic checks for an ANOVA model, the Shapiro–Wilk statistic provides a formal statistical test for normality.

Deletion Diagnostics There are numerous so-called "delete-one" diagnostics such as studentized residuals, Cook's D, DFFITS, and DFBETAS, that provide the analyst with information about the *influence* of a particular observation. This is achieved by removing the ith observation and refitting the regression model to the remaining $n - 1$ observations. If that particular point is "influential," then the regression fitted to the $n - 1$ observations *should* be different from the orginal regression fit. Definitions of these diagnostics can be found in most regression texts. See, for example, Besley et al. (1980).

In practice, Cook's D(istance) provides a useful indicator of observations that may be influential since it represents a compromise between large residuals and *leverage*. Thus, if the response is a clear outlier with respect to the Y values, but it occurs near the "center" of the X domain, the "influence" of this observation may be minimal. Similarly, an observation may be an outlier with respect to the X domain, but the Y response is close to the fitted regression (hyperplane). In both cases, neither observation would have a large Cook's D value which suggests that the observation is not influential. A plot of Cook's D is shown in Figure 3.7 which clearly identifies the first and last observations as being highly influential.

Choosing the "Best" Regression

The explanatory variables are sometimes called the independent variables which is misleading since the X_js often overlap in terms of their predictive information about Y. More important, some explanatory variables may not contribute significantly when combined with other explanatory variables. Statistically, it makes sense to avoid using redundant information and to try to find the smallest subset of explanatory variables that all contribute significantly to the prediction of the response variable. This is referred to as the Principle of Parsimony. The following diagnostics are useful for choosing a parsimonious regression model.

***T*-Values** The T-value statistics indicate whether each X_j contributes significantly to the regression model. The problem is that the t-

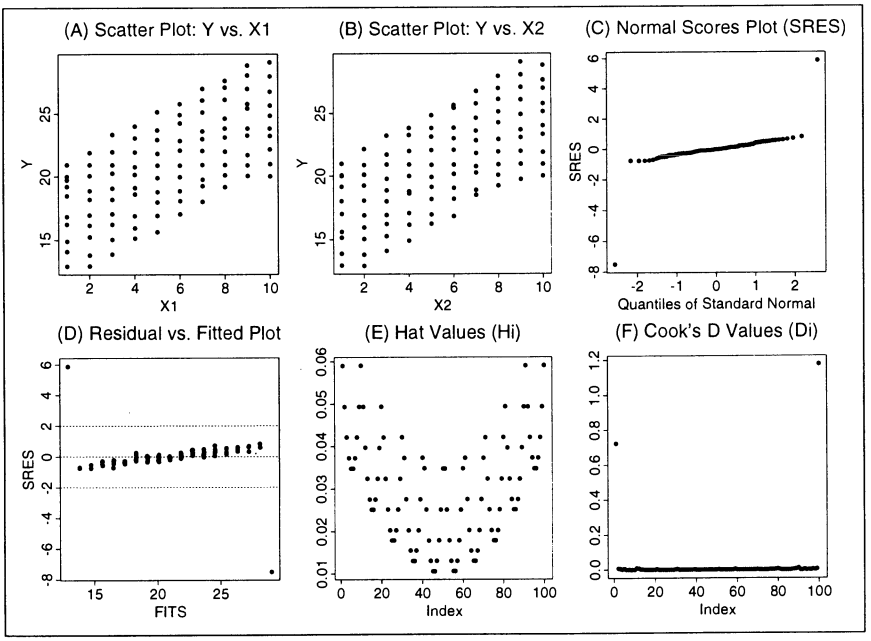


FIGURE 3.7. Regression Diagnostics

statistics are based on the FULL regression model. That is, the significance of "each" X_j is evaluated on the basis that ALL the other explanatory variables are in the model. Hence, eliminating all the nonsignificant X_js at once can be misleading since the t-statistics do not account for multiple dependencies among the explanatory variables.

Adjusted R^2 The quantity $R^2 = \text{SSR/SST}$ provides a measure of the quality of fit for a regression model. However, this can *always* be increased by simply adding another explanatory variable. The "adjusted" R-square statistic defined by $R^2_{\text{adj}} = 1-(n-1)*\text{SSE}/(n-k)*\text{SST}$ provides a measure of the quality of fit *relative* to the complexity of the regression model. Thus, R^2_{adj} can be employed to compare models fitted with different subsets of the explanatory variables.

Stepwise Procedures A stepwise procedure is an iterative regression fitting method that is available in most statistical software packages. At each iteration it either adds or eliminates an explanatory variable according to the significance of that variable in the current regression model. This continues until no further changes occur or the same X_j is being added and then immediately removed.

Remarks

1. "And what, exactly, do you mean by nonrandom pattern?" asks the client. (Isn't "random pattern" an oxymoron?)

   What we are really trying to say is that there should be no obvious relationship present in the *vertical* spread of the residuals when plotted against the fitted value or an explanatory variable. This concept is not an easy one to convey and in our experience, sketching some "nonrandom" patterns usually helps clarify the issue.

2. The stepwise procedure is simply an algorithm that uses prespecified significance levels for entry and exit. It does not guarantee that the final subset model is necessarily the most meaningful one. Other information criteria such as Mallows C_p statistic and AIC, can also be employed for subset selection.

3.6 General Methods

The methods presented in the previous section may be sufficient for completing the statistical analysis required in some projects, but certainly not all. Often, the statistical consultant needs to employ a more specialized technique to perform the appropriate analysis and in some cases, this can be the first time a client has ever heard of such a method. In this situation, the onus clearly shifts to the consultant to provide an adequate description of the methodology without embroiling the client in the technical details.

Since the methods we discuss below are familiar to any reader with statistical training, we have kept our presentation necessarily terse as a general exposition on these methods can be found elsewhere. Moreover, details concerning the application of a particular method are presented in the context of a case study in Part II. References to other texts and the relevant case study are provided here.

Although our presentation of these methods is terse, our aim is to provide the reader with the *basis* of an "adequate description" of the general methodology. To expect to achieve this aim is quite unrealistic of course, but the client is waiting for our explanation. ...

Nonparametric Tests

Nonparametric tests provide *distribution-free* alternatives to traditional methods of analysis. While methods such as the t-test and one-way ANOVA assume the sample has an underlying normal distribution, many nonparametric tests are based on analyzing the **ranks** associated with the data sample. Thus, strict distributional assumptions of the sample are not required in nonparametric tests.

Nonparametric procedures are particularly useful for analyzing ordinal data such as those generated from market research surveys. Responses of the form "never — rarely — sometimes — $\cdots$" would be numerically encoded as 1 — 2 — 3 — $\cdots$ say, but employing a t-test explicitly assumes the *distances* between these numbers are meaningful. This may be unrealistic and a rank test, which uses only the inherent order of the responses, is more appropriate. With metric data, however, there are some limitations associated with nonparametric procedures due to the loss of distributional information. Some well-known nonparametric tests and procedures are:

Spearman's Correlation This provides a measure of linear association between the ranks of two ordinal variables. It has the same interpretation and properties of a correlation coefficient, but avoids treating the actual interpoint distances of an ordinal scale as meaningful.

Sign Test Alternative to the one-sample t-test. This test simply considers the number of values above the hypothesized median and employs the binomial distribution to test $H_o : p = 1/2$. It is very conservative and can be sensitive to round-off error.

Wilcoxon Signed-Rank Test Alternative to the one-sample t-test. Assumes underlying distribution of the sample is symmetric.

Mann–Whitney U-Test Alternative to the two-sample t-test. Also referred to as the Wilcoxon rank sum test. Ties in the data ranks need to be treated carefully.

Kruskal–Wallis Test Alternative to one-way ANOVA.

Further information on nonparametric tests and procedures can be found in Sprent (1993) and Conover (1980).

Remarks

1. Tests based on means and variances are likely to be more intuitive to a client than one based on ranks. Interpretation and specification of the nonparametric hypothesis itself may be difficult to convey to the client.

2. Beware of nonparametric tests employed simply to avoid a "significant" result such as in litigation cases.

3. Nonparametric tests may not be useful with small samples, which is precisely when the assumption of normality is likely to be an issue. Resampling methods may need to be considered.

4. Kernel density estimation and k-nearest-neighbor methods can be employed in procedures such as discriminant and cluster analysis, and provide an alternative to the assumption of multivariate normality.

General Linear Models

Regression and ANOVA models are special cases of the **general** linear model. Other examples of the general linear model are:

Response Surface Models In factorial experiments, interest is often concerned with finding the optimal response with respect to quantitative design factors. When interaction is present, response surface models incorporate second-order effects in the regression.

Response surface analysis is used in variety of applications such as chemical engineering, agriculture, and food science where the yield from a response variable is "known" to depend on certain input variables. That is, response surface methodology is a *sequential* procedure, with each new experiment exploiting the results obtained from the previous one. Since this type of iterative approach is expensive in practice, more specialized designs are used in response surface experiments.

The eventual objective of response surface methods is to determine the optimum operating conditions of the system under investigation. The definition of "optimum" is pursued in Case Study 7.2. Box et al. (1978) consider the application of response surface methodology in a chemical example and provide references to several other studies in a variety of different areas. For a more comprehensive presentation of response surface methodology, see Myers and Montgomery (1995).

Analysis of Covariance In certain experimental situations, additional quantitative information may be available in the form of so-called *concomitant* variables. To incorporate this "regression" information into an analysis of variance design, the analysis of covariance model can be employed under certain assumptions. The main additional assumption is that the concomitant variables are not affected by the "treatment" factors in the ANOVA design. For example, a patient's age will not be affected in a clinical drug trial, but weight changes could easily occur. If weight were to be included as a covariate, it would need to refer to the patient's pretrial weight. Many texts on regression and design also discuss analysis of covariance. Here, we mention Cochran (1957) which appears in the edition of *Biometrics* that is devoted almost entirely to the analysis of covariance.

MANOVA The ANOVA procedure was described in terms of a general technique for partitioning the overall variability of a *single* response variable into components associated with treatment effects. In many experiments, however, several response characteristics may be of interest to the investigator and applying the same ANOVA procedure to each response variable separately will not account for correlations between the response variables. To examine treatment effects on more

than one response variable *simultaneously,* multivariate analysis of variance (MANOVA) is required. Two special cases of interest are:

Repeated Measures Consider the situation where two groups of people — a control group and a treatment group — are studied over a period of time, during which a particular characteristic (Y) is recorded at three regular intervals. Since Y is being measured several times for each subject (person), these measurements will be correlated. Furthermore, the responses from different subjects can also vary greatly making it difficult to detect a meaningful difference between the two groups. To account for the correlation in the response, a multivariate repeated measures ANOVA model can be employed. Everitt (1995) presents a practical review of the analysis of repeated measures. See Crowder and Hand (1990) for a text on this subject.

Profile Analysis Psychology experiments often involve administering a battery of tests to certain groups of people, and creating a response "profile" that can be compared across groups. The tests are usually standardized to a common scale and the "time series" graph obtained by connecting the points: (j, Y_j), $j = 1, 2, \ldots, k$, where k is the number of scales (tests), is called a response profile. An estimate of a group profile would be based on the average scores within the group.

If there are only two groups involved, with no additional factors, then various tests such as parallelism (profiles have similar shape), can be based on Hotelling's T^2 statistic. With more than two groups or additional design factors involved, profile analysis leads to the MANOVA procedure. Profile analysis is discussed in Seber (1984).

Remarks

1. We should add a word of caution here. The output from a MANOVA procedure is complex and, in our experience, clients often have great difficulty understanding the results. While it may be tempting to carry out the client's earnest request, "Can't we just run an ANOVA on each response?" these results will **not** replace the MANOVA results. As we mentioned at the beginning of this section, it is the consultant's job to introduce clients to new procedures as necessary.

2. In some cases, where there are only two response variables involved, the objectives of the project "may" be achieved by examining the *difference* between the responses in a univariate ANOVA model. Again, the MANOVA analysis should also be performed and the results from the two procedures compared. If we are fortunate, only the ANOVA results may need to be presented, otherwise ...

3. Presenting the results from a MANOVA analysis to a client takes courage, and possibly some sleight-of-hand. For example, if asked, "What's a Pillai Trace?" how should we respond? One approach is to simply say it's an F-test like the ones seen in ANOVA tables — not exactly true of course, but we have moved the emphasis of the discussion onto interpreting the results, rather than getting bogged down on technical details involving ratios of matrix determinants. Remember, there are lies, damn lies, and ... statistics!

Multivariate Methods

The multivariate methods presented here are concerned with some type of dimension reduction (variables) or homogeneous grouping (observations) of an $n \times p$ data matrix, X say, consisting of n observations on which p characteristics (variables) were measured for each observation. Seber (1984) is a good general reference covering both the theoretical and practical issues involved in multivariate analysis. Muirhead (1982) is strictly theoretical, but does contain several tables that the statistical consultant may find useful. Other relevant texts are: Gnanadesikan (1997), Mardia et al. (1979) and Rencher (1995). A combination of several multivariate methods is explored in the data mining example of Case Study 8.4.

Principal Components Analysis Principal components are the eigenvectors associated with an orthogonal decomposition of the $p \times p$ sample variance (or correlation) matrix of X. The corresponding eigenvalues therefore provide a measure of the proportion of total variance (trace of the variance matrix) explained by each principal component. Thus, the "effective" number of variables in the data matrix X can be reduced if the $k < p$ largest eigenvalues account for a substantial proportion of the total variance. This is the aim of principal components analysis (PCA). Note that these "effective" variables are the principal components[18] which are linear combinations of the original variables with coefficients determined by the k eigenvectors. Use of the correlation matrix implies each of p variables in X is equally important and different results may be obtained depending on which matrix is used. The following example also shows that the largest eigenvalues do not necessarily provide the most "interesting" statistical features of X.

Example 3.6 *Male and female incomes by months of experience.*

The largest eigenvalue will usually correspond to the overall variability in `income`. In this case, the smaller eigenvalue will correspond to

[18] They are also called latent variables.

the variability associated with any *difference* between incomes due to `gender`. Hence, projecting the data onto the second eigenvector will show whether `gender` has an effect on income (which is clearly of more interest).

Factor Analysis Factor analysis (FA) essentially treats the observations of X as satisfying a regression of the form $x = \mu + \Gamma\eta + \epsilon$, but where everything on the right is *unknown*. The vector η is assumed to consist of $k < p$ (common) "factors" with "loadings" given by the elements of the $p \times k$ matrix Γ. Notwithstanding the undetermined nature of the model, statisticians tend to be wary of the results obtained by factor analysis for the reason that any orthogonal rotation of η will also satisfy the model. Similarly, k must be determined and even if the model is true, an incorrect choice of k can render poor results. Hence, the results from a factor analysis need to be interpreted cautiously. However, factor analysis is used quite extensively in market research and we present the results of a Factor analysis in Case Study 8.2.

We should mention that the difference between factor analysis and principal components analysis is often confused since the method of principal components is often used to fit a factor analysis model. ... Confused? Seber (1984; p. 215) provides a clear comparison between the Factor analysis and principal components analysis methods.

Discriminant Analysis Discriminant analysis is concerned with the problem of allocating an observation to one particular "group" (out of a fixed number of groups) based on the p characteristics that were measured. To perform the allocation optimally when groups overlap requires that the discriminant function yield the smallest possible misclassification rate. However, this depends on how much is known (or can be assumed) about the distribution of the observations in each group, how likely it would be for an observation to come from a particular group, and what the costs of misclassification are.

Cluster Analysis Cluster analysis tries to classify observations into groups when the number of groups is unknown. There are two general methods for doing cluster analysis: hierarchical clustering which requires the definition of an interpoint distance and an intercluster distance, and centroid methods where k seed points are chosen and the data are distributed among k clusters. The k clusters with possible merging are then slowly optimized using some criteria such as R^2. Cluster analysis is one of the key methods in data mining and is considered in more detail in Case Study 8.4.

Partial Least Squares Partial least squares (PLS) is widely used in process industries to relate a matrix of explanatory or process vari-

ables to another matrix of output variables. For example, to investigate reactor temperature flow rates, the PLS method can be used to relate the process variables, X = feed times, agitation rates, and metered amount, to a variety of Y = quality output measures.

PLS is commonly used in chemometrics which is essentially the study of statistical applications to chemical data analysis. For a review of chemometric regression tools (including PLS and principal components regression), see the article by Frank and Friedman (1993). An example of an application using principal components regression is given in Manchester et al. (1999).

Time Series Analysis

When data are indexed over time, sequential observations in a time series are likely to be correlated. To capture this underlying feature two approaches are used: time domain models and spectral analysis. Time domain modeling was popularized by Box and Jenkins (1970) and is more commonly used, particularly when forecasting is of interest. Spectral analysis is appropriate when periodicities are of interest. Both methods may be used for vector processes. Some useful references on time series analysis are Brockwell and Davies (1991), Box et al. (1994), and Priestley (1981). An application of time series analysis is given in Case Study 7.4.

ARIMA Models In the Box–Jenkins approach, correlogram estimates of the autocorrelation function (ACF) and partial autocorrelation function (PACF) are plotted and used to select suitable candidates from the autoregressive integrated moving average (ARIMA) class of models. Multiplicative ARIMA models can be used to model seasonal time series.

Spectral Analysis Spectral analysis methods are based on the discrete Fourier transform (DFT) which can be used to represent any series in terms of sinusoids. In practice, the series will need to be detrended since the DFT assumes a finite series repeats itself. The periodogram can be computed from the DFT, and is the primary diagnostic for detecting important frequencies in the series. Smoothing the periodogram provides an estimate of the spectrum which uniquely defines a continuous stationary process.

Categorical Time Series An example of a categorical time series is a DNA sequence that can be indexed in terms of the base pairs (A,T,C,G) which make up the double helix. An interesting approach to the analysis of categorical time series is the spectral envelope methodology given in Stoffer et al. (1993). An alternative approach which has applications to speech recognition is the use of hidden Markov chains.

Specialized Methods

The statistical consultant will frequently deal with many projects where little more than a standard analysis is required. It is rather easy, therefore, to become somewhat complacent about keeping up with current research and the development of new methodology. Do **not** fall into this trap. A statistical consultant is expected to be an expert and this rarefied status cannot be mantained unless the consultant makes a positive effort to publish, attend conferences, and perhaps most important, keep up to date on statistical techniques by reading articles in relevant journals. Appendix A lists some scholarly journals that would be of interest to the statistical consultant. Assuming the reader now feels sufficiently chastised, we briefly discuss some statistical methodology that may be required in more research-oriented projects.

Generalized Linear Models In the general linear model, $y_i = \mu_i + \epsilon_i$, μ_i is assumed to be a *linear* function of the explanatory variables. That is, $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ where $\boldsymbol{x_i}$ is the vector associated with the ith observation of k characteristics. The error ϵ_i distribution is usually assumed to be $N(0\sigma^2)$.

In a *generalized* linear model, the error distribution is allowed to be more general (but within the exponential class of distributions) and there exists some **link** function such that, $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$. Two examples of generalized linear models that are commonly encountered in practice are indicated below. A classic reference on this subject is McCullagh and Nelder (1989). For a less intensive treatment, see Dobson (1990). Agresti (1990) provides a thorough treatment of log-linear models. Collett (1991) and Morgan (1992) are two texts that deal with the analysis of binary and quantal response data.

Logistic regression When the response variable Y is categorical, many prediction models transform the problem into a regression model for predicting $\pi_j = P[Y_i = j]$, where $j = 1, 2, \ldots, k$ represent the possible levels or states of Y. Logistic regression assumes that the probabilities of Y can be modeled by the logistic distribution $f(w = x'\beta) = e^w/(1 + e^w)$.

Log-linear Models When the response variable is count data, then the Poisson distribution combined with the log transform leads to the study of *log-linear* models for contingency tables. For clients who require more than a simple test of independence, log-linear modeling is analogous to ANOVA for contingency tables.

Nonlinear Models With certain phenomena such as biological growth and chemical reactions, prior knowledge of a nonlinear relationship

may be well established. In these situations the use of an appropriate nonlinear model can be considered. See Seber and Wild (1989) for a detailed treatment of the theory and applications of nonlinear regression methods.

Modern Regression Methods A *generalized additive model* has the form $Y = \beta_0 + \sum_{j=1}^{k} f_j(X_j) + \epsilon$, where the $f_j(X_j)$ are unknown, but assumed to be "smooth" functions of the explanatory variable X_j. This extends the notion of linear regression and has led to the development of several so-called modern regression methods: *smoothing splines, projection pursuit regression, ACE, AVAS, and MARS algorithms, and neural networks.* Case Study 7.3 presents an application of smoothing splines and neural networks are used in Case Study 8.2. We refer the reader to Venables and Ripley (1994; Chapter 10) for further details (and relevant references) on all the methods listed above. The statistical aspects of fitting generalized additive models are considered in Chambers and Hastie (1993). Tree-based models are also discussed in the last reference based on the work of Brieman et al. (1984).

Robust Methods Robust methods allow statistical procedures to be applied in the presence of contaminated data such as outliers, rounding effects, or departures from distributional assumptions. For this reason, it is worthwhile employing a robust equivalent of, for example, least squares regression to check whether the results are compatible. If not, we need to investigate the cause of the discrepancy.

S-PLUS provides a wide variety of functions associated with robust and resistant methods. A "resistant" method provides protection against gross errors (e.g., the median versus the sample mean). A "robust" method maintains reasonable efficiency when the data are *not* contaminated (e.g., a trimmed mean). The series of books by Hoaglin et al. (1983, 1985, 1991) provides information on a wide range of resistant and robust methods. Rousseeuw and Leroy (1987) is a practical text on robust regression methods.

We should note that robust methods do **not** extend easily to the multivariate situation (which "vector" is larger?), and some multivariate robust techniques (e.g., S-estimators) can be computationally intensive. The S-PLUS function `cov.mve()` provides a robust estimate of the covariance matrix based on the minimum volume ellipsoid.

Resampling Techniques The primary tool underlying many classical statistical inference procedures is the central limit theorem. But what do we do when the sample size does not permit the asymptotic normality theory to hold? Nonparametric methods can be applied, but for small samples these methods lack power and tend to provide results that are too conservative. In this situation, it may be useful to

employ resampling methods such as the jackknife or bootstrap procedure. Resampling techniques enable the investigator to estimate the standard error, confidence intervals, and distributions for any statistic. For more detailed descriptions of bootstrapping, see Efron and Tibshirani (1993) and Shao and Tu (1995).

Structural Equation Models Path analysis is concerned with the problem of assigning causal effects that lead to phenomena of interest to the investigator. This methdology is often employed in the field of psychometrics, but relies on the investigator to specify the relationships between variables (which is clearly subjective in many cases). Fuller (1987) discusses the statistical theory associated with measurement error models. Additional references may be found in the `PROC CALIS` section of the SAS/STAT manual (SAS 1990).

3.7 Design of Experiments

The principles of statistical design of experiments were introduced in the context of data collection methods (Section 3.1). Here, we briefly discuss specific designs that the statistical consultant is likely to encounter. Our list of designs is certainly not complete and the problem of *unbalanced* data can arise in practice. We return to this issue at the end of this section.

The purpose of experimental designs is to provide an efficient analysis of variance where known sources of variation are accounted for. These designs also maintain reasonable sample sizes. Two classic texts worth mentioning are Cochran and Cox (1957) and Cox (1958). Montgomery (1997) is a standard text that provides good overall coverage of experimental designs. Box et al. (1978) emphasize the practical application of experimental designs.

Randomized Block Designs

Batches, people, and time are common sources of variability that can be controlled through blocking designs. These designs enable better comparison of treatment differences by eliminating the variability between "blocks." If the variability due to blocks is *not* controlled for, as in a completely randomized experiment, this variability will be absorbed in the overall experimental error which can mask meaningful treatment differences.

Randomized Complete Block Design (RCB) This design generalizes the concept of the paired t-test and allocates *all* levels of a treatment to each **block** identified as a potential source of variability in the study. There is one observation per treatment level in each block and the order in which the treatments are performed is randomized within each block.

Latin Squares If two blocking factors have the same number of levels as the treatment factor, I say, then a Latin square design can be employed which requires I^2 observations. The name of this design refers to the use of *Latin* letters A, B, ... to denote the levels of the treatment factor. The design layout can be represented as a square table with the row and column denoting the levels of the two blocking factors. In order to separate the treatment effect from the two extraneous sources of variability, the design must be *orthogonal*. This is achieved by having each Latin letter (treatment level) appear only once in every row and column of the square table.

Row	Column Factor			
Factor	C_1	C_2	C_3	C_4
R_1	Aα	Bβ	Cγ	Dδ
R_2	Bδ	Aγ	Dβ	Cα
R_3	Cβ	Dα	Aδ	Bγ
R_4	Dγ	Cδ	Bα	Aβ

The example of two superimposed 4×4 Latin squares presented above shows that it is possible to introduce a third factor (with levels denoted by *Greek* letters) into the design. This is called a Greco–Latin square design which can be used to analyze the effect of *two* treatment factors, or to control for a third source of extraneous variability. Again, an orthogonal design is needed to separate the effect of each factor and this can be achieved by superimposing two (orthogonal) Latin squares such that each Greek letter appears exactly once with each Latin letter. Greco–Latin squares exist for all $I \geq 3$ except $I = 6$.

Incomplete Block The RCB design requires all levels of the treatment factor to be allocated in each block. In some situations this may not be possible, or even desirable. The classic example of an incomplete block design is the so-called wine tasting experiment: several wines are to be tested by a panel of judges. Since the ability to distinguish between tastes diminishes as each new wine is sampled, it is undesirable to have a RCB design. Instead, an incomplete block design can be employed where only certain treatment levels (wines) are allocated to each block (judge). The design is *balanced* if any two treatment levels appear together in a block an equal number of times.

Split-Plot The RCB and Latin square designs are effective techniques for dealing with randomization restrictions relative to a single treatment factor. These designs can also be used to analyze more than one treatment factor (see the Greco-Latin square design example above) *provided* all the treatment level combinations are present in each block. However, this assumes the *order* of experimentation within

each block can be completely randomized. For example, if A has two levels ($I = 2$) and B has three ($J = 3$), then we should be free to toss a die to see which particular level combination will be run first.

In some multifactor experiments, it may not be economically feasible or practical to collect the data using complete randomization within blocks. Using the example above, suppose A represents two recipes for a cake, which will be baked at three different heights (B), in four ovens (blocks). The weight of the cake will be measured to determine whether it was cooked properly. If each oven has room for three cakes, it is more practical to use one recipe at a time and make enough cake mixture to bake three cakes. With complete randomization, we are not permitted to dictate how we make our cake *and* eat it!

The split-plot and split-split plot designs are generalizations of the randomized block design that can be employed in multifactor experiments where restricted randomization is needed. The terminology derives from their application in agricultural field trials where a "plot" represented a physical parcel of land which was "split" into subplots. (In a "split-split" design, subplots are further divided into subsubplots). Because randomization only occurs within each level of the *main* treatment (A in the cake example), main effects are said to be *confounded* with blocks in a split-plot design. That is, the effect on the response from any other uncontrolled factor that varies with the levels of the main treatment cannot be distinguished from the effect of the main treatment. Since the subplot treatment is not confounded with blocks, the treatment factor of most interest in the study is usually assigned to subplots.

Remarks

1. If the block differences account for a large portion of the variability not explained by the treatment (levels), then the completely randomized block design will have high relative efficiency compared to the one-way ANOVA. Relative efficiency is essentially a measure of how many replicates would be needed in a one-way ANOVA to achieve the same sensitivity (MSE) as the RCB design. If the relative efficiency is 3, for example, a one-way ANOVA would need to be run with 3 *times* as many observations to achieve the same sensitivity as the RCB design experiment.

2. Most statistical software will provide the usual two-way ANOVA table, including the F-test: $F_B = \text{MSB}/\text{MSE}$, associated with the block effect. This test should **not** be regarded as valid since the randomization of treatment levels was *restricted* within blocks.

3. Additivity of the treatment and block effects should be checked by examining the residual-versus-fitted plot, or constructing an interac-

tion plot (see two-way ANOVA). In some cases, nonadditivity can be caused by outliers, or an underlying multiplicative relationship. A power transformation may be useful for restoring additivity if the relationship is multiplicative.

4. One disadvantage of blocking designs is that the error degrees of freedom are reduced. For example, in a Latin square design with $I = 4$ treatment levels, there are only 6 degrees of freedom for the error component. This means the estimate of variability associated with experimental error (i.e., σ^2 from the presumed $N(0, \sigma^2)$ error distribution) is effectively based on a sample of size 6 — which is rather small. To increase the precision of the estimate for σ^2, complete replicates of the design can be employed.[19] As suggested by our first remark, however, including a potential blocking effect is generally recommended.

5. There are many other blocking designs we have not mentioned such as Youden squares, partially balanced incomplete block designs, and lattice designs. The purpose of these designs is provide the investigator with "good" alternatives when certain practical constraints prevent a balanced design to be employed.

Factorial Experiments

In a factorial experiment, each replicate contains all combinations of the levels of the treatment factors. The factors in a factorial design are therefore said to be *crossed* and at least two replicates are needed to consider all possible **interaction** effects. In the simplest case where only two design factors are involved, the analysis follows the two-way ANOVA procedure discussed earlier. When several factors are involved, however, the number of interaction effects increases dramatically ($2^k - k - 1$ for k factors) and implementing a full factorial design with multiple levels per treatment is often impractical. During the initial phase of a project, however, several factors may be of interest to the investigator and some special cases of the general factorial design are particularly useful.

2^k **Factorial** The 2^k factorial design provides the smallest number of treatment combinations for analyzing k factors. Each factor has only two levels, referred to as "high" and "low," which may indicate the presence or absence of a treatment, or two representative settings of a quantitative factor such as temperature, or simply two different qualitative states (e.g., two operators). Since only two levels are available for each factor, it is necessary to assume the response is linear over

[19] A special case of a replicated 2×2 Latin square design is called a **crossover** design. See Montgomery (1997) for further details.

the range of levels selected. Extreme values should **not** be used for the levels of a factor. (See Case Study 7.2.) Note that no estimate of error will be available from a single replicate of a 2^k design unless higher-order interactions are pooled to form an estimate of the "error" variability.

Confounding in 2^k Designs A complete replicate of the 2^k factorial design will include all higher-order interactions, but it may be impractical to run the entire experiment at once. A common procedure is to run the 2^k experiment in 2^p blocks so that each block contains 2^{k-p} runs. This means certain effects will be confounded with blocks and so the choice of which p independent effects are to be confounded is important. Note that a total of $2^p - 1$ effects will be confounded with blocks, these being (generalized) interactions associated with the initial p effects chosen. In general, higher-order interactions are selected to be confounded with blocks since these are of little interest to the investigator (how do we interpret a four-factor interaction?). However, care is needed to ensure that important effects (i.e., main effects and two-factor interactions) are *not* included in the additional interactions that will be confounded with blocks. Montgomery (1997) provides a table of suggested blocking designs and the effects chosen to generate the blocks.

Fractional Factorial For *screening* experiments, the 2^k fractional factorial design is widely used for the purpose of identifying which of the k factors have large effects. These designs only contain 2^{k-p} runs of a complete 2^k design which results in the treatment effects being *aliased* with each other. This means that the estimate of an effect of interest actually represents the combination of the effect and its aliased effect(s). The term **resolution** is used to describe the type of aliasing present in a 2^{k-p} design:

Resolution III main effects are aliased with two-factor interactions;

Resolution IV two-factor interactions are aliased with each other;

Resolution V two-factor interactions are aliased with three-factor interactions.

3^k Factorial The 3^k factorial design extends the 2^k design and enables k treatments to be studied where each factor has three levels of interest. The concepts of confounding and fractional replication can also be extended to the 3^k factorial design.

Composite Designs After the screening designs have been run and the results assessed, composite designs may be used to investigate important factors more thoroughly. These designs allow the variability to be estimated by adding replicate points at certain positions in

the design space. Case Study 7.2 provides an application of a central composite design.

Nested Designs

In a two-stage nested or hierarchical design the levels of treatment B are similar but not identical across the levels of treatment A. For example, two different suppliers (factor A) each provide three randomly selected batches (factor B) of raw material which is tested for purity. In this case, the batches are specific to an individual supplier and the levels of B are "nested" within each level of A. The situation is depicted in Figure 3.8 where the prime notation emphasizes that there is no connection between the batches from different suppliers — B1 and B1′ are *not* the same batch.

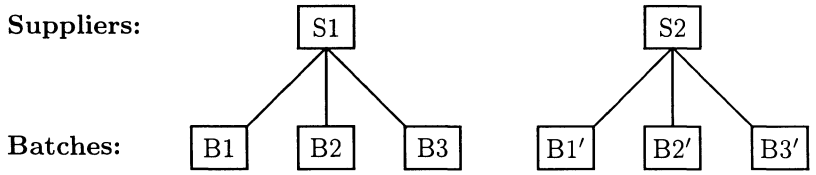


FIGURE 3.8. Layout for a Two-Stage Nested Design

The nested effect "B within A" is denoted as $B(A)$ with the parent factor appearing inside parentheses.[20] Since not every level of B appears with each level of A there is no interaction between these factors and the sum of squares decomposition has the form: SST = SSA + SSB(A) + SSE. The extension to multistage nested designs is straightforward and it is also possible to have both nested and crossed factors in an experiment.

Remarks

1. In some cases we may be uncertain as to whether a factor is crossed or nested based on the client's initial description of the project. This needs to be ascertained since the interpretation of a significant "interaction" may be misleading if a nested factor is incorrectly analyzed as if it were crossed. As in the case of the split-plot design, nested factors can often be determined by probing the client on the precise method of data collection that was used in the experiment.

2. Nested designs can be employed to analyze the variance components associated with a manufacturing process. Identifying sources that

[20] In S-PLUS nested models are defined by the slash operator / with the nested factor appearing **after** the slash: `A/B`.

contribute most of the variability in the output is an important step towards quality improvement.

3. In Figure 3.8, the nested factor $B(A)$ will often be a **random** effect so $F_A = \text{MSA/MSB(A)}$ is the appropriate statistic for testing the effect of factor A (Supplier). In general, the appropriate F-statistics will depend on whether the factor is fixed or random and particular care is needed when the design involves both nested and crossed factors. We refer the reader to Montgomery (1997) for an overview of the analysis of nested designs.

Unbalanced Designs

An *unbalanced* design arises whenever there are an unequal number of observations per treatment effect. This can occur because of problems in the data collection (e.g., some specimens died during the experiment), resulting in a loss of observations in what may have been a balanced experimental design initially. On the other hand, an unbalanced design may be intentional. Certain treatment combinations may be more expensive to run so fewer observations are taken or more replications may be performed for factors that are of greater interest in the study.

The main problem with unbalanced designs is that the usual analysis of variance techniques associated with balanced designs does not apply. This follows since certain treatment combinations can be estimated more precisely than others in an unbalanced experiment. As a result, the orthogonality property of treatment effects that is present in balanced designs does not carry across to the unbalanced case, complicating the analysis. Furthermore, the hypotheses being tested for a treatment effect are different from those of a balanced design and may not be easily interpretable. In some cases, the hypothesis being tested may not be unique, or is of questionable merit. This was discussed with regard to the different types of SS and hypotheses associated with ANOVA models. (See two-way ANOVA.)

In general, a comprehensive statistical software package, or one that specializes in experimental designs, should be employed to analyze unbalanced designs. (Statistical software is discussed in the following section.) However, if the data are "close" to being balanced, there are several approximate methods that can be used. These methods convert an unbalanced problem into a balanced one which can be analyzed by most statistical software packages.

Remarks

1. Apart from the computational ease of analysis, there are two important advantages balanced designs have over unbalanced designs:

    - The power of the test is maximized when the number of observations per treatment effect is equal.

- The test is less sensitive to departures from the assumption of equal variances across the treatment levels when the sample sizes are equal.

2. Approximate methods will not work when there are missing cells (treatment combinations). In this case, an exact analysis is required and the results need to be carefully interpreted. Note that different parameterizations are used by different software packages. Our role as a statistical consultant will be important in this type of analysis. The text by Searle (1987) deals with the analysis of unbalanced data.

3.8 Statistical Software

There is an extensive range of packages and applications that can perform statistical analyses. The statistical consultant needs to become familiar with the capabilities of several software packages and to be proficient in the use of at least one comprehensive package. In this book we use SAS and S-PLUS (Mathsoft 2000) in the case studies and a brief discussion of these packages is given in the following sections. Here, we provide an overview of some statistical software currently available and their levels of statistical analysis.

Level 1

Menu-driven. May not include certain statistical procedures. Quality of analysis can vary. Should provide good application-dependent graphics. Suitable for novice/student user. Relatively easy to learn. Low to moderate cost depending on quality and procedures available.

Examples: JMP, Statgraphics, Statview, Excel.

Level 2

Menu-driven versions may include a command-line window. Should provide a sufficiently comprehensive suite of statistical procedures. Quality of analysis is generally good. Allows user-defined graphics. Often used by nonstatisticians. Command-line syntax may take time to learn. Moderate cost for full versions.

Examples: JMP (Version 4), SPSS, MINITAB, Systat.

Level 3

Designed for the expert user. Command-line driven. Some may be specialized for certain types of analysis only. Quality of analysis should be excellent. Output not always presented in standard form. Requires time and effort to become a proficient user. Some of the comprehensive packages can be (very) expensive.

Examples: SAS, S-PLUS, R, GLIM, GENSTAT, BMDP, Enterprise Miner, Clementine, BUGS.

Remarks

1. We should emphasize that our "Level" categories are primarily based on how easy it would be for a client, student, or consultant to use the statistical software we listed. There is, for example, a considerable difference between JMP and Excel in terms of their statistical capabilities. No statistical consultant would employ Excel to perform a statistical analysis, but Excel is used extensively for data entry and *can* perform certain statistical procedures in addition to providing summary statistics and graphs. Hence, the consultant needs to be aware of what Excel can provide a client.[21]

2. The distinction between Level 1 and 2 is also somewhat blurry. JMP 3.x is entirely menu-driven and essentially lacks only a module for time series analysis. (This is present in JMP 4.0.) However, it is a package that we would recommend to a client who wanted to invest in some "real" statistical software. It runs on both Macintosh and PC platforms and is very easy to use. The PC versions of SPSS and MINITAB are also very easy to use, and one of these may be a more appropriate choice for some clients. (MINITAB tends to be used in business schools, SPSS in humanities.)

3. Level 3 software is clearly what the consultant needs to perform statistical analyses. SAS and S-PLUS are certainly the dominant statistical packages, but the full versions are expensive. PC versions of SAS and S-PLUS are available and R is public domain software that is very similar to S-PLUS. The main difficulty that often arises is that clients may feel very uncomfortable looking at the output produced by these packages. They can feel overwhelmed and even become quite frustrated by the information. "Ugh, I don't understand any of this printout!" The consultant may need to reduce the clutter in output before presenting the results to a client. We address this very issue in the next chapter.

SAS and S-PLUS

The statistical software that we use in all the case studies are SAS and S-PLUS. Both are expert-level applications that require a lot of time and effort in order to become a proficient user. Appendix B does provide an introduction to SAS and S-PLUS, but this is unlikely to take a novice user

[21] Clients are sometimes quite surprised to find out just how much statistics Excel can do for them.

very far towards proficiency. So how much SAS and S-PLUS code do we expect the reader to know? We address this question shortly. First, we begin with a brief description of these applications.

SAS Can be run interactively, but is primarily a batch-processing application that can handle large amounts of data efficiently. The syntax associated with the so-called "DATA" step is really the hard part of SAS. Extensive data manipulations may be required before a statistical procedure ("PROC" step) can be employed. Fortunately, most of the PROC statements in SAS have names that clearly indicate their purpose. Readers familiar with statistical programming applications such as SPSS, GENSTAT, or GLIM should not have too much difficulty understanding the general purpose of the code in a SAS procedure. For readers in this category, Appendix B provides a useful reference. SAS has recently introduced a new data mining software called SAS Enterprise Miner that is more suitable for analyzing and "mining" large data sets. In Case Study 8.4 we introduce methodology for statistical data mining.

S-PLUS Should always be run interactively. S-PLUS is really a programming language environment based on a multitude of built-in "functions." Some are quite basic: `q()` to quit from S-PLUS; others can produce a complete statistical analysis: `lsfit(x,y)` to perform a least squares regression. Creating customized functions is one of the main strengths of S-PLUS. A statistical analysis in S-PLUS typically consists of the user interacting directly with the data at each step of the analysis, making small changes and modifications to a model or statistical procedure, and creating customized functions as required. S-PLUS also allows the user to exert precise control over a graphics object which makes it ideal for producing presentation quality graphics. R is a public domain statistical software that from the user's point of view is almost identical to S-PLUS. The main differences are that S-PLUS provides a graphical user interface (GUI), and it contains some additional software packages. As a commercial product S-PLUS provides the users with online support. All the S-PLUS code included in the examples in this book is also R compatible.

Comparing SAS and S-PLUS

S-PLUS is very different from SAS. Indeed, we would be hard-pressed to find a more redundant statement. The fundamental difference is in their approach to statistical analysis:

SAS provides *the statistical procedure* — for users.

S-PLUS provides *the user* — with statistical procedures.

This is really the important issue. A list of specific differences would be rather pointless since these are inherently due to the different approaches. The statistical consultant needs to accept this difference and learn how to best exploit the advantages of each application.

Other Expert-Level Software

SAS and S-PLUS are not the only expert-level statistical software available, but they are widely known and used internationally.

SPSS (statistical package for the social sciences) is predominantly used by, surprisingly, social scientists. Statisticians generally avoid SPSS and tend to regard SPSS output produced by a nonexpert with suspicion. While suspicious output from a nonexpert user may not have changed, the quality of SPSS has certainly improved and this criticism is somewhat unfair now. Because SPSS is menu-driven, it has gained its place in introductory statistics courses. Clementine is a very successful data mining software produced by SPSS.

GENSTAT and GLIM are well known outside the U.S. and both are highly regarded statistical packages. GENSTAT is a general purpose application whereas GLIM is specialized for generalized linear (interactive) modeling.

XlispStat (Tierney 1991) and BUGS are two public domain software with a strong following in the academic community. BUGS is for Bayesian statistical analysis and XlispStat has great interactive graphics.

Presenting SAS and S-PLUS Code

Presenting SAS and S-PLUS code means we are assuming that the reader is familiar with SAS and S-PLUS programming statements as well as their standard conventions and syntax. Of course, "assuming" the reader is familiar with SAS and S-PLUS code does not necessarily mean we "expect" the reader to be familiar with this code. In order to accommodate a cross-section of readers, certain compromises have been made as follows.

Nonuser The context of the problem presented in the case study and description of the analysis is your primary guide. Most of the program code will probably be meaningless to you. Ignore it! If you've read this far into the book, you didn't pick it up by mistake; your interests clearly lie with other aspects of the consultation process.

Novice-User The introduction and details in Appendix B will be of some help. However, it probably won't be enough for you to reconstruct a complete program for analyzing the case study. You will need to get assistance, but the effort will be well worthwhile.

ItsBeenAWhile-User Good software is constantly evolving and changing. You may not recognize some parts of the code, but it's hoped

that it hasn't been that long since you used the software. The description of the analysis may need to be your primary guide in some cases.

ExpertOther-User You should not have too much difficulty understanding the general purpose of the code in a SAS procedure. The description of the analysis and context of the problem in a case study will be sufficient for you to replicate the analysis in your preferred choice of statistical software. The S-PLUS code may still be cryptic for you.

Expert-User You will note that our program code (as written) often does more than we mention, and less than it should. At best, the program code may provide you with a helpful reminder or hint, but you probably know more about these applications than we do!

An Overview of Computational Tools

Statistical consultants have access to a large number of computational tools that are available on different platforms and over the Internet. The combination of these tools opens new possibilities but at the same time creates many new difficulties. For example, we may have generated presentation quality graphs in PostScript format, but now need to convert them to slides for a PowerPoint presentation. Similarly, suppose we start with an Excel data file. We would like to do the exploratory data analysis in S-PLUS, but we need to generate some analysis using a SAS procedure. Situations like this often arise in many statistical consulting projects and can involve shifting data between platforms. Sometimes it is necessary to save the data in a plain text (ASCII) file and then, if necessary, perform some manual editing in order to convert the data to a certain format. However, this can be very time consuming and it is often better to use automatic procedures provided by other software.

Shifting Data Among Platforms and Applications

Most applications can produce plain text data files. Delimiters are usually necessary to separate data fields, but they need to be chosen carefully. For example, the default delimiter may be a tab stop, blank space, or comma. But if the dataset contains a field that uses blank spaces and commas as characters (e.g., "Smith, James B."), then using a comma or blank space as a delimiter is clearly going to cause problems. In these cases, a special character such as `":"` or `";"` needs to be used as the delimiter. Both SAS and S-PLUS allow the user to set the delimiter to be any character.

Nontext formats such as Excel or DBaseIV are often used since many applications understand these formats. Excel is very popular and for most projects, entering and saving the data as an Excel file are sufficient. If the dataset contains variable names, tags, and special formatting information,

then it is probably best to enter and save the data into a format such as DBaseIV. The DBaseIV file can then be transferred to the new system and read into the software application using the new format, thus preserving all the variable and case information. A permanent SAS dataset is a good example of a "database" that contains this type of information and this can be created from data stored in other database management system (DBMS) software. Transporting a large SAS dataset between platforms **directly** is usually more practical than trying to reconstruct the database from a plain text formatted file. The introduction to SAS and S-PLUS in Appendix B provides examples of reading and writing various types of data files in SAS and S-PLUS.

Exporting Graphics Files

Statistical consulting reports often need to include graphical displays. One option is simply to print off hardcopies of all the graphs and insert them in the report. However, this can be rather inefficient and time consuming since it invariably means that the analysis must be reproduced, possibly several times. A much better option is to save a graphical display to a graphics file. There are several ways to do this and the main issue is what format to use.

In the Windows or Macintosh environments, it is easy to copy and paste among the different applications. Hence, if we decide to keep our graphs in a PowerPoint file, whenever we generate a graph in S-PLUS or SAS we can copy and paste it into a PowerPoint file. From PowerPoint, we can transport the graphs to many other applications and across platforms.

Both SAS and S-PLUS provide the facility for exporting graphs to files in a variety of formats. We recommend exporting graphics files in GIF or JPEG format because these formats are available on most platforms. Encapsulated PostScript (EPS) format can be used when the graph is primarily to be used for printing. One advantage of EPS format is that the graph is saved in ASCII code. That is, a plain text file that can be edited. PostScript graphics files can be incorporated in LaTeX (as in this book) and it is possible to covert any printable output by sending the output to a printer and "saving" (to EPS format) instead "printing." Be aware, however, that converting some types of printable output, such as a GIF or JPEG file, can create *very* large PostScript files.

Software Combinations and Interfaces

The combination of two (or more) software applications can sometimes generate better results: let each do what it can do best. Xgobi can be used as a graphics interface between S-PLUS and R, thus providing S-PLUS with advanced dynamic graphics capabilities. The interfaces of S-PLUS with Mathematica, SAS, and Excel also give S-PLUS added functionality.

In the next chapter, we exploit the ease of using the JMP statistical software for presenting preliminary results to our client, SAS for the final statistical analysis, and S-PLUS to produce the presentation quality graphics the client wanted.

StatLib

An important source of information for statistical software is StatLib: This can be accessed using the Web or "anonymous" ftp to the site:

```
Web:   http://lib.stat.cmu.edu/
FTP:   ftp  lib.stat.cmu.edu
```

Among the software available in StatLib are S-PLUS and SAS programs and packages contributed by users, the public domain software R, Xgobi and much more. Anonymous ftp is required to download certain files and applications which (uncompressed) may be quite large.

Anonymous ftp simply means we use the word: `anonymous` as our "guest" username when accessing the ftp site. The `"password"` field is completed by entering our: `email-address` . After we have successfully logged on (sometimes there are too many users) and located the files we wish to download, we make sure the transfer mode is set to "binary" for downloading compressed files and applications. "ASCII" mode should only be used for downloading plain text files such as `"README"` files. The archive or compression format is usually indicated by an extension at the end of the filename. Some common extensions are listed below.

Archives

`.sea` self-extracting archive (Windows and Macintosh). Bundled files should self-extract when activated.

`.tar` tape archive (UNIX). Use `tar xf filename.tar` to extract. (The reference to "tape" is historical.)

Compression

`.Z` (UNIX) Use `uncompress filename.Z` .

`.gz` gzip file (UNIX). Use `gunzip filename.gz` . Can be unzipped on Windows and Mactintosh platforms.

`.zip` zip file (Windows and UNIX). Use `unzip filename.zip` on UNIX. Can be unzipped on Macintosh platform.

`.sit` StuffIt file (Mactintosh). Can be unstuffed on Windows.

4
A Consulting Project from A to Z

In this chapter our aim is to try to reproduce the entire consultation process for a particular project from initial contact with the client to the final written report and postcompletion followup. The actual consultation took place in a university environment which we refer to as the SCP (see Section 1.4.5). We have modified certain aspects of this project to maintain anonymity of the client, and to keep the length of this presentation within reasonable limits.

4.1 Prior Information

Arrangements for the time and place of the initial consultation session were made through telephone contact with the client. The project was briefly discussed and we instructed the client to bring relevant information, such as printouts of the data, to our forthcoming meeting. This contact provided us with the opportunity to obtain some prior information about the project:

- The project was postexperiment.
- The client wanted us to perform the analysis and provide graphs.
- The project was the client's dissertation study.
- It had something to do with teaching methods and learning styles.
- The experiment involved $n = 87$ teachers.

Remarks

An obvious question is, "Why didn't we gather more information?" One answer is simply that we will be asking the client to start from the beginning, which effectively means reiterating anything that was discussed prior to the initial consultation session. In our experience, trying to obtain too much information prior to the first session tends to introduce more chance for misunderstandings. This not only detracts from the objectivity needed by the statistical consultant; it can also be quite difficult to readjust preconceived notions about the data and direction of analysis.

Establishing contact with the client prior to the initial consultation session does have certain advantages, of course. It starts the communication process and provides us with the opportunity to become acquainted with the general nature of the project. In some cases, we may find that the project does **not** warrant setting up an initial consultation session. Obviously we want to extend our discussion with the client to make sure we have not misinterpreted the information that was provided by the client. We therefore need to obtain enough information to form a good opinion of the extent or level of (statistical) sophistication involved in the project. Some of the reasons why we might decide to defer or terminate setting up an initial consultation session are:

- In legal cases, a conflict of interest may arise with respect to the client's project simply because of current or previous work we performed for another client.
- The sample size is too small to justify a statistical analysis.
- It would better if the client performed certain tasks before setting up a meeting: collecting, entering, or formatting the data.
- We do not have enough resources, or expertise, with respect to the statistical analysis that would be required for the client's project. Could we design and implement a clinical trial? Do we know how to interpret a market model?

For this particular project, the prior information we obtained from the client warranted setting up a consultation session and enabled us to form the following opinions about the project.

1. The design and implementation of the experiment would need to be carefully diagnosed to ensure it satisfied the principles of experimental design: control, randomization, and replication.
2. The analysis would probably require standard ANOVA and t-test procedures. If there are problems with the data, nonparametric procedures may be necessary. The sample size may be of concern if too many factors are involved.

3. Quality graphics may be required since the project was the client's dissertation study. It was also likely that the project could be regarded as a pilot study — significant results will need to be interpreted with caution.

4. We will need to learn what "learning styles" really means.

4.2 Financial Issues

We can also approach the question, "Why didn't we gather more information?" from a purely economic point of view. An equally obvious answer is that the statistical consultant expects to be *paid* for their time and professional advice! While this may not necessarily be the answer the reader expected, it would certainly be prudent for a client to establish what costs are involved *before* committing to an initial consultation session. Thus, we will need to address the client's question:

"How much do you charge?"

Providing the client with a standard hourly rate will usually suffice, but for short-term projects the client is often looking for an overall cost. They may also want to know how "soon" we can complete the analysis. To realistically deal with these issues would require that we know the full details of the project, details which we clearly want to defer until the actual consultation session. So how do we respond? The key is to get the client to agree to set up an initial consultation session without encumbering ourselves with unrealistic cost or time estimates. Some possible strategies are:

- Provide the first consultation hour free of charge, but under a no-obligation clause. That is, we reserve the option to decline the client's project during this period, but the client is not charged for that hour whether or not we exercise this option.

- Offer a contract option to client. After the initial consultation session we provide the client with a fixed total amount for the project: data processing, analysis, report writing, and graphics as required.

In a few isolated cases, the best option may be to simply adopt a *take-it-or-leave-it* approach. We may lose the client, but our instincts suggest that this might be for the best.

Adopting this approach may become necessary whenever the previous two approaches have failed. If the client is not willing to discuss the project face-to-face, then it's probably not worth getting involved.

Remarks

While it is possible that we may completely solve a client's problem within an hour, the *first-hour-free* option is well suited to the situation where we have prior information about the client's project. This approach is often a good way to get the client to quickly agree to setting up an initial consultation meeting, avoiding the need to have a protracted discussion on specific rates and charges. These can be discussed in detail at the first consultation session.

Estimating the total time (hence the total cost) to complete all aspects of a client's project is not easy. Data processing and report writing often take much longer than we expect. Employing graduate students certainly helps to defray the overall cost and provides the student with the opportunity to gain experience, but we are ultimately responsible for the analysis. Contracts work well for small-scale projects where all the requirements of the project can be specified explicitly. The terms and conditions of the contract will need to be documented and signed by both parties.

Although university consulting programs have the ability to involve graduate students in a wide variety of projects, clients sometimes assume graduate students can be exploited on the basis that they are providing the student with "experience." This is certainly true and some flexibility is usually required on our part; students are often prepared to work short-term for minimal rates as long as the experience is beneficial. Our job is clearly to filter out the more extreme cases.

4.3 Session I: The First Meeting

We are about to meet our client face-to-face for the first time and have made appropriate preparations for the meeting. Specifically, we should:

- Relax! The client is coming to us for advice and probably feels even more nervous than we do.
- Make sure the meeting will take place in an environment that is conducive to a focused verbal interaction with the client. The student center, scheduling the meeting during teaching-related office hours, or having no space on the desk to look at printouts, clearly do not provide good environments for consulting.
- Be attired in a manner that reflects a professional standard of service. The fact that we provide consulting services within a university environment does not alter the usual business protocol.
- Have the client's file in front of us. This should contain our notes from any prior contact, sign-in form for compiling the project summary

(see Figure 4.2 later), as well as paper, pens, and pencils[1] ready for taking notes during the consultation session.

- Be punctual. Our doctorate is only in statistics.
- Make any additional preparations in advance. For example, the presence of other personnel who might be involved in the consultation: graduate student, coconsultants, or experts from other disciplines.

Based on the prior information we obtained from the client, it seemed quite likely that a graduate student would be able to perform most of the analysis involved in the client's project. We therefore arranged to have the student present for this initial consultation session. This provided the student with the opportunity to participate in the "consulting process" and also allowed the client to see "who" the student was. We address the reason why this is important later.

In addition to the routine preparations listed above, there is one more important item that we need to prepare. Us! That is, we should always try to approach the consultation meeting with some type of agenda in mind. A good way to formulate an agenda is to ask ourselves the question:

What do we expect to achieve in this session?

With practice, it will become easier to anticipate the general pattern of our consultation sessions and our agenda may simply consist of mental footnotes. Otherwise, we should write down our proposed agenda and have it with us during the meeting.

So what do we expect to achieve in this example? Since the client's study was relatively small (only $n = 87$ observations), it was possible that we could be in a position to begin performing the statistical analysis for the project by the end of the session. To achieve this best-case scenario, the following items on our agenda would all need to be properly dealt with in sequential fashion:

- A clear definition of the problem and variables associated with the client's project.
- The objectives of the study can be supported by a statistical analysis.
- The specific contributions required for this project can be clearly stated.
- The time frame and terms of payment are mutually acceptable.

---

[1] Make sure to ask clients for permission *before* writing on any of their printouts, and always use pencil. We once made the mistake of not doing this and the client almost got up and left! Fortunately they didn't, but this did make the remainder of the consultation session somewhat tense.

Remark

In this particular example, we managed to cover all of the above items in one consultation session. That is, by the end of this session we had achieved our objective of being in a position to perform the statistical analysis for the project. This will not always be possible, and we should certainly not expect to be able to resolve every client's project within a single consultation session. Indeed, the discerning reader may already be somewhat suspicious about our claim in this example. Did we *really* get through everything that we present below in one session? We answer this question at the end of the presentation.

Initial Contact

The client arrived on time for our consultation session. What happens if they don't? If the client is early (as is common), introductions can be made and, if appropriate, we can ask the new client (A) to stay while we finish up with our current client (B). The purpose of this is to allow client A to hear and see the type of interaction that they will be encountering shortly. If the client is late (by more than 20 to 30 minutes), we should still try to accommodate them on that day. It is usually better to obtain at least some information about the client's project rather than simply arrange a new meeting time. Our schedule should therefore allow some flexibility: stacking up too many clients on a single day is more likely to cause rescheduling problems. In this example, our dialogue begins with the initial introductions:

client: Hi, I'm Another Client.

cons: Hi. I'm Zee Consultant and this is Affine Student who will be helping us with your project. Now, I understand you have brought some information for us to look at. But first, perhaps we could start by having you describe your project ... ?

client: Okay. ... We wanted to show that incorporating LSI preferences in technology training helped long-term retention. In the experiment we augmented traditional methods by auditory, visual, kinesthetic and tactile preferences of the participants. We used the same instruction formats in each session and our instruments were pretest, posttest, and SDS score. Most of the workshop participants were female elementary school teachers. ... We've collected the data and I have some printouts for you to see. ...

cons: (interrupts client) ... Good, but before we do that I'd just like to go back over some details about ...

client: ... my advisor also said something about "analysis of variance"? — which is why I'm here !

cons: Well, let's find out if ANOVA is needed.

Remarks

Choosing *when* to interrupt a client is not always easy. If we jump in too soon, the client may feel we are not giving them a fair opportunity to explain their project. It is worth remembering that we will expect the client to listen carefully to our explanation of certain statistical issues later. On the other hand, it is important to keep the discussion focused and we may need to interrupt the client to avoid backtracking too far. That is, we need to start processing "batches" of information. These verbal "cues" serve the dual purpose of not letting the client get too far ahead of our questions and understanding of the project, and it helps the client learn something about what we *really* need to know.

In the above dialogue, we were able to interrupt the client at a convenient point: the client's description of the project had already raised several questions that needed to be resolved; they had moved on to the subject of data and printouts. In practice, this opportunity is quite often presented by clients who, understandably, try to shift quickly from old news (project description) to the current status of the project (data analysis).

Before presenting the questions that we need to resolve with the client, it is worth considering the client's reference to ANOVA. This provided some indication of their statistical knowledge: they knew enough to seek our help, but we need to be careful with our response. To embrace a client's methodological suggestion at this early stage is generally unwise: we may end up explaining why their "great" suggestion is **not** appropriate and we *both* come out feeling somewhat foolish. In this type of situation, we should try to sound encouraging, but avoid committing ourselves to the statistical methods suggested by the client. If their suggestion turns out to be correct, they feel good; if not, then that's why they came to us in the first place. Now back to our questions.

1. Several terms were mentioned by the client which need further explanation: LSI, traditional methods, kinesthetic and tactile preferences, SDS score.

2. How does long-term retention relate to this study and how was it measured? Details concerning the pretest and posttest instruments will be needed.

3. To what extent are the factors *gender* and *school level* of interest in this study? We will need to know details about the sample sizes involved in these categories.

4. Our main concern is that the usual "control" versus "treatment" design setup is not obvious from the client's description. Furthermore, the reference to *session* formats adds a potentially complicating design factor into the analysis. We will need the client to carefully describe the design format and implementation methods that were employed in this study.

Defining the Problem

After presenting these questions to the client we obtained the following information. During the course of this discussion we also examined the printouts brought by the client.

Design The sample consisted of high school and elementary school teachers who were randomly allocated into two groups: Control and Experiment. The traditional instruction format (textbook) was used for the Control group. For the Experiment group, traditional instruction was augmented by activities specifically suited to the preferred learning styles of the participants. Both groups were split into four sessions (subgroups) and each session received the same instruction format.

Implementation The Learning Style Inventory (LSI) instrument was first used to classify the preferences of the participants for both groups. For the purposes of this study, a single preference was assigned to a participant based on his or her highest LSI score obtained in the (A)uditory, (V)isual, (K)inesthetic, or (T)actile categories, provided the LSI score exceeded 50; otherwise the participant was considered to have (N)o preference.

In the Experiment group, specific sound (A), sight (V), role playing (K), and construction (T) activities were employed for **all** the participants, relative to their assigned preference. Thus, the N preferenced participants were also involved in activities specific to their highest LSI score in the A,V,K,T categories.

Variables For assessment purposes, three quantitative measures were employed in this study. They consisted of: a pretest (PRE) given at the start of the instruction period, an attitude scale score based on the semantic differential scale (SDS) instrument which was given at the end of the instruction period; and a posttest (POST) given one month after the instruction. Other factors recorded were GENDER, SESSION, and school level (SLEVEL). Details concerning these variables are summarized below.

Quantitative Measures

PRE	Pretest:	Maximum mark = 100

POST	Posttest:	Maximum mark = 100
SDS	Attitude Score:	Maximum mark = 60

Categorical Factors

GROUP	Control / Experiment
GENDER	Male / Female
SLEVEL	Elementary / High School
SESSION	S1, S2, S3, S4 (within each GROUP)
PREF	Auditory, Kinesthetic, None, Tactile, Visual

Sample Sizes ($N = 87$)

Control	43	Experiment	44
Female	70	Male	17
Elementary	59	High school	28
Preferences:	A = 12, K = 4, N = 32, T = 23, V = 7		
Session:	*Unavailable — to be provided by client*		

Overall Issues and Objectives

So far, the direction and purpose of the consultation session has been primarily for our benefit. We have identified the important components of the study and established properties of the variables involved. Now we need to return to the purpose of the client's study. A formal statement of the research hypothesis[2] developed by the client as the objective of this study is given below.

H_1 Teachers in technology training sessions that utilize a processing activity that matches their perceptual learning-style preferences will demonstrate significantly greater long-term retention of content than teachers in a traditional setting that has not utilized that processing activity.

This essentially translates to $H_o : \mu_{\text{Cont}} = \mu_{\text{Expt}}$, in our terminology. t-tests and ANOVA procedures may be used to investigate this hypothesis. Our first task is to explain these statistical methods to the client.

t-Tests These are used to assess whether a significant difference exists between the means associated with two independent samples. For example, we are interested in whether the Experiment group had a *significantly* higher average POST score than the Control group.

ANOVA To test the effects of several factors simultaneously, an *analysis of variance* procedure can be used. For example, we can test whether

[2]For brevity, we confine our attention to the issue of long-term retention (PRE and POST). A similar research hypothesis was also developed for attitude (SDS) differences.

GENDER, SLEVEL, and SESSION also have an effect on POST, after the GROUP effect has been accounted for. Perhaps high school teachers performed better within each group?

It is worth noting that our "explanation" is really just a statement of how we can use the method, illustrated by an example in the context of the client's project. Remember, we haven't actually performed the analysis yet. In our experience, introducing the abstract concepts of statistical inference in the absence of results tends to produce little more than a lengthy, but rather vacuous, discussion.

The rosy examples we employed for illustrating the statistical methods will need to be tempered with a realistic appraisal of the assumptions and potential issues that may affect the statistical validity of the analysis. In particular, we **must** emphasize that significance does **not** imply causality. This is not always quite as simple as it sounds. Dealing with the issue of causality often represents the boundary between what the client wants to conclude versus what the statistical analysis will actually provide. Some of the issues that we need to discuss in more detail with the client are:

1. Assuming a significant t-test result for POST by GROUP, what is our interpretation (conclusion)?

2. How should we proceed if PRE by GROUP is significant?

3. The imbalance in the GENDER and PREF levels (class sizes) may adversely affect the ANOVA procedure.

4. What happens if SESSION turns out to be a significant factor?

5. Was the assignment of a *single* preference realistic?

Remarks

As can be seen from the client's research hypothesis, *long-term retention* is asserted as the outcome associated with a significant GROUP effect. Now comes the hard part. Should we insist that the client remove this reference to long-term retention on the basis of causality? In this example we did not.

We can certainly argue that a statistically significant result only provides evidence of an *association* between the "treatment" effect (teaching methods utilizing learning style preferences) and posttest performance; it does not *prove* that long-term retention occurred. But long-term retention is the client's contextual interpretation of what posttest performance implies, which seems reasonable given the experiment's design and purpose. Whether we necessarily agree with this type of subjective interpretation, we must be careful not to impose our expertise on the client's field of study.

Our responsibility is to make sure the client clearly understands that statistical evidence is strictly that; it is not proof. The subjective nature of a contextual interpretation is ostensibly the client's responsibility.

Since the study is postexperiment, the main focus of our appraisal of the statistical issues associated with the client's project is on identifying potential sources of nonrandom error. That is, can we find any type of bias effect that may seriously compromise the statistical validity of the analysis? In this situation, where we are reliant on the client's description of the design and implementation of the experiment, it can be useful to employ the interrogation approach.

The client provides their interpretation of a hypothetical result

which we pose for them, such as a significant t-test for POST versus GROUP, and then we ask "what if ..." scenarios.

The purpose of this exercise is **not** to try to "trip up" the client — no experiment is going to be perfect — but to provide an independent and objective assessment of the client's study. In many cases, both the client and consultant gain a much better understanding of the issues surrounding the investigation: we may hit on an issue that resolves a contextual problem for the client, and the client may remember details about the study that turn out to be statistically important.

So what did we both learn in this example? The following items correspond to the issues that we discussed in detail with the client.

1. Some of the issues involved with interpreting a significant result for the POST by GROUP t-test are:

    - As is common in small-scale studies, the sample profile tends to limit the extent to which a significant result can be applied to a larger population. Our conclusions will therefore need to based on the suggestive value associated with a "pilot" study.
    - The implication of long-term retention based on a significant POST versus GROUP result is clearly subject to debate. This is a contextual interpretation of the GROUP effect; the statistical result does not *prove* this assertion is necessarily correct. A secondary issue is whether one month really constitutes evidence of "long-term" retention. Both issues were deferred to the client's judgement.
    - This t-test only evaluates the knowledge of the participant at the one month time point. Does the client really know what the participants did during the intervening month? For example, suppose the participants in the Experiment group were so turned off by those ridiculous preference activities, they all enrolled in a

fast-track technology course utilizing traditional teaching methods.[3]

- The POST versus GROUP result does not account for different knowledge levels (baselines) that may exist between the participants prior to the instruction. The pretest will be employed to account for possible baseline differences.

2. For the client, a significant PRE by GROUP result would have presented some difficulties. It suggested their design was somehow flawed and complicated the interpretation of posttest performance. Neither of these is necessarily true, of course, but it is not unusual for clients to employ pretests with "crossed fingers": they know pretests are important; they just hope they're not significant in their study! In this situation, we should spend some extra time with the client to help clarify the following statistical aspects.

   - Randomization is used to avoid selection bias. For example, we would never consider $\{1, 2, 3, 4, 5\}$ to be a "random" lotto drawing, but this is just as likely as any other set of five numbers! The client's design is not flawed simply because they obtained two groups with different pretest performances; it's just the *luck of the draw.*
   - Posttest performance can be assessed by employing a *paired* t-test procedure on DIFF = POST – PRE. That is, we consider the *difference* between a participant's PRE and POST scores. This actually makes more sense since we would generally expect a positive relationship between these scores for each participant (i.e., high PRE $\Rightarrow$ high POST). The paired t-test accounts for this baseline effect.
   - The DIFF by GROUP t-test also has another advantage. If there is a wide range of POST scores within each group, this may mask the GROUP effect. By accommodating the participant's baseline, a meaningful GROUP effect can be more easily detected.

3. The imbalance in the GENDER and PREF levels (class sizes) may adversely affect the ANOVA procedure. It is possible that we may obtain spurious results which are simply an artifact of the small sample size. Note that the number of K and V participants will be further split by GROUP in this analysis. A more likely outcome is that these factors will not be significant. This does not necessarily mean that these factors are unimportant; there just isn't enough information

[3]We were assured that this did not happen. The point here is that the consultant needs to be creative in generating "what if" scenarios.

in this study to make a meaningful determination. For the PREF factor, it may be worthwhile combining certain levels. The client recommended A + V and K + T.

4. Although the intention of this study was to provide identical instruction formats in the four sessions within each group, it is quite possible that the performance levels may vary significantly between SESSION. This was not an important concern for the client.

5. Two issues associated with the assignment of a *single* learning-style preference were:

   - Some participants could be multipreferenced. The decision to assign a specific preference to these participants therefore introduced some subjectivity into the study.
   - Nonpreferenced participants should be analyzed separately.[4]

Specific Contributions

Time to wrap things up. We have used up most of the time that was scheduled for this consultation and now need to set up the postsession agenda for the client. This is very important: the client should *always* be provided with a sense of what has been accomplished at the end of a consultation session. In this example, we are in a position to address the specific contributions of the client's project:

- What we will be doing for the client;
- What we may need from the client;
- Determining the time frame and costs.

Consultant From our discussion of the overall issues and objectives, the client knows how we intend to approach the statistical analysis of the project. Thus, the main purpose of this agenda is to provide the client with an outline of the steps involved in the analysis and to establish any specific requirements associated with the project. Based on this information, we will then determine a realistic time frame for completing the work.

1. Data processing: Database transfer and error-check analysis. Corrections to be provided by the client.
2. Exploratory data analysis: Summary statistics for all the variables in the database, overall and by GROUP.

[4]This was an important part of the project which we have left as an exercise.

3. Statistical analysis:
   t-tests: PRE, POST, and DIFF by GROUP
   ANOVA: PRE, POST, and DIFF with all explanatory factors: GROUP, SESSION, PREF, GENDER, and SLEVEL.
4. Written report: Will include an explanation of the statistical methods employed (requested by the client).
5. Presentation quality graphics: bar charts and histograms. Scatter plot (or equivalent display) for presenting the *t*-test results. Final selection to be made by client.

Client In this example, the client provided us with the database already in a suitable electronic format (Excel file on a disk). The printouts we had examined previously were self-explanatory, but did not include the SESSION codes for the participants. Since this was a relatively small sample, a simple approach was to have the client just email us the list of the SESSION codes (with the participant's ID code).

In general, clients are responsible for the data collection and data entry phase of the study. They will also need to be able to provide corrections for possible errors identified by the consultant during the error-checking analysis of the database. To assist with this initial phase of the analysis, the database should adhere to the following.

Software Excel spreadsheet (disk) or plain text document (email).

Format Rectangular. Each row is an observation, with an entry in each column (the variables).

Missing A special code (e.g., "−9") to be used for missing values. There should be **no** blank entries in the database.

Encoding: A key listing all the codes used in the database.

Time and Costs Involving graduate students in projects is one of the main aims (and advantages) of university consulting programs. This particular project was ideally suited for a graduate student and our primary role was to oversee the progress and provide assistance as necessary. Of course, our intention to employ a graduate student needs to be discussed with the client. Some of the main issues are:

1. We are ultimately responsible for the project; not the student.
2. The cost benefit to the client needs to be weighed against a more flexible time frame. The student will have other commitments

such as coursework, exams, assignments, and teaching duties which are equally important.

3. Students often find the written report to be the most difficult task and several drafts may be required. The client should **not** be charged for this learning experience!
4. Creating presentation quality graphics can be very time consuming and some students may not (yet) be proficient with high-level graphics in statistical software such as SAS and S-PLUS.
5. Confidentiality or other constraints associated with the project would extend to the student.

Session Summary

In this example, we were in a position to wrap things up and address the specific contributions for the project. That is, we had essentially resolved what needed to be done for the client's project and any further communication before our next meeting would be performed indirectly — by phone, fax, or email. Our next meeting would take place when the results and preliminary report were complete. These would be sent to the client prior to that meeting which would focus on the interpretation of the results.

Did We Really Do All This in One Session?

The answer to this question is yes ... and no. In reality, more than one consultation session would have been required to address the issues of this project in the detail presented here. So how can we claim only one session was needed? ... We took *extensive notes* during the session. That is, we relied on our notes to reconstruct certain details about the project *after* the consultation session was over. Our presentation of the consultation session is therefore the result of merging these two sources of information:

- Information we obtained during direct interaction with the client;
- Details that we reconstructed from our notes.

The main advantage of this approach is that it allows us to gain a good overview of the client's project. The obvious benefit for the client is that they only need to describe the details of their project once.

Additional Sessions

Even in small-scale projects, it may not be possible to resolve the issues associated with a client's project within a single consultation session. The discussion of the overall issues and objectives of the project may still be incomplete. Keeping track of time is therefore important since we should always try to reserve a sufficient amount of time at the end of a consultation

session to outline an agenda of future activity for the client. This should **not** just consist of scheduling another meeting with the client; they first deserve to know what was achieved from *this* meeting.

Briefly summarizing the main points of the client's project, or outlining the statistical approach that we are considering, are simple and effective ways for making the client feel the session was productive. In this situation, we should also try to set up a "task" for the client to perform before the next consultation session. Although this will depend on the particular nature of a project, some common examples are:

- Starting to create the database (or template for one)
- Getting copies of relevant references that the client cited
- Revising or creating a draft of a questionnaire for review
- Writing a protocol draft for an experiment
- Reporting pertinent issues back to an advisor/supervisor.

Students

Clients are sometimes apprehensive about having students involved in their project. If the client's demeanor suggests that this might be the case, act promptly to reassure the client in this situation. Their main concern — the quality of the analysis — may potentially lead to more negative feelings: they are being palmed off; that we do not have a vested interest in their project, and so on. Having the graduate student present during the consultation session is certainly helpful in this regard. The student obviously benefits by having the opportunity to participate in the "consulting process" and, the client knows "who" the student is. Whether the student is present or not, we will still need to emphasize our responsibility for the analysis to the client.

Graphics

Reasonable diagnostic graphics are usually provided as part of the output from a statistical procedure. Adding titles, labels, and other annotation may take a little more effort, but any serious modification for presentation purposes will be time consuming. This is really the key issue and our decision to produce presentation quality graphics for a client needs to be considered carefully.

In this example, the student had the opportunity to gain some experience with S-PLUS and so what we charged the client did not reflect the true costs (time) involved in creating the requested graphics. The ability to undercharge a client is not a luxury we can always afford, however, and some issues to consider with regard to presentation quality graphics are:

- What is the level of quality required?
- How complex are the graphics?
- Is color required?
- Aesthetics?
- How many graphs are needed?

Standard output from the statistical software with titles, labels, and other appropriate annotation, may actually be sufficient. If not, an example should be used to establish the level of quality required. In this situation, we should emphasize that simple displays can be quite effective and are easier to modify. Multivariate displays involving grouping variables, contours, or 3D representations require far more effort to "get it right." Note that grey-scale should be employed for displays that are likely to be photocopied; color requires additional resources and what we see on the screen is **not** necessarily reproduced on hardcopy.

What happens if the client doesn't like the design or aesthetics of our resulting display? We will need to be able to convey to the client what the graph will look like *before* creating it. Finally, creating a special (unique) graph takes time. The number of special graph "formats" should be kept to a minimum and each used for more than one display.

4.4 Documentation

The consultation session is over. The client has left with the outline of the work to be performed and we provided the client with a contract estimate for the cost of our services. As with any exchange of services on a contract basis, the agreement should be formalized in writing and signed by both parties. Do **not** perform contracted services solely on the basis of a verbal agreement. Otherwise, be prepared to experience that wonderful feeling of putting in a great effort, making the deadline — and not getting paid! A simple contract outline is presented in Figure 4.1.

We should emphasize that this type of contract is really just a "gentleman's agreement" and certain details will need expanding (e.g., the amount of the payment, when the payment should be remitted, etc.). It should **not** be used as a substitute for a proper legally binding contract. Such contracts are required where confidentiality, intellectual property, and liability issues need to be carefully addressed in legal terms. Small business consulting firms may also want to protect their clientele base when subcontracting projects to independent consultants.

The project summary outline shown in Figure 4.2 serves two purposes. It provides the client with a list of the main tasks we will perform, and it provides us with the type of information we can use for reference purposes. This

STATISTICAL CONSULTING PROGRAM
SCP Letterhead Information

CONTRACT

Date
Consultant's Name and Address

Client's Name and Address

Project Title
Project Description

Client agrees to the services and conditions to be provided by the SCP as detailed in the Project Summary attached. On completion of SCP services client agrees to remit payment ...

Signed

Z. Consultant	A. Client

SCP/Client Additions:

- The SCP analysis is based on the information and database provided by the client. To the best of the SCP's knowledge the integrity of the database, and information provided by the client, is without prejudice.
- The statistical computing will be performed on a subcontracted basis by a graduate student in statistics under the supervision of the SCP consultant.
- The SCP is to provide the client with presentation quality graphical summaries.
- *Other items as necessary.*

FIGURE 4.1. Sample Contract Outline

STATISTICAL CONSULTING PROGRAM
SCP Letterhead Information

PROJECT SUMMARY

Date
Consultant
Student Assistant

Client
Client's Contact Information

Summary:

Project Title

Summary statistics as requested by client

Statistical analysis of project hypotheses

Written report to be provided to client

Graphical summaries as requested by client

Contract estimate: [ ]

Important:
The SCP assumes you accept the contract estimate with the understanding that the final SCP invoice may include additional charges. You will be informed of the need for any increase prior to the SCP performing ...

FIGURE 4.2. Outline of the Project Summary

documentation contains the pertinent information about the client and the project which can be easily retrieved from our "Client File" database. The project summary would also be included in the client's "Main File" which will eventually contain all the documentation associated with the project.

4.5 Project Analysis

The signed contract and missing SESSION codes were received from the client and a backup copy of the original database was made. Most of the project analysis was performed by the graduate student and consisted of the following:

- Data processing
- Exploratory data analysis (EDA)
- Statistical analysis
- Draft version of the written report
- Example of a presentation quality graph.

Data Processing

The first task was to download the client's data from Excel format into something more useful for statistical purposes. In this example, we employed the statistical software package JMP, which can import Excel files.[5] The SESSION codes were merged into the JMP database and a plain text (ASCII) version was output for later use in SAS and S-PLUS.

The numerical and graphical summary statistics provided by JMP did not indicate any obvious errors and agreed with the printout information that we had examined during the the consultation session. Thus, we were able to proceed directly to the exploratory phase of the analysis.

Remarks

Usually we will not be this fortunate. Our client just happened to be particularly thorough and avoided the type of common mistakes that can occur in practice (see Section 3.2, *Data Processing*). We had also seen a printout of the client's data which motivated our choice of JMP. Minitab and Statgraphics are other examples of sufficiently comprehensive, menu-driven packages that could be used in place of JMP. So what made us choose JMP?

[5]Version 4 of JMP can import an Excel file directly. JMP 3.x requires a "Save as text" version of the Excel file.

1. The sample size involved was relatively small.
2. The database was in rectangular format with no missing values.
3. The point-and-click, menu-driven interface of JMP would make it very easy to perform the exploratory and initial statistical analyses.
4. The standard diagnostic displays would provide the client with a useful basis for deciding on the presentation quality graphics.

Clearly, we were rather fortunate in this example. What about a more general situation? We consider the following aspects.

Plain Text In most cases, we would convert the client's database into plain text format and use SAS or S-PLUS to perform the data processing. One of the main advantages of plain text files is that they can be imported into (almost) any application irrespective of the computing platform being used. In particular, using plain text attachments in email ensures that both parties will be able to read the information.

An obvious disadvantage is that we lose the special formatting of the original database. Working efficiently with plain text files therefore requires a good editor such as Emacs — and, of course, knowledge of how to use it.

Client Database In the "unlikely" event[6] that the client provides us with a plain text file, what database format should we ask the client to provide us? Excel is widely available, can be used by a novice, and from our perspective, is easily converted into plain text format:

File → *Save as ...* → *Text (Tab delimited)*

If you, or your client, prefer to use something other than an Excel spreadsheet, make sure both parties have the desired application.[7] Translators do not always work well, especially across different platforms.

Coding Many of the data processing problems arise from improper coding. For example, survey questions that allow the respondent to select more than one option should be recoded with separate dummy variables for each option. Similarly, a dummy variable can be used to encode write-in comments or responses to "Other, explain" since these are usually of nonstatistical interest. We return to this issue in the case studies presented in Part II.

---

[6] Clients often use spreadsheets, such as Excel, but do not really understand that all that wonderful formatting usually needs to be discarded by the statistical consultant.

[7] One of our worst experiences was a case where the client had entered a large amount of data into an application that was completely outdated. Eventually, we were able to write a C program to extract the database directly from the ASCII octal code!

EDA

The numerical and graphical diagnostics from JMP indicated the presence of two potential outliers in the Control group for POST (low scores). However, both these participants also had low PRE scores and no outliers were present in the derived variable, DIFF = POST − PRE. For the categorical variables, PREF and GENDER showed the most disparity in their class sizes. CPREF was therefore created by combining the PREF classes, A + V and K + T, with CPREF = N making up the third class.

We used JMP to perform *t*-tests and to fit various ANOVA models (see *Statistical Analysis* below). The main purpose of this exercise was to gain some insight into the results that would be investigated in more detail using SAS, and to evaluate the assumptions underlying these statistical procedures. As can be seen from Figure 2.3 in Chapter 2, the distribution of the POST scores by Group exhibits nonnormality. However, the two-sample *t*-test is robust against departures from distributional assumptions.

Significant results with respect to the GROUP factor were only obtained for POST and DIFF. None of the other factors were significant in the ANOVA models (including CPREF). The JMP results and diagnostic graphics were printed and saved for review at our forthcoming meeting with the client.

Remarks

We could have actually completed the project analysis using JMP alone since the results above essentially formed the basis of the written report. However, a disadvantage of menu-driven systems such as JMP is that regenerating output usually requires repeating all the point-and-click interface interactions that we had performed previously. This is fine for exploratory purposes, but rapidly becomes rather tedious and inefficient when an entire analysis needs to be replicated.[8]

For report writing, a more serious disadvantage is that the output from JMP can not be saved as a plain text file. Hence we can't edit, delete, or modify parts of an output "file" to suit our purposes, nor can we email (readable) output from JMP to our client. Given all this, why did we even consider using JMP? Our answer is simply that JMP was the right tool for what we wanted to achieve in the exploratory phase of this analysis.

1. It is easy to use (and doesn't take long for a new student to learn).
2. The diagnostic summary statistics and graphics are good.
3. There is a comprehensive range of statistical methodology available.

---

[8]We once made the mistake of providing a client with our only copy of some JMP output. Unfortunately, we became a little enthusiastic in our explanation of the results and wrote numerous helpful comments on this version. At the end of the consultation session, the client requested a clean copy of the output!

There are other statistical software applications that also satisfy the three requirements above and could be used in place of JMP: Minitab and Statgraphics being two examples we have mentioned previously. The main point is that the statistical consultant really needs to be fluent in more than one application. Using the strengths and advantages of different software will help make the consulting process more productive.

Statistical Analysis

The t-tests and ANOVA models considered above were rerun using SAS. The results we obtained previously did not change, of course, but the SAS output could now be incorporated in the draft version of the written report. The following SAS code gives the basic steps of the analysis.

SAS code for analyzing client's data

```
data a ;
  infile 'client.dat' ;
  input id grp $ session $ pref $ gender $ slevel $
        pre post sds ;

proc freq ;
  tables grp session pref gender slevel
         (session pref gender slevel)*grp ;

proc means ;
  var sds pre post ;

proc ttest ;
  class grp ;
  var sds pre post ;

proc glm ;
  class grp session pref gender slevel ;
  model sds pre post = grp session(grp) pref gender slevel;
```

Although not shown in the SAS program above, the variables DIFF and CPREF were created and several variations of the ANOVA models were investigated. As in the JMP analysis, GROUP was found to be the only significant factor in the POST and DIFF models. Mean comparisons and residual diagnostics provided further support of this result. Satisfied that the statistical analysis was now complete, we obtained two versions of the output from the SAS program:

SCP Our version which contained the full range of summary statistics, diagnostic checks, and additional variables such as DIFF and CPREF in the t-tests and ANOVA models.

Client The client's version which only contained the pertinent results. This version was edited and incorporated in the preliminary report.

Remarks

As we indicated in Chapter 3 (see Section 3.8, *Statistical Software*), we are assuming that the reader is familiar with SAS programming statements as well as standard conventions (such as the need for `grp $` when the `grp` column contains character values). While we may get away with assuming what we like about the reader, this is not the case with our clients; they are certainly not expected to understand SAS code! However, they will need to be able to understand the output generated from this SAS code. For the benefit of the unassuming reader we briefly describe what this SAS program does.

SAS program summary:

`data a;` Reads 9 columns of data from the file `"client.dat"`

`proc freq;` Generates frequency tables (4 two-way tables all versus `grp`)

`proc means;` Computes means, variances, etc. for quantitative variables

`proc ttest;` Performs t-tests of SDS, PRE, and POST by GROUP

`proc glm;` Fits ANOVA models for SDS, PRE, and POST.[9]

In this example, the SESSION factor was "nested" within GROUP which is denoted by `session(grp)` in the model statement of the general linear models procedure, `proc glm`. This follows from the fact that session S1 of the Control group is unrelated to session S1 of the Experiment group. To compare different session classes across groups is clearly meaningless since the session "labels" were arbitrarily chosen.

Preliminary Report

Our intention is to email the client a draft report containing our conclusions based on results contained in their version of the SAS output. To avoid potential complications and confusion on the client's part, we performed additional editing of the output. Although the following items are specific

[9] As written, `proc glm` will perform a MANOVA analysis in addition to the (univariate) ANOVA models requested. Although we examined the MANOVA analysis, it was edited from the client's version of the output.

to SAS, they show that trying to make things easy for our client can take some effort. (See Section 4.7, *Final Report,* Tables 3 and 4.)

1. The F-test for homogeneous variances given below each t-test result was deleted.

2. The Type I sums of squares for the individual factor effects in the ANOVA models were removed.

Having taken care of the results presented in the output, we were left with the task of completing the report and addressing the issues below. These were dealt with as shown in the Final Report (Section 4.7).

- Significance. What results were significant and why? (P-values)

- `SESSION(GROUP)` appears in the ANOVA output. What was the reason for using a *nested* session effect?

- Type III sums of squares. Why are these used to test an individual factor?

The draft report was completed and emailed to the client along with a request for scheduling our next meeting. The client responded to this request and also added that they had several questions concerning P-values and significance that they wanted to ask at the forthcoming meeting. Prior to this meeting, our one remaining task was to produce an example of a presentation quality graph for the client.

Remarks

Why did we bother with editing the output? Two reasons:

1. If the client doesn't understand something in the output, we will need to explain it.

2. We needed to do this for the Final Report anyway.

Let's go back to the first reason for a moment, and assume that we didn't edit this information. What will our explanation be? "Don't worry about it. It's not important!" — [Client] "Then why do I need to have it?" ... Perhaps we could try our first response again? The point is that clients **do** worry. They have no way of knowing whether something is irrelevant. That's why they come to us.

Presentation Quality Graphics

We used S-PLUS to produce the presentation quality graphs for the client. Although we had hoped to be able to email a PostScript version of the graph to the client, they did not have the resources (PostScript printer) to be able to print the file. The following S-PLUS code provides a very rough outline of how Figure 2.3 in Chapter 2 was produced.

S-PLUS code for generating Figure 2.3

```
yy <- read.table("client.dat")    # read in data to S-PLUS
y  <- yy$post                     # extract the POST scores
ys <- yy$grp                      # extract the GROUP codes
g.hist2.fun(y,ys,...)          # call our customized graphics
                               # function (options not shown)
```

Outline of our customized graphics function

```
g.hist2.fun <- function(y,ys,signif=T,...)
{
  yb <- split(y,ys)                # split POST by GROUP
  xm <- c(mean(yb[[1]]),           # get POST means by GROUP
          mean(yb[[2]]))

  par(new=F,mar=c(6,4,3,4)+0.1)  # Figure Region Margins

  # ------------------------ This is the key step:
                                   # Create 2 sub-Figure
  fg <- list(c(0,1,0.4,1),         # Regions. One takes 60%
             c(0,1,0,0,6))         # of the top; the other
                                   # takes 60% of the bottom

  for(i in 1:2){                   # Loop through the two
       par(new=T,fig=fg[[i]])      # histgram plots, using
       hist(yb[[i]], ... )         # the sub-Figure Regions.
       legend( ... )               # Add Legend (indexed) and
       abline(v=xm[i])             # put vertical line at mean
  }
                                   # If significant, put this
  if(signif)                       # text on plot (position
      text( ... ,                  # arguments not shown)
           "** Means Differ Significantly ** ")

  polygon( ... )                   # Draw the bowtie polygon

  box()                            # Box around everything
  title( ... )                     # Title and subtitles

  invisible()                      # Make plot invisible ??
}
```

This function was also used to create histogram displays for the non-significant PRE and SDS by GROUP t-test results. Setting `signif=F` in the options line: `g.hist2.fun(yy$pre,ys,signif=F,...)` would skip the

`text()` step when creating the PRE by GROUP display. The default option was to print the text as shown in Figure 4.3. Many other options and arguments were required to produce this display, of course, but the key steps are shown above. Note that the statement `invisible()` does not really make the plot invisible. S-PLUS complains if a function doesn't return some value; `invisible()` returns a special "nothing" value which stops it from complaining.

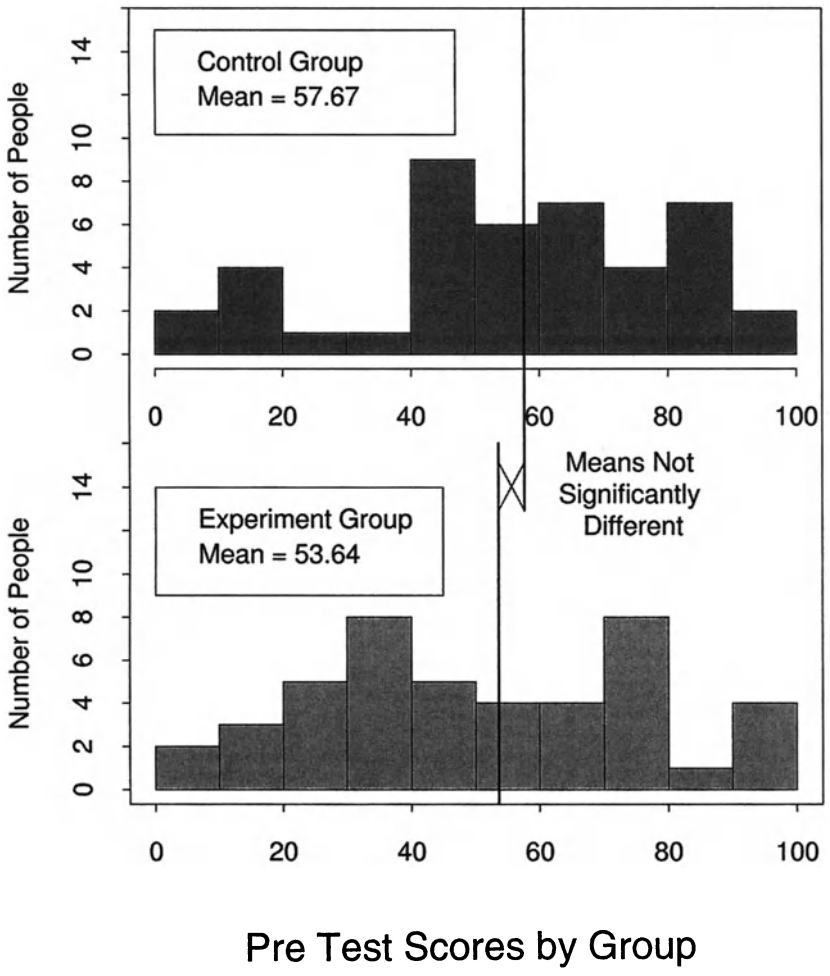


FIGURE 4.3. The Nonsignificant t-Test Display

Since the PRE and POST by GROUP displays would provide us with a very effective way of visually illustrating the t-test results to the client, we also produced Figure 4.3 for the forthcoming consultation meeting.

4.6 Session II: Presenting the Results

The client arrived on time for the consultation session, preliminary report in hand, and questions armed and ready. ... (Somewhere along the way we did actually exchange greetings!)

client: I've read through the report and I'm afraid I have a lot of questions to ask you ...

cons: (jumping in quickly and not pausing) Really? I must have sent you the wrong report! No, seriously. I would have been very surprised if you *didn't* have a lot of questions. So what I thought we could do is show you some of the output that Affine Student has prepared for us first, and see if that answers some of your questions as we go along. Then we can look at the report that you brought with you. How does that sound?

client: Okay ... great. But can I just ask one question before we start?

cons: Sure.

client: I'm really confused about the "Equal/Unequal" parts in Table 3 ... and the ANOVA tables too; which P-value am I supposed to be looking at? And ...

cons: (interrupts client) ... Wow, long question! I thought you might find those parts a bit confusing. But I think if we start with the easier parts first, the t-test and ANOVA results won't seem nearly as bad as they look when we get to them. Any other questions before we begin?

client: Well, um ... No.

Phew! Getting the client to relax and hold back their questions at the beginning of the meeting is often the hardest part of the (interpretation) consultation session. We need to be reasonably firm about setting the agenda for the meeting, but not overpowering. Hence, we must let the client ask the inevitable "just one ..." question. Our response to this question should again be a reiteration of the agenda for the meeting. Clients usually only need one round of this to realize that we seem rather determined about this agenda thing. Adding a little humor (if appropriate — don't force it) can help take the edge off our determination.

The Agenda

Our agenda for this meeting is fairly straightforward: we present the results and tell the client what they mean. Well ... yes and no. The agenda is certainly obvious, but presenting the results so the client actually understands them requires good preparation. Simply telling the client whether a certain result is significant does little to enhance the client's understanding of the analysis. They will more than likely contact us in a day or so and ask why such-and-such a result was significant. Result: a wasted consultation session for both us and the client.

So how should we prepare? As might be expected, visual explanations provide the best mechanism for explanation and ease of understanding. Where possible, use graphs instead of tables of statistical output. The advantage with graphs is that even complex displays (such as mosaic plots) can be interpreted much more easily than tables. Detecting signature patterns in a table takes experience; something we clearly do not have time to properly develop in a client new to the game.

Graphs are great, but only part of the story. ANOVA tables, for example, are rather difficult to visualize graphically. We will need to refer to tables at some point and so an important part of our preparation is to provide a logical and cohesive presentation. In this example, the main aim of our agenda was to progress slowly from graphical displays to the numerical output. The agenda therefore consisted of the following sequence.

1. Univariate summary statistics. Bar charts and histograms.
2. Bivariate statistics. The two-way frequency table.
3. t-tests. The presentation quality graphics we produced for the PRE and POST by GROUP results.
4. t-tests. The SAS output explained with reference to these displays.
5. The ANOVA tables. No graphics.

Since the student had worked on the project, this was a good opportunity for the student to gain some experience with presentations and learn to interact directly with the client. In this case, it also has the serendipitous effect of taking the client's focus away from us for a while, given that we have just railroaded them into accepting our agenda for the meeting.

Student's Presentation

The student started by showing the client the univariate summary statistics: bar charts, histograms, and frequency tables. As expected, the client had no difficulty following this part of the analysis, the main purpose being to initiate discussion about which displays the client thought should be converted to presentation quality.

The student needed to briefly explain about the cell row, column, and total percentages in the two-way tables, but the client decided this format was more than they needed. We suggested merging the GROUP breakdown into the one-way tables (see Table 1, page 185). Perfect! The client decided on bar charts split by GROUP for the PREF and GENDER variables. (The PREF by GROUP bar chart is shown in Figure 3.1, Chapter 3.)

The only problem with the means output was deciding how to present, graphically, the average scores for SDS, PRE, and POST by GROUP. We suggested three separate bar charts: one for each variable, with two bars on each plot, the heights of which would correspond to the mean scores from the two groups. The client agreed with this suggestion for SDS (since it was on a different scale), but was interested in the possibility of combining PRE and POST. This was certainly possible (and had the same "graph format" as the PREF, GENDER by GROUP bar charts above), but we were concerned that the POST bars would visually dominate such a plot. That is, the impact of the important *nonsignificant* difference between the PRE bars would be lost. The client understood our point but was not totally convinced. Both versions would therefore be produced.

t-Test Results

At this stage, the client was quite relaxed and seemed to have no problems with understanding the student's presentation. We now introduced the histogram displays of the *t*-test results for POST by GROUP (Figure 2.3, Chapter 2), and PRE by GROUP (Figure 4.3). The client was suitably impressed and no longer seemed concerned as we discussed the PRE and POST *t*-test results in Table 3, page 186, in relation to these two graphs.

The notion of *P*-value was initially discussed in terms of the likelihood that, given there was really no advantage in incorporating learning style preferences, the client's experiment produced a bowtie (difference between the means) of width observed in the histograms. The POST result clearly did not appear to support this hypothesis, whereas the PRE result did. The conventional value of 5% (P-value < 0.05) was introduced as the standard criterion for assessing significance. While the existence of this "magic number" for assessing significance was certainly not new to the client, interpretating the PRE and POST results in statistical terms was clearly the hard part. Thus, we slowly and carefully walked our way through the "statistical" conclusion that applied to each of these results. To check whether we had been successful, we asked the client to interpret the SDS result. The client was surprisingly good with the contextual interpretation and only needed a little prodding to check whether the standard deviations were comparable.

ANOVA Results

The slow but steady transition from graphical to numerical output seemed to have done the trick. The client was now ready for the ANOVA results (Table 4, page 187). We first discussed the issue of SESSION being a nested factor in the model:

```
POST = GROUP + SESSION(GROUP) + PREF + GENDER + SLEVEL
```

The client agreed that session S1 of the Control group bore no relation to S1 of the Experiment group, hence that session differences were only relevant within each group. This explained the presence of `SESSION(GROUP)` in the ANOVA output. The next step was to introduce the idea of ANOVA as a two-step process.

1. Did any of the factors have an effect? (The overall model F-test)
2. If so, which ones? (The Type III F-tests)

Step 1 was reinforced by the PRE and SDS results; we needn't go any further. That left the POST result. The model was significant (P-value = 0.0270) so it made sense to examine the Type III F-tests. We simply told the client that these so-called "Type III" F-tests were required for two reasons: none of the factors had an equal number of participants in their respective classes (the design was unbalanced); and we wanted to see whether a factor was still important even after accounting for the other factors (partial sums of squares). We suspected that the client didn't really understand the second reason, but they had no difficulty identifying GROUP as being the only significant factor in the POST model.

Conclusions

The client was somewhat disappointed that the PREF factor didn't show up to be significant (which added to our suspicion above). We quickly pointed out that the absence of a significant PREF (or CPREF) effect did *not* necessarily mean that this factor was unimportant. Rather, that there was insufficient evidence from the client's experiment to support this conclusion with regard to *individual* preferences. Since GROUP was significant, it followed that we could reasonably conclude that incorporating learning styles *did* have an impact on POST scores. Which particular preferences were more important was not indicated by this study. This seemed to help the client and eased their disappointment about PREF.

We tried (gently) to remind the client about the issues involved in associating the POST result with long-term retention, but this was a very short-lived discussion. The client's interpretation of posttest performance would be in terms of long-term retention. Instead, we switched the discussion towards the limitations of the inference and the extent to which our conclusions could be applied to a larger population profile. The client

agreed that their sample should not be considered as representative of the general population of elementary and high school teachers, but this concurred with our suggestion that the client's experiment be regarded as a pilot study.

Closure

The client was very pleased with what had been achieved in this meeting and made one further request: "Would we be able to review their methods and results chapter draft?" (for their dissertation). This still needed to be written by the client. We agreed to perform a review of the statistical aspects[10] of that chapter. We closed out the meeting by briefly discussing and summarizing what remained to be done for this project.

- Reiterating the specific graphs that were to be produced in presentation quality format
- Revising the preliminary report as per the client's requests
- Sending the Final Report, invoice, and graphics to the client
- The terms of payment
- Followup review of the client's methods and results chapter.

Remarks

And so another project draws to a close. There are still some loose ends to tie up, but we will have no further direct contact with this client. In our experience, this particular stage of the consulting process is where the "let-down effect" is most likely to occur. (The exact medical term probably has the word syndrome somewhere in it, but for our purposes the above will suffice.) What are we referring to here? Shouldn't we feel pleased with, perhaps even proud of, our efforts? Of course. But it's over, and sometimes it may be hard to just "detach" ourselves from this reality. In this situation, a good strategy is to try to divorce ourselves from that particular project, for example, by working on something completely different, or simply taking a break. On other hand, the "let-down effect" (if any) can also be a very positive experience: at last, we're almost finished.

The main point we are trying to make is that interacting with clients can take a lot of emotional energy. How that plays out after the client has gone obviously depends on many things and we may feel no effect at all. Just don't be surprised if there is.

---

[10]The emphasis on "statistical" makes it clear that we will be checking this aspect only. We will **not** rewrite the client's chapter.

Finishing up the Project

The report was revised and all the presentation quality graphs were produced. Apart from our promise to review the client's methods and results chapter when they had written it, the following documentation would finish off the project. Examples are shown in the figures indicated.

- Cover letter (Figure 4.4)
- Invoice (Figure 4.5)
- Title page (Figure 4.6)

STATISTICAL CONSULTING PROGRAM
SCP Letterhead Information

Date of Letter

Client's Name and Address

RE: SCP Report, invoice, and graphs

PROJECT: **Analysis of Dissertation Project:** Effects of Accommodating Perceptual Learning-Style Preferences on Long-Term Retention and Attitudes Toward Technology of Elementary and Secondary Teachers in Professional Development Training.

Dear A Client,

The SCP Report, invoice, and graphical summaries for the above project are enclosed. Please make your check payable to ...

If you have any questions, please do not hesitate to contact me.
Sincerely,

Z. Consultant
Title information

FIGURE 4.4. Example of a Cover Letter

Remarks

Adding our "title information" (academic degree, position) to the cover letter makes it clear that we are qualified to perform this type of statistical analysis. The client may not be the only person who reads our report.

The invoice example is provided for illustration purposes only. It does **not** reflect a complete accounting of the charges (hours) associated with the actual project, nor do we advocate this as a "standard" format for invoicing. Notwithstanding the "as-is" label, we note the following:

- The hourly rates (not shown) would include any overhead cost, unless this *must* be listed separately.
- NC (no charge) is included for the client's benefit. They can see that we honored the first-hour-free appointment and did not charge for the five minutes it took to convert and process their Excel file.
- We have charged for services in hour units which will not always be practical. So to avoid (vulgar?) fractions or double decimals, some consultants employ tenths-of-an-hour (six-minute) units — and a good watch, presumably.

4.7 The Final Report

STATISTICAL CONSULTING PROGRAM

CONSULTANTS:	Z.Consultant and A.Student
DATE:	January 1, 2000
CLIENT:	A. Client
PROJECT:	Dissertation Project

1. Introduction:

The aim of this study was to investigate whether instruction based on a person's learning-style preferences would improve retention of the material taught.

1.1 Study Design:

The sample consisted of high school and elementary school teachers who were randomly allocated into two groups: Control and Experiment. The traditional instruction format was used for the Control group. For the Experiment group, traditional instruction was augmented by activities specifically suited to the preferred learning styles of the participants. Both groups were split into four sessions and each session received the same instruction formats.

STATISTICAL CONSULTING PROGRAM
SCP Letterhead Information

INVOICE

Date *Client*
Consultant *Project*

Services	Hours	Amount
Subcontracted:		
Data Preparation	1	NC
Statistical Computing	2	
Documentation of Results	3	
Subcontract Total @ [ ] per hour	5	[ ]
SCP Consultant:		
Appointment: *Date*	1	NC
Statistical Analysis	1	
Report Preparation	1	
Appointment: *Date*	1	
Presentation Quality Graphics	2	
SCP Consultant Total @ [ ] per hour	5	[ ]
SCP Contract Total	10	[ ]

FIGURE 4.5. Example of an Invoice

Statistical Analysis of Dissertation Project

Effects of Accommodating Perceptual Learning-Style Preferences on Long-Term Retention and Attitudes Toward Technology of Elementary and Secondary Teachers in Professional Development Training

Report prepared for A. Client

by

Z. Consultant and A. Student
Statistical Consulting Program (SCP)

January 1, 2000

Executive Summary

The results of the SCP analysis suggest that augmentation of the teaching methods based on learning style preferences improves long-term retention of material taught. Further study is necessary to properly evaluate the longitudinal effect of the retention.

FIGURE 4.6. Title Page for the Final Report

1.2 Variables:

The Learning Style Inventory (LSI) instrument was first used to classify the preferences (PREF) of the participants for both groups (GROUP). For assessment purposes, three quantitative measures were employed in this study. They consisted of: a pretest (PRE), an attitude scale score (SDS), and a posttest (POST) given one month after the instruction. Other factors recorded were GENDER, SESSION, and school level (SLEVEL). Details concerning these variables are summarized below:

Quantitative Measures:

PRE	Pretest:	Maximum mark = 100
POST	Posttest:	Maximum mark = 100
SDS	Attitude Score:	Maximum mark = 60

Categorical Factors:

GROUP	Control / Experiment
GENDER	Male / Female
SLEVEL	Elementary / High School level
SESSION	S1,S2,S3,S4 Four sessions within each GROUP
PREF	Auditory, No preference, Tactile, Kinesthetic, Visual [coded as: A,N,T,K,V]

Note: For people with PREF = N in GROUP = Experiment, the teaching method preference was assigned based on the highest LSI score with respect to the categories A, T, K, and V.

2. Methodology:

A statistical analysis of this experiment was performed by the SCP using the statistical software package SAS [1]. Three statistical procedures were used in this analysis: exploratory data analysis (EDA) was used to summarize the data, t-tests were used to detect differences between the average test scores between the Control and Experiment groups, and analysis of variance (ANOVA) was used to detect significant factor effects. Details concerning the methodology and interpretation of these statistical procedures are briefly discussed below. Further information is given in [2].

2.1 Exploratory Data Analysis:

EDA techniques are used to summarize the data. Frequency tables, bar charts, and histograms effectively display the distribution of the variables under consideration. Bar charts were employed for the categorical variables: GROUP, SESSION, SLEVEL, PREF, and GENDER; histograms for the quantitative variables: PRE, POST, and SDS.

2.2 t-Tests:

t-tests are used to assess whether a significant difference exists between the means associated with two independent samples (e.g., the average POST scores of the Experiment vs. Control Groups). Significance is based on the P-value associated with the t-test. By convention, a P-value < 0.05 (5%) is considered to provide

sufficient evidence of a significant difference. A P-value less than 1% would suggest strong evidence of a statistically significant result.

2.3 ANOVA:

To test the effects of several factors simultaneously, an analysis of variance model can be used. For example, the appropriate model to test whether any of the pertinent factors had an effect upon POST would be:

```
POST = GROUP + SESSION(GROUP) + PREF + GENDER + SLEVEL ,
```

where POST is the response variable, and GROUP, SESSION(GROUP), PREF, GENDER, and SLEVEL are all factors that could potentially affect the response. In this study, SESSION is said to be a "nested" factor since its levels were allocated *within* each GROUP. The statistical effect of SESSION is therefore assessed as the nested factor, SESSION(GROUP).

The overall significance of the model is first used to determine whether any of the factors had a significant effect on the response. To determine the effect of an individual factor, the Type III sums of squares (SS) is used. These show the individual effect of a factor when the contributions from all the other factors have already been taken into account.

3. Results:

Table 1 consists of several frequency tables giving the breakdown of the participants with respect to the categorical factors. Tables 2A and 2B present summary statistics for the PRE, POST, and SDS variables overall, and with respect to GROUP.

The t-tests are presented in Table 3. It was found that there is no significant difference between the Experiment and Control groups when SDS and PRE are considered (P-value > 0.05). This suggests that there was no meaningful difference between the Control group and the Experiment group as far as prior knowledge and attitude in concerned. However, POST is strongly significant which suggests that the means of the two groups are significantly different as far as knowledge retained is concerned.

The analysis of variance results for the variables PRE, SDS, and POST are presented in Table 4. The same factors were employed in each model as follows.

```
(1)  PRE  = GROUP SESSION(GROUP) PREF GENDER SLEVEL
(2)  SDS  = GROUP SESSION(GROUP) PREF GENDER SLEVEL
(3)  POST = GROUP SESSION(GROUP) PREF GENDER SLEVEL
```

Of these models, only (3) provided a significant result with a P-value of 0.0270. From the P-values ($Pr > F$) associated with the Type III SS, only GROUP was found to be significant with a P-value of 0.0006.

4. Conclusions

The SCP found that there were no significant results for the variables PRE and SDS which suggests that the participants as a whole did not differ significantly with respect to these variables. There was a significant GROUP effect for the variable POST, suggesting that augmentation of teaching methods through the use of learning-style

preferences would improve retention of the material for a similar seminar. It is the conclusion of the SCP that the investigation performed by the client has produced some significant results that would be worth pursuing in a larger-scale study.

References:

[1] SAS Institute Inc. (1990) *SAS/STAT User's Guide.* Version 6. Cary, NC.
[2] Moore, D.S. and McCabe, G.P. (1993) *Introduction to the Practice of Statistics.* 2nd ed., Freeman Press, NY.

Appendix 1: Tables

List of Tables

Table 1:

Frequency Tables for Categorical Variables
Total Number = 87

GROUP	Number	Percent
Control	43	49.4
Experiment	44	50.6

SESSION	Control	Expt	Number	Percent
S1	11	11	22	25.3
S2	11	12	23	26.4
S3	11	11	22	25.3
S4	10	10	20	23.0

PREF	Control	Expt	Number	Percent
A	12	9	21	24.1
K	2	2	4	4.6
N	17	15	32	36.8
T	11	12	23	26.4
V	1	6	7	8.0

GENDER	Control	Expt	Number	Percent
F	33	37	70	80.5
M	10	7	17	19.5

SLEVEL	Control	Expt	Number	Percent
E	31	28	59	67.8
H	12	16	28	32.2

Table 2A:

Summary Statistics of Test Scores
Total Number = 87

	PRE	POST	SDS
Min.	5.00	36.00	21.00
1st Qu.	40.00	76.00	42.50
Median	55.00	80.00	50.00
Mean	55.63	80.87	48.05
3rd Qu.	75.00	90.00	55.00
Max.	100.00	100.00	60.00
Std. Dev.	25.31	12.11	8.56

Table 2B:

Summary Statistics of Test Scores by GROUP
Total Number: Control = 43 , Experiment = 44

	Control			Experiment		
	PRE	POST	SDS	PRE	POST	SDS
Min.	5.00	36.00	21.00	5.00	64.00	21.00
1st Qu.	45.00	72.00	42.00	35.00	80.00	44.25
Median	60.00	80.00	49.00	50.00	88.00	51.50
Mean	57.67	76.00	47.12	53.64	85.64	48.95
3rd Qu.	80.00	84.00	55.00	75.00	96.00	44.25
Max.	95.00	96.00	59.00	100.00	100.00	60.00
Std. Dev.	24.67	11.58	8.48	26.04	10.75	8.64

Table 3:

t-Test Results
Difference in Average Test Scores across GROUP
Total Number 87: Control = 43, Experiment = 44

Variable: SDS

GROUP	N	Mean	Std Dev	Std Error	Min	Max
Control	43	47.116	8.477	1.293	21	59
Experiment	44	48.954	8.637	1.302	21	60

Variances	T	DF	Prob> $\|T\|$
Unequal	-1.0018	85.0	0.3193
Equal	-1.0016	85.0	0.3194

Variable: PRE

GROUP	N	Mean	Std Dev	Std Error	Min	Max
Control	43	57.674	24.673	3.763	5	95
Experiment	44	53.636	26.045	3.926	5	100

Variances	T	DF	Prob> $\|T\|$
Unequal	0.7425	84.9	0.4598
Equal	0.7421	85.0	0.4601

Variable: POST (*** Significant)

GROUP	N	Mean	Std Dev	Std Error	Min	Max
Control	43	76.000	11.580	1.766	36	96
Experiment	44	85.636	10.751	1.621	64	100

Variances	T	DF	Prob> $\|T\|$	
Unequal	-4.0202	84.2	0.0001	***
Equal	-4.0237	85.0	0.0001	***

Table 4:
ANOVA Results
General Linear Models Procedure
Number of Observations = 87

Model 1 : SDS
Model 2 : PRE
Model 3 : POST
} = GROUP + SESSION(GROUP)+ PREF + GENDER + SLEVEL

Model 1: SDS (No Significant Factors)

Source	DF	Sum of Squares	Mean Square	F-Value	$Pr > F$
Model	13	876.6456	67.4343	0.91	0.5492
Error	73	5423.1705	74.2900		
Corrected Total	86	6299.8161			

Source	DF	Type III SS	Mean Square	F-Value	$Pr > F$
GROUP	1	65.0331	65.0331	0.88	0.3526
SESSION(GROUP)	6	707.4197	117.9033	1.59	0.1631
PREF	4	80.4978	20.1245	0.27	0.8958
GENDER	1	5.4309	5.4309	0.07	0.7876
SLEVEL	1	10.6014	10.6014	0.14	0.7067

Model 2: PRE (No Significant Factors)

Source	DF	Sum of Squares	Mean Square	F-Value	$Pr > F$
Model	13	12281.0661	944.6974	1.61	0.1017
Error	73	42809.1638	586.4269		
Corrected Total	86	55090.2299			

Source	DF	Type III SS	Mean Square	F-Value	$Pr > F$
GROUP	1	23.3249	23.3249	0.04	0.8425
SESSION(GROUP)	6	4339.8085	723.3014	1.23	0.2993
PREF	4	4465.2543	1116.3136	1.90	0.1189
GENDER	1	6.4008	6.4008	0.01	0.9171
SLEVEL	1	689.2741	689.2741	1.18	0.2819

Model 3: POST (** GROUP ** Factor Fignificant)

Source	DF	Sum of Squares	Mean Square	F-Value	$Pr > F$	
Model	13	3389.7547	260.7504	2.06	0.0270	**
Error	73	9231.8545	126.4638			
Corrected Total	86	12621.6092				

Source	DF	Type III SS	Mean Square	F-Value	$Pr > F$	
GROUP	1	1644.6794	1644.6794	13.01	0.0006	**
SESSION(GROUP)	6	880.1432	146.6905	1.16	0.3372	
PREF	4	295.3252	73.8313	0.58	0.6753	
GENDER	1	50.6251	50.6251	0.40	0.5289	
SLEVEL	1	12.4934	12.4934	0.10	0.7542	

Appendix 2: The Data

GROUP = Control							GROUP = Experiment						
Order: SESSION, PREF, GENDER, SLEVEL, PRE, POST, SDS													
S1	V	F	H	25	64	51	S1	A	F	H	40	96	52
S1	K	F	E	65	64	21	S1	T	F	E	5	64	29
S1	A	F	E	5	60	41	S1	N	F	E	20	64	54
S1	A	M	E	60	76	56	S1	N	M	E	95	92	57
S1	T	M	E	65	76	51	S1	T	F	H	75	92	53
S1	T	M	E	95	88	55	S1	A	F	E	70	100	46
S1	T	F	H	85	84	50	S1	T	F	E	40	84	21
S1	N	F	E	85	72	42	S1	N	M	H	80	96	37
S1	A	M	E	65	76	34	S1	N	F	E	25	76	52
S1	N	F	H	50	68	42	S1	V	F	E	20	92	38
S1	N	M	E	5	36	36	S1	V	F	E	5	96	44
S2	T	F	H	65	80	56	S2	A	F	E	15	76	54
S2	T	F	E	15	76	47	S2	N	F	H	50	80	48
S2	A	M	E	20	80	44	S2	V	M	H	50	88	57
S2	A	F	E	85	76	51	S2	N	F	H	25	88	41
S2	A	F	E	20	80	56	S2	T	F	H	50	96	48
S2	T	F	E	45	80	58	S2	N	F	E	80	92	49
S2	N	F	E	55	76	54	S2	N	M	E	60	68	43
S2	A	F	H	50	80	42	S2	T	F	E	40	92	52
S2	A	M	H	15	44	39	S2	N	F	E	35	96	32
S2	N	M	E	90	88	52	S2	N	F	E	90	92	60
S2	A	F	E	55	72	44	S2	V	F	H	25	88	52
							S2	A	F	H	50	80	42
S3	N	F	H	50	64	41	S3	A	F	H	70	72	56
S3	N	F	H	45	80	50	S3	T	F	H	35	68	49
S3	N	F	E	90	88	40	S3	N	F	H	75	96	48
S3	N	F	E	70	80	49	S3	A	M	H	25	100	45
S3	T	F	E	85	84	59	S3	N	F	H	40	84	56
S3	T	F	H	70	76	46	S3	K	F	E	75	96	60
S3	N	F	E	80	68	59	S3	N	F	E	80	96	51
S3	T	F	H	45	88	52	S3	A	F	E	45	92	45
S3	N	F	H	80	80	42	S3	T	F	E	95	80	57
S3	N	F	E	75	84	56	S3	T	F	E	60	88	37
S3	N	F	H	70	84	32	S3	T	F	E	100	96	59
S4	N	F	E	50	80	50	S4	A	F	E	80	92	56
S4	N	M	E	45	64	43	S4	T	M	H	65	92	58
S4	N	F	E	85	96	58	S4	A	F	H	30	80	52
S4	A	F	E	95	80	56	S4	N	F	E	60	64	49
S4	A	F	E	75	84	44	S4	K	F	E	70	64	60
S4	T	M	E	45	60	35	S4	N	M	E	40	80	54
S4	T	F	E	55	88	58	S4	V	F	E	35	80	52
S4	K	F	E	60	76	45	S4	T	F	E	100	96	51
S4	N	F	E	55	88	49	S4	V	F	E	60	84	43
S4	A	F	E	35	80	40	S4	T	F	E	75	80	55

4.8 Postscript

We received the client's methods and results chapter that we had agreed to review, along with a check for payment. (Clients can pull the right strings too!) In these types of projects, making ourselves available for certain post-completion services is good "PR" (public relations) and usually does not take much time or effort. (If it does, we can always charge for it.) Below is a modified transcript of the informal, first-name basis, email review we provided the client. The reference to Table 1 pertains to the client's document.

A,
Some notes/suggestions for your methods chapter ...
[page/line numbers as per the document you sent me]
Hope this is helpful.
Regards,
Z

---

page 1: *The dependent variables were the subjects' mean ...*

The dependent variables are really the "actual" scores since you only gave one POST/SDS-test; i.e., you didn't average the scores from "several" POST tests.

page 3: *... and (c) Analysis of Variance (ANOVA) ...*

What you have written seems fine. Remember that ANOVA is employed when:

- a factor has more than two levels
- the simultaneous effect of several factors is to be considered.

Note that the actual procedure we used was: general linear models (GLM) — the difference between GLM and ANOVA is simply that GLM takes account of the "unequal cell sizes" (i.e., GLM accounted for the different number of Male/Female, Elem/High School, etc.). ... In effect, GLM was nothing more than a "correct" ANOVA analysis.

page 3: [line 4 ↑] ... *experimental groups (P-value > 0.05).*

I would suggest expanding the concept of a "P-value" ... perhaps end the first sentence with:

" ... experiment group." Then add something like:

To assess the statistical significance (or lack of) associated with a particular test, such as the two-sample t-test employed in Table 1, we may use the "P-value" criterion. Standard statistical practice has adopted 0.05 as the cutoff criterion for assessing significance (P-value < 0.05). As can be seen from Table 1, both (*) values listed under "Prob> $|T|$" exceed 0.05. This implies that the Control and Experiment groups did not differ (significantly) in terms of ....

(*) [A possible footnote to explain why there are two t-tests.]

The two-sample t-test can be conducted under the assumption that the variability of pretest scores within the Control and Experiment groups are the same (denoted by "Equal" in Table 1). Although this assumption can be considered reasonable in this study (and is supported by comparable sample standard deviations from the two groups), the two-sample t-test can also be conducted without making the assumption that the variability is the same ("Unequal"). In either case, the P-value criterion can be applied to assess statistical significance.

page 4: [line 10] ... *different* ($p < 0.0004$).

The P-value is equal to 0.0004 (not less than 0.0004). Rephrase as:

" ... different (P-value of 0.0004 (*))."

(*) [Possible footnote.]

This result is considered "strongly" significant. ... (since it is smaller than 0.01 which is often used to qualify significance beyond the standard 0.05 criterion).

page 5: [line 12] typo: (P-value of 0.0031) $\longrightarrow$ (P-value of 0.0013).

page 6: Seems fine.

In the next chapter you could postulate "further research directions" to investigate the SDS variable. Clearly, you did not have enough data in this study to perform a meaningful analysis of SDS by PREF (too few people in the K, V groups and not really enough in the A group once you split SDS by GROUP with PREF = N excluded). ... Your next thesis maybe?

Questions

1. One of the main issues we excluded in our presentation of this case study example concerned the PREF = N paticipants.

   IMPORTANT: You may want to read the next question before starting any part of this question.

   (a) Discuss the statistical issues associated with the PREF = N class with the client. Recall that these participants were "assigned" a preference in the Experiment group.

   (b) An obvious approach is to analyze the response variables with respect to the separate subsets: PREF = N and PREF $\neq$ N. What problems arise? Is there a better way to approach the analysis?

   (c) Conduct the analysis suggested above. Are your conclusions different for the two subsets? Should they be?

2. Suppose you were asked to create a detailed invoice for the client. Your analysis of the PREF = N issue above will be very helpful here: document every task you perform, and note the time it takes to complete each task. Alternatively, just use our presentation of the case study entirely as your basis for creating the invoice. Assume both consultation sessions took a full hour.

   (a) The first step is to itemize every cost-related component of the consultation process. Where did we actively engage in working on the project? What task did we do?

   NOTE: Do not include capital costs. The client isn't going to pay for your new printer **and** your consultation fee!

   (b) Attach a time unit (use tenths of an hour) to each component. Don't forget little things like the 12 minutes we spent arranging the initial consultation session and getting the prior information. You will need to estimate times for many of these components.

   [Hence the value of documenting each step (and time taken) in your analysis of the PREF = N issue.]

   (c) Allocate your hourly rate, assuming you did all the work. Your rate would normally include an overhead which absorbs incidental costs and can be put towards capital improvements. Calculate the total cost to the client.

   Is this amount realistic for $n = 87$ observations?

   (d) Adjust this total cost by providing the first consultation hour free of charge and allocating appropriate tasks to an "assistant" (with a lower hourly rate). Divide the adjusted total cost by

$n = 87$. When might this per observation cost be useful? Not useful?

3. In view of our Remarks on page 171, it may seem surprising that we retained the "Unequal Variance" entries in the Preliminary and Final Reports. Note that the `Std Dev` values in Table 3 for the Experiment and Control groups are very similar within each response variable.

   (a) Is there any reason why we should **not** have deleted the Unequal Variance entries?

   (b) No test for normality appears to have been performed with regard to the t-test or ANOVA analyses. How would you explain the results of this test to the client?

      (i) What are the consequences of a significant result? (That is, the normality assumption is rejected.)

      (ii) How would you incorporate the issue of "Robustness" in your explanation?

4. For S-PLUS users.

   (a) Analyze these data using S-PLUS. Note that Tables 2A and 2B were actually produced using the S-PLUS `summary()` function. What are the differences in the output of the t-test and ANOVA results from S-PLUS as compared to SAS?

   (b) Complete the function `g.hist2.fun()` outlined on page 172. The objective is to be able to produce either the PRE by GROUP (Figure 4.3), or the POST by GROUP (Figure 2.3) histogram display using only the arguments provided in the options line of your `g.hist2.fun()` function.

   (c) The client requested presentation quality histograms for several variables. On each plot, the client wanted the value of the mean and standard deviation of the variable printed on the graph (within the plot region) with the labels: "Mean = `<value>`" and "Standard Deviation = `<value>`" Consider the following options.

      (i) Create a generic function that automatically determines a "good" place to put this text within the plot region of any histogram display.

      (ii) Position the text inside the plot region of each histogram by trial and error.

      Which approach should we take?

Part II

Case Studies

5
Introduction to the Case Studies

The best way to learn about statistical consulting is to do it! There is really no substitute for this and making mistakes is an important part of the process. We have already mentioned several of our mistakes in the previous chapter; don't worry, there are still plenty left for you to make. This brings us to the case studies. These are intended to provide the reader with the opportunity to gain some experience — at least in terms of learning about the statistical analyses required in different types of consulting problems.

The purpose of this introduction is to provide the reader with an overview of the procedure that we use to present the case studies in Part II. The case studies have been divided into three groups, uniquely labelled as Group I, II, and III, which constitute the next three chapters. The presentation format and details concerning the general approach that are employed for these case studies are described in the next section. The level of complexity and the type of statistical analysis required in each group are also discussed. The last chapter of Part II consists of a collection of ungrouped case studies that are presented as exercises.

5.1 Presentation Format for the Case Studies

The title of the case study appears in the section heading. This is immediately followed by a box containing pertinent information about the data, variables, and statistical methods used for analysis. This box may some-

times contain additional information. Specific details associated with the case study are then presented in the subsections that follow.

An outline of the presentation format is given below. For illustrative purposes only, we use data collected from a Battery-Failure study conducted by NASA. The data can be found in Johnson and Wichern (1998, Table 7.4) The order and basic content of the subsections indicated in this outline are generic to all the case studies.

Case Study Format

Title NASA BATTERY-FAILURE EXPERIMENT

Box

Source:	Johnson and Wichern (1998)
Methods:	Exploratory Data Analysis Response Surfaces Analysis of Covariance
Data:	Designed Experiment with 5 Control Variables and Response: Cycles to Failure

Subsections:

1. *Context* *"A planned experiment"*
   The context of the problem. Questions to address and the main statistical issues.
2. *Data* Description and format of the database. Properties of the relevant variables.
3. *Methods* Details on the statistical procedures that can be used for analysis. In this example:
   - Exploratory Data Analysis
   - Response Surfaces
   - Analysis of Covariance.
4. *Analysis* Some preliminary results. SAS and S-PLUS output.

Questions Completing the analysis. The report.

5.2 Case Study Details

This presentation format is clearly not intended to reproduce a case study in the style of Chapter 4. While there are some disadvantages in presenting a "condensed" version of a case study, the important advantages are:

1. Far more case studies can be presented.
2. The context of the problem can be described efficaciously.

3. Data processing can be dealt with in more general terms.

4. Appropriate statistical methodology can be discussed in detail.

5. The preliminary analysis can focus on pertinent results.

6. The questions provide a pedagogical basis for the case study.

The main disadvantage is that the reader is not exposed to the dialogue which resulted in the condensed version of the case study. However, the guidelines and information we provided in Part I should compensate for this compromise, and allow readers to perform the same type of "condensation" in their consultation projects.

Context

The statistical application associated with the case study is indicated by the heading of this subsection. *"A planned experiment"* is not particularly imaginative (and somewhat redundant in this example), but it can be helpful when the *title*[1] of the case study seems a little cryptic!

The main purpose of this subsection is to describe the context of the problem that is addressed in the case study. Since all the case studies are based on real projects, certain modifications were made to the actual problems for obvious reasons. However, we have tried to maintain the spirit of the original investigation in our description of the case study, and we have provided the necessary background information to make the context of the problem sufficiently complete and meaningful.

In this example, early satellite applications motivated the development of silver–zinc battery cells. Failure-time tests were conducted to assess the reliability of these cells under various conditions.

The Data

The datasets for all the case studies can be downloaded from the Websites listed in Appendix B (*Datasets*).

The presentation of the data for each case study varies. Small datasets are reproduced in this subsection, but with larger datasets, we only provide a description of the variables involved.

It is important to note that most datasets do **not** come in the "clean" state that we present. Many hours may have already gone into making corrections and adjustments to the data in order for the dataset to arrive at

[1]The title appears as a section heading in the Table of Contents, hence the reason for choosing a suitable "application" heading for this subsection.

this stage. Again, this is part of the compromise we have made. Nevertheless, an exploratory analysis should always be conducted since we have not removed potential outliers or otherwise "sanitized" the data.

The main questions of interest in each case study are normally discussed in the *Context* subsection above. Sometimes it may be more appropriate to consider these questions after the dataset and variables have been presented. For example, statistical issues are often easier to address with respect to a particular variable than in a general context. In this example, we would suggest that a transformation of Y is necessary, Y having already been defined by the following tables.

Table I.1: Variable Definitions

X_1	Charge Rate (amps)
X_2	Discharge Rate (amps)
X_3	Depth of Discharge (% of rated ampere-hours)
X_4	Temperature (°C)
X_5	End of Charge Voltage (volts)
Y	Cycles to Failure (*response variable*)

Table I.2: Battery-Failure Data

Obs	X_1	X_2	X_3	X_4	X_5	Y
1	0.375	3.13	60.0	40	2.00	101
2	1.000	3.13	76.8	30	1.99	141
3	1.000	3.13	60.0	20	2.00	96
4	1.000	3.13	60.0	20	1.98	125
5	1.625	3.13	43.2	10	2.01	43
6	1.625	3.13	60.0	20	2.00	16
7	1.625	3.13	60.0	20	2.02	188
8	0.375	5.00	76.8	10	2.01	10
9	1.000	5.00	43.2	10	1.99	3
10	1.000	5.00	43.2	30	2.01	386
11	1.000	5.00	100.0	20	2.00	45
12	1.625	5.00	76.8	10	1.99	2
13	0.375	1.25	76.8	10	2.01	76
14	1.000	1.25	43.2	10	1.99	78
15	1.000	1.25	76.8	30	2.00	160
16	1.000	1.25	60.0	0	2.00	3
17	1.625	1.25	43.2	30	1.99	216
18	1.625	1.25	60.0	20	2.00	73
19	0.375	3.13	76.8	30	1.99	314
20	0.375	3.13	60.0	20	2.00	170

The statistical problem of interest is now clear:
Find the values of $X_1, \ldots, X_5$ *that optimize* Y.

Methodology

The statistical methods employed for analysis will dictate the group level of the case study. From a consultant's perspective, these are defined as

GROUP I Simple case studies requiring only standard (see Chapter 3) statistical methods for analysis. Obtaining the answer is straightforward. However, there may be some interesting statistical and scientific issues to be discussed.

GROUP II More complicated case studies that will require statistical methods such as logistic regression, time series modeling, and factorial designs. The complexity may be partly due to the size of the dataset. The statistical problem is generally well defined, but broader in scope than the Group I case studies. Several solutions may need to be evaluated.

GROUP III Research-oriented case studies that often require multivariate methods or specialized statistical methods for analysis. Several stages of analysis may also be needed to obtain suitable results. There may not necessarily be an "answer" to the statistical problem.

There is, of course, a considerable difference between what the *consultant* needs to know in order to resolve the statistical issues presented by the case study — and what is required from a teaching perspective. In our experience, the level of statistical sophistication implied by these groupings does **not** present a barrier for students. Indeed, there is something rather special about learning to use an unfamiliar statistical method for the first time. More important perhaps, the exploratory analysis of a complex case study can often provide a more realistic learning experience.

The purpose of this subsection is therefore to address pertinent aspects of the statistical methodology that can be employed to analyze the case study. In addition to providing a detailed discussion of the appropriate statistical techniques needed to perform a comprehensive analysis, we also focus on specific components associated with the exploratory analysis of the data. In some cases, variations and alternatives to the comprehensive approach may be suggested.

In this example, response surface methodology would be discussed in detail. We only briefly outline some of the main issues of this discussion below. It is worth pointing out that the initial focus of the discussion could address the interesting problem of performing an exploratory graphical analysis of multivariate data.

- *Exploratory Data Analysis:*

  Variance stabilizing transformations (e.g., $\log Y$)

  Contour plot of Y vs. X_1, X_2 (design variables)

  Trellis plots: Y vs. X_1, X_2 (contour) by X_3, X_4, X_5 (categorized)

- *Response Surface Analysis:*
  Polynomial regression
  Canonical analysis and composite designs
  Interpreting the optimization.

Preliminary Analysis

In this subsection we present pertinent results obtained from applying the statistical methods discussed above. An outline of the SAS and S-PLUS code used to generate these results is normally included here. In general, we do **not** present a complete analysis of the case study. There are several reasons for this, but the most important one is:

Statistical consulting is best learned by doing it.

So after all this, it's a sink-or-swim approach. Not quite. Our aim is to provide the reader with a *progressive* series of case studies. That is, the preliminary analyses associated with previous case studies augment the particular methods that are the focus of the results presented in the current case study.

In this example, we assume the reader is familiar with the general concepts and results that would be obtained from applying a standard regression analysis (regression methods would have been presented in an earlier case study). Thus, the main focus of the preliminary analysis presented here would concern the results from fitting a response surface model. A summary discussion of these results would be provided.

Questions

One purpose of the summary discussion is to indicate the general direction for further analysis. There may also be particular features of the case study that are of interest and warrant further consideration. To address these issues, we provide a series of questions. This provides a pedagogical basis to the case study and enables the reader to complete the statistical analysis of the case study. Ideally, the completed analysis should be compiled in the form of a written report with particular emphasis on the contextual interpretation of the results, namely, the "conclusions" of the report. We leave that decision to the discretion of the reader.

Without further ado, let's go consulting!

6
Case Studies from Group I

In this chapter we present four case studies where the method and approach to the analysis are straightforward, but the context of the problems raises some interesting statistical and scientific issues. In some of the case studies, a meaningful interpretation of the statistical results may prove quite difficult. The case studies are listed below by section.

6.1 Job Promotion Discrimination

Methods:	Contingency Table Analysis Chi-Square Test Fisher's Exact Test
Data:	Two Contingency Tables Promotions versus Candidate Support

6.1.1 A Claim Based on Statistical Evidence

The Department of Social Events of the country of Oz[1] is facing a discrimination complaint. The plaintiff is a high-ranking employee of this department who was not promoted after the last election for prime minister. He believes that the reason for his nonpromotion is that only candidates who contributed to the election campaign of the winning candidate for prime minister were promoted. Other candidates like him who did not contribute to the campaign of the winning candidate were not promoted. A summary of the pertinent facts is as follows.

Result:

Including the plaintiff, 10 employees of the department were up for promotion. Seven were promoted and of these seven, six had made financial contributions to the campaign of the winning candidate for prime minister. The remaining four did not contribute to the prime minister's campaign.

Promotion Procedure:

The promotion procedure is based on the scores from a standard civil service test that was taken by all 10 candidates. The scores are ranked and the promotion procedure requires that for each promotion slot, the successful candidate must be selected from those who are currently among the top three ranked candidates (including ties). This is applied sequentially until all the available promotion slots have been filled.

Plaintiff's Claim:

The plaintiff was ranked number 4 and believes that this ranking should have been more than sufficient to obtain a promotion from the seven that were available. Since campaign contributions are not a requirement for promotion the plaintiff claims that discrimination occurred.

---

[1] In discrimination cases the actual identities involved are confidential and need to remain anonymous, hence the fictitious Department title.

Additional Information:

Employees who were not candidates for promotion were asked a question regarding whether they felt a positive or negative change in their job conditions after the election was held. They were also asked whether they made financial contributions to the campaign of the winning candidate.

The Data

The data are presented in three tables. Table 6.1 summarizes the promotion results by campaign contributions for the 10 candidates. As can be seen from Table 6.2, the plaintiff (D) only becomes eligible for promotion after the first slot has been filled. Table 6.3 summarizes the responses obtained from the employees who were not candidates for promotion. The category "Unknown" means that no response to that question was provided.

TABLE 6.1. Contribution by Promotion

	Promoted	Not Promoted
Contributed to Winner	6	0
Did Not Contribute to Winner	1	3

TABLE 6.2. Candidate Ranking (1 = highest)

Candidate	A	B	C	D	E	F	G	H	I	J
Rank	1	2	3	4	5	6	7	8	8	8

TABLE 6.3. Contribution by Change in Job Condition

	Positive	Negative	Unknown
Contributed to Winner	4	0	2
Did Not Contribute to Winner	0	7	0
Unknown	1	9	2

Methodology

6.1.2 Contingency Table Analysis

Tables 6.1 and 6.3 are commonly referred to as contingency tables and the standard method of analysis is to perform a test of *independence*. This was discussed in Chapter 3 (see *Contingency Tables*) and concerns the problem

of assessing whether there is evidence of a statistical association between the levels of two categorical variables. In this case study we follow the usual approach which is to compute Pearson's chi-square statistic or, in the case of small samples, Fisher's exact test. Both these procedures were discussed in Chapter 3. The main points are summarized below.

Chi-Square Test This can be found in most introductory statistics texts (see, for example, Ott (1993)) and provides a test for the null hypothesis of no association (independence) between the two variables. The chi-square test should only be used when the sample size is sufficiently large. In practice, the usual criterion is that no more than 20% of the table cells should have **expected** counts less than 5.

Fisher's Exact Test This was originally applied in the case of a 2×2 contingency table: two categorical factors, each at two levels. For small samples, however, the P-values will be highly discretized and there is always the question of whether a statistical analysis of a small sample is necessarily meaningful. Fisher's exact test for independence can be generalized to any $n_{\text{row}} \times n_{\text{col}}$ contingency table, albeit at a rather substantial computational cost in some cases. This is **not** a test to run on several arbitrary sized contingency tables just for exploratory purposes. Fisher's exact test is discussed in Agresti (1990).

Other Tests There are many other tests and measures of association that can be employed for contingency tables. For a more detailed summary, the reader is referred to the discussion presented in the `PROC FREQ` section of the SAS manual (SAS 1990).

6.1.3 Interpretation of Significance

The difficulty that arises in discrimination cases is that even if a result is deemed to be statistically significant,[2] it does not "prove" discrimination occurred. We emphasize this important point with the following well-known paradigm of statistical inference:

> A statistical association between two factors does **not** imply a causal relationship necessarily exists between these factors.

In contingency table analysis, the context and interpretation of the problem can easily affect the analysis, and in some cases may suggest associations that are misleading or even meaningless. A significant result based

---

[2] Hazelwood School District vs. United States [433 U.S. 299, 311 n.17 (1977)], in the Supreme Court: *If the difference between the expected value and observed number is greater than* two or three standard deviations, *then the hypothesis that employees were hired without regard to race would be suspect.*

on the P-value from either test above must therefore be interpreted with considerable caution. This is particularly important in situations such as discrimination cases where a legal judgment can have serious consequences. From a statistical perspective, some of the key issues that need to be considered when performing a contingency table analysis are:

1. Are the two factors meaningfully related?
2. What are the potential confounding factors involved?
3. Does an association between two factors still exist when other factors are included in the analysis?
4. An interpretation of a significant association may not be appropriate if a factor includes a level such as "Other."
5. What type of study is involved? Prospective, retrospective, or cross-sectional studies need more than just a test for independence.
6. Do the counts have equal weight? The two factors may not account for important differences between the participants in the study.
7. How sensitive is the analysis to small changes in the cell counts?
8. What criteria were used to determine the breakpoints when converting a continuous variable to a categorical factor?

Example 6.1 *Simpson's Paradox*

The following example illustrates Simpson's Paradox wherein a reversal of direction of association occurs when a third factor is included in the analysis. In an article by Radelet (1981) (See Agresti (1990), Section 5.2.2), the effect of racial characteristics on whether individuals convicted of homicide receive the death penalty is studied. The results are given in Table 6.4.

TABLE 6.4. Simpson's Paradox

Defendant's Race	Victim's Race	Received Death Penalty	
		Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

Overall, it can be seen that $19/160 = 12\%$ of white defendants received the death penalty as opposed to $17/166 = 10\%$ of black defendants. However, if the association between defendant's race and death penalty verdict

is now considered with respect to the victim's race, we obtain the opposite conclusion. That is, when the victim is white, $11/63 = 17.5\%$ of black defendants received the death penalty as opposed to $19/151 = 12.6\%$ of white defendants. The same direction of association also holds when the victim's race is black (0% versus $6/103 = 5.8\%$ for black defendants).

6.1.4 Preliminary Analysis

The following SAS program can be used to analyze Table 6.1. Output from this code is given below. Here, `promote=1` denotes candidates that were promoted and `promote=2` denotes those that were not. Similarly, `contrib=1` denotes candidates that contributed to the winner, `contrib=2` that they did not. The variable `kount` gives the frequencies for the combinations of the `promote` and `contrib` levels.

SAS Program for Analyzing Table 6.1

```
data a;
   input promote contrib kount ;
   cards;
   1 1 6
   1 2 1
   2 1 0
   2 2 3
   ;
proc freq ;
   tables promote*contrib / chisq exact ;
   weight kount ;
endsas ;
```

Output from the SAS Program

```
TABLE OF PROMOTE BY CONTRIB

Frequency|
 Percent |
 Row Pct |
 Col Pct | 1       | 2       |  Total
---------+--------+--------+
       1 |      6 |      1 |      7
         |  60.00 |  10.00 |  70.00
         |  85.71 |  14.29 |
         | 100.00 |  25.00 |
---------+--------+--------+
       2 |      0 |      3 |      3
         |   0.00 |  30.00 |  30.00
```

```
         |   0.00 | 100.00 |
         |   0.00 |  75.00 |
---------+--------+--------+
   Total        6        4       10
           60.00    40.00   100.00

STATISTICS FOR TABLE OF PROMOTE BY CONTRIB

Statistic                       DF   Value   Prob
-------------------------------------------------
Chi-Square                       1   6.429  0.011
Likelihood Ratio Chi-Square      1   7.719  0.005
Continuity Adj. Chi-Square       1   3.353  0.067
Mantel-Haenszel Chi-Square       1   5.786  0.016
Fisher's Exact Test (Left)                  1.000
                    (Right)                 0.033
                   (2-Tail)                 0.033
Phi Coefficient                     0.802
Contingency Coefficient             0.626
Cramer's V                          0.802
```

6.1.5 Summary

This case study illustrates the difficulties of finding objective criteria by which to assess discrimination issues. There can be many conflicting results that are not easily resolved by statistical methods alone. The following questions indicate some of the issues that a detailed report would need to address.

Questions

1. The P-value from the output above would suggest that there is evidence to reject the null hypothesis of independence. What is your interpretation of this statistically significant result?
2. Although Table 6.3 can be analyzed in similar fashion, a significant result would not provide evidence of discrimination. Why?
3. With reference to Table 6.2 and by assuming each promotion is chosen at random among the top three ranked candidates currently available, what is the probability that candidate D would not be one of the seven promoted?
4. Since this probability is larger than the minimum standard of 5%, the plaintiff needs to argue that the "random" assumption is unreasonable. Is the plaintiff's argument valid, and what does it imply about the selection procedure used for promotion?

5. What are some of the potential confounding factors (other variables) that have not been accounted for in the analyses of Tables 6.1 and 6.2? Was the description of the promotion procedure complete?

6. Irrespective of the statistical test results, it is of interest to note the high frequencies in a couple of the *key* cells in Table 6.3. Does this suggest there may be a difference in the contextual interpretation of the two Unknown categories?

6.2 The Case of the Lost Mail

Methods:	Sample Survey Population Profiles
Data:	List of Zip Codes Survey Data Post Office Data

6.2.1 Sample Survey Analysis

For one particular magazine distributor, which we refer to as *Miss N Magazine,* the addition of an extra zero to the four-digit routing number used in U.S. postal zipcodes (the so-called "Zip+4" number) went unnoticed until a complaint was received from someone about an article contained in the magazine. The person also happened to object to receiving 20 copies of the same issue, particularly when they never even subscribed to it! ... Obviously something was not right.

As it turns out, the routing number change resulted in many issues being sent (incorrectly) to essentially the same address. Since the distribution was nationwide, it was of interest to the magazine distributor to obtain a realistic estimate of the number of magazines that were not delivered. To do this, a telephone survey was conducted to determine what happens to the mail when the routing number is incorrect. From this information, an estimate of the proportion of undelivered mail was able to be obtained.

Methodology

Designing the survey is really the critical stage in this type of analysis. The quality of the data that will be collected clearly depends on the survey design and hence, any statistical analysis of these data. It is therefore important to get it right the first time. Although the actual questionnaire in this case study is very simple, many of the issues considered in Chapter 3 (see *Sample Surveys*) apply here. The following discussion is presented in the form of a report containing the protocol and questionnaire design for

the proposed survey. The objective of the telephone survey is to obtain enough information to be able to evaluate the proportion of undelivered mail of the magazine in question.

6.2.2 The Survey Proposal

1. PLAN FOR SURVEY

The survey is a telephone survey directed at the carrier managers of post offices where the mail with incorrect address was delivered. Since zipcodes correspond to post offices, the units of the survey are post offices. The following protocol is proposed.

1.1 Stratification

A stratification of the list of postal zipcodes will be generated. This will serve the purpose of avoiding oversampling post offices with a small number of magazines, which are the majority. In addition, the grouping will make the subpopulations more homogeneous since post offices with a similar number of items will be more alike.

GROUPING VARIABLE: The grouping variable will be the number of items with incorrect address that were delivered to each post office.

GROUPING METHOD: The groups will be chosen so the total number of items per group will be approximately the same.

NUMBER OF GROUPS: The number of groups will depend on the data on items per post office but in any case it should be between 5 and 10.

1.2 Random Sampling

Random samples of zipcodes will be generated for each group, using the statistical software S-PLUS. For each group, the sample will consist of a list of zipcodes in random order.

SAMPLE SIZE: The sample size for each group will be approximately the same. This minimizes the number of zipcodes that need to be sampled. For each stratum we observe a sample of proportion delivered that has a mean and a variance. If the variances among the strata are very different then it is better to sample proportional to the square root of the variances but if there is no prior knowledge or no evidence of this then the sample sizes should be the same for all strata.

1.3 Standard Errors

We recommend that the goal of the survey is to measure the proportion of undelivered mail with a standard error of 2.5 to 5%. This means that the sample size should be between 100 to 400 depending on the value of the proportion we are trying to estimate. 100 corresponds to 10% undelivered mail whereas 400 corresponds to 50%.

1.4 Two-Stage Sampling

Once we obtain the first 100 questionnaires, we will evaluate the responses and decided if more sampling is needed.

2. QUESTIONNAIRE

2.1 Who Should Answer the Questionnaire

The questionnaire should be answered by the person who is responsible for managing the carriers or the senior carrier.

2.2 Introduction to the Questionnaire

The interviewer will introduce the purpose of the survey as follows.

> StatCon Enterprises is conducting a survey on behalf of Miss N Magazine about a problem with undelivered mail which occurred during ...

2.3 Questionnaire

Question 1.
Are you familiar with the specific problem? Yes/No
If the answer to Question 1 is No go to Question 4.
If the answer to Question 1 is Yes go to Question 2.
Question 2.
Do you remember what you did with the mail? Yes/No
If the answer to Question 2 is No go to Question 4.
If the answer to Question 2 is Yes go to Question 3.
Question 3.
What proportion was ... ?
(A) delivered to the correct addresses. _____
(B) returned to the sender. _____
(C) disposed of. _____
End.
Question 4.
What would you have done if you found such mail?
(A) Deliver it to the correct addresses.
(B) Return it to the sender.
(C) Dispose of it.
(D) I do not know.
End.

6.2.3 Preliminary Analysis

The data from this survey are contained in two files: `c62.survey.dat` contains the responses to the survey and `c62.surinfo.dat` contains information about the sample used. The variables present in `c62.surinfo.dat` are:

`group` = grouping variable for post office zipcode
`id` = post office zipcode
`x` = number of magazines sent to `id`
`state` = state code of post office zipcode

The first task is to process the data obtained from the survey. Missing responses were encoded by a period "." in the file `c62.survey.dat` which will need to be distinguished from actual missing values. In the following SAS code, each option of Question 3 is entered as a separate variable. This allows the options with missing responses to be correctly recoded as 0 or deleted in the case of an actual missing value. After processing the data, a frequency tabulation of the answers is performed. A summary of the results is presented below.

SAS Code for Processing the Survey Data

```
data survey ;
  infile 'c62.survey.dat';
  input id q1 q2 q3a q3b q3c q4 $;
  if q3a eq . and ( q3b ne . or q3c ne . ) then q3a = 0;
  if q3b eq . and ( q3a ne . or q3c ne . ) then q3b = 0;
  if q3c eq . and ( q3a ne . or q3b ne . ) then q3c = 0;
  if q3b eq . and q3a eq . and q3c eq . and q4 eq " "
      then delete;
data infosur;
  infile 'c62.surinfo.dat';
  input group id x state $;
```

Summary of Answers to the Questionnaire

Question 1: Are you familiar with the specific problem?

Responses	Frequency	Percent
No	361	86.6
Yes	56	13.4
TOTAL	417	100

Question 2: Do you remember what you did with the mail?

Responses	Frequency	Percent
No	23	41.1
Yes	33	58.9
TOTAL	56	100

Question 3: What percentage was ...

(A) delivered to the correct addresses?

Responses	Frequency	Percent
0	10	30.3
50	4	12.1
75	2	6.1
80	1	3.0
95	2	6.1
99	1	3.0
100	13	39.4
TOTAL	33	100

(B) returned to the sender?

Responses	Frequency	Percent
0	29	87.9
5	1	3.0
33	1	3.0
100	2	6.1
TOTAL	33	100

(C) disposed of?

Responses	Frequency	Percent
0	16	48.5
1	1	3.0
5	1	3.0
20	1	3.0
25	2	6.1
50	4	12.1
66	1	3.0
100	7	21.2
TOTAL	33	100

Question 4: What would you have done if you found such mail?

(A) Deliver it to the correct addresses.

(B) Return it to the sender.

(C) Dispose of it.

(D) I do not know.

Responses	Frequency	Percent
A	309	80.5
B	11	2.9
C	53	13.8
D	11	2.9
TOTAL	384	100

The above tables give a summary of the responses to the survey. The responses to Questions 1 and 2 indicate that only 56 carrier managers were aware of the problems and of those, only 33 remember what they did with the mail. As a result of this, the information provided by Question 4 becomes very important. On the other hand the information from Question 4 is likely to be less accurate than the one provided by Question 3. This must be taken into consideration in the final analysis.

6.2.4 Summary

In this case study we described the process of conducting a survey starting with survey and questionnaire design and following with the data collection and analysis. The following questions provide some of the steps that may be followed for generating the final report.

Questions

1. The next step is to compute the proportion of delivered mail. To do this, a variable `qtot` can be created as

```
if q3a ne . then  qtot = q3a ;
else if q4 eq 'a' then qtot = 100;
else if q4 eq 'b' or q4 eq 'c' then  qtot = 0;
else if q4 eq 'd' then  qtot = .;
```

2. After merging the two datasets, it follows that the estimated percentage of delivered magazines (per zipcode) is `x * qtot / 100`.

3. Let `ntot` denote the result above. From this, an estimate of the total number of magazines can be calculated.

4. Doing this by `group` enables one to see the variation among delivery percentages.

5. A report would contain a summary table constructed as above and also discuss differences between the overall `ntot` and `qtot` delivery percentages.

6. Another way to approach the study is to calculate separate estimates of the percentage delivered using the answers to Questions 3 and 4, respectively. How would you combine both estimators to produce a final estimator?

6.3 A Device to Reduce Engine Emissions

Methods:	t-Tests Analysis of Variance
Data:	13 Days Testing on Van

6.3.1 Testing a Manufacturer's Product Claim

Manufacturers often "claim" their product will achieve a certain result (if used as directed), based on laboratory tests. In some cases, however, there may be questions about the testing methods used by the manufacturer. This may be due to conflicting results from independent tests, or there may be insufficient statistical evidence to support the manufacturer's claim.

In this case study, we present an investigation of a claim made by a manufacturer that its device, when placed on a motor vehicle engine, would cause hydrocarbon and carbon monoxide (CO) emissions to steadily decrease (to zero), and carbon dioxide (CO_2) emissions to steadily increase. In order to investigate these claims, tests were conducted by StatCon Enterprizes using the company's device on:

1. A Yugo limosine belonging to the manufacturer;

2. A Bentley belonging to the independent investigation team.

Here, we focus on the analysis using the manufacturer's car. The second experiment is left as a case study exercise in Chapter 9.

Car Data Description

In this investigation, hydrocarbon (ppm), carbon dioxide (% of volume), and carbon monoxide (% of volume) emissions were measured for the car owned by the manufacturer. The car was operated under the same conditions on 13 different days and 4 replicates of the experiment were performed on each of the testing days. The device was not placed on the engine until after the second day's measurements and remained on the engine while measurements were made on 9 additional days. The device was then removed from the engine and 2 additional sets of measurements were made. The study took approximately six months to complete. Table 6.5 summarizes the variables of interest in the `c63.dat` dataset.

Methodology

There are several approaches that can be employed to analyze these data. The main purpose of the study is to see if there is an effect due to `DEV` on any or all of the three emissions (response variables). A simple approach

TABLE 6.5. Variables in `c63.dat`

Variable	Definition
HC	Hydrocarbon (ppm)
CO	Carbon Monoxide (% of Volume)
CO2	Carbon Dioxide (% of Volume)
DAY	Test Day: $1, 2, \ldots, 13$
REP	Replicate: $1, 2, 3, 4$
DEV	Device: 1/0 (Present/Absent)

would be to employ a two-sample t-test to analyze the effect of DEV for each emission. However, this assumes the factors DAY and REP are not statistically significant. An analysis of variance model can be employed to account for the effects of these factors. Details on t-tests and ANOVA models were given in Chapter 3.

6.3.2 t-Tests

The effect of DEV can be analyzed using a two-sample t-test for each emission. This assumes that the effects of DAY and REP can be ignored which will need to be checked.

SAS will automatically provide the equal and unequal versions of the two-sample t-test. The unequal result should be considered when the standard deviations differ markedly (by a magnitude of four say). Since the t-test is only approximate in the case of unequal variance, the results should be interpreted with caution and compared with a nonparametric test such as the Wilcoxon Signed Rank test. Appendix C provides details on these tests.

6.3.3 Analysis of Variance

The effects of the factors DAY and REP can be investigated by incorporating them into an ANOVA model for each emission. Interactions could be added, although spurious results may be obtained if the error degrees of freedom is reduced too much.

Since the study took over three months to complete, a significant DAY effect would suggest a some type of "time" component was involved. A significant REP effect could potentially be due to the order of the replicate experiments that were conducted on each test day. Posthoc analyses will need to be performed to determine the nature of a significant effect.

6.3.4 Preliminary Analysis

We are interested in estimating the effect of the device on the mean emissions. That is, are there differences between emissions on days in which

TABLE 6.6. Mean Emissions for Manufacturer's Car

Device	Day	Hydro(ppm)	CO_2(%)	CO(%)
Absent	1	15.30	15.83	0.25
	2	10.75	14.41	0.11
Present	3	5.46	14.85	0.11
	4	6.04	13.48	0.16
	5	8.66	14.62	0.08
	6	13.18	13.38	0.13
	7	25.89	13.75	0.21
	8	4.45	15.95	0.20
	9	13.18	15.52	0.10
	10	1.11	15.42	0.11
	11	12.71	14.85	0.14
Absent	12	11.31	15.61	0.13
	13	13.54	14.24	0.15
Absent	Avg	12.72	15.02	0.16
Present	Avg	10.07	14.65	0.14

the device is present and days in which the device is absent. Table 6.6 summarizes the average emissions data for the car and Table 6.7 contains the results of the two-sample t-test for each emission. The output for these tables was generated from the following SAS code.

SAS Program for t-Test Analysis

```
data a ;
      infile 'c63.dat' ;
      input day hc co2 co ;
      dev=0 ;
      if days < 3 or day > 11 then dev=1 ;
proc means ;
      var hc co2 co ;
      by day ;
proc ttest ;
      class dev ;
      var hc co2 co ;
```

The results from the two-sample t-test suggest that there is not enough evidence to conclude that a statistically significant difference in mean emissions exists with respect to the presence and absence of the device.

TABLE 6.7. t-Test Results for Manufacturer's Car

Variable: HC

DEV	N	Mean	Std Dev	Std Error
Absent	16	12.724	4.172	1.043
Present	36	10.076	7.767	1.294

Variances	T	DF	Prob> $\|T\|$
Unequal	1.593	48.0	0.1178
Equal	1.279	50.0	0.2068

Variable: CO_2

DEV	N	Mean	Std Dev	Std Error
Absent	16	15.022	1.211	0.303
Present	36	14.647	1.262	0.210

Variances	T	DF	Prob> $\|T\|$
Unequal	1.016	30.0	0.3178
Equal	1.000	50.0	0.3222

Variable: CO

DEV	N	Mean	Std Dev	Std Error
Absent	16	0.159	0.099	0.025
Present	36	0.138	0.074	0.012

Variances	T	DF	Prob> $\|T\|$
Unequal	0.774	22.8	0.4468
Equal	0.865	50.0	0.3910

6.3.5 Summary

The manufacturer claimed that emissions should decrease (increase) monotonically for hydrocarbons and CO (CO_2) after the device is placed on the engine, as the engine is allowed to run (or mileage is accumulated), after an initial stabilization period (10% of the original mileage of the car). However, it seems clear from looking at Table 6.6 that these data do not substantiate that claim. The manufacturer also claimed that once the device is placed on an engine, the engine is irrevocably changed! If the device is removed, the effects of the device remain.

We should note that the variability in the emissions measurements analyzed here may be due to other factors such as weather and length of time the engine remained idle between tests, which were not controlled in the experiments. Furthermore, the magnitude of the measured hydrocarbon

emissions are, for the most part, less than the accuracy of the measuring device. Some other issues are presented below.

Questions

1. Use an ANOVA model to check whether the factors DAY and REP are important.
2. The accuracy of the measuring instrument is ±12 ppm for hydrocarbons, ±0.5% for CO_2, and ±0.06% for CO. What impact does this have on the statistical analysis?
3. Discuss what impact the following may have on the experimental results.
   - The test days were not necessarily contiguous.
   - The car was driven for approximately 700 to 2,000 miles between tests.
   - Testing began in November and was completed the following May.
   - New spark plugs were placed on the car after Day 7.
4. What is the potential consequence arising from the fact that the measuring instrument was recalibrated each day the measurements were taken.
5. To investigate the manufacturer's claim that once the device is placed on an engine, the engine is irrevocably changed, we can do a two-sample t-test for equality of mean emissions using the data from Days 1 and 2, versus those of Days 12 and 13. Is there evidence to support this claim?
6. An alternative approach is to consider all three emissions simultaneously. This leads to Hotelling's T^2 statistic or analyzing the data using a MANOVA model. The potential advantage of this approach is that the correlation between the responses is accounted for, which is expected according to the manufacturer's claim. Does the multivariate approach make any difference?

6.4 Reverse Psychology

Methods:	Rater Survey Analysis Correlation
Data:	PANSS Instrument 30 Rating Categories 72 Respondents

6.4.1 An Observational Experiment

The results of a large suicide prevention study provided evidence that a new antipsychotic drug was effective in helping prevent suicide in certain patients with schizophrenia. For this treatment to be effective, however, a physician would need to be able to assess accurately the status of a patient's psychosis and intention of suicide. Several types of rating scales can be used as measurement instruments for this purpose. In this case study, we consider the results obtained from a one-day workshop that was conducted for the PANSS instrument (positive and negative syndrome scale).

Workshop Outline:

A total of 72 physicians from the international community participated in the workshop. The main components of the training session were:

1. The physicians were first familiarized with the PANSS instrument which listed 30 psychological symptoms (see Table 6.8): 7 were designated as *negative,* 7 as *positive,* and the remaining 16 classified as *generic.* A scale of 1 (low) to 7 (high) was used to rate each symptom according to degree of psychosis exhibited by a patient for that particular symptom.

2. The next stage of the workshop involved all the physicians watching a 30-minute video of a patient and using the PANSS instrument to assess the patient's psychotic symptoms.

3. The PANSS instrument was provided in three languages and at the conclusion of the video, the physicians' ratings were compared to an "Expert's" rating, referred to as the *key.* After a short break, the workshop concluded with a presentation of the results.

The Data

Table 6.8 lists the 30 psychological symptoms that were rated during the video segment of the workshop. Each symptom was rated on a scale from 1 (low) to 7 (high) by the physicians who were distinguished only by a `RATER` code and their choice of language (`LANG`). The PANSS workshop data and key ratings provided by the Expert (identified as `RATER = 0`) are contained in the dataset `c64.dat`. The format of this dataset is indicated below.

RATER	LANG	P1	P2	...	P7	N1	N2	...	N7	G1	G2	...	G15	G16
0	E	4	2	...	2	5	4	...	2	3	4	...	3	1
101	F	4	1	...	1	5	3	...	3	3	5	...	2	5
:	:													
172	E	3	2	...	2	5	3	...	2	2	4	...	4	3

TABLE 6.8. PANSS Variable Definitions

	Variable	Definition
	RATER	Rater's identification code RATER = 0 is the Expert Rater
	LANG	Language used E = English F = French I = Italian
Positive Symptoms	P1	Delusions
	P2	Conceptual disorganization
	P3	Hallucinatory behavior
	P4	Excitement
	P5	Grandiosity
	P6	Suspiciousness/Persecution
	P7	Hostility
Negative Symptoms	N1	Blunted affect
	N2	Emotional withdrawal
	N3	Poor rapport
	N4	Passive/Apathetic social withdrawal
	N5	Difficulty in abstract thinking
	N6	Lack of spontaneity
	N7	Stereotyped thinking
Generic Symptoms	G1	Somatic concern
	G2	Anxiety
	G3	Guilt feelings
	G4	Tension
	G5	Mannerisms and posturing
	G6	Depression
	G7	Motor retardation
	G8	Uncooperativeness
	G9	Unusual thought content
	G10	Disorientation
	G11	Poor attention
	G12	Lack of judgment and insight
	G13	Disturbance of volition
	G14	Poor impulse control
	G15	Preoccupation
	G16	Active social avoidance

As can be seen from this excerpt, the first row of `c64.dat` is actually a "column header" with character entries corresponding to the variable names given in Table 6.8. This particular row will be useful for reading the data into S-PLUS since setting the option `header=T` in the S-PLUS function `read.table()` enables these names to be assigned to the columns of the resulting data frame. With other statistical software, this row may need to be skipped or deleted prior to reading in the data.

Methodology

The type of statistical analysis required in observational studies often involves little more than providing a good descriptive summary of the results. Summary tables, graphical displays, and simple statistical analyses are effective methods that can be employed by the consulting statistician for this purpose. That is, of course, assuming we have sufficient time to process the data, perform the analysis, and also compile the results into a suitable presentation format for our client ... let's say, around 20 minutes? In other words, during the "short" break at the PANSS workshop!

6.4.2 Statistics $\mathcal{R}$ Us

Providing the physicians with "on-the-spot" feedback about their performance with the PANSS instrument requires good planning and advance preparation. From the statistical consultant's perspective, the key difference in this analysis is that everything — from data processing to presenting the results — needs to be preplanned. But, without the data! Clearly, this is **not** the sort of approach we should try to attempt without a good understanding of the problem. We address some of the main issues involved in preplanning a statistical analysis.

Data Processing The structure and format of the data to be collected need to be relatively simple in order for the data processing phase to be completed efficiently. From our discussion in Chapter 3, Section 3.2, it follows that

- Establishing a specific format for data entry is essential. Missing values need to be appropriately identified so they can be properly dealt with during the analysis phase. Similarly, date[3] and time entries (if present) need to be interpreted correctly.

[3]While the Y2K problem turned out to be minimal, statistical software packages such as SAS now employ a cutoff date for two-digit year formats. Hence, someone born in 1921 may seem quite young for their age, particularly since they haven't actually been born yet!

- The major disadvantage of a preplanned analysis is that the data quality is unknown. This is where a preplanned analysis is likely to fail, especially if the usual error and data quality checks are skipped in the analysis program.

In this example, the dataset consists of a categorical grouping variable (`LANG`), and 30 ordinal variables corresponding to the PANSS ratings. Since this type of data can be easily entered through a scanning device, data quality should not present a problem.

Program Testing The analysis program needs to be carefully tested using dummy data. In particular, the effect of missing values needs to be checked since these can cause unexpected problems and errors to occur in what would otherwise be a correctly coded analysis program. Here, we confine our attention to S-PLUS and assume that the scanning device will produce a dataset which can be read by S-PLUS. Further information on S-PLUS may be found in Appendix B.2.

- The option `na.rm=T` is required in functions such as `mean(x)` and `sum(x)` to remove missing values *before* evaluating the vector `x`; otherwise these functions will return `NA` (missing) if any element of `x` is missing.
- Missing values can also make "logicals" invalid. These appear in statements such as `if(a && b) { ... }` which is executed whenever the logical `"a && b"` (`a` and `b`) is evaluated as `TRUE`. If `"a"` is missing, the logical cannot be evaluated as either `TRUE` or `FALSE` and an error occurs.
- An S-PLUS session typically involves the user creating functions that perform certain tasks related to the analysis. Where practical, these functions should be written in *generic* form and should employ default options instead of *hard-coded* quantities.

S-PLUS provides a comprehensive suite of statistical and graphical functions, so most of the functions we create will usually be concerned with data manipulations and setting up options for a particular statistical procedure or graphical display. In the Preliminary Analysis section of this case study, we have outlined some of the components that would be included in an S-PLUS analysis program for the PANSS dataset.

Generating Output The output generated by the analysis program needs to meet the objectives of the study. Normally, the objectives of the study would be investigated using an iterative approach to the analysis. That is, a preliminary analysis is conducted to identify important or interesting results which in turn, lead to further analysis. With a preplanned analysis the opportunity to pursue further

analysis is limited and the objectives of the study essentially need to be met by the initial output generated by the analysis program. Two important decisions are required with respect to the output generated by the analysis program.

- What are the specific results that need to be generated?
- What components of the program need to remain flexible?

For the PANSS study, deciding on the specific results to generate is important since there are 30 ratings and a grouping variable `LANG`. It follows that a rather substantial number of multiway frequency tables, correlations, t-tests, and nonparametric tests, can easily be generated from these variables. The same proliferation of graphical displays could also be achieved. While all these results can be examined in detail later, only the most appropriate summary statistics and displays should be generated to achieve the objective of providing feedback to the physicians.

The second question raises the issue of how the S-PLUS analysis program should be set up. This may depend on who actually performs the analysis at the workshop. For example, the entire program could be incorporated in a function, `run.anx()`, which generates all the output and formats it (discussed next). This approach provides no flexibility, but could be used to generate all the initial output that someone could start preparing for the presentation. Meanwhile the workshop "analyst" could rerun certain components of the program with different options such as the pass/fail comparison of the physicians' ratings with those of the Expert. We return to this last point in the Preliminary Analysis section.

Presentation The output generated by the analysis program will invariably require some form of "repackaging" before the results can be presented to a client or, as in this case study, to the workshop audience of physicians. S-PLUS provides minimal formatting of the numerical results it generates and the output from some S-PLUS statistical procedures would be quite cryptic to a nonstatistician. For presentation purposes, functions such as `table()` and `cat()` can be used to reformat some of the numerical output.

While reformatting and editing numerical output may be necessary, the default graphical displays produced by S-PLUS are usually quite good and should be exploited in the presentation. One approach would be to use PowerPoint (see Chapter 2, *Using PowerPoint for Presentations*) which can incorporate S-PLUS graphics as well as text excerpts from the numerical S-PLUS output. Again, these PowerPoint slides could be set up and tested with dummy data prior to the workshop.

6.4.3 Ordinal Data

Measurement instruments based on scales are widely used in many fields of study and the term *psychometrics* is associated with the statistical analysis of such instruments. Like the PANSS ratings, the data produced by scale-based instruments are *ordinal*, with integer-encoded "rankings" ($1, 2, \ldots, 7$ in the PANSS example). As ranks, this numerical encoding makes sense; it also allows summary statistics to be computed such as the mean which can be interpreted as the "average" rank.

In practice, the rank interpretation is often discarded (or not even considered) and the ordinal data are treated as so-called "raw scores." The rapid transition to continuous data comes when these raw scores are converted to "standardized scores" (the standardization procedure somehow (?) makes the scores better). In any case, after all the t-tests and (product) correlations have been exhausted, no analysis would be complete without the application of some multivariate technique such as factor analysis. . . . Clearly, the authors do not agree with this approach.

The problem we have with treating ordinal data as scores is that a fixed metric is imposed on what was essentially an arbitrary choice for the integer encoding. That is, the ranking system $A, B, \ldots, G$ could also have been used to represent the PANSS ratings. While this encoding is not numerically convenient, it does emphasize that the "distances" between the level codes are not necessarily well defined: $B - A$ is the same as the numerically coded version $2 - 1$, but is this really "1"? The letter ranking also illustrates the duality of data types possessed by ordinal data — categorical and ranked. From the methods discussed in Chapter 3, this suggests the following.

- In categorical mode, the PANSS ratings become count data: frequency and summary tables, large sample tests for proportions, and tests of association can be performed.
- In ranked mode, nonparametric rank tests, Spearman's correlations, and area-under-curve comparisons (t-tests) can be performed.

6.4.4 Preliminary Analysis

The purpose of this study was to assess the training program for the PANSS instrument. The workshop organizers were also interested in whether "language" had any impact on the performance of the physicians.

The following code illustrates some of the steps that would be involved in producing a summary analysis of the PANSS data using S-PLUS. The purpose of the code presented in each step should be apparent to most readers. Appendix B.2 provides further details on S-PLUS.

S-PLUS Analysis

```
#-------------------------------------------------------------
# Read in the data from "c64.dat" using 1st row as column
# labels. Treat periods as missing values.

  panss <- read.table("c64.dat",header=T,na=".")
  print(panss)          # Output the data
  panss$P1              # Output the "P1" ratings
  stem(panss$P1)        # Stemplot of P1 (Check the data ?)
```

The function `read.table()` reads the PANSS dataset `"c64.dat"` and creates the S-PLUS object `"panss"` which is a special type of list object called a "data frame." This allows the column of `P1` ratings to be referenced as a list component: `panss$P1`. A Stemplot provides a simple method for checking the data for one variable; checking all the data should be performed by a user-created `"chk(panss)"` function.

```
#-------------------------------------------------------------
# Recalibrate the physicians' ratings to those of the Expert

  a <- panss[-1,]     # Copy of panss, excluding the 1st row

# Simple loop. "0" entries now indicate a correct rating.
  for(i in 1:72)
     a[i,-c(1,2)] <- panss[i+1,-c(1,2)] - panss[1,-c(1,2)]
```

Since our comparisons need to be made with respect to the scores of the Expert Rater (`RATER = 0`), recalibrating the physicians' ratings to those of the Expert will be useful. The recalibration is achieved by the `"for"` statement shown above. A "correct" rating for any symptom is now indicated by a `"0"` entry in the data frame `"a"`. The remainder of the analysis will be performed with respect to the recalibrated ratings.

```
#-------------------------------------------------------------
# Graphical summaries.

  X11()           # Specify a graphics device (UNIX)
  hist(a$P1)      # Histogram of the recalibrated P1 ratings.

# Parallel Boxplots of the PANSS ratings.
  boxplot(a[,-c(1,2)])

# Boxplots of mean across all symptoms by Language
  plot.factor(apply(a[,3:32],1,mean,na.rm=T) ~ a$LANG)
```

A graphical device must be specified to produce the graphical displays above. `X11()` is a standard UNIX graphics device for interactive use. To out-

put these plots to a PostScript file in ESP format (encapsulated PostScript), specify the graphics device `postscript()`. Several options are available to control the graphics output, such as `onefile=T` which creates a separate PostScript file per plot.

Since `"a"` inherits all the attributes of `panss` (including column labels), `boxplot()` produces a parallel boxplot display of the recalibrated ratings labelled by symptom. The result is shown in Figure 6.1. The columns of a data frame can also be employed as variables in a "model" specification. (S-PLUS uses the `"~"` character in the model formula.) Hence the function `plot.factor()` produces parallel boxplots of the average (mean) rating across all 30 symptoms by `LANG` grouping. The `apply(*,1,mean,*)` function is needed to obtain the mean rating *per physician*. That is, by `1` = "rows." The resulting display is shown in Figure 6.2.

```
#-------------------------------------------------------------
# Numerical summaries.

# Overall summary ("RATER" column excluded). Includes mean,
#   median, min, max, std dev, and quartiles.
# Same summary, but "split" by Language (hence exclude the
#   "LANG" column as well as the "RATER" column)

  summary(a[,-1])
  lapply(split(a[,-c(1,2)],a$LANG),summary)
#-------------------------------------------------------------
# Spearman Correlation between P1 and P2 ratings.
# Overall and within the "E" group only

  cor.test(a$P1, a$P2, method="s")
  cor.test(a$P1[a$LANG == "E"], a$P2[a$LANG == "E"],
           method="s")
```

Graphical displays provide an overview of the results. For specific details, numerical summaries are needed. The `summary()` function provides simple statistics (indicated above) for each column of a data frame. The `lapply()` function can be used to obtain summaries for more complicated list objects, such as `split(a[,-c(1,2)], a$LANG)`. Simple descriptive statistics such as Spearman (rank-based) correlations can also be obtained as indicated. Both these methods would be appropriate in a more detailed analysis of the PANSS data.

```
#-------------------------------------------------------------
# Overall summary table.

  table(rep(names(a)[-c(1,2)],72),
        as.vector(t(as.matrix(a[,-c(1,2)]))))
```

```
     -4 -3 -2 -1  0  1  2  3 4 5 6
  G1  0  0 36 26  5  5  0  0 0 0 0
 G10  0  0  0  0 64  7  1  0 0 0 0
 G11  0  0  0 26 24 17  3  2 0 0 0
 G12  0  0  0 19 20 27  6  0 0 0 0
 G13  0  0  9 12 28 18  5  0 0 0 0
 G14  0  0  0  0 44 19  8  1 0 0 0
 G15  0  0 11 21 13 17  9  1 0 0 0
 G16  0  0  0  0  2  3 40 11 9 6 1
  G2  0  1  4 43  8 16  0  0 0 0 0
  G3  1  1  6 16 32 14  2  0 0 0 0
  G4  0  0 24 31 14  2  1  0 0 0 0
  G5  0  0 44  8 20  0  0  0 0 0 0
  G6  0  0  0  5 36 11 16  4 0 0 0
  G7  0  1  2 20 35  7  6  1 0 0 0
  G8  0  0 31 28 13  0  0  0 0 0 0
  G9  0  0  7  7 28 10 18  0 1 0 0
  N1  1  1  4  9 37 12  8  0 0 0 0
  N2  0  0  3 14 39 14  2  0 0 0 0
  N3  0  3 12 37 13  6  1  0 0 0 0
  N4  0  0  0  2 12 17 32  9 0 0 0
  N5  0  0  1  9 28 25  8  1 0 0 0
  N6  0  2  2 17 20 27  4  0 0 0 0
  N7  0  0  0  9 24 27  9  3 0 0 0
  P1  0  4  3 13 28 14 10  0 0 0 0
  P2  0  0  0 16 25 18 10  2 1 0 0
  P3  0  1  4 13 43 11  0  0 0 0 0
  P4  0  0  0  0 59  9  3  1 0 0 0
  P5  0  0  0 27 27  7  4  5 1 1 0
  P6  0  1 11 20 20 13  7  0 0 0 0
  P7  0  0  1 51 15  4  0  1 0 0 0
```

For presentation purposes a more compact type of "summary" format is preferable. The `table()` statement above provides an initial summary tabulation of the overall recalibrated PANSS ratings. As can be seen from the unedited output, the default alphanumeric ordering of the PANSS symptoms needs to be sorted into a more "reader-friendly" format. Tables could also be generated for each `LANG` group as in the correlation case. For example, substituting `a[a$LANG == "E",]` for `"a"` above, will generate the `"E"` summary table.

```
#-------------------------------------------------------------
# Did they pass ?
```

```
pass.fun<-function(a, np=5, nn=5, ng=10) {

# Default options pass/fail criteria:
# Pass = at least 5 of 7 within abs(1) of key ("0") for
#        both the N and P symptoms and at least 10 of 16
#        within abs(1) of key for the G symptoms
#        (Missing values count against the physician)

  b <- matrix(0,72,4)
  r1 <- 3:9       #   "Positive" symptoms (columns)
  r2 <- 10:16     #   "Negative"
  r3 <- 17:32     #   "Generic"

  for(i in 1:72) {
    b[i,1] <- sum(abs(a[i,r1])<= 1, na.rm=T)
                - sum(is.na(a[i,r1]))
    b[i,2] <- sum(abs(a[i,r2])<= 1, na.rm=T)
                - sum(is.na(a[i,r2]))
    b[i,3] <- sum(abs(a[i,r3])<= 1, na.rm=T)
                - sum(is.na(a[i,r3]))
    b[i,4] <- 1
    if( b[i,1] < np ) b[i,4] <- 0
    if( b[i,2] < nn ) b[i,4] <- 0
    if( b[i,3] < ng ) b[i,4] <- 0
  }
  return(b)
}
```

As with any "test," people want to know whether they passed and some criterion needs to be set to determine this. Here, flexibility in the analysis program will be required. The passing standard set *prior* to the workshop may turn out to be too stringent[4] and the workshop organizers may want to modify the passing criterion on the spot.

The default version of this function `"pass.fun(a)"` will use the specified options to compute the number of physicians who passed. To raise the "Generic" symptoms standard to 12 out of 16, we would run this function as: `"pass.fun(a,ng=12)"`. In either case, the pass/fail result is returned in last column of the matrix `"b"`, where 0 = fail, 1 = pass.

```
#-------------------------------------------------------------
# Fisher's exact test.
```

[4]Failure is not an easy subject to address. If nobody fails, was the workshop really necessary? If a few people fail, the onus is placed on those individuals to improve; if too many fail, the workshop itself is perceived to have failed.

```
table(a$LANG,b[,4])
   0  1
E 13 35
F  3 11
I  5  5

fisher.test(a$LANG,b[,4])

         Fisher's exact test

data:  a$LANG and b[, 4]
p-value = 0.3147
alternative hypothesis: two.sided
```

The final part of our analysis program addresses the question of whether language had an impact on the performance of the physician. Fisher's exact test is employed in case there were only a small number of physicians using one of the language formats for the PANSS instruments.

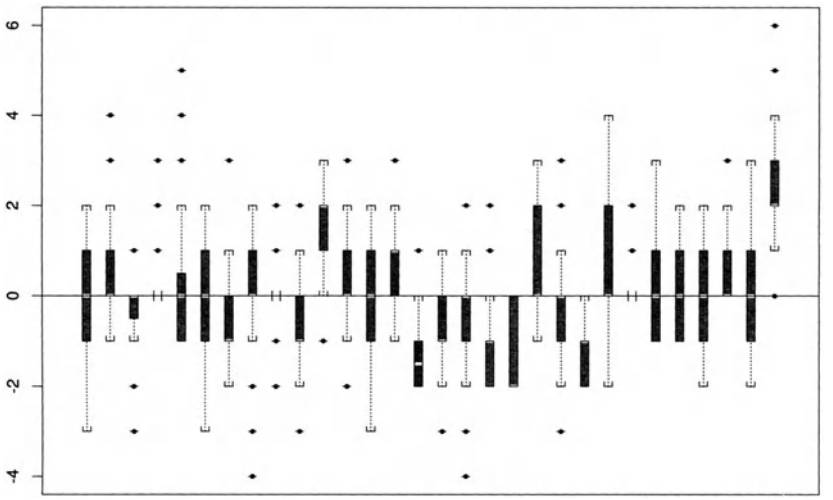


FIGURE 6.1. Parallel Boxplots of the PANSS Ratings

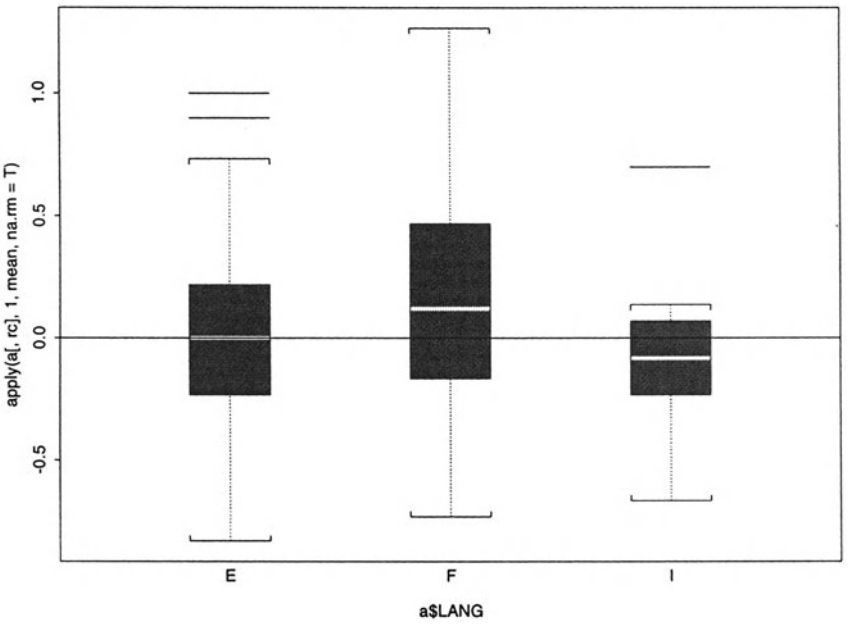


FIGURE 6.2. Average PANSS Rating by Language

S-PLUS Results

From the parallel boxplots in Figure 6.1, it can be seen that most of the physicians were within ±2 of the Expert Rater. However, there were some large disparities on some symptoms with the most noticeable exception being the G16 symptom ("Active social avoidance") which was rated higher by the majority of physicians. Since some symptoms may be more important in determining the intent of suicide, this information is clearly useful.

Figure 6.2 would seem to suggest that the French and Italian participants performed differently from their English counterparts. Specifically, the French ratings were, on average, higher than the Expert's; the Italian ratings lower. One confounding factor is that the Expert Rater was English. Furthermore, there were only 14 French and 10 Italian participants, as opposed to 48 English participants.

The result from Fisher's exact test suggests that language did not have an impact on the performance of the physicians. However, this may be partly due to the small sample sizes of the French and Italian groups. Despite the lack of significance, the distributional differences clearly suggest that the language factor warrants further investigation.

SAS Analysis

To perform this summary analysis with SAS is a little more complicated since we can't directly access the row elements of the variables. Some dexterous data manipulation is required to compare the Expert Rater's scores with the other physician's scores. Assuming the "column header" (row 1) has been deleted[5] from the `c64.dat` dataset, the following SAS code provides one approach to this problem.

```
data a ;
  infile 'c64.dat' ;
  if _N_ = 1 then
     do ;
        input key $ klang $ kp1-kp7 kn1-kn7 kg1-kg16 ;
        retain  ;
     end ;
  else
     input rater $ lang $ p1-p7 n1-n7 g1-g16 ;
```

The SAS dataset `"a"` now contains 64 columns: the Expert Rater's id code, language, and 30 rating scores, followed by the same variables (with different names) for the other 72 raters. The purpose of the `"retain"` statement is to keep the Expert Rater's results and replicate them in the first 32 positions (columns) before each new physician's ratings are read in.

We can now perform the usual column operations in SAS to obtain differences between the Expert Rater and the scores of other raters. This is performed in the data step below. The use of the special name `"_LAST_"` just means that the current contents of the SAS dataset `"a"` will be replaced by the results of this data step. The first row is excluded before the difference variables are created and only the rater code, language, and differences are kept.

```
data _LAST_ ;
  set a ;
  if _N_ > 1 ;

  dp1 = p1 - kp1 ;  ...  ; dp7 = p7 - kp7 ;
  dn1 = n1 - kn1 ;  ...  ; dn7 = n7 - kn7 ;
  dg1 = g1 - kg1 ;  ...  ; dg16 = g16 - kg16 ;

  keep rater lang dp1-dp7 dn1-dn7 dg1-dg16 ;
```

The dataset is now ready for analysis. Frequency tables and correlation tests can be performed in the usual way via `proc freq` and `proc corr`. The `"by lang ; "` statement can be used to obtain summaries by language.

---

[5] Alternatively, add `firstobs=2` to the `infile` line

6.4.5 *Summary*

Time constraints clearly limit the extent to which the statistical consultant can pursue an investigation and compromises often need to be made. Although the circumstances were rather special in this case study, good preparation and a thorough understanding of the problem were essential components in this preplanned analysis that facilitated the "on-the-spot" analysis of the PANSS workshop results.

We have emphasized the importance of background knowledge since some clients assume the consultant can provide them with a "complete analysis" of their project on the spot. "It's just a simple survey with 20 questions. ... Can't we do it now?" The answer to this question may seem obvious (No !), since we first need to understand the purpose of the survey and establish how the data were collected before "doing" the analysis. A more difficult situation arises when the client just wants to know "how" to program the analysis — they can do the actual analysis; they just need the program. Again, this case study illustrates that creating an analysis program is not a trivial exercise.

Questions

1. A subjective aspect of the PANSS study is the pass/fail criterion. Investigate whether a *significant* language effect can be achieved by modifying this criterion. If so, how would you explain the contradictory results?

2. An alternative way to process the PANSS dataset would be to "stack" the rating columns. That is, create a single `RESP` column consisting of all the ratings with `SYMP` identifying the actual symptom. The `RATER` and `LANG` codes would now be replicated 30 times for each physician. What are the advantages of doing this?

3. Complete the S-PLUS and/or SAS analysis program. It should include summary tables broken down by language grouping and provide the number and percentage of physicians below, above, and within ± 1 of the Expert Rater.

7
Case Studies from Group II

In this chapter we present four case studies that require more general statistical methods such as logistic regression, time series modeling, and factorial designs. The statistical problem is still well defined, but broader in scope than the Group I case studies. Several solutions may need to be evaluated. The case studies are listed below by section.

7.1 The Flick Tail Study

Methods:	Probit Analysis: ED50
Data:	Study of Drug treatment on Mice Response: Count Tail Movements: "Flicks"

7.1.1 Preclinical Statistics

This is a study about the potency of drug interactions between morphine and marijuana. In Chapter 1 we discussed the drug development process and one of the important steps is to determine the minimum dosage amount of the drug that achieves efficacy. Sometimes this may involve estimating combinations of drugs that achieve maximum efficacy with a minimum drug amount.

The experiment in this case study consists of administering a combination of two drugs to mice at different doses. The drugs are morphine and marijuana and are used as pain relief medication. In order to measure their effect "flick tail" tests are performed for two drugs and their combination. The objective is to detect whether an interaction effect, or *synergy*, exists between these two drugs. In addition, we are interested in calculating confidence intervals for the dose combination that will produce a 50% response; that is, the *effective dosage* at which 50% of the subjects would be expected to respond. This is commonly known as the $ED50$ value.

The Data

The data consist of a response variable P which gives the proportion of mice (out of 10) that flicked their tails after being administered a heat stimulus for a given drug combination. The predictor variables are the `MORPHINE` and marijuana (`DEL9`) dosages in milligrams (mg).

TABLE 7.1. Variable Definitions

Variable	Definition
`OBS`	Index number
`SET`	One of four possible combinations
`MORPHINE`	Dose of morphine sulfate (mg/kg) injected into study mice. The range is 0 to 8.0
`DEL9`	Dose of Delta9-THC (mg/kg) injected into study mice. The range is from 0 to 16
`FLICK`	The number of mice that flick the tail after being applied a heat stimulus from beneath
`REPS`	Number of repetitions (with different mice)
`P`	`FLICK / REPS`

The experiment was repeated for 20 different drug combinations and the results are shown in Table 7.2. For six of the experiments the mice were injected with a combination of both drugs, whereas for 13 experiments the mice were administered only one drug. One of the experiments was a control for which no drug was given to the 10 mice.

TABLE 7.2. The Flick Tail Data

OBS	SET	MORPHINE	DEL9	FLICK	REPS	P
1	O	0.0	0.0	0	10	0.0
2	A	1.0	0.0	0	10	0.0
3	A	2.0	0.0	2	10	0.2
4	A	3.0	0.0	4	10	0.4
5	A	4.0	0.0	4	10	0.4
6	A	5.0	0.0	6	10	0.6
7	A	6.0	0.0	9	10	0.9
8	A	7.0	0.0	9	10	0.9
9	A	8.0	0.0	8	10	0.8
10	B	0.0	1.0	0	10	0.0
11	B	0.0	2.0	0	10	0.0
12	B	0.0	4.0	9	10	0.9
13	B	0.0	8.0	9	10	0.9
14	B	0.0	16.0	10	10	1.0
15	AB	0.5	0.5	0	10	0.0
16	AB	1.0	1.0	2	10	0.2
17	AB	1.5	1.5	7	10	0.7
18	AB	2.0	2.0	9	10	0.9
19	AB	2.5	2.5	9	10	0.9
20	AB	3.0	3.0	9	10	0.9

Methodology

7.1.2 Logistic Regression

Experiments for studying the efficacy or toxicity of a drug produce a response that is often a binary variable (1: Response, 0: Nonresponse) or binomial (k out of n are Response). For this type of data it is customary to fit the logistic regression model:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Dose \ ,$$

where p is the probability of Response at a given $Dose$. The variable $Dose$ may be represented in the log scale as $\log(Dose)$ since drug doses are typically chosen in an exponential progression. This is normal of experiments

where a reasonable range of doses is not known and the experimenter tries to capture responses for doses in several scales.

For each dose value the experiment would be repeated n_i times and the observed number of responses is r_i. The model parameters (β_0, β_1) can be estimated by maximum likelihood estimation (MLE) which we denote by $(\hat{\beta}_0, \hat{\beta}_1)$. It can be shown that $(\hat{\beta}_0, \hat{\beta}_1)$ is approximately normally distributed with covariance matrix V where $V_{00} = Var(\hat{\beta}_0)$, $V_{11} = Var(\hat{\beta}_1)$, and $V_{01} = V_{10} = Cov(\hat{\beta}_1, \hat{\beta}_0)$.

One quantity of interest is the median effective dose $ED50$ or the dose necessary to obtain $p = 50\%$ efficacy. From the model equation we obtain

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{1/2}{1-1/2}\right) = 0 = \beta_0 + \beta_1 ED50$$

and hence $ED50 = -\beta_0/\beta_1$. The MLE of $ED50$ is $-\hat{\beta}_0/\hat{\beta}_1$. For other values of p we define $ED100p$ as the dose of $100p\%$ effectiveness. The MLE is: $\widehat{ED100p} = \left[\log(p/(1-p)) - \hat{\beta}_0\right]/\hat{\beta}_1$.

Once we obtain the estimates of $ED50$s we calculate confidence intervals for them. For this purpose we apply Fieller's theorem which enables us to obtain confidence intervals of ratios of Gaussian random variables. Let $\gamma = z_{\alpha/2} V_{11}/\hat{\beta}^2$; then the $100(1-\alpha)\%$ confidence interval for $ED50$ is

$$\widehat{ED50} + \frac{\gamma}{1-\gamma}(\widehat{ED50} + \frac{V_{10}}{V_{11}}) \pm \frac{z_{\alpha/2}}{(1-\gamma)\hat{\beta}_1} K \ ,$$

where

$$K^2 = V_{00} + 2\widehat{ED50} V_{10} + \widehat{ED50}^2 V_{11} - \gamma(V_{00} - V_{10}^2/V_{11}) \ .$$

The matrix of $V_{(ij)}$s is obtained from the logistic regression output. This equation can also be extended to the case of general p by replacing $ED50$ by $ED100p$ in the equation.

7.1.3 Multiple Logistic Regression

In this case study we also consider the logistic model with two predictors and one interaction term.

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Dose1 + \beta_2 Dose2 + \beta_3 Dose1 * Dose2 \ .$$

The dose response is now a surface of two variables, $Dose1$ and $Dose2$, and the $ED50$ is not a point but a contour on the surface at height 0.5. The $ED50$ curve can be calculated by giving a value of $Dose1$ for each fixed value of $Dose2$, which is

$$Dose1 = \frac{-\beta_0 - \beta_2 Dose2}{\beta_1 + \beta_3 Dose2} \ .$$

If this curve is convex, it is a sign that there is synergy.

7.1.4 Preliminary Analysis

To analyze these data we recommend the use of PROC PROBIT in the SAS software. PROC PROBIT calculates maximum likelihood estimators for all our parameters and confidence intervals for $ED100p$ for any p using Fieller's theorem. A more sophisticated analysis can be obtained from S-PLUS using the `glm()` procedure as follows.

S-PLUS Notes

Here are some of the S-PLUS functions that will be used.

`glm`	basic function for generalized linear models (this includes logistic regression).
`contour`	function for graphing a contour plot.
`predict.glm`	function for calculating predicted values.

Analysis for the Example Dataset

```
# Create the dataset flick from the file "c71.dat"
flick <- read.table("c71.dat",head=T)

# Run the logistic regression for the model with
# interaction term
flick.glm <- glm(p ~ morphine*del9,w=rep(10,20),
                     data=flick, family=binomial)

# Create the dataset with the grid points for the
# contour plot
morphine <- sort(unique(flick$morphine))
del9 <- sort(unique(flick$del9))
pred <- data.frame(morphine=rep(morphine,10),
                   del9=rep(del9,rep(12,10)))

# Evaluate the estimated model at the grid points
phat <- predict.glm(flick.glm, pred, type="response")
dim(phat) <- c(12,10)

# Graph the ED50 and ED75 curves.
contour(morphine,del9, phat, levels=c(0.5,0.75))
```

If there is synergy between morphine and Del9 the graph should show convex contours.

7.1.5 Summary

This case study illustrates the methodology for dose response experiments. Fieller's theorem is very useful here for calculating confidence intervals for ratios, but it has many other applications not covered here.

Questions

Some interesting questions that we must answer in this study are:

1. Calculate ED50s for each of the two drugs and for the combination of the two. One possible alternative is to model the response using $log(Dose)$ or $log(1 + Dose)$. Which model produces the best fit?

2. Calculate confidence intervals for ED50s using Fieller's theorem.

3. Look at the drug combinations. Do they exhibit synergy?

   As stated in the section above, one way to show synergy is to produce a contour plot of the surface of values of P predicted by the model, over the dose region $0 < MorphineDose < 8$ and $0 < Del9 < 16$. Specifically one could draw the contour corresponding to the $ED50$. Convex contours would indicate synergy.

4. Give the linear combination of drugs that produces the highest efficacy.

7.2 Does It Have Good Taste?

Check List

Methods:	Factorial Design Centerpoint Design
Data:	5 Design Factors 4 Response Variables

7.2.1 Factorial Designs in Food Science

This particular case study represents a good example of an ongoing research project where the *educational* role of the statistical consultant played an important part during the initial stages of the investigation. The initial consultation with our client actually occurred several years prior to the experiment presented here. For illustrative purposes, let us briefly retrace our steps back to the initial problem.

The Beginning

The client's research project was concerned with the study of measurable characteristics of a food product such as color, density, and moisture content. These response variables were obtained by forcing the food product (corn) through an "extruder" machine which could be set up to operate under different conditions. For example, the temperature level of each oven chamber, the amount of water and dye added to the corn mixture, and the rate of extrusion could all be varied by the operator. For each control variable or factor, there were at least three potential levels of interest to the client. The client was well aware that this would lead to an experiment involving several thousand (!) factor combinations — hence the reason for seeking statistical advice.

Even at this early stage of the consultation session, it was clear that we would need to introduce the client to the statistical approach towards experimentation. However, rather than immediately divert the focus of the discussion to statistical issues, our first step was to work with the client and eliminate the factors and levels that were not considered a priority. This enabled us to gain a better understanding of the details associated with the client's project. After completing the initial elimination, the remaining factors still left us with 1000 combinations and we were now ready to start convincing our client about the need to consider alternative experimental designs. This didn't take very long.

It was at this point that the client indicated they would be doing well to perform two trials per week! Needless to say, an alternative design was immediately considered which reduced the study to 16 trials. As it turned out, 20% of these initial fractional factorial trial combinations were unstable and did not produce results.

The Present

The snack food industry is highly competitive and marketing a product successfully requires more than just a "tastes nice" characteristic. Since the nutritional value of some products may be rather dubious, enhancing the appearance, smell, and other "qualities" of the food product are important. Slick packaging and glitzy advertising certainly help, but nobody will want to eat something that smells bad!

The purpose of our client's research project was to try to optimize the conditions under which certain food quality characteristics would be optimized. Like many clients, moving away from the one-factor-at-a-time approach to experimentation can be difficult and in this situation, the educational role of the statistical consultant becomes important. The aim is to provide the client with guidance and a better understanding of the **iterative** nature of statistical experimentation.

The initial fractional factorial "screening" design we employed above clearly showed that high moisture content and high temperature levels led

to the failures. Based on these results, the client managed to convince the experimentation team of the merits of the statistical approach. The next "iteration" of the project employed a fractional factorial design with three levels per factor. This was later refined to the use of *central composite* designs. The results from a central composite experiment are presented in this case study.

The Data

The control variables (factors) and four response variables (outputs) that were measured in this extrusion experiment are listed in the table below. Different combinations of the factor levels were employed for each run where Base Rate refers to the quantity of food product (corn) placed in the extrusion machine. Water and Dye (measured as a percentage) were added to the corn and the resulting mixture was slowly "cooked" (Temperature) as it was pushed through the extruder machine. The pushing mechanism was basically a long winding "screw" plunger and hence Screw Speed is a measure of how fast the mixture was being extruded.

Variables in the Central Composite Design

Outputs: Density, Thickness, Moisture, Flavor							
Factors			*Factor Levels*				
A	Screw Speed	(rpm)	550	590	625	665	700
B	Base Rate	(lbs)	1.5	1.7	1.9	2.1	2.3
E	Water Added	(%)	14	17	20	23	26
G	Dye Added	(%)	0.00	0.75	1.5	2.25	3.00
H	Temperature	(°F)	240	255	270	285	300

The results from this experiment are contained in the dataset `c72.dat`. As indicated in Table 7.3, the central composite design employed consisted of a half replicate of a 2^5 factorial (16 runs), augmented by 10 so-called *axial* points and 4 *center* points giving a total of 30 runs. Now let's find out what all this really means!

Methodology

The objective of this study was to try to determine the combination of operating conditions associated with the quantitative factors A,B,E,G, and H, that will optimize the output of the four response variables: density, thickness, moisture, and flavor. For analyzing this type of problem, response surface methodology can be employed.

TABLE 7.3. Central Composite Design Experiment

Factorial (2^5 Half Rep.)					*Outputs*			
A	B	E	G	H	Density	Thickness	Moisture	Flavor
590	1.7	17	0.75	285	197.725	0.18350	11.300	337
665	1.7	17	0.75	255	187.940	0.19565	11.175	340
590	2.1	17	0.75	255	198.290	0.21460	11.125	337
665	2.1	17	0.75	285	184.620	0.22895	11.050	339
590	1.7	23	0.75	255	303.340	0.16855	14.000	326
665	1.7	23	0.75	285	280.615	0.16540	13.600	329
590	2.1	23	0.75	285	280.265	0.19550	13.525	326
665	2.1	23	0.75	255	269.440	0.19385	13.375	329
590	1.7	17	2.25	255	218.215	0.19575	11.575	333
665	1.7	17	2.25	285	202.620	0.22005	11.125	338
590	2.1	17	2.25	285	220.495	0.22495	11.150	337
665	2.1	17	2.25	255	214.700	0.22175	11.425	339
590	1.7	23	2.25	285	312.415	0.15585	13.975	321
665	1.7	23	2.25	255	306.810	0.16190	13.775	322
590	2.1	23	2.25	255	315.135	0.18230	13.900	320
665	2.1	23	2.25	285	290.955	0.19585	13.300	325
Axial Points								
550	1.9	20	1.50	270	265.405	0.18465	12.625	329
700	1.9	20	1.50	270	238.715	0.19395	12.200	334
625	1.5	20	1.50	270	245.055	0.15345	12.525	331
625	2.3	20	1.50	270	243.790	0.20715	12.225	331
625	1.9	14	1.50	270	167.655	0.22575	9.750	341
625	1.9	26	1.50	270	336.835	0.16625	14.575	324
625	1.9	20	0.00	270	224.930	0.18630	12.025	336
625	1.9	20	3.00	270	261.510	0.18990	13.125	331
625	1.9	20	1.50	240	255.015	0.19100	12.625	330
625	1.9	20	1.50	300	246.800	0.21305	12.275	331
Center Points								
625	1.9	20	1.50	270	253.840	0.18490	12.625	331
625	1.9	20	1.50	270	248.495	0.18800	12.675	333
625	1.9	20	1.50	270	237.240	0.19405	12.325	339
625	1.9	20	1.50	270	247.310	0.20325	12.600	332

7.2.2 *Response Surface Methodology*

The notion of a *response surface* follows from the model representation:

$$y = g(x_1, x_2, \ldots, x_k) + \epsilon \ , \tag{7.1}$$

where y is the observed response, ϵ is the random error component, and $g(x_1, x_2, \ldots, x_k)$ is a function of the levels of the quantitative control variables $x_1, x_2, \ldots, x_k$. The expected value of the response is therefore defined by the "surface" $g(x_1, x_2, \ldots, x_k)$. Note that when $k = 2$, the response surface can be graphically represented by a contour plot.

In general, the true form of the functional relationship between y and the control variables will be unknown. Thus, the first stage of a response surface analysis is to find a suitable approximation for the function $g(x_1, x_2, \ldots, x_k)$ that appears in (7.1). Since the main objective is to determine the optimum operating conditions for a process under study, a linear or quadratic polynomial approximation is often sufficient for this purpose. These approximations are referred to as, respectively, *first-order* and *second-order* models:

First-Order $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$

Second-Order $y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \sum_{j=1}^{k} \beta_{ij} x_i x_j + \epsilon \ .$

Clearly, these approximations can only be expected to work well within a "local" region of the space spanned by the control variables and do not aim to provide the investigator with an understanding of the actual $g(x_1, x_2, \ldots, x_k)$ underlying the process. However, they can be employed effectively for the purposes of optimizing a process in "iterative" fashion. That is, response surface methodology is a *sequential* procedure: the results from one experiment point the investigator towards the region where the "next" experiment should be conducted.

Method of Steepest Ascent

The quickest route to the top of a hill is to follow the path of *steepest ascent.* It is also the most physically exhausting route! Fortunately, our computer will do the climbing and in the context of analyzing a response surface, adopting this strategy provides an economically efficient procedure for finding the optimum.[1]

During the initial stages of the investigation, the location of the region containing the optimum is unknown. To simplify interpretation of the results, a first-order model is often used. In this case , the direction of steepest ascent corresponds to the direction in which $\hat{y}$ increases most rapidly. It

[1] Minimizing the expected response is equivalent to walking down to the base of a valley. The method of steepest *descent* then applies.

can be shown that steps along this path are proportional to the vector comprised of the fitted regression coefficients $\hat{\beta}_j$. The actual step size to employ needs to be determined by the investigator based on their experience of the process under study. In practice, experiments would be conducted along this path until a decrease is obtained, or the limit of feasible operating conditions is reached.

Once the region containing the optimum has been identified, more elaborate models may be used to determine the actual optimum. The quadratic approximation is usually reserved until this stage since the experimental designs employed for these models require more trials to be conducted.

7.2.3 First- and Second-Order Designs

The linear and quadratic approximations employed in a response surface analysis could be fitted as a standard "regression" problem: the investigator simply selects enough values for the control variables in order to uniquely estimate the regression parameters. A more efficient approach is to select the values of the control variables according to an experimental design.

Coded Factors

The control variables or factors associated with a response surface analysis are assumed to be quantitative and directly *controllable* by the investigator. This means that any level can be achieved within the feasible range of a factor. While calibration effects may impose some limitations in practice, it follows that the original levels of a factor can be "coded." The usual form of the coding transformation applied to a factor, ξ_j say, is

$$\text{coded value} = x_j = \frac{\xi_j - \frac{1}{2}\left(\min \xi_j + \max \xi_j\right)}{\frac{1}{2}\left(\max \xi_j - \min \xi_j\right)} \tag{7.2}$$

which maps the levels of the factor ξ_j to the interval $[-1, 1]$. The term *natural* variable is sometimes used to distinguish between the original measurement units of a factor and its (unitless) *coded* version. In the following discussion, the $x_1, \ldots, x_k$ denote the coded factors.

First-Order Model

The class of orthogonal designs that minimize the variance of the regression coefficients of the first-order model includes the 2^k factorial and 2^{k-p} fractional factorial designs in which main effects are not aliased with each other. These designs enable k factors to be investigated with the fewest possible runs since only two levels per factor are involved. In terms of our coded factors, the two levels are denoted by ± 1. These designs were briefly discussed in Chapter 3 (see *Planned Experiments*), and are extensively dealt

with by many statistical texts. In particular, we refer the reader to Montgomery (1997) and Box et al. (1978) for further details on the 2^k factorial and related designs.

A single replicate of a 2^k design, however, does not afford an estimate of the experimental error unless some runs are repeated. Instead of repeating runs, a standard method is to augment the 2^k factorial with several observations at the "center" of the design. In terms of the coded factors, this is the point $x_j = 0,\ j = 1, 2, \ldots, k$. The addition of center points does not alter the orthogonality property of the design, nor do they influence the regression coefficients $\hat{\beta}_j$ associated with the factors x_j.

Another orthogonal design that requires only $k + 1$ runs for k variables is the *simplex*. This is a regular polyhedron with $k + 1$ vertices in k dimensions. For example, with $k = 2$ factors the simplex design is an equilateral triangle; with $k = 3$ it is a regular tetrahedron.

Second-Order Model

To fit a second-order model, each factor must have at least three levels. This is clearly satisfied by the 3^k factorial design and is also achieved by the addition of center points to the 2^k design. However, the number of runs required in a 3^k design increases rapidly with k and neither of these factorials are *rotatable* designs. We define what "rotatable" means in our discussion of a central composite design below. Before doing so, we need to address what is meant by the "canonical analysis" of a response surface.

Canonical Analysis

The analysis of the fitted second-order model for a response surface is called a *canonical analysis* which is the main reason for using coded factors. This is because the *canonical analysis* of the response surface is **not** invariant with respect to location and scale changes. Thus, the coding defined by (7.2) gives equal weight to each factor and makes the levels of different factors comparable.

The results of a canonical analysis characterize the nature of the response surface at the stationary solution associated with the quadratic model. To see this, we note that the fitted second-order model can be written in the form:

$$\hat{y} = \hat{\beta}_0 + \boldsymbol{x}'\boldsymbol{b} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} \ ,$$

where $\boldsymbol{x}$ is the $k \times 1$ vector of coded factors, $\boldsymbol{b}$ is the $k \times 1$ vector of fitted first-order regression coefficients, and $\boldsymbol{B}$ is a $k \times k$ matrix of second-order coefficients. Optimizing the predicted response with respect to $\boldsymbol{x}$ gives the stationary solution

$$\boldsymbol{x}_0 = -\tfrac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b} \quad \Rightarrow \quad \hat{y}_0 = \hat{\beta}_0 + \tfrac{1}{2}\boldsymbol{x}_0'\boldsymbol{b} \ ,$$

where $\hat{y}_0$ is the predicted response at the stationary point.

The stationary point is characterized by the sign and magnitude of the *eigenvalues* associated with the matrix $\boldsymbol{B}$. Letting $\lambda_1, \lambda_2, \ldots, \lambda_k$ denote the eigenvalues of $\boldsymbol{B}$, the nature of the response surface at the stationary point can be determined as

Maximum If all the λ_j are *negative,* the response surface has a local maximum at the stationary point.

Minimum If all the λ_j are *positive,* it has a local minimum.

Saddle Point If the λ_j have *mixed* signs, the stationary point is called a "saddle point." This means the response surface increases in one direction (as we move away from the stationary point), but decreases in another direction.

Stationary Ridge If one or more of the λ_j are close to *zero,* then the response surface is "flat" over some region or has a stationary ridge. This means the response variable is insensitive to changes in a certain combination of the factor levels over that region or along the ridge.

Central Composite Design

The most widely used design for fitting a second-order response surface model is the *central composite design.* Five levels per factor are required in these designs which consist of the following.

Factorial A 2^k or 2^{k-p} fractional factorial design for two of the factor levels. These are coded by the usual ± 1 notation. Main effects should not be aliased with each other in the 2^{k-p} design.

Axial Points A total of $2k$ additional observations is made by assigning two runs for the two extreme levels of each factor. In each case, the two levels are coded as $\pm\alpha$ with all the remaining factors set to their center point (coded) level: $x_j = 0$. The choice of α is discussed below.

Center Points A total of n_o *center point* runs is made with all the factors set at their middle level. This corresponds to the center point code of $(0, 0, \ldots, 0)$.

The central composite design can be easily built up from a first-order 2^k factorial or 2^{k-p} design. By appropriate choice of α and n_o, the following properties hold.

Rotatable A design is said to be rotatable if the variance of $\hat{y}$ at some $\boldsymbol{x}$ depends only on the distance of that point from the design center, and not on its direction.

Let $n_f = 2^{k-p}$ denote the number of points used in the (fractional) factorial portion of the design. Then setting $\alpha = \sqrt[4]{n_f}$ (fourth root of n_f) makes the central composite design rotatable.[2]

Orthogonal In an orthogonal central composite design, the number of center points is chosen so that the second-order parameter estimates are minimally correlated with other parameter estimates. The choice of n_o depends on the number of factors k and whether *blocking* is employed.

Uniform Precision In a uniform precision design, the variance of $\hat{y}$ at the origin is the same as the variance at a unit distance from the origin. A central composite design can be made uniform precision by the appropriate choice of n_o.

A table of some central composite designs is presented in Appendix C, Table C.10. The entries in this table were obtained from the statistical software package JMP. In addition to the choice of n_o and blocking schemes, we have included values of α that will make the effects uncorrelated (orthogonal). These values are denoted by α_{or} and apply to the three types of central composite designs: orthogonal, block, and uniform precision.

Box–Behnken Design

An alternative to the central composite design is the *Box–Behnken* design where only three levels per factor are required. This is essentially a 3^k factorial design, but only the midedge points are actually run. Thus, there is higher uncertainty near the vertices of the hypercube in a Box–Behnken design as compared to the central composite design.

7.2.4 Practical Considerations

We conclude our discussion of response surface methodology with some practical considerations and comments.

- The mathematical sense of the term "optimum" refers to maximizing (or minimizing) a response function with respect to the control variables. In practice, the true form of the functional relationship is unknown and approximations may only be useful over a relatively small region.

---

[2] Geometrically, the value of α corresponds to the radius of the hypersphere that circumscribes the hypercube defined by the 2^{k-p} factorial. In the simplest case ($k = 2$), the axial points lie at the compass points (N, S, E, and W) on a circle of radius $\alpha = \sqrt{2} = \sqrt[4]{4}$ that circumscribes the square with vertices at $(\pm 1, \pm 1)$. Since the vertices of the square correspond to the 2^2 factorial points, the design is clearly rotatable.

- The optimum obtained from a "local" approximation may fall outside the feasible operating region. This imposes constraints on the optimization procedure and nonlinear programming methods may be required to determine the optimum factor level combination.

- A "sequential" approach is required to analyze these types of optimization problems. That is, the investigator needs to be prepared to perform several sets of experiments in order to identify the region of optimality.

- The optimum factor combination found in a small-scale experiments may not be optimal in the "scaled-up" process due to the greater variability inherent in a full-scale operation.[3]

- The optimum factor combination for one response variable may result in a poor response for another output variable. The investigator must then search for nonoptimal conditions that provide acceptable responses for all output variables.

7.2.5 Preliminary Analysis

The RSREG procedure in SAS is specifically designed for analyzing response surface experiments. In the following code, we have taken moisture as our response variable Y.

SAS Program

```
data a ;
      infile 'c72.dat' ;
      input a b e g h density thick moist flavor ;
      y = moist ;

proc rsreg data=a  ;
      model y = a b e g h  ;
```

The RSREG procedure is then applied to this dataset and the output contains the results from fitting a *second-order* model to Y. That is, the quadratic and crossproduct terms are automatically generated by SAS from the list of the factors specified in the model statement. Prior to fitting the quadratic model, SAS employs (7.2) to code each factor. The complete output from this SAS program is presented below.

---

[3]These variations can also cause the process to "drift" away from the optimum. The technique of *evolutionary operation* (EVOP), proposed by Box (1957) and Box and Draper (1969), provides a method for maintaining optimum operation conditions.

SAS Output

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
a	625.000000	75.000000
b	1.900000	0.400000
e	20.000000	6.000000
g	1.500000	1.500000
h	270.000000	30.000000

Response Surface for Variable y

Response Mean	12.451667
Root MSE	0.201209
R-Square	0.9902
Coefficient of Variation	1.6159

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	5	36.573050	0.9795	180.67	<.0001
Quadratic	5	0.205313	0.0055	1.01	0.4626
Crossproduct	10	0.197187	0.0053	0.49	0.8611
Total Model	20	36.975551	0.9902	45.67	<.0001

Residual	DF	Sum of Squares	Mean Square
Total Error	9	0.364366	0.040485

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	-11.525846	28.875689	-0.40	0.6991	12.526067
a	1	0.011770	0.045124	0.26	0.8001	-0.209533
b	1	1.536970	7.412558	0.21	0.8404	-0.192992
e	1	1.240421	0.475572	2.61	0.0284	2.436631
g	1	1.679162	1.834098	0.92	0.3838	0.274575
h	1	0.023274	0.115973	0.20	0.8454	-0.168621
a*a	1	-0.000009458	0.000027669	-0.34	0.7403	-0.053200
b*a	1	0.005113	0.006702	0.76	0.4651	0.153384
b*b	1	-0.566433	0.974196	-0.58	0.5752	-0.090629
e*a	1	-0.000538	0.000447	-1.20	0.2592	-0.242141
e*b	1	-0.085937	0.083837	-1.03	0.3321	-0.206250
e*e	1	-0.008420	0.004330	-1.94	0.0837	-0.303129

g*a	1	-0.000663	0.001787	-0.37	0.7192	-0.074612
g*b	1	0.135417	0.335349	0.40	0.6958	0.081250
g*e	1	-0.004861	0.022357	-0.22	0.8327	-0.043750
g*g	1	0.048609	0.069276	0.70	0.5006	0.109371
h*a	1	-0.000002589	0.000089360	-0.03	0.9775	-0.005825
h*b	1	-0.005729	0.016767	-0.34	0.7404	-0.068750
h*e	1	0.000034722	0.001118	0.03	0.9759	0.006250
h*g	1	-0.005139	0.004471	-1.15	0.2801	-0.231250
h*h	1	-0.000017366	0.000173	-0.10	0.9223	-0.015629

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
a	6	0.363290	0.060548	1.50	0.2817
b	6	0.306766	0.051128	1.26	0.3612
e	6	35.722186	5.953698	147.06	<.0001
g	6	0.534401	0.089067	2.20	0.1385
h	6	0.229543	0.038257	0.94	0.5091

Canonical Analysis of Response Surface Based on Coded Data

	Critical Value	
Factor	Coded	Uncoded
a	7.022171	1151.662799
b	4.738991	3.795596
e	-0.257964	18.452219
g	-2.134848	-1.702273
h	-1.383836	228.484934

Predicted value at stationary point: 10.842389

	Eigenvectors				
Eigenvalues	a	b	e	g	h
0.191566	-0.018442	0.205180	-0.078416	0.836458	-0.501747
0.073780	0.733600	0.541105	-0.374166	-0.168158	-0.027548
-0.084988	0.011803	0.111807	-0.095366	0.481977	0.863692
-0.162026	0.611921	-0.766940	0.005267	0.192547	-0.015948
-0.371550	0.294796	0.253789	0.919091	0.051815	0.035686

Stationary point is a saddle point.

The analysis of variance results would appear to indicate the response surface has no significant curvature and that moisture content is essentially

determined by factor E: the percent of water added.[4] From the regression diagnostics associated with individual terms in the model, there appears to be weak evidence (`Pr > |t| = 0.0837`) of a quadratic effect due to factor E. Thus, the relationship between factor E and moisture is not necessarily linear.

The canonical analysis reveals that the "optimum" solution is a saddle point. It also shows that the optimum is impossible to achieve since factor G would require adding a *negative* amount of dye. (Not the easiest thing to do in practice.) In this case, of course, the analysis of variance results have already shown that moisture appears to be insensitive to all the factors except E and so the optimum conditions suggested by the canonical analysis should be regarded as a mathematical artifact of the optimization.

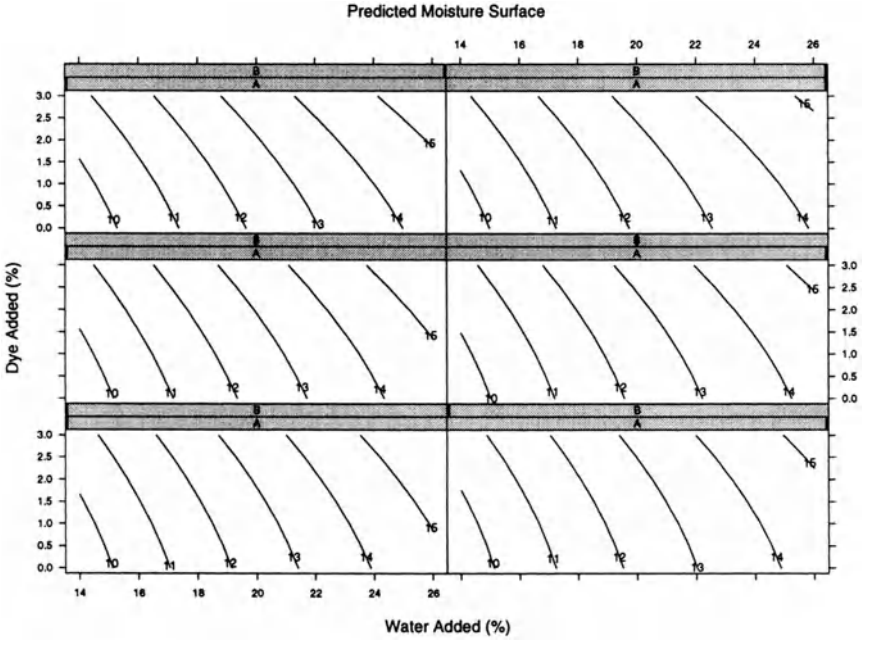


FIGURE 7.1. Contour Plot of Moisture Surface Using Trellis Graphics

A response surface can be conveniently displayed using a contour plot. In this example, we first used the fitted model to generate predicted moisture values over a grid of points defined by the factors, and then employed the "trellis" graphics system in S-PLUS to produce a sequence of *conditional*

[4]While this which makes perfect sense intuitively, the statistical approach not only provides support for this "intuitive" conclusion, it also suggests that moisture content can be *entirely* controlled by factor E. The key difference is that "statistical" conclusion is based on objective methods.

contour plots. The result is shown in Figure 7.1. As might be expected, the surface appears to change very little from panel to panel which agrees with our previous conclusions from the numerical output. Details concerning this plot are discussed next.

Contour Plot Details

Factors E and G were chosen for generating the main grid of points for each "individual" contour plot. Two levels of A (590, 660) and three levels of B (1.7, 1.9, 2.1) were used to provide the sequence of conditional "Panel" plots. Factor H was fixed at the optimum value of 228 which is an achievable operating temperature and also allowed easy exporting of the predicted responses over this grid. The following SAS code assumes the dataset "a" we created previously, is available.

```
data b ;
     set a end=eof ;  * copy the dataset "a" into "b"  ;
     output ;         * when eof = end-of-file reached ;
                      * start the grid of points ;
     if eof then
        do ;
           y = . ;    * set the response Y as missing  ;
           h = 228 ;  * fix factor H at the value 228  ;

                      * begin the A, B Panel loop      ;

           do a = 590 to 660 by 70;    * 2 levels of A ;
           do b = 1.7 to 2.1 by 0.2 ;  * 3 levels of B ;

                      * begin the E, G main grid loop  ;

             do e = 14 to 26 by 1 ;
             do g = 0  to 3  by 0.5 ;

                output ;  * output the current value of ;
                          * Y,H,A,B,E for every G loop  ;

     end ; end ; end ; end ; end ;
*-------------------------------------------------------;
      * the nonmissing observations are needed to fit   ;
      * the RS model which is then used to predict the  ;
      * missing Y's over the grid. The predicted values ;
      * are saved to the dataset "b2"                   ;

proc rsreg data=b out=b2 noprint ;
      model y = a b e g h / predict ;
```

```
*-------------------------------------------------------;
data _LAST_ ;          * set up dataset "b2" for exporting ;
  set b2 ;             * to the file "f0.dat" (extension   ;
  if h = 228 ;         * added by SAS). Only output the    ;
  file f0 ;            * predicted values (when H=228)     ;
  put a b e g y ;      * hence no need to output factor H  ;
```

The file "f0.dat" can now be imported directly into S-PLUS and the following function was created to perform the trellis contour plots. Note that a graphics device needs to be specified before g.cont() can be run. Appendix B.2 provides further details on S-PLUS. (The reader may also find the discussion presented in Case Study 6.4, *Reverse Psychology,* helpful.)

```
g.cont <- function(file="f0.dat") {
   xx <- read.table(file,col.names=c("A","B","E","G","Y"),
                    row.names=NULL)
   attach(xx)
   ans <- contourplot(Y ~ E * G | A * B , data = xx,
               xlab = "Water Added (%)",
               ylab = "Dye Added (%)",
               main="Predicted Moisture Surface")
   detach()
   ans
}
```

To use trellis graphics, a data frame needs to be created and then "attached" so the column names can be used like variables in a model statement. In this example, we used the trellis graphics function contourplot() with a model statement that specifies E and G as the grid variables to use for the main contour plot for Y. This contour plot is conditional on the number of combinations associated with the variables (A and B) appearing after the "|" symbol. The completed display is then returned as the S-PLUS graphics object "ans" for plotting.

In Figure 7.1, the levels of factors A and B are indicated by a small line in the two shaded panels. Although the indicator line is somewhat hard to see, the three left-side plots all correspond to A at the low level (590); the right-side plots are all for $A = 660$. Reading the plots from top to bottom on each side, the B panel indicator line corresponds to the levels of B going from high (2.1), to middle (1.9), to low (1.7).

7.2.6 Summary

The next step is to analyze the other response variables in similar fashion and to see whether a suitable *overall* optimum exists. If not, then the statistical consultant can use these results to advise the client where the next experiment should be conducted. Although this is left as an exercise for

the interested reader, we should emphasize that the experience of the client will play a critical role in determining precisely where the next experiment takes place.

In ongoing research projects such as the one presented in this case study, the educational role of the consultant can be just as important as providing an efficient design. In this example, we were able to convince the client of the merits of adopting an *iterative* approach to the investigation. As a result, the client was able to conduct an evolving sequence of small-scale experiments using efficient specialized designs to achieve and refine their research objectives.

Response surface methodology is a very useful statistical technique, but it does demand considerable attention to details in the design and implementation of the experiments. In retrospect, the mistake we made in developing the initial fractional factorial screening design for the client was to refer to the levels of each factor as "high" and "low." The 20% of failures was clearly the result of our misleading the client into thinking these levels needed to be near the *extremes* of the range for each factor. Education works both ways ... next time we won't use terms like "high/low" or "± 1" until the factor range is well established by the client.

Questions

1. The default coding (7.2) employed by SAS rescales the whole design so that the axial points are at ± 1. This is referred to as an *inscribed* coding which the user can override with the option:

   ```
   model y = ... / nocode ;
   ```

   In this experiment, the actual levels of factor A differ slightly from the $\pm 2, \pm 1, 0$ coding expected in a five-factor (half rep.) central composite design. Create a new file containing the expected coding for the factors and rerun the analysis using the `"/nocode"` option.

   Does this make any difference in the results or conclusions?

2. Investigate the other response variables. Is there a suitable overall optimum within the region spanned by the factors in this experiment? If not, where should the next experiment be conducted? (The statement `ridge max ;` can be added after the model specification for this purpose.)

7.3 Expenditures in NY Municipalities

Methods:	Multiple Linear Regression
Data:	Property Tax

7.3.1 Regression Modeling

The objective in this case study is to provide an estimate of the cost impact on municipal expenditures resulting from the proposed construction of new housing projects in three New York State towns. Since many of the services provided by a municipality are funded largely through property taxes, it is clearly of interest to try to determine whether these projects will produce an increase in expenditures. We approach this problem by developing a suitable regression model for predicting *per capita* expenditures, based on a common set of demographic and income-related predictors. These variables are defined shortly, but it is worth noting that using per capita expenditures as our response variable allows us to make direct comparisons between different towns.

However, we still need to address the issue of predicting *future* expenditures since this is the basis on which the decision will be made whether to approve permits for the new housing project. For purposes of this case study, we confine our attention to the modeling process. This involves the following steps.

1. Data from all New York State municipalities were collected on the response and predictor variables for the year of 1992. The 1992 data are provided in `c73.dat` and are used to obtain a suitable regression model for predicting per capita expenditures.

2. Projected values for the predictor variables of the three towns in question were estimated for the years 2005 and 2025. The projections are provided in Table 7.6 and are employed in the regression model to produce the predicted future expenditures.

The challenge, of course, is getting past the first step.

Before looking at the data and variables in more detail, we emphasize that the years 2005 and 2025 should be regarded simply as labels of the new observations for which we plan to calculate our predicted values. It would be impossible to foresee all the changes and eventualities that will actually take place in a municipality over the next 30 years. We are not really trying to predict expenditures into the future. Rather, we are only interested in the specific changes that are likely to be a product of the new housing as of today. It is this information that was incorporated into the projected values of the predictors; anything else which was not directly affected by the new housing projects was ignored.

The Data

The data consist of a response and six demographic predictors. There are also three housekeeping variables: State, County, and ID number. The response is Expenditure measured in per capita in dollars. The predictors are

total Population, Density, Percent Intergovernmental, Wealth per capita, Income per capita, and Population Growth Rate.

TABLE 7.4. Variable Definitions

Variable	Definition
ID	Identity Number (for matching)
ST	State Code
CO	County Code
EXPEN	Expenditure per Person
WEALTH	Wealth per Person
POP	Population
PINTG	Percent Intergovernmental
DENS	Density
INCOME	Mean Income per Person
GROWR	Growth Rate

TABLE 7.5. Towns of Interest (1992 Data)

	Warwick	Monroe	Tuxedo
ST	36	36	36
CO	33	33	33
EXPEN	237	159	926
WEALTH	78908	55067	155034
POP	16225	9338	2328
PINTG	24.7	8.8	6.1
DENS	170	599	52
INCOME	19044	16726	30610
ID	8730	5420	8400
GROWR	30.3	30.0	2.5

The variable Density is defined as Population/Area and is strongly correlated with Population. This suggests that perhaps we should look at adding the Area variable to the dataset and dropping one of the other two. Wealth per capita only measures wealth related to real estate property values and is correlated with Income but they measure different things.

The variable Percent Intergovernmental represents the percentage of revenue that comes from state and federal grants or subsidies. This variable is correlated with Wealth and Income but it also reflects other factors such as public services provided by the municipalities to other communities (such as a large railroad station for commuters located at a small municipality).

The purpose of this study is to predict expenditures for the years 2005 and 2025 for the towns listed in Tables 7.5 and 7.6.

TABLE 7.6. Projected Data for Towns of Interest

Town	Year	POP	WEALTH	PINTG	DENS	INCOME	GROWR
Warwick	1992	16225	78908	24.7	170	19044	30.3
	2005	20442	85000	24.7	214	19500	35.0
	2025	31033	89000	26.0	325	20000	40.0
Monroe	1992	9338	55067	8.8	599	16726	30.0
	2005	10496	58000	8.8	695	17100	35.0
	2025	13913	60000	10.1	959	18000	35.0
Tuxedo	1992	2328	155034	6.1	52	30610	2.5
	2005	10685	116000	6.1	249	28300	300.0
	2025	29246	115000	7.0	656	25000	100.0

Methodology

7.3.2 Regression Analysis

A regression model can often provide a suitable approximation to the prediction problem. The general approach to fitting a regression was described in Chapter 3 (*Regression*). There, we discussed the important role that residual diagnostics play in assessing the quality and validity of the fitted model. In this case study, we also need to evaluate the quality of the predictions. We focus on these issues in the topics discussed below.

Transformations

A few scatter plots and Q–Q plots are sufficient to justify the need for transforming the variables. They will show long tails towards the large positive values for most variables including the response. These large values act as leverage points in the regression but because they are part of a long tail, if they are removed others will pop up, even if the process is repeated several times. The reader may find it useful to do this as an exercise.

From Figure 7.2 it can be seen that a log transformation appears to have adequately accounted for the strongly skewed Expenditure data. Although the transformed values still exhibit some right skewness, the interpretation of a log transformation is easier to justify to a client than say, a reciprocal square root.

The log transformation and other power transformations are not appropriate when the data contain zeros or a few negative numbers. One possibility in such cases is to add a constant and calculate $\log(x + c)$. For nonnegative numbers if we set $c = 1$ the transformation will map 0 into 0. Larger values of c may be used when there are a few negative values, and there may be cases when we subtract a number c because the data have no small numbers.

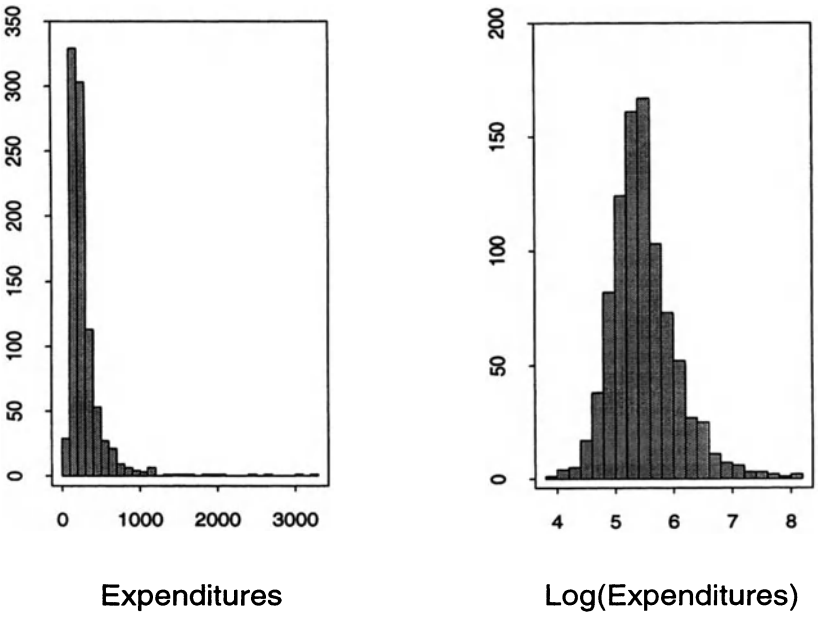


FIGURE 7.2. Original and Log Transformed Distributions

There are also cases such as the variable Growth Rate that exhibit very large tails on both positive and negative values. The popular power transformations are only defined for nonnegative values of x. In order to eliminate the problem we use a transformation $t(x)$ for $x > 0$ and another $-t(-x)$ for $x < 0$, as long as the joint transformation of $t(x)$ and $-t(-x)$ is monotonic. For example, take $\log(x+1)$ for $x > 0$ and $-\log(-x+1)$ for $x < 0$. This transformation keeps the zero at zero and it is monotonic.

Outliers

Outliers should be detected and investigated to determine if they are true mistakes or they are real valid values. For modeling purposes is useful to omit them or downweight them when they greatly influence the output model. A linear model fitted by least squares without omitting outliers would not satisfy the standard normality assumptions on the errors and the model output related to confidence intervals and test statistics on parameter values would be invalidated.

The regression procedure available in SAS has implemented methodology for detecting outliers called *influence diagnostics*, but it lacks any robust regression methodology. If one or more outliers are found, the procedure should be repeated omitting the outliers until no more outliers are found. Other software such as S-PLUS has included robust regression procedures

that can be applied in order to avoid outliers and influential observations. One practical rule is to perform the regression using both least squares and a robust regression method and compare the parameter estimates from both methods. If they are very similar then we proceed with the standard analysis but if they are very different we have to find the outliers that are responsible for the difference and then proceed.

Nonlinear Relationships

Sometimes we find nonlinear relationships that are not easy to fit using a low degree polynomial and require more complicated nonparametric regression techniques. In addition, this nonlinear methodology makes it difficult to calculate confidence intervals and P-values and the models often lack good interpretation. Again, a practical rule of thumb is to compare the R^2s obtained from fitting a low degree polynomial and a nonlinear nonparametric model such as that provided by the `gam()` function in S-PLUS.

When datasets are sufficiently large a simple way to get around the problem is to select a subset around the point where we want to make the prediction and to fit a linear or low degree polynomial model.

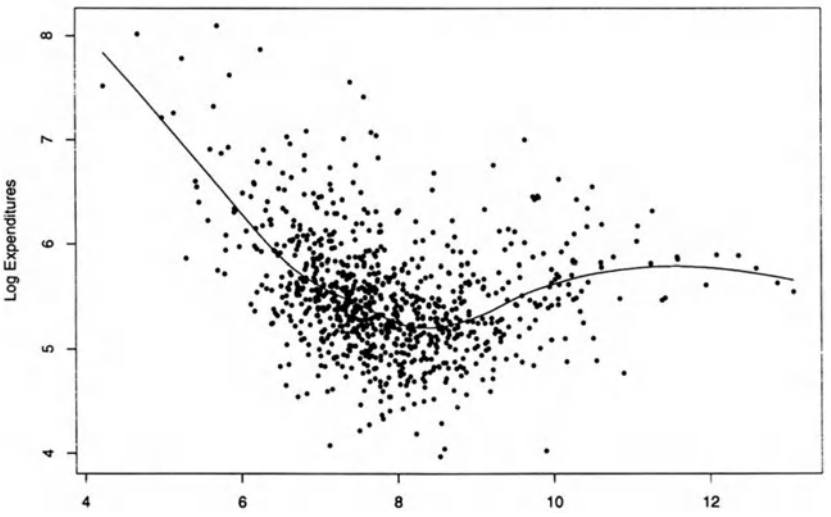


FIGURE 7.3. Log(Expenditures) versus Log(Population) with a Fitted Smooth Curve

The dataset under study contains examples of these nonlinear relationships. The graph of Log(Expenditure) versus Log(Population) shows a non-

linear relationship that can not be fitted with a polynomial of degree 2 or 3. Hence it is reasonable to take a subset of the data by restricting Population to an interval around the value where we are going to make our prediction.

Figure 7.3 shows that the relationship between Log (Expenditure) and Log(Population) starts with high Log(Expenditure) values that decline sharply until they reach a minimum around a population of 10,000 and then increase again at a more moderate rate. It would be interesting to know if this pattern holds for other states suggesting that there are ranges of population size that optimize town expenditures.

Prediction Intervals

We need to estimate the effect of the new housing projects on the per capita expenditures of these three towns. One peculiarity of this study is in the interpretation of the model's error term. If we compare year-to-year residuals from modeling per capita expenditures we observe that they are very correlated. Economists believe that this is because the error term mostly measures the peculiarities of that particular town that are not explained by the model. In addition, there is a random error component that measures year-to-year independent random fluctuations. These two terms are impossible to separate without historical data. To avoid this problem, one alternative option is to estimate the effects of new housing projects by calculating the difference of predicted values from the current values of the predictor to the values for 2005 and 2025, and describe the change rather the actual predictions.

A second issue is what to do about calculating predicted values since we applied a log transformation to the data. We use a result about the mean of the lognormal distribution that says $E(e^x) = e^{\mu+\sigma^2/2}$. In our case this means that to obtain expected values in the original scale we use the above formula replacing μ by $\widehat{Log(Expenditure)}$ and σ by the standard error of $\widehat{Log(Expenditure)}$, both easily obtainable from the regression output produced by SAS or S-PLUS.

The same type of calculation applies when we use the square root transformation. To calculate predicted values on the original scale we use the formula $E(x^2) = \mu^2 + \sigma^2$.

These calculations are not recommended with small sample sizes because we are using estimates of μ and σ that have error themselves. Suppose that we fit a model to the log response and that we overestimate the standard error of the predicted value. Then we square it and add this to the exponent resulting in very large overestimation of the predicted value for the untransformed response. The same will happen with confidence intervals and prediction intervals derived from small samples. When we exponentiate the intervals after a log transformation the results may be unreliable because of the error of the estimates.

7.3.3 Preliminary Analysis

Following the ideas outlined above we propose one possible analysis of the data.

- Choose transformations.
- Select a subset of the data containing the region where we are making the projections.
- Do a search for outliers and check for collinearity.
- Fit our best regression model.
- Calculate the predicted values and their standard errors for the points at which we are making the projections.
- Transform the projections to the original scale to produce the final estimates.

SAS Program

The following SAS code provides the implementation for the initial analysis of the data in the file `c73.dat`.

```
* In this data step we read the data and do the ;
* preliminary transformations ;
data a;
   infile 'c73.dat';
   input st co expen wealth pop pint
         dens income id grower ;

* These are the transformations that were chosen ;
*  after a few trials ;
   lex = log(expen);
   wealth = log(wealth);
   plop = log(pop);
   lens = log(dens);
   income = log(income);
   pint2 = pint**2;
   pint3 = pint**3;
   plop2 = plop**2;
   lens2 = lens**2;
   plop3 = plop**3;
   lens3 = lens**3;

* Example of transformation when both the positive ;
* and negative tails are very large ;
```

```
   if grower < 0 then lgrowr = - log(-grower);
   if grower > 0 then lgrowr =  log(grower);
   if _N_ ne 887;
   if _N_ ne 475;

* To check that the transformations work ;
proc univariate plot;
   var  st co expen wealth pop pint
        dens income id grower
        lex wealth plop lens income;

proc plot;
   plot  expen*(grower wealth pop dens income
                pint grower);
   plot lex*(lgrowr wealth plop lens income);

* Finally do the regression ;
proc reg;
   model lex = wealth plop plop2 plop3 pint
               pint2 pint3 lens lens2
               lens3 income lgrowr
               / method=stepwise ;
```

S-PLUS Notes

If you use S-PLUS here are some basic comments about modeling.

`lm`	basic function for linear models
`lmsreg`	function for least median squares robust regression
`gam`	function for generalized additive models.

Example

```
# Read the data
ny <- read.table("c73.dat")
# Run the linear regression model
lm.ny <- lm(lex~lpop+ldens+lwealth+lincome+
                    pintg+growr,data=ny)
# Run the LMS robust regression model
rr.ny <- lmsreg(lex~lpop+ldens+lwealth+lincome+
                    pintg+growr,data=ny)
# Run the nonlinear regression model
gam.ny <- lm(lex~s(lpop)+s(ldens)+s(lwealth)+lincome+
                    pintg+growr,data=ny)
# Check the residual sum of squares from the summary output
#     to determine if the nonlinear model did much better.
```

```
summary(lm.ny)
summary(rr.ny)
summary(gam.ny)
```

Objects such as `lm.ny` can be used to print the model statistics:

`summary(lm.ny)`	prints coefficients , t-values, R^2
`anova(lm.ny)`	prints the ANOVA table.

Models

The symbol `"~"` is like the equal sign. It separates the response from the predictors. You can not use `"="` because it has a different meaning in the S-PLUS language. You can add a term to the model formula such as

```
+ pop^3   or
+ pop^2 + pop^3
```

Look at the R^2 to decide if it improves very much.

More on Objects

The object `lm.ny` can also be used to get residuals, predicted values, or plots:

`resid(lm.ny)`	returns the residuals
`predict(lm.ny)`	returns the predicted values
`predict(lm.ny,new)`	returns predicted values at new points
`plot(lm.ny)`	plot of residuals vs. predicted values.

7.3.4 Summary

The relationship between expenditures and the predictor variables is clearly complicated. At best, the regression model we develop will provide a useful approximation for the purposes of the investigation. Figure 7.3 also shows that we may need to restrict attention to certain subsets of the data in order to apply the regression model. The S-PLUS function `smooth.spline()` was employed to add the smoothing spline to this scatter plot. More complex models involving splines could be investigated using the `gam()` function, but we leave that up to the reader.

Questions

The main question of interest here is to calculate projections for per capita expenditures on one of the given municipalities. Following the ideas outlined above, one possible analysis of the data is the following.

1. Choose transformations.
2. Select a subset of the data containing the region where we are making the projections.
3. Do a search for outliers and check for collinearity.
4. Fit our best regression model.
5. Calculate the predicted values and their standard errors for the points at which we are making the projections.
6. Transform the projections to the original scale to produce the final estimates.

7.4 Measuring Quality Time

Methods:	Time Series Analysis ARIMA Models
Data:	4 Years of Monthly Data 2 Time Series

7.4.1 Time Series Analysis

The quality of a product and service provided by a business is very important and many companies have a full-time Quality Assurance Manager; larger companies often have a team of "QA" personnel. The development of quality control (QC) was briefly discussed in Chapter 1 and while QC methods form the statistical basis of quality assurance schemes, the latter terminology emphasizes the importance of combining management protocols with statistical methods.

Background

In this case study, the company manufactures a range of products and quality control methods are employed throughout the manufacturing process to assess quality of the inputs. This is often referred to as statistical process control (SPC) and one of the aims of SPC is to enable the company to pursue the goal of "continuous improvement" of its products.

A Quality Index

The company also developed their own index of quality (IQ) such that IQ = 100 represented the highest possible value. The IQ score provided the company with the ability to combine individual quality scores and to redefine what constituted maximum quality without changing the IQ scale. The IQ score could therefore be used to forecast realistic improvement goals for products that were subject to seasonal variation in sales.

The Data

The objective of this analysis is to obtain a good time series model for modeling the IQ score. Monthly data for the last four years of the IQ score for a certain product are given in Table 7.7 below. The `BATCH` count corresponds to the amount of this particular product manufactured by the company. This dataset is also available in the file `c74.dat`.

Methodology

Let y_t denote the response value of interest at time t. Here, the actual unit of time is months and our interest is in forecasting y_t = IQ. In a regression situation, we could let X denote the time index, Y the response values, and then use the fitted regression model to obtain $\hat{y}_{t+1}$. However, a regression model assumes that the error components of the model are independent and this is not a reasonable assumption for a process that evolves over time. Thus we introduce the ARIMA class of models.

7.4.2 ARIMA Models

Box and Jenkins (1970) popularized the class of autoregressive integrated moving average (ARIMA) models for fitting time series. Their approach is commonly referred to as the Box–Jenkins method and involves three steps:

Identification Differencing is applied to make the series y_t stationary

Estimation An ARMA(p, q) model is fitted to the differenced series

Forecasting The ARIMA model is used to forecast future values of y_t.

Identification

Before the serial correlation can be modeled, deterministic features of a time series such as seasonality and trends need to be accounted for. In practice, these features can often be removed by applying the difference operators:

$$\nabla y_t = y_t - y_{t-1} \qquad \nabla_s y_t = y_t - y_{t-s}$$

TABLE 7.7. IQ Data

Month	Batch Count	IQ Score	Month	Batch Count	IQ Score
Jan94	2339	86.63	Jan96	2971	89.25
Feb94	2275	84.60	Feb96	3083	90.54
Mar94	2881	87.04	Mar96	3504	89.89
Apr94	2780	87.19	Apr96	3580	90.28
May94	3227	87.91	May96	3855	89.46
Jun94	3291	87.99	Jun96	3894	89.42
Jul94	2944	88.09	Jul96	3772	89.28
Aug94	3163	88.25	Aug96	3705	89.17
Sep94	2770	87.62	Sep96	3364	90.42
Oct94	2827	87.43	Oct96	3341	90.46
Nov94	2392	86.74	Nov96	2680	88.63
Dec94	1973	84.86	Dec96	2418	89.74
Jan95	3006	87.44	Jan97	2963	90.48
Feb95	2924	87.77	Feb97	2890	89.76
Mar95	3592	88.09	Mar97	3455	90.20
Apr95	3460	88.53	Apr97	3747	90.68
May95	3807	88.11	May97	3685	90.19
Jun95	3753	88.59	Jun97	3672	89.78
Jul95	3648	88.67	Jul97	3865	89.72
Aug95	3698	88.87	Aug97	3729	90.78
Sep95	3166	89.92	Sep97	3205	90.32
Oct95	3159	88.93	Oct97	3158	90.64
Nov95	2545	87.17	Nov97	2552	90.97
Dec95	2208	89.07	Dec97	2135	90.23

to y_t until the resulting series, w_t say, appears to be stationary. That is, the mean and variance of w_t are approximately constant. A simple difference should be applied first and no more than two simple differences ($\nabla^2 y_t = \nabla y_t - \nabla y_{t-1}$) should be applied to a time series.[5] If needed, a seasonal difference can also be applied. For example, $s = 12$ would be used to remove annual seasonality in a series observed monthly.

Estimation

Given w_t, the serial correlation may now modeled by fitting an ARMA(p, q) model of the form:

$$w_t + \sum_{j=1}^{p} \alpha_j w_{t-j} = \epsilon_t + \sum_{j=1}^{q} \beta_j \epsilon_{t-j} \ , \tag{7.3}$$

where the ϵ_t denote the unknown random components or innovations, which are usually assumed to be independent and identically distributed.

The orders p, q of the ARMA model first need to be determined and this can be achieved by examining the autocorrelation (ACF) and partial autocorrelation (PACF) correlograms. This would be combined with the use of an information criterion such as AIC or BIC which enables the most parsimonious model to be selected. After the orders p, q are determined, estimation of the ARMA parameters can be performed.

ACF This is a plot of $\rho_k = Corr(y_{t-k}, y_k)$ versus k. For an AR(p) process, the correlation between y_t and y_{t-k} decreases in exponential fashion as the time "lag" k increases. In contrast, $\rho_k = 0$ for all $k > q$ in a pure MA(q) process.

PACF The *partial* autocorrelation between y_t and y_{t-k} is the correlation **after** the effect of $y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$ have been "regressed" out. This is denoted by ϕ_k and for an AR(p) process, the "link" between y_t and y_{t-k} no longer exists ($\phi_k = 0$) when $k > p$. In contrast, an MA(q) can be represented as an AR(∞) process so the ϕ_k decrease exponentially. The PACF is a plot of ϕ_k versus k.

AIC/BIC The combination of ACF and PACF correlograms can be quite effective for determining the order of pure AR or MA processes. They can also be employed as a diagnostic for the residuals from a fitted ARMA model. AIC was proposed by Akaike (1974) and provides a useful criterion for comparing different time series models. BIC is another information criterion (also called SBC), and both criteria

[5] Overdifferencing can introduce so-called unit roots that will make the fitted ARMA(p, q) model unstable. This is similar to the problem of multicollinearity in regression.

are usually provided by statistical software with time series modules. They are defined as

$$\text{AIC}(k) = \log \hat{\sigma}_k^2 + \frac{2k}{n} \qquad \text{BIC}(k) = \log \hat{\sigma}_k^2 + \frac{k \log n}{n} ,$$

where $\hat{\sigma}_k^2$ denotes the residual variance associated with fitting a model of order k. The second term is the penalty for increasing k, so the minimizing value of these criteria provides an estimate of the most parsimonious model.

Forecasting

For forecasting, the "integrated" model may be reconstructed from the fitted ARMA model and differences applied in the Identification step, to provide a linear model in terms of the original time series y_t. For example, the ARIMA(1,1,1) model has the form:

$$y_t - y_{t-1} + \alpha_1(y_t - y_{t-1}) = \epsilon_t + \beta_1 \epsilon_{t-1} .$$

Seasonal ARIMA Models

The ARIMA class of models can also include seasonal effects by fitting an ARMA(P, Q) model to observations separated by the seasonal period. That is, j in (7.3) is replaced by sj and w_t denotes the resulting series from a seasonal difference (if applied). The multiplicative seasonal ARIMA model is often denoted as $(P, D, Q)_s \times (p, d, q)$ where

D = number of seasonal differences
P, Q = seasonal ARMA model orders
s = seasonal lag

d = number of simple differences
p, q = deseasonalized ARMA model orders.

Stationarity

ARIMA models can often provide a useful starting point and may be sufficiently accurate for the purposes of the project at hand. However, these models require that the original series can be reduced to a stationary series (w_t above) wherein the mean and variance are constant. In general, mean stationarity can be achieved by differencing, but to obtain constant variance, a nonlinear transformation may be required.

7.4.3 Preliminary Analysis

Figure 7.4 exhibits seasonality as well as an upward trend — which is good! These structural features need to be removed and the following SAS code

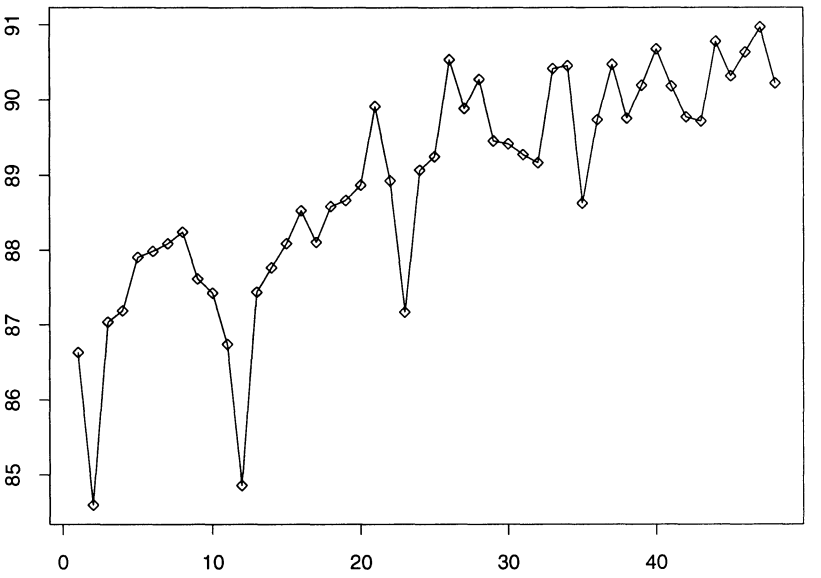


FIGURE 7.4. Plot of Monthly IQ Score

presents some generic statements for ARIMA model specifications in PROC ARIMA. We also illustrate the use of the date informant " monyy5." which enables SAS to correctly interpret the character entries for the mon (such as "jan94").

```
filename in 'c74.dat' ;
data a ;                          * "monyy5." is similar to "$".    ;
  infile in ;                     * It tells SAS how to read the    ;
  input mon monyy5. count iq ;    * date variable "mon" (mon = xxx ;
  y = iq ;                        * and yy = 2 digit year)          ;
  format mon monyy5. ;   * use for printing purposes. otherwise   ;
                         * "jan94" is printed as a Julian date!   ;
proc arima ;
  * ------------------------------------------------------------ ;
  * Identifying an arima model                                    ;

  i var=y ;          * acf/pacf for time series y(t)              ;
                     * to detrend use (one of) the following      ;
                     *      instead                               ;
  i var=y(1) ;       * simple difference w(t) = y(t) - y(t-1)     ;
  i var=y(12) ;      * seasonal difference w(t) = y(t) - y(t-12)  ;
```

```
i var=y(1,12) ;  * simple difference giving w(t), then          ;
                 *   seasonal difference giving w2(t):          ;
                 *       w2(t) = w(t) - w(t-12)                 ;
* ------------------------------------------------------------ ;
* Fitting an arima model                                        ;

e p=(1) ;          * fit ar(1) to w(t)                          ;
e p=(1) plot ;     * fit ar(1) and plot the acf/pacf of         ;
                   *      residuals                             ;
e p=(1) q=(1) ;    * fit arma(1,1)                              ;
e p=(1)(12) ;      * fit multiplicative ARIMA:                  ;
                   *      ar(1) x sar(1) (season lag=12)        ;
* ------------------------------------------------------------ ;
* Forecasting from the fitted arima model                       ;

f out=r lead=1 back=0 id=mon ;* only forecast 1 step ahead.    ;
                             *    output to data=r             ;
f out=r lead=12 id=mon ;     * forecast 12 steps ahead.       ;
                             *    include forecasts within     ;
                             *    series. output to data=r     ;
* ------------------------------------------------------------ ;
```

The identification, estimation (fitting), and forecast steps are normally applied sequentially. That is, the `proc arima` procedure needs to be run several times with the first task being to achieve stationarity through differencing or other means. We have kindly left this as an exercise for the reader and instead, briefly indicate the S-PLUS approach.

S-PLUS Code

In S-PLUS, the parameters of an ARIMA model are specified as a list. In the code below, the first list specifies the parameters of the "within-season" ARIMA model in the order `list=c(p,d,q)` with $p = 1, d = 1$, and $q = 0$. That is, an AR(1) model will be fitted to the ∇y_t. The second list specifies the parameters of the "seasonal" ARIMA model in the order `list=c(P,D,Q)` with $P = 1$ and the seasonal period specified by the `"period=12"` option. In this case, the multiplicative model

$$\nabla y_t - \alpha \nabla y_t - \Phi \nabla y_{t-12} + \alpha \Phi \nabla y_{t-13} = \epsilon_t$$

will be fit by the `arima.mle()` function. The fitted model can be assessed by examining the residual ACF and PACF plots which are provided by the diagnostic function, `arima.diag()`.

```
z <- read.table("c74.dat")
y <- z$iq
m1 <- list(list(order=c(1,1,0),
           list(order=c(1,0,0),period=12)))
fit <- arima.mle(y, model=m1)
```

```
arima.diag(fit)
```

7.4.4 Summary

Quality control methods and management protocols have proved to be very effective tools for manufacturing companies. The influence of Deming, for example, provided the Japanese automotive industry with the opportunity to gain a substantial portion of the U.S. market. In this example, we have seen that continuous improvement also needs to reflect the reality of a manufacturing process. That is, the effect of seasonal variations, while they exist, should be incorporated in QA schemes. From Figure 7.4, it would appear that the seasonal variation in quality no longer existed over the final year, even though the amount of product still has a strong seasonal trend. Perhaps just having the right tools to look at a problem can make it go away!

Questions

1. From the initial time series plot it would appear that a transformation is required to stabilize the variance. What sort of transformation would you apply?

2. Fit an ARIMA model to the original and transformed data. What are your conclusions?

8
Case Studies from Group III

The four case studies presented in this chapter are research-oriented and require multivariate methods or specialized statistical methods for analysis. Several stages of analysis may also be needed to obtain suitable results. There may not necessarily be an "answer" to the statistical problem. The case studies are listed below by section.

8.1 A Tale of Two Thieves

Methods:	Analysis of Variance Random Effects Mixed Models
Data:	Two Experiments Response: Assay of Active Ingredient 2 Sampling Instruments Postmixing Quality Analysis

8.1.1 Analysis of Variance with Mixed Effects

Prescription and over-the-counter drugs contain a mixture of both active and inactive ingredients, with the dosage determined by the amount of active ingredient in each tablet. Making sure the tablets contain the correct dosage is an important problem in the drug manufacturing industry and in this case study, we consider an experiment conducted by a pharmaceutical company to investigate sampling variability and bias associated with the manufacture of a certain type of tablet.

Outline of the Problem

Tablet Manufacture The tablets were manufactured by mixing the active and inactive ingredients in a "V-blender," so-named because it looks like a large **V**. (See Figure 8.1.) Mixing was achieved by rotating the V-blender in the vertical direction.

After the mixture was thoroughly blended, the powder was discharged from the bottom of the V-blender and compressed into tablet form.

Uniform Content The most important requirement of this manufacturing process was that the tablets have uniform content. That is, the correct amount of active ingredient must be present in each tablet.

The content uniformity of the mixture within the V-blender will need to be assessed.

Thief Sampling A "thief" instrument was used to obtain samples from different locations within the V-blender. This was essentially a long pole with a closed scoop at one end, which was plunged into the powder mixture by a mechanical device. At the appropriate depth for a given location, the scoop was opened and a sample collected.

Considerable force was needed to insert a thief into the powder mixture and it was of interest to compare two types of thieves.

- The **Unit Dose** thief collects three individual unit dose samples at each location.

- The **Intermediate Dose** collects one large sample which is itself sampled to give three unit dose samples.

Experiment Procedure

The objective of this experiment was to study bias and variability differences between the two thieves and to compare the thief-sampled results with those of the tablets. The experiment was implemented as follows.

1. Blend the mixture in the V-blender for 20 minutes.
2. Tie the thieves together and use them to obtain samples from six locations within the V-blender. A schematic of the V-blender and sampling locations is shown in Figure 8.1.
3. Discharge the powder from the V-blender and compress it to form tablets. Load tablets into 30 drums.
4. Select 10 drums and sample three tablets from each of these drums.
5. Assay all samples to determine the amount of active ingredient in each sample. The specified assay value is: 35 mg/100 mg.

In this example, applying the analysis of variance technique to the two sets of assay results (Thief and Tablet) from this experiment is more complicated since *random* and *fixed effects* are involved. When both types of factor are present, this is called a *mixed model* analysis of variance.

The Data

The datasets `c81.thief.dat` and `c81.tablet.dat` contain the results from this experiment which we have presented in a slightly more compact format in Tables 8.1 and 8.2, respectively. The variables and factors of interest for analysis purposes are summarized in the following table.

Variable	Levels	Definition
Y	(mg)	Assay Value per 100 mg
METHOD	INTM	Intermediate Dose Thief
	UNIT	Unit Dose Thief
LOC	1,2,...,6	Sampling Location
REP	1,2,3	Replicate per Location
DRUM	mod(n, 3)	Randomly Selected Drum
TABLET	1,2,3	Tablet Sample (per Drum)

The locations shown in Figure 8.1 represented the "desired" sampling positions for the thieves. In the actual experiment, these "fixed" positions were subject to a certain amount of variability. The samples collected by the thieves can be regarded as random within each location.

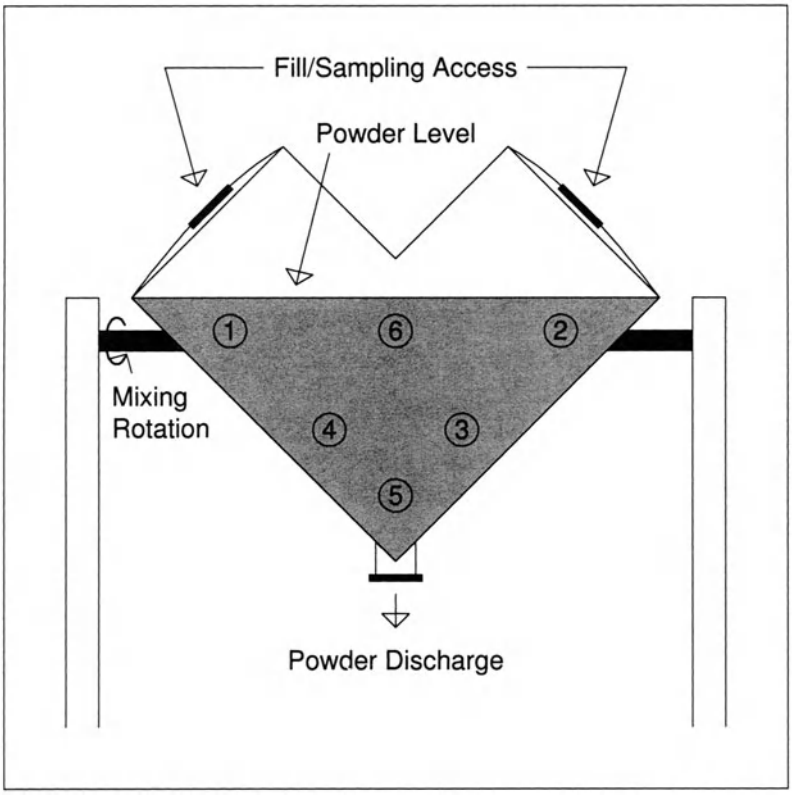


FIGURE 8.1. V-Blender Schematic and Sampling Locations

In the Tablet experiment, the order in which the drums were filled was recorded and this information was incorporated into the random selection procedure. Specifically, one drum was randomly selected from each triple sequence: $\{1,2,3\}$ $\{4,5,6\}$ ... $\{28,29,30\}$. The factor DRUM could therefore be used to test for a "time" effect in the Tablet data.

Methodology

To help motivate the discussion, recall the balanced one-way ANOVA model introduced in Chapter 3 (*Standard Methods*):

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases}, \tag{8.1}$$

where the τ_i represent the levels of a factor, TRT say, and the ϵ_{ij} are assumed to be independent $N(0,\sigma^2)$ innovations. The key difference between a fixed effects model and a *random* effects model lies in the interpretation of the inference associated with these models.

TABLE 8.1. Thief Data

Location	Replicate	Assay Method	
		Intermediate	Unit Thief
1	1	34.38	33.94
	2	34.87	34.72
	3	35.71	34.10
2	1	35.31	39.11
	2	37.59	37.51
	3	38.02	37.79
3	1	36.71	37.46
	2	36.56	34.12
	3	35.92	35.94
4	1	37.80	38.05
	2	37.41	34.82
	3	38.00	35.42
5	1	36.28	36.52
	2	36.63	38.60
	3	36.62	38.16
6	1	38.89	39.16
	2	39.80	32.77
	3	37.84	36.95

Fixed Effect Inference is conditional on the "fixed" levels of the TRT factor. In this example, the factor METHOD is clearly fixed since any inference associated with the effect of METHOD applies specifically to the two types of thief involved.

Random Effect Inference is made about a "population" of TRT levels. The factor TRT is considered *random* if the τ_i were randomly selected from the TRT levels. Hence, τ_i is assumed to be a $N(0, \sigma_\tau^2)$ random variable (independent of ϵ_{ij}) for testing purposes. The model (8.1) is also called the *components of variance* model.

Since the τ_i are randomly selected, testing individual effects is meaningless and the hypotheses of interest are: $H_0 : \sigma_\tau^2 = 0$ versus $H_1 : \sigma_\tau^2 > 0$. Under H_0, the test statistic

$$F_o = \frac{\text{SS(TRT)}/(a-1)}{\text{SSE}/(N-a)} = \frac{\text{MS(TRT)}}{\text{MSE}} \sim \mathcal{F}_{a-1,N-a}$$

is constructed in exactly the same manner as in the fixed effects case. However, the expected mean squares associated with the random effects model are different and are needed to construct estimators of the variance components. For the model (8.1), it can be shown that:

$$E[\text{MS(TRT)}] = \sigma^2 + n\sigma_\tau^2 \qquad E[\text{MSE}] = \sigma^2 \ .$$

TABLE 8.2. Tablet Data

Drum	Tablet	Assay	Drum	Tablet	Assay
	1	35.77		1	35.43
1	2	39.44	17	2	33.80
	3	36.43		3	35.15
	1	35.71		1	34.56
5	2	37.08	19	2	35.33
	3	36.54		3	37.69
	1	35.08		1	35.82
7	2	34.25	22	2	35.67
	3	33.09		3	35.06
	1	35.21		1	35.75
11	2	34.36	25	2	37.32
	3	35.94		3	35.06
	1	35.17		1	38.58
14	2	36.54	28	2	36.63
	3	36.45		3	35.60

Equating the observed and expected mean squares provides the following estimates of the variance components

$$\hat{\sigma}_\tau^2 = \big(\text{MS}(\texttt{TRT}) - \text{MSE}\big)/n \qquad \hat{\sigma}^2 = \text{MSE}\ .$$

8.1.2 Mixed Model Analysis

The two-way ANOVA model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \end{cases} \tag{8.2}$$

is said to be a *mixed model* analysis of variance when one of the factors is fixed and the other is random. Here, we consider the situation where factor A is fixed, factor B is random, and $n > 1$ so that the interaction term can be estimated. The case where **both** A and B are fixed, or both are random, was discussed in Chapter 3.

Statistical Analysis

The interaction term AB is always assumed to be a random effect, but different assumptions can be imposed on the random components. This leads to different versions of the mixed model:

A Fixed effect: $\sum_{i=1}^{a} \alpha_i = 0$.

B Random effect: $\beta_j \sim N(0, \sigma_\beta^2)$.

AB Random effect:

standard model: $\gamma_{ij} \sim N\big(0, (1-(1/a))\sigma_\gamma^2\big)$ and $\sum_{i=1}^{a} \gamma_{ij} = 0$.

alternative model: γ_{ij} are uncorrelated with common variance σ_γ^2 .

The main difference is that in the standard model (which is more commonly used), interaction effects are **not** uncorrelated at different levels of the fixed factor: $Cov(\gamma_{ij}, \gamma_{i'j}) = -(1/a)\sigma_\gamma^2$ for $i \neq i'$. Under the assumptions above, the expected mean squares provide the appropriate test statistics and variance component estimates as follows.

$H_0 : \alpha_i = 0$ versus $H_1 :$ *some* $\alpha_i \neq 0$

$$F_o = \text{MS(A)}/\text{MS(AB)} \sim \mathcal{F}_{(a-1),(a-1)(b-1)}$$
$$E[\text{MS(A)}] = \sigma^2 + n\sigma_\gamma^2 + \tfrac{bn}{n-1}\textstyle\sum \alpha_i^2$$
$$\hat{\sigma}^2 = \text{MSE}$$

$H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$

standard model:

$$F_o = \text{MS(B)}/\text{MSE} \sim \mathcal{F}_{(b-1),ab(n-1)}$$
$$E[\text{MS(B)}] = \sigma^2 + an\sigma_\beta^2$$
$$\hat{\sigma}_\beta^2 = \big(\text{MS(B)} - \text{MSE}\big)/an$$

alternative model:

$$F_o = \text{MS(B)}/\text{MS(AB)} \sim \mathcal{F}_{(b-1),(a-1)(b-1)}$$
$$E[\text{MS(B)}] = \sigma^2 + n\sigma_\gamma^2 + an\sigma_\beta^2$$
$$\hat{\sigma}_\beta^2 = \big(\text{MS(B)} - \text{MS(AB)}\big)/an$$

$H_0 : \sigma_\gamma^2 = 0$ versus $H_1 : \sigma_\gamma^2 > 0$

$$F_o = \text{MS(AB)}/\text{MSE} \sim \mathcal{F}_{(b-1),ab(n-1)}$$
$$E[\text{MS(AB)}] = \sigma^2 + n\sigma_\gamma^2$$
$$\hat{\sigma}_\gamma^2 = \big(\text{MS(AB)} - \text{MSE}\big)/n \ .$$

8.1.3 Preliminary Analysis

The objective of this experiment was to study bias and variability differences between the two thieves and to compare the thief-sampled results with those of the tablets. The three main components of our preliminary analysis are as follows.

- Preliminary data screening
- Comparing the two thieves
- Thief versus Tablet samples.

Data Screening

The preliminary data screening involves checking for outliers and assessing the distributional properties of the assay values. This is important since our intention is to apply the analysis of variance technique which is sensitive to the presence of outliers and departures from normality. Of course, with a sample size of only 18 per thief, we will need to exercise a certain amount of caution in interpreting distributional features.

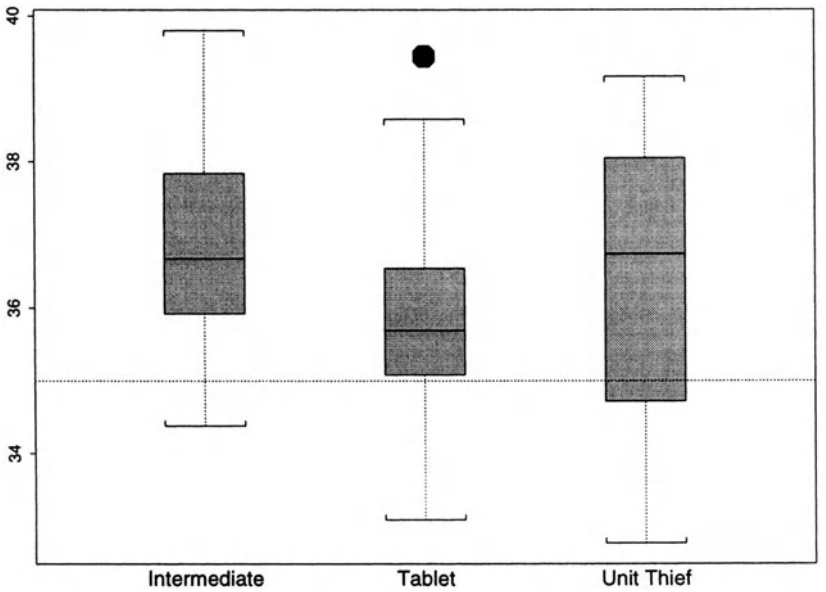


FIGURE 8.2. Parallel Boxplots of the Assay Values

From the parallel boxplot display in Figure 8.2, it can be readily seen that the methods were all positively biased relative to the specified assay value of 35 mg. However, there are some differences in the variability and distribution of the assay values among these methods. In particular, the "average" assay value of the tablet sample is noticeably lower. Where did the active ingredient go? Summary statistics for the three methods are presented in the table below.

	Thief		Tablet
Method	UNIT	INTM	DRUM
N	18	18	30
Mean	36.40	36.91	35.79
Std Dev	1.98	1.40	1.36
Max	39.16	39.80	39.44
Q_3	38.08	37.88	36.54
Median	36.74	36.67	35.69
Q_1	34.57	35.87	35.08
Min	32.77	34.38	33.09

Although the Shapiro–Wilk test and Q–Q plots with Lilliefor's confidence bounds (Conover 1980) did not indicate any significant departure from normality for any of the methods, we should regard these results as inconclusive due to the small samples involved. Similarly, a histogram of the Tablet data shows that the "outlier" indicated in Figure 8.2 is consistent with a heavy right-tail trend. This observation is also less than $Q_3 + 2 * \text{IQR}$ which is commonly used as a more conservative (upper) criterion for determining potential outliers.

Comparing Thieves

Since the preliminary analysis results did not indicate any obvious violation of the assumptions underlying the analysis of variance procedure, we proceed with using the mixed model analysis of variance model to compare the two thieves. Several SAS procedures can be used to perform a mixed analysis which we have presented in the program code below.

SAS Program Code

```
*-------------------------------------------------------------;
  filename in1 'c81.thief.dat' ;  * input the two datasets  ;
  filename in2 'c81.tablet.dat' ;
  data a ;
    infile in1 ;
    input method $ loc rep ya ;

  data b ;
    infile in2 ;
    input method $ drum tablet yb ;
*-------------------------------------------------------------;
                                  * repeated measures ANOVA  ;
proc glm data=a ;                 * with "loc" as random     ;
  class method loc rep ;
  model ya = method loc(method) rep rep*method ;
  random loc(method) / test  ;
```

```
  test h=methda e=loc(method) ;
*-------------------------------------------------------------;
                                  * get variance components  ;
proc varcomp data=a method=ml ;   * (subsumed by proc mixed) ;
  class method loc rep ;
  model ya = method rep rep*method loc(method) / fixed=3 ;
*-------------------------------------------------------------;
                                  * set up the data for the  ;
proc sort data=a ;                * multivariate version of  ;
     by method loc ;              * repeated measures ANOVA  ;

proc transpose out=a2(rename=(_1=y1 _2=y2 _3=y3)) ;
     by method loc ;
     id rep ;

proc glm data=a2 ;
  class methda loc ;
  model y1 y2 y3 = method ;
  repeated rep / summary ;
*-------------------------------------------------------------;
* Output from the "mixed" procedure is presented below.      ;
* ...........................                                ;
                                 * use the "mixed" procedure ;
proc mixed method=ml data=a ;    * which allows more general ;
  class method loc ;             * covariance structures to  ;
  model ya = method ;            * be specified. (SAS 6.08+) ;
  random loc ;
  repeated / subject=loc group=method ;
*-------------------------------------------------------------;
                                 * use the "mixed" procedure ;
proc mixed method=ml data=b ;    * to test for a time effect ;
  class drum ;                   * with respect to DRUM.     ;
  model yb = ;                   * ar(1) = autoregressive    ;
  random drum / s type=ar(1) ;   * covariance structure      ;
```

The `proc mixed` procedure is available in the SAS release version 6.08 and subsumes the `proc varcomp` procedure. It generalizes the `proc glm` procedure and allows a wider class of mixed models to be fitted. It also provides a more flexible "repeated" measures statement than the univariate or multivariate versions employed in `proc glm`.

In this example, we have declared `LOC` to be a random effect and treated the three `REP` levels as repeated measures. The output from `proc mixed` is presented below and shows no significant `METHOD` effect. This implies that either thief could be employed for sampling during the mixing process and for practical purposes, the Intermediate Dose Thief was found to be easier

to handle. However, there were significant differences among sampling locations which suggests the powder mixture may not have uniform content.

SAS Output from Proc Mixed

The Mixed Procedure
Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
loc			0.9977
Residual	loc	method intm	0.7068
Residual	loc	method unit	3.4001

Fit Statistics

Log Likelihood	-64.4
Akaike's Information Criterion	-67.4
Schwarz's Bayesian Criterion	-67.1
-2 Log Likelihood	128.8

Solution for Random Effects

Effect	loc	Estimate	Std Err Pred	DF	t Value	Pr > |t|
loc	1	-1.6389	0.5498	29	-2.98	0.0058
loc	2	0.2959	0.5498	29	0.54	0.5946
loc	3	-0.4341	0.5498	29	-0.79	0.4362
loc	4	0.5308	0.5498	29	0.97	0.3422
loc	5	-0.0792	0.5498	29	-0.14	0.8864
loc	6	1.3255	0.5498	29	2.41	0.0225

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
method	1	29	1.14	0.2934

To examine the issue of bias, we can use the *bootstrap* procedure (Efron and Tibshirani 1993). The idea of the bootstrap is to resample the observed data *with replacement* and use the resampled estimates, $\hat{\theta}_i$ say, to assess the variability of $\hat{\theta}$ about some unknown true θ. Hence the *bias* of $\hat{\theta}$ may be estimated by the taking the mean of the differences: $\hat{\theta}_i - \hat{\theta}$.

Bootstrap Estimates of P-Values

	UNIT vs. INTM	UNIT vs. DRUM	INTM vs. DRUM
Observed P-Value	0.3770	0.2135	0.0090
Bootstrap Mean	0.3932	0.3204	0.0587
Bias	0.0162	0.1069	0.0498
Standard Error	0.2958	0.3092	0.1244

In this example, we have employed the bootstrap to assess the reliability of the t-test for comparing the individual differences among the three methods. Note that for simultaneous comparisons, the Bonferroni inequality provides a conservative estimate of the critical level to use for each test: $0.0167 = 0.05/3$. The results presented in the table above are based on 1000 resamples and suggest that the significant difference originally observed between the Intermediate Dose Thief and Tablet assay values may not necessarily be valid.

8.1.4 Summary

Practical considerations and time constraints often limit the size and extent to which diagnostic experiments can be performed in the manufacturing sector. Careful planning and efficient sampling procedures are needed to ensure the experiment will satisfy the objectives of the study. We should point out that having *more* than the specified level of active ingredient in the tablet would not present a problem for the consumer. From the drug manufacturer's perspective, however, achieving the specified value is clearly desirable. The results from this experiment can be analyzed further and in the questions below, we have indicated some of the areas and approaches that the interested reader may wish to pursue.

Questions

1. Are the assay values generally well behaved? Note that when we treat the Thief samples as repeated measures, the issue of correlation needs to be incorporated in the criteria for determining outliers. Find out what procedures or tests are available and apply them to these data.

2. Is there any evidence of a location effect? Our preliminary analysis suggests there is, which would be of concern to the production management. What recommendations should we make in our report to management regarding this issue?

3. Do the tablet data show any drum or time effect? We employed an autoregressive AR(1) covariance structure in the last `proc mixed` procedure for this purpose. Is this the same as using the Durbin–Watson test?

4. Are the thief-sampled values comparable to the tablet values? One approach would be to consider the concordance correlation coefficient proposed by Lin (1989): quantify the agreement between two readings from the same sample by measuring the variation from the 45° line through the origin.

8.2 Plastic Explosives Detection

Source:	Authors
Methods:	Discriminant Analysis
Data:	21 Variables Last Variable is 0/1 Response

8.2.1 Pattern Recognition

The importance of plastic explosives detection was made clear by the tragedy of Pan Am Flight 103 which exploded over Lockerbie, Scotland on December 21, 1988. One of the most effective devices for detecting plastic explosives is a type of X-ray scanner that produces a profile of the chemical composition of a small area inside the suitcase. If the profile shows a pattern similar to one of the known plastic explosives then the suitcase is classified as a bomb.

The emphasis in this case study is on the reliability of plastic explosive detection based on an early X-ray machine prototype. That is, how well can plastic explosive be detected? What type of classification method should be used? Before continuing, we must first point out that:

> The analysis presented here does NOT in any way have implications concerning the reliability of the actual inspection machines used at security points such as in airports.

The Data

In order to obtain a classification rule we performed an experiment where 2500 profiles were obtained, of which 1250 corresponded to explosive substances and the remaining 1250 were from typical substances found in suitcases. Each profile is a vector of 20 numbers $x_1, \ldots, x_{20}$ which are a summary of the signal absorbed by the material. The response y is 1 if the suitcase has a bomb and 0 otherwise. The profiles and response variables are provided in the dataset `c82.dat`.

Methodology

There is a rich industry of methodology for classification. We give here a brief summary and we describe our experience with classification methods.

8.2.2 A Quick Review of Discriminant Analysis

Suppose we have samples from k multivariate populations or groups and we observe a new x for which the corresponding population or group is unknown. We would like to define a classification rule that assigns x to one of the groups. In some cases we may know or may be able to assign prior probabilities π_is that x belongs to group i. A Bayesian approach is favored in such cases.

Fisher's Linear Discriminant Functions

Suppose we have samples from k populations and we calculate the sample means and covariance matrices $\overline{x}_1, \ldots, \overline{x}_k$ and $S_1, \ldots, S_k$. Suppose that the population covariance matrices are equal. Then we can estimate the common population covariance matrix with $S_p = 1/(N-k)\sum_{i=1}^{k}(n_i-1)S_i$, where $N = \sum_{i=1}^{k} n_i$ and n_i is the size of the sample obtained from the ith population. The linear discriminant functions are of the form

$$L_i(x) = \overline{x}_i{}' S_p^{-1} x - \overline{x}_i{}' S_p^{-1} \overline{x}_i = a_i' x + a_{i0} \ .$$

The classification rule is to assign x to the ith group if $L_i(x)$ is the maximum.

Quadratic Discriminant Functions

Suppose that we do not assume that the population covariance matrices are equal. Then we maximize the quadratic discriminant function.

$$Q_i(x) = -\log|S_i| - 1/2(x-\overline{x}_i)' S_i^{-1}(x-\overline{x}_i) \ .$$

The procedure is otherwise the same as the linear discriminant analysis.

When prior probabilities are assigned to the groups the classification rule becomes: $Q_i^*(x) = \log \pi_i + Q_i(x)$.

Classification Trees

An introduction to classification trees is presented in Case Study 8.4.

Nonlinear Classification Methods

The use of nonlinear classification methods such as flexible discriminant analysis (FDA) are becoming more and more popular because fast computers are available and there are many implementations of them in modern software such as S-PLUS or R. The FDA procedure consists of the following steps.

1. Nonlinear fit. Fit a nonlinear model to the data using a binary numeric representation of the response.

2. Linear discriminant step. Fit a linear discriminant classification rule to the fitted values from step one as predictor variable and the same response variable.

For the first step use any nonlinear nonparametric regression estimator such as the generalized additive models, lowess, projection pursuit regression, or MARS. For further reference on FDA and nonlinear models see Hastie et al. (1995) and Chambers and Hastie (1992).

Neural Nets

Neural nets — we are told over a slide of the human brain — work in a similar way to the brain (and we add, "If that is the case we are in real trouble"). In reality they tend to be heavily overparameterized nonlinear models or classification rules. This goes against the conventional statistical thinking that identifies overparameterization with overfitting.

However, in many of the modern classification problems the overparameterization may be an advantage because the data may have some or all of the following characteristics.

1. There is no overlap between the groups.

2. The separation surface is highly nonlinear.

3. There are massive amounts of data so the separation is well defined.

For these types of problems the risk of overfitting may be worthwhile if at the same time the fit is able to capture very high nonlinearities.

A basic model for neural networks appears under the name "feed-forward single hidden layer neural nets" which consists of an input layer, output layer, and a hidden layer in between. Each node has one or more inputs and one output. The neural net model is built as follows.

1. The input layer (or first layer) consists of as many nodes as predictor variables are available for the fit. The output of each node is the corresponding value of the predictor variable assigned to it.

2. Each node inputs the outputs of the nodes of the prior layer and it outputs a fixed function of the linear combination. The function, sometimes called a "transfer function," is usually a logistic function or any other sigmoidal-shaped function for classification problems and the identity function for regression problems.

3. The output layer (or last layer) has as many nodes as responses are available for the fit.

The process of fitting the parameters of the neural net uses a very complicated "backfitting" algorithm that does not always find the optimal parameter values. But a very good feature of the neural net model is that it can be implemented in hardware in microchips, which is considered very valuable in engineering applications. It would be a mistake for statisticians to ignore the neural net methodology because it does produce excellent results in many applications. The difficulties mentioned above will likely be overcome with new research and greater computational resources.

k Nearest Neighbors

Nearest neighbor methods are also useful for the cases where neural nets are preferred. The idea is quite simple. We need to start with a definition of interpoint distance $d(x_1, x_2)$ such as the ones mentioned in Case Study 8.4 as part of the cluster analysis methodology. The Mahalanobis distance is usually applied.

To define the classification rule for a new x we find the closest k points to x in the appropriate distance $x_1, \ldots, x_k$ and calculate p_i the proportion of points belonging to the ith group. The classification rule will be to assign x to the group of maximum proportion within the nearest neighbor set. If we have a prior distribution, the Bayes classification rule will maximize the product $p_i \times \pi_i$.

One simple way to summarize the results of a classification analysis is to display a two-way table of the classification predicted by the decision rule versus the true classification. The proportion of misclassifications is used to evaluate the procedures. In some setups it may be better to divide the data into training and testing sets and use the testing set to compare the classification procedures.

8.2.3 Preliminary Analysis

First we divide the data at random between a *training* set of 1000 observations of each group and a *testing* set of 250 of each group. The training set is used to estimate the classification rules and the testing set is used to compare the performance of the rules.

We fitted the basic linear and quadratic discriminant analysis. In addition the FDA model was fitted using projection pursuit regression in the function estimation step.

Finally a feed-forward neural net with a hidden layer of 10 nodes (half the predictors) was fitted. Both projection pursuit and neural nets were fitted in S-PLUS using the default settings. The results are shown in Table 8.3 where the number in parentheses gives the result for the training set while the main number is the result obtained from the testing set.

The change from linear to quadratic fits is large. The three nonlinear fits produced similar results, the neural net doing slightly better in the training set and a tiny bit worse in the testing set.

TABLE 8.3. Results of Linear Discriminant Analysis

	Discriminant Function						Neural Net	
	Linear		Quadratic		Flexible			
False Alarms	8	(37)	2	(3)	1	(3)	3	(2)
Bombs Missed	2	(1)	1	(3)	0	(5)	1	(1)
Total	10	(38)	3	(6)	1	(8)	4	(3)

8.2.4 Summary

This case study illustrates an example of a dataset with almost no statistical error. The various competing methodologies are very comparable and there is no clear overall winner. The following questions indicate some of the issues that a detailed report would need to address.

Questions

1. The main objective here is to select the method that outperforms the others for this dataset. The preliminary analysis is very basic. There may be a need to tune up the parameters of the various methods to be able to improve the performance.

2. Can you devise a better way to compare the methods?

3. Which one gives the best results?

8.3 A Market Research Study

Methods:	Factor Analysis
Data:	26 Variables 5 TV Show Ratings 6 Shopping Categories

This case was presented to us by Maria Drelich, from the MBA program at Rutgers University, who generated the ideas in this report. This example corresponds to a very typical situation where the available data have already been processed and it is the result of tabulations performed from the raw data. The raw data are not available and only the tables are available.

The objective of the study was to identify consumer segments with similar "purchasing from catalogue" profiles to those of the viewers of popular TV shows. Each profile consists of a vector of six numbers, representing the proportion of people in a particular subset of the population who bought products from each of six categories.

By this method we can find patterns of association between segments of the population and viewers of particular shows and try to produce profiles of specific viewers for our products.

The Data

The data in Table 8.5 consist of shopping profiles from 29 population subgroups. These results are available in the dataset `c83.dat`. The variable *Subgroup Description* gives an accurate characterization of the subgroup. The shopping pattern consists of six measurements of the number of people of a subgroup who purchased certain types of merchandise from catalogues in the last 12 months as a percentage of the total U.S. population for that subgroup. Hence the variables V_3 to V_8 of the dataset in Table 8.5 are percentages.

At the same time similar profiles were calculated for the viewers of five popular TV shows. The underlying assumption here is that if the shopping profile of a segment of the population is close to that of the viewers of a show we can identify that subgroup as part of the viewers of that show. This may not be necessarily true but as in many market research studies the results may provide interesting associations that may make sense from the marketing point of view.

TABLE 8.4. Variable Definitions

V_0	ID Number
V_1	Subgroup Description
V_2	Subgroup Population Total
	Percentage of V_2 Subgroup Who Bought ...
V_3	Clothing
V_4	Electronics
V_5	Home Furnishings
V_6	Houseware Products
V_7	Sporting Goods
V_8	Toys and Games

Methodology

8.3.1 A Quick Review of Principal Components Analysis

Principal components analysis is a method for dimension reduction. Here are some examples of statistical applications where dimension reduction is important.

TABLE 8.5. Types of Merchandise Bought from Catalogues in the last 12 Months by Demographic Variables and the TV Shows as a Percentage of the Group

V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
1	18–24	23965	15.68	2.34	5.05	4.21	4.19	5.27
2	25–34	42832	22.57	3.98	7.99	5.74	4.20	8.99
3	35–44	39908	31.02	4.94	11.63	6.85	6.12	10.97
4	45–54	27327	31.10	5.72	9.43	8.39	4.99	6.54
5	55–64	21238	28.26	3.22	9.83	8.39	4.17	7.40
6	65–over	30552	24.36	3.09	7.27	4.66	1.43	4.07
7	College Grad.	36463	36.34	5.97	10.16	6.39	6.24	10.41
8	Attend College	44294	29.08	4.55	10.92	6.86	4.37	8.28
9	High School	66741	23.70	3.27	8.10	7.02	4.20	7.32
10	No High Sch.	38324	15.14	2.75	5.77	4.30	2.38	4.56
11	Total Male	88956	18.91	4.76	5.38	4.34	4.99	6.00
12	Total Female	96866	31.93	3.30	11.74	8.09	3.60	9.04
13	Employed Male	65500	19.46	4.93	5.42	4.25	5.51	6.45
14	Employed Fem.	55910	36.42	3.62	14.02	9.01	4.54	10.51
15	Full Time	110363	26.65	4.46	9.20	6.31	8.00	5.03
16	Part Time	11047	33.51	2.98	10.74	7.78	5.47	11.51
17	Unemployed	64412	22.73	3.37	7.41	6.01	2.76	6.20
18	Single	41284	16.86	3.60	5.17	4.24	3.60	3.75
19	Married	109023	29.94	4.51	10.65	7.23	5.12	9.61
20	Div/Sep/Wid	35515	22.95	2.90	6.79	5.83	2.41	5.84
21	Parents	62342	29.22	4.76	10.75	6.40	5.65	12.37
22	75K–more	24165	35.59	4.42	11.09	8.03	5.00	11.19
23	60K–more	40979	32.80	4.85	9.87	7.20	5.12	9.80
24	50K–more	57996	32.66	4.57	10.46	7.07	5.55	9.59
25	40K–more	80078	31.56	4.65	10.54	6.94	5.60	9.27
26	30K–more	106838	30.62	4.64	10.15	6.71	5.45	8.97
27	20K–29K	30669	21.32	3.91	8.14	6.01	3.14	6.10
28	10K–20K	29083	19.23	2.39	6.17	6.36	2.65	5.84
29	under 10K	19232	15.12	2.96	5.35	4.37	1.94	4.92
30	*ER*	19640	31.65	4.22	10.65	7.22	4.97	9.58
31	*Friends*	16000	23.92	3.52	7.54	5.86	4.50	5.64
32	*Frasier*	14840	31.03	3.98	10.51	8.25	4.66	8.51
33	*Jesse*	13550	19.14	3.77	6.49	4.97	4.24	5.28
34	*Ally McBeal*	10190	21.80	2.40	6.42	6.32	4.58	6.67

- Data mining: This is employed in applications that deal with datasets with many variables and cases and one of the main objectives is data reduction. Principal components can be a tool for data reduction by reducing the number of variables. For details on data mining see Case Study 8.4.

- Regression Analysis: When the number of predictors q is comparable to the error degrees of freedom n_ϵ. We need to reduce the number of variables to $q* \ll n_\epsilon$.

- MANOVA: When the number of responses p is comparable to the error degrees of freedom n_ϵ. We need $p \ll n_\epsilon$.

Suppose we observe n random vectors $x_i = (x_{i1}, \ldots, x_{ip})'$ $i = 1, \ldots, n$ and assume that the x_is have mean zero and sample covariance matrix S_x. Let A be an orthogonal transformation such that $z_i = Ax_i$ and the z_is are uncorrelated. Since A is orthogonal $z_i'z_i = x_i'x_i$ so the z_is have zero mean and the sample covariance matrix is the diagonal matrix $S_z = AS_xA'$. The rows $\{a_i\}$ of the matrix A are called the eigenvectors of S_x and the diagonal elements $\{\lambda_i\}$ of the matrix S_z are called the eigenvalues of S_x.

The eigenvectors represent a set of new variables called "principal components" that correspond to the "natural" coordinate system of the data. The eigenvalues are the variances of their corresponding principal component variables and the principal components are usually ordered by the size of their eigenvalue, from largest to smallest. The eigenvalues are often described as the variances explained by the corresponding principal component. The first principal component represents the direction of maximum variability, the second principal component represents the direction of maximum variability in the space orthogonal to the first principal component, and so on.

The methodology of principal components is applied when many of the variables in the data set are redundant; that is, they provide no new information that is not already accounted for by other variables. Geometrically this means that the data lie approximately on a subspace of dimension $k < p$. If this is the case we expect the k largest principal components to explain almost all the variability in the data.

Two quantities of interest are $b_k = \lambda_k / \sum_{i=1}^{p} \lambda_i$ and $c_k = \sum_{i=1}^{k} b_i$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. These represent, respectively, the proportion of the variance explained by the kth principal component, and the proportion of the variance explained by the first k principal components. We propose three simple criteria to determine k.

- The smallest number of components that explain over 80% (or any other magic number) of the variance; that is, $c_k > 0.8$.

- They explain more than the average of the variances; that is, $b_k > c_p/\mathrm{p}$.

- Graph of the factor variances versus the order; that is, b_k versus k. The last large jump downwards shows the cutoff for the number of principal components.

Ideally the three methods should point out a single number of principal components. If there are any doubts it is better to take a conservative approach.

8.3.2 A Quick Review of Factor Analysis

We follow the same setup as in the *Principal Components* section. The idea of factor analysis is to express the observed variables $x_1, \ldots, x_p$ as linear combinations of a few unknown quantities called factors $f_1, \ldots, f_k$ $(k < p)$. For the following discussion it is important to note that:

1. Factors are unknown quantities.

2. The number of factors is unknown.

3. For simplicity we assume a model that represents our observed variables as linear combinations of the unknown factors with some added noise.

There is much controversy around the idea of factor analysis because factors can not be measured; they may or may not exist. However, in this book we take a pragmatic approach to this methodology and we recommend factor analysis whenever it is useful.

In order to distinguish principal components analysis from factor analysis we note that principal components are linear combinations of the variables but variables are linear combinations of the factors. Principal components try to explain the variability but factors try to capture the correlation structure.

Factor Model. As in the section before we assume that the x_is are mean centered. A model for factor analysis is as follows: $x = Af + \epsilon$, where f is a random vector with zero mean and covariance matrix equal to the identity I_p, the errors are uncorrelated, and $Cov(f, \epsilon)$ is a $k \times p$ matrix of zeros.

Factor Loadings. The elements of A are called *factor loadings* representing the covariances of the variables with the factors $a_{ij} = Cov(x_i, f_j)$. The ith row of A gives the coordinates of the variable x_i in factor coordinates.

Communalities. We break down the variance of a variable between a component due to the common factors (its communality) and a variable

specific component:

$$Var(x_i) = h_i^2 + \sigma_i^2 = \sum_{j=1}^{p} a_{ij}^2 + Var(\epsilon_i) \ .$$

Estimation. To determine the number of factors we may use the same three-way procedure that was recommended for choosing the number of principal components. We illustrate the procedure in the preliminary analysis in the next section.

To estimate the parameters in the factor model we may assume that the errors are normally distributed, in which case we can estimate A and f by maximum likelihood.

If we do not want to assume the normal model we may estimate the factor by iterative principal components, where at each step of the iteration we fit the factor model by applying principal components to the residuals of the prior iteration.

Rotations. One of the characteristics that makes factor analysis more controversial is that the unobservable factors are nonunique under orthogonal transformations. Let H be an orthogonal transformation; then if $A^* = AH$ and $f^* = H'f$ then $Af = A^*f^*$ and A^* are the corresponding factor loadings for the transformed factors f^*. For this reason the initial factors are usually rotated for the purpose of identifying each variable with a single factor.

Manual Rotations. When there are only two or three factors it is useful to graph the rows of A as 2D or 3D graphs and determine if a rotation would be useful to identify the factors with variables or groups of variables. If that is the case it is easy to eyeball the rotation angles and perform the rotation *manually.*

Varimax Rotations. The rotation that maximizes the variances of the squared loadings of each factor is called "varimax" rotation. The idea is that each factor has loadings either close to zero or to one (for the correlation matrix).

8.3.3 Preliminary Analysis

In order to determine the shopping profile of our specified customer who watches a particular TV show we start by performing a factor analysis using the software S-PLUS.

The method of choice for the factor analysis is principal components followed by a varimax rotation. We pick two factors because the three criteria we proposed point towards two factors:

- They explain over 80% of the variance.
- They explain more than the average of the variances (16.7%).
- A graph of the factor variances versus the order shows the last large jump downwards from points 2 to 3.

TABLE 8.6. Factor Analysis Results

(a) Variance Explained by Rotated Factors

	Factor 1	Factor 2
SS Loadings	2.9403433	1.8754577
Proportion Var	0.4900572	0.3125763
Cumulative Var	0.4900572	0.8026335

(b) Factor Loadings

Loadings	V3	V4	V5	V6	V7	V8
Factor 1	0.857	0.156	0.921	0.860	0.331	0.696
Factor 2	0.448	0.753	0.327	0.111	0.831	0.546

The factor loadings suggest that factor 1 is essentially a combination of sales of clothes, home furnishings, housewares, and toys/games, whereas the second one is the combination of sales of electronic and sporting goods. Figure 8.3 displays the projection of the original variables into the plane defined by the two factors. In addition the observations are shown in the factor coordinates.

In simple examples like this there is no need to perform a cluster analysis since the graph in Figure 8.3 is simple enough to derive the clusters by hand. In more complex datasets it may be suitable to proceed by performing a cluster analysis.

At the bottom left corner of the graphs we see the group made of 10, 20,28, 29,..., which corresponds to the viewers who did not graduate from high school, are divorced, separated, or widowed, and the groups of lowest income. They are low on both factors. On the other hand, groups 11 and 13 that correspond to males and employed males are high in factor 2 but low in factor 1 while employed females and females overall (12 and 14) are high in factor 1 and low in factor 2. All this may sound like a stereotype, but it is suggested by the numbers.

In terms of the TV shows our preliminary conclusions are as follows.

The viewers of the *ER* and *Frasier* shows have similar shopping patterns, moderate to high income (above 40K), moderate age (30 to 40s), and are parents. So according to this analysis we conclude that many viewers of *ER* and *Frasier* have a moderate to high income (above 40K), moderate age (30 to 40s), and many are parents.

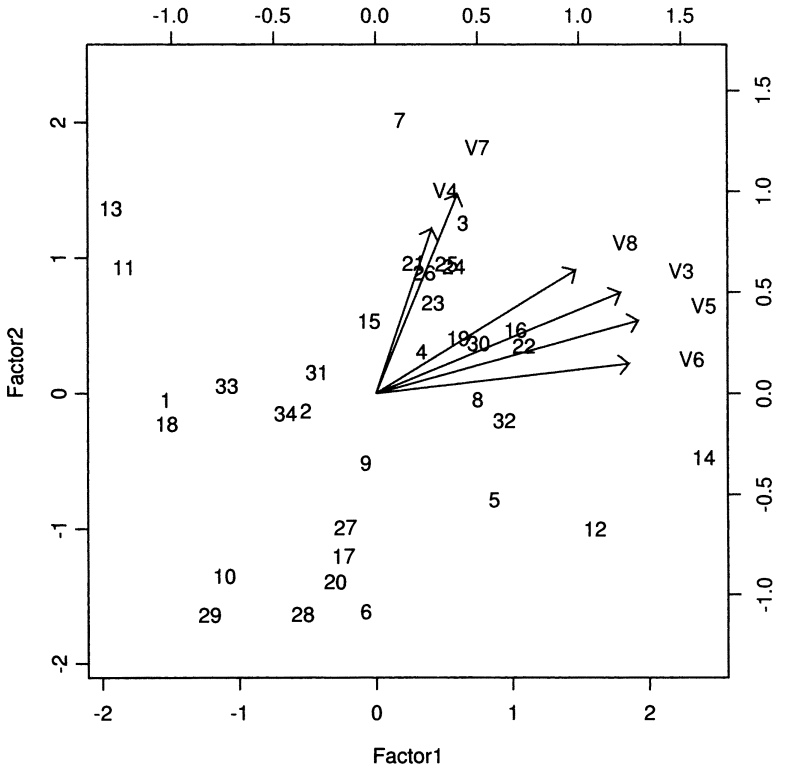


FIGURE 8.3. Factor Loadings and Datapoints in the Factor Basis

The audience of the *Ally McBeal* show is composed of viewers with slightly lower income (30K+), more mature age, many married and female, and possibly with college education.

Finally the viewers of *Jesse* are young people mainly between 25 to 34 years old, full-time employed, and with a high school education.

One additional very strong cluster is composed of employed females, college graduates, and people with income above 75,000. The catalogue shopping pattern for this group is the strongest. Although they are not directly associated with any of the five TV shows they are only moderately related to *ER, Friends, Frasier,* and *Ally McBeal* but not to *Jesse.*

The remaining group is made up of youngsters (less than 24), or seniors (65 or over), the unemployed, divorced, or separated, with low income. They

do not follow the viewer's profile for any of the above TV shows. Along the way, as we keep adding more shows we will perform similar analyses for every one of them in order to get an adequate viewer's profile.

8.3.4 Summary

This case study illustrates some of the methodology available for market research studies. One of the main advantages of market research analysis is that it is exploratory. In many applications it is important to obtain information quickly and the findings need not be supported by formal procedures such as experimental design and hypothesis testing. The objective is to obtain new information faster than your competitors even at the risk of not always being correct. Please address the following questions in your detailed report.

Questions

1. Which transformations, if any, are appropriate for these data?
2. Consider several alternatives for estimation and rotation; are the results consistent?
3. Provide a criticism of the preliminary analysis and alternative analyses. Interpret the results in relation to the objectives of the study.

8.4 Sales of Orthopedic Equipment

Methods:	Data Mining Exploratory Data Analysis Discriminant Analysis Cluster Analysis Regression Analysis
Data:	Two Responses 8 General Demographics Predictors 8 Predictors Specific to Orthopedics

8.4.1 Data Mining Applications to Market Research

Our client in this case study is a company that is a large manufacturer of orthopedic equipment in the United States, with a customer base that consists of almost all hospitals over the 50 states. The products that are involved in this study range from orthopedic parts and equipment to medications administered in the process of surgery, rehabilitation, and recovery.

From the point of view of sales our company believes that the market consists of three groups (or segments) of hospitals:

1. Hospitals where we have high sales and we have reached the desired market positioning. We are not concerned with this group.
2. Hospitals where we have moderate sales but there is still further sales potential.
3. Hospitals where we have little or no presence and there is a substantial potential gain.

From the point of view of the activities we classified hospitals as:

1. Small general hospitals where few orthopedic activities take place. In these hospitals most orthopedic operations are referred to other more specialized centers. These hospitals have little or no interest for us since potential sales are marginal.
2. Large general hospitals that frequently have trauma or rehabilitation units or that perform some orthopedic surgery. They are not necessarily very specialized in orthopedics but because of their size they need some facilities for these types of operations. They are of moderate interest, especially the bigger ones: because of their size they can contribute significantly to the company's market share.
3. Specialized hospitals who perform many orthopedic procedures and have units dedicated to traumatology and rehabilitation. These hospitals are our main target group because they are a large part of the market and of market growth.

The objective of this study is to find ways to increase sales of our products to hospitals in the more desirable groups. We will do this by identifying hospitals that are likely clients, but for unknown reasons currently buy at a lower level than we would expect. Then we will study them individually to try to determine the reason for the low sales and if we believe that there is growth potential we will invest in the marketing, so we will focus only on a few hospitals where we believe we can maximize our return.

Another objective of the study can be to determine if the above hospital classification is realistic or if there are other ways to classify hospitals that can be used for marketing purposes.

The Data

The dataset `"c84.dat"` contains the list of variables that are more interesting to us so we already selected a subset that is reasonable. Each case corresponds to a hospital and all U.S. hospitals are in the database. Table 8.7 describes the groups of variables associated with each hospital and Table 8.8 gives a precise definition of each variable.

TABLE 8.7. Variable Groups

Groups	Variables
Responses	SALESY, SALES12
Geographic	HID, CITY, STATE, ZIP
General	BEDS, OUT-V, ADM, SIR
Orthopedic	RBEDS, HIP95, KNEE95, TH, TRAUMA, REHAB, HIP96, KNEE96, FEMUR96

Methodology

8.4.2 Data Mining

This is an interesting example of what has now become known as data mining. This case was prepared for a Spring 1993 class well before data mining was invented so in a way we can say that we were doing data mining before it was invented officially. This claim is probably shared by many people since data mining is something that statisticians and computer scientists have been doing for a long time. As a result, data mining has developed from a variety of applications without being specific to statistics. From the user's point of view the software applications for data mining (such as Enterprise Miner, Diva, Spotfire) provide a list of procedures (data visualization tools, recursive partitioning, cluster analysis, neural nets, and maybe some AI methods) that work well for reasonably large data sets (100 variables $\times$ 100K observations). The procedures are not necessarily that new but the software capabilities are more appropriate for large datasets. For this reason data mining has been driven by computer scientists and engineers.

Data analysis has been represented in the past as detective work where the objective is to find interesting structures in the data. This is a good approach for many scientific problems where there is such a structure to be found and maybe not too much or too badly behaved noise. On the other hand, in many modern problems there are massive datasets where there is little information and almost no structure at all.

The objective of data mining is to identify nuggets, small clusters of observations in these data that contain unexpected, yet potentially valuable, information. The definition of valuable is generally reflected by a large response value of a specific category of a qualitative response. Sifting through a large volume of data that is noisy, badly behaved, and that may have many missing values, or that may just be irrelevant is the main challenge of data mining. Another is the merging of diverse databases in the hope that snippets of information in each may synergize into an identifiable nugget in the whole. This is what can be found in typical market research datasets, supermarkets, retail stores, phone call information, high throughput screening, with DNA microarrays, and in the clinical trials area. In all

TABLE 8.8. Variable Definitions

Variable	Definition
SALESY	Sales of Rehabilitation Equipment Since Jan 1
SALES12	Sales of Rehabilitation Equipment for the Last 12 Months
ZIP	US Postal Code
HID	Hospital ID
CITY	City Name
STATE	State Name
BEDS	Number of Hospital Beds
OUT-V	Number of Outpatient Visits
ADM	Administrative Cost (in $1000s per year)
SIR	Revenue from Inpatient
RBEDS	Number of Rehabilitation Beds
PRHIP	Number of Hip Operations (total for previous year)
PRKNEE	Number of Knee Operations (total for previous year)
TH	Teaching Hospital? 0 = no, 1 = yes
TRAUMA	Trauma Unit? 0 = no, 1 = yes
REHAB	Rehabilitation Unit? 0 = no, 1 = yes
HIP	Number Hip Operations (total for current year)
KNEE	Number Knee Operations (total for current year)
FEMUR	Number Femur Operations (total for current year)

of these problems there are large amounts of data collected systematically, for example, all the phone calls that are made by businesses, or all the purchases in retail stores. The data are stored in data warehouses that are available for analysis.[1]

There are standard database query tools developed by computer scientists for the purpose of retrieving subsets of certain characteristics, and on the other hand, statisticians have developed tools to find interesting features of data. What we really want here is some way to combine both of these ideas. How to produce intelligent queries? How to select interesting subsets where the meaning of "interesting" is not defined very precisely as in a database query? These types of questions are the ones that data mining tries to answer. A structured approach to data mining would involve, in sequence:

1. Dimension (variable) reduction.
    - Principal components.
    - Factor analysis.

[1]For a price! Databases have become an increasingly valuable commodity which is another reason why data mining has developed so rapidly.

2. Data segmentation and selection.
    - Cluster analysis.
    - Tree methods.
    - Neural nets.
3. Data analysis of interesting segments.

To a large extent, this is a recipe for data reduction, where most of the non-informative data are systematically eliminated, allowing one to concentrate on a few interesting pieces.

We can regard the data being mined as an array of cases by variables. The variables can be numeric, categorical, or even simple data structures, so this setup is quite general. One can start by selecting a subset of variables that may be informative with regard to our objective. Usually further dimension reduction is necessary. This can be accomplished using standard statistical tools for variable reduction such as principal components analysis and factor analysis.

Case reduction may be performed using cluster analysis to partition the dataset into homogeneous clusters or segments. Another way of performing the segmentation is by using a tree method such as CART or similar methods of recursive partitioning (for a description of these methods see the subsection below). The idea there is a bit different because it requires involving response variables, so it may not yield a set of natural clusters but a more efficient segmentation in terms of the responses. On the other hand, there may not be a set of natural segments in the population and the tree method will give a more relevant answer. In the best of all worlds, the results from both procedures will be similar, but don't count on it.

Once subsets of potential interest have been identified, they can be examined, summary statistics calculated, sufficient analysis performed, and scientific interpretability assessed to decide which segments merit further investigation. This may be just a simple operation such as calculating the average sales per segment or it may require more sophisticated modeling.

From this point on comes the analysis of the small subsets and the extraction of the final pieces of relevant information which are dependent on the problem. In the example that we have here a good strategy would be to perform regression analysis on the segments and to extract those hospitals that are well below their expected sales. This is covered in the data analysis section below.

8.4.3 A Quick Review of Cluster Analysis

We have a dataset and we want to group the data into k distinct natural groups. There are many approaches to cluster analysis and this is very noticeable in the software implementations of cluster analysis in the usual

statistical packages, which implement 10 or 15 different methods. Although the ideas of cluster analysis are quite intuitive, it is surprisingly difficult to formalize the idea of a "cluster" in a general unique sense. There are two popular methods for doing cluster analysis.

Hierarchical Clustering This method requires the definition of interpoint distance and intercluster distance. The interpoint distance is normally taken to be the Euclidean distance:

$d_E(x_1, x_2) = \sqrt{\sum_{j=1}^{p}(x_{1j} - x_{2j})^2}$.

Sometimes we may use Manhattan distance:

$d_M(x_1, x_2) = \sum_{j=1}^{p} |x_{1j} - x_{2j}|$.

The intercluster distance between two clusters is defined as a function of the interpoint distances between pairs of points where each point comes from a different cluster. The popular definitions of intercluster distances are:

Single Linkage:	distance between the closest two points
Complete Linkage:	distance between the farthest two points
Average Linkage:	average distance between every pair of points
Ward:	R^2 change.

We build a hierarchical tree starting with a cluster at each sample point, and at each stage of the tree the two closest clusters join to form a new cluster.

Once we finish building the tree the question becomes, "How many clusters do we choose?" One way of making this determination is by inspecting the hierarchical tree and finding a reasonable point to break the clusters. We can also plot the criteria function for the different number of clusters and look for unusually large jumps.

Centroid Methods K-means algorithm.

K seed points are chosen and the data are distributed among K clusters. The clusters are then slowly optimized using some criterion such as R^2. At each stage of the algorithm one point is moved to the cluster that will optimize the criteria function. This is iterated until convergence occurs. The final configuration has some dependence on the initial configuration so it is important to make a good start. One possibility is to run Ward's method and use the outcome as the initial configuration for k-means.

8.4.4 Recursive Partitioning and Classification And Regression Trees (CART)

Recursive partitioning, and tree-based methods such as CART, are other ways of performing the segmentation of a multivariate dataset that contains

a response variable and several predictors. The response variable as well as the predictors can be categorical or numerical.

Although these methods appear very similar in the sense that they expand a tree by optimizing some criteria over all possible splits, there are some differences. Recursive partitioning methods use a stopping rule based on the P-value of some test statistic such as the χ^2- or F-statistics. Tree methods expand the tree until the node sizes are smaller than some number and then proceed to prune it. The pruning is necessary to avoid overfitting.

8.4.5 Preliminary Analysis

We concentrate our study on a subset of related states of about 300 to 500 hospitals and find those that have high consumption of such equipment but where our sales are low. The objective is to come up with a selected group where we think our efforts will be rewarded.

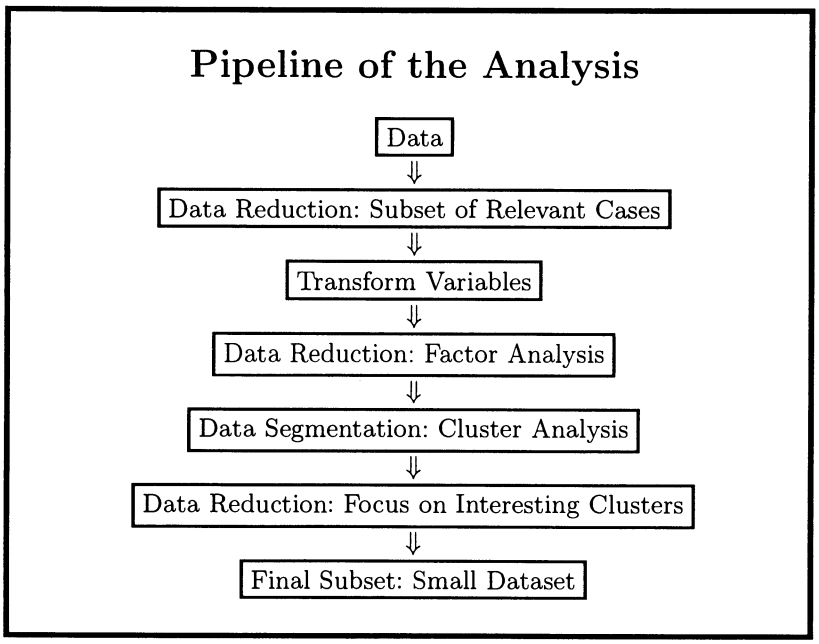


FIGURE 8.4. Schematic for Data Mining Analysis

The names of the variables are given in Table 8.8. In Figure 8.4 we display a pipeline of a proposal for analyzing these data. The focus of the graph is on steps of data mining and variable reduction rather than on standard statistical analysis. SAS code is included to perform this preliminary analysis.

Part 1. Select your market segments.

Data reduction. Select subset of relevant cases. We select a state or group of states for the study, but we would like to have at least 300 hospitals in the state or group of states.

Graph each individual variable and decide whether a transformation is appropriate. Separate the variables into two groups: responses and demographics.

Demographic variables are used to divide the list of hospitals (all possible clients = the market) into subsets that we call market segments.

Variable reduction. If there are too many demographic variables, apply a principal components analysis or factor analysis to summarize the demographic variables into a few components or factors.

Market segmentation. Use cluster analysis or tree methods to find the market segments or clusters.

Data reduction: Focus on interesting clusters. Once the clusters are chosen, study the summary statistics for each cluster and try to describe their content. Interpretation is very important at this stage.

Final subset: A small dataset. Select the cluster(s) that agree with the objectives. In this study we are looking for segments (hospitals) with high overall sales but where our company's sales are low. Some segments may have mostly zeros or low numbers for sales. This would indicate hospitals with only a few patients who would need our products so for marketing purposes, we are not interested in these segments.

Part 2. Estimating potential sales.

To finish the analysis we perform a regression analysis for each of the selected segments. Notice that since the segments are very homogeneous we may expect small R-squares SO DO NOT BE CONCERNED WITH LOW R-SQUARES.

The hospitals with large negative residuals are the ones that have low sales but their characteristics suggest that they are below their potential sales (use predicted values as potential sales). To finish we make a list of the hospitals in the segment where sales can be improved and give an estimate of the potential gains.

This model of analysis is only one of many ways to analyze these data. Think of it as a list of suggestions that can be applied to this or other similar datasets. For the sake of simplicity we did not use a very large dataset here but the above steps become more and more important as the dataset size gets bigger.

The SAS code below can be used to perform the computations described above. The factor analysis step was omitted and the cluster analysis method implemented was Ward's.

SAS program for analyzing the sales of orthopedic equipement.

```
options ps=55 ls=78;

data ortho;
infile 'c84.dat' delimiter=',';
input zip $ hid $ city $ state $ beds rbeds outv adm sir
      salesy sales12 hip95 knee95 th trauma rehab hip96
      knee96 femur96;

/* 1. Select subset of interest */
if state eq 'fl' or state eq 'ga';
array x {12} beds rbeds prhip prknee hip knee femur
             outv adm sir salesy sales12;

/* 2. Transform the variables */
do i=1 to 7 ;
   x{i} = sqrt(x{i}) ;
end;
do i=8 to 12 ;
   x{i} = log(1+x{i}) ;
end;
run;
/* 3. Factor analysis step is omitted */

/* 4. Data segmentation using cluster analysis */
proc cluster method=ward;
   var beds rbeds outv adm sir prhip knee th trauma
       rehab hip knee femur;
   copy beds rbeds outv adm sir prhip prknee th trauma
       rehab hip knee femur sales12 salesy ;

/* Select 8 clusters or market segments. Assign the
   cluster variable to the cases */
proc tree noprint ncl=8 out=txclust;
   copy beds rbeds outv adm sir prhip prknee th trauma
        rehab hip knee femur sales12 salesy ;
run;

/* 5. Data reduction. Examine the composition of the
      clusters and pick the most interesting ones. */
```

```
proc sort data=txclust; by cluster;
proc means data=txclust;
   var beds rbeds outv adm sir prhip prknee th trauma
       rehab hip knee femur sales12 salesy ;
   by cluster;
   output out = c
   mean = mbeds mrbeds moutv madm msir mprhip mprknee
          mth mtrauma mrehab mhip mknee mfemur
          msales12 msalesy ;
proc print;
   var cluster mbeds mrbeds moutv madm msir mprhip
       mprknee mth mtrauma mrehab mhip mknee mfemur
       msales12 msalesy ;
run;

/* 6. Final subset. Do the standard analysis of the small
      sample */
```

In the above data mining example we are able to examine the clusters carefully because we have only a moderate number of observations. However, in data mining problems with very large datasets it may not be possible to carefully examine most of the clusters, but only a few that appear very interesting. Hence the objective of data mining may not be to analyze "all" the data but to focus on interesting subsets.

8.4.6 Summary

The above analysis is specific to the questions presented by this study but it can be used as the basic part of many data mining problems. The following questions are perhaps specific to the problem at hand but can be adapted to many other situations.

Questions

1. Find the variables that need to be transformed and the corresponding transformations.
2. Explore the possibility of using principal components to summarize the variables. Is there any interpretation of the first few principal components?
3. Graph the main principal components. Are there any visible clusters?
4. Perform a cluster analysis with the best principal components and with the raw variables. Compare and interpret the analyses. Should you include the response in the cluster analysis?

5. Once you have a partition that is satisfactory select a few interesting clusters and analyze the data using regression analysis or any statistical method. Remember that the objective here is to find hospitals whose characteristics suggest high sales but the observed sales are low.

9
Additional Case Studies

Now it's your turn! You were highly recommended by Another Client as Zee Consultant to see. So here are a few projects we've been working on. ... See you next week for the results?

9.1 Improving Teaching

Description

The purpose of this study was to investigate whether incorporating activities tailored to the learning style preferences of students provided evidence of an improved "performance" over traditional (textbook) teaching methods. The Learning Style Inventory (LSI) instrument[1] provides a measure of a person's strength with respect to their tactile (hands-on construction), kinesthetic (role playing), auditory, and visual perceptive abilities. A person who scores 60 or more in one of these categories is regarded as having a strong preference for a learning style which addresses their particular perceptive ability. From a teaching perspective, this suggests that students will more easily absorb and retain new information if their learning style preferences are incorporated into the teaching process.

Design

A total of 59 elementary students (9 to 10 years old) in three classes (`CLASS`) were taught three successive science units (`UNIT`) over a period of three weeks. Before commencing the experiment, the students' learning style preferences were assessed using the LSI instrument. To avoid bias, the students were not made aware of their learning style preferences during the experiment.

To establish prior knowledge of the student, a pretest (`PRE`) consisting of multichoice questions was given at the beginning of each unit. At the completion of the unit, students were assessed by means of a multichoice format posttest (`PST`), a semantic differential scale (SDS) or "attitude" test (how the students felt about learning science), and a higher-learning test (`HLT`) which required a worked-out solution to a problem. The maximum score for `PRE` and `PST` was 100%. For `SDS`, 60 represented the highest possible (positive) attitude score. The `HLT` test was marked on a scale of 1,2,3,4 with 4 representing a correctly worked-out solution and 1,2,3 assigned to a partially correct solution accordingly.

A different class was selected as the Control group (`GROUP=C`) for each unit. This was taught using traditional teaching methods only. The other two classes were designated as the Experiment groups (`GROUP=E`) which augmented traditional instruction with tactile and kinesthetic activities for the students. Auditory and visual activities were not considered in this experiment. The control group assignments were as follows.

```
UNIT = 1 : C = CLASS 3
UNIT = 2 : C = CLASS 1
```

[1]The LSI instrument is also discussed in Chapter 4.

```
UNIT = 3 : C = CLASS 2
```

The Data

The results from this experiment are contained in the dataset `c91.dat` which has the following format. The variables are defined in Table 9.1.

```
                    UNIT=1          UNIT=2          UNIT=3          PREF
ID  CLASS SEX  P1  F1 S1 H1  P2 F2 S2 H2  P3 F3 S3 H3   T  K  A  V
s01   2    M   70 100 60  4  48 88 52  3  72 64 60  2  61 44 57 55
s02   2    F   80  85 60  2  50 76 60  3  70 72 60  1  35 44 57 34
 :              :               :               :               :
s59   3    F   78  78 58  3  18 70 60  3  26 90 60  4  61 51 48 50
```

TABLE 9.1. Variable Definitions

Variable	Description
ID	Identity Code Assigned to Student
CLASS	Class Group: 1, 2, 3 ($n_1 = 21$, $n_2 = n_3 = 19$)
UNIT	Science Unit Taught: 1, 2, 3
GROUP	C = Control: Traditional Teaching Method E = Experiment: T and K Activities Included

Assessment Variables		*Column Coding*
PRE	Pretest (for each UNIT)	P1, P2, P3
PST	Posttest	F1, F2, F3
SDS	Attitude score	S1, S2, S3
HLT	Higher-learning test	H1, H2, H3

Learning Style Preferences		
PREF	T = Tactile	A = Auditory
	K = Kinesthetic	V = Visual

Assignment

The objective of the project was to assess the statistical significance of the following four hypotheses which are stated in *directional* form.

H1 There will be a difference between PST by GROUP.

H2 There will be a difference between HLT by GROUP.

H3 There will be a difference between SDS by GROUP.

H4 There will be an interaction between PST and PREF by GROUP.

9.2 Random Sampling?

Description

Sampling schemes are commonly used by state and federal agencies to monitor compensation claims submitted by companies that provide services for government subsidized programs such as health care and housing assistance, at a reduced cost to qualified recipients. In this case study exercise, a company is disputing the results of a government agency's review of its claims on the basis that the sampling method employed by the agency was inappropriate.

The Review

The report provided by the agency stated that a statistical random sample from the 40,000 Type I and II claims submitted by the Inexs company in the previous year resulted in $n_1 = 400$ Type I claims and $n_2 = 350$ Type II claims being selected for review. The compensation for Type II claims is generally larger than for Type I claims. Based on the agency's review of these sampled claims, 50% of the Type I claims and 30% of the Type II claims were found to be ineligible for compensation under the conditions stipulated for the Yapreven subsidized program. Since these claims were paid automatically, the agency determined an overpayment estimate based on the sample results, rather than perform a full audit of every financial transaction.

The Appeal

Inexs attempted to replicate the statistical sampling procedure purportedly used by the agency and obtained very different results. To eliminate the inherent variability associated with statistical sampling procedures, the entire database of claims was analyzed to obtain a true "population" profile of Type I and II recipients as defined by the agency overseeing the Yapreven program.

Results

Database Profile

Total Number of Claims		40000
Total Number of Recipients		8000
Type	Total	Proportion
I	35000	0.875
II	5000	0.125

Assignment

1. Compute a conservative estimate of the probability of obtaining a statistical random sample consisting of 400 Type I claims, as reported in the agency review.

2. When Inexs confronted the agency with the above result, the agency stated that it had employed a *stratified* random sample for estimating the overall **proportion** of eligible claims. Are the results reported by the agency consistent with this sampling method? Does this have any bearing on the overpayment estimate?

9.3 Left or Right?

Description

The data in this case study exercise were collected over several years and consist of various measurements taken from "flakes" that were used as cutting tools by early hominids in the Pleistocene[2] period. These artifacts were collected from many sites and one of the main purposes of the project was to try to determine some common characteristics among the observations. In addition, previous studies had also suggested that certain flake patterns were consistent with knapper "handedness" and hence that early hominids may have possessed cognitive reasoning abilities. This conclusion is subject to debate, of course, but leads to the interesting question of whether hand preference is discernible from this archeological record.

The Data

The flake measurements from several sites are all contained in the dataset `c93.dat`. Most of the variables in this dataset are discrete and the key variable for discerning hand preference is `FT` (flake type). Specifically, the levels of `FT` coded as 2.2, 5.2 are considered to be characteristics associated with right-handed knappers and the levels 2.3, 5.3 associated with left-handed knappers. A complete list of the variables is given in Table 9.2. These are presented in the same order as they appear columnwise in the dataset `c93.dat`.

Assignment

Do the data provide evidence of hand preference?

[2]That is, flakes which were used 2 million years ago.

TABLE 9.2. Variable Definitions and Codes

FN	Site Identifier	
ID	Artifact Identifier	
L,W,T	Length, Width, Thickness (mm)	
MD	Maximum Dimension: max(L,W,T)	
FS	Flake Shape	1, 2, ..., 7
FT	Flake Type	2.2, 5.2 (**right**)
		2.3, 5.3 (**left**)
		1, 2.1, ..., 9 (other)
PC	Platform Character	1, 2, ..., 6
PL,PW,PA	Platform Length, Width, Angle (mm or degrees)	
TQ	Technique	1, 2
RM	Raw Material	1, 2, ..., 99
FA	Flake or Angular Fragment	1, 2, 3
CN	Flake Condition	1, 2, ..., 6
H	Hinge	1 = yes, 0 = no
X	Cortex	1 = yes, 0 = no
R	Ridge	1 = yes, 0 = no
E	Hinge Elimination	1 = yes, 0 = no
B	Dorsal Battering	1 = yes, 0 = no
A	Thermal Alteration	1 = yes, 0 = no

9.4 Making Horse Sense

Description

An experiment was conducted on seven horses to assess the effect of two treatments on glucose and insulin levels in the horses' blood over time. Both response variables were measured at each of 12 time points which occurred at equally spaced intervals. A crossover design was employed which conformed to the following schematic.

	horse						
trial	h1	h2	h3	h4	h5	h6	h7
1	a	a	a	a	b	b	b
2	b	b	b	b	a	a	a

```
treatment = a b
responses = glucose, insulin
time = 1,2,...,12
```

Assignment

The purpose of this case study exercise is to create a SAS or S-PLUS program that will perform the analyses described below. Since the data are not available, a dummy dataset will need to be created for testing purposes.

1. The first step is to create a new variable "`seq`" that indicates the `treatment + trial` sequence:

```
seq = 1 if (trial = 1 & treatment = a)
seq = 2 if (trial = 1 & treatment = b)
seq = 2 if (trial = 2 & treatment = a)
seq = 1 if (trial = 2 & treatment = b)
```

2. Compute the following summary statistics associated with each response curve and apply the appropriate analysis of variance procedure to test for a significant "`treatment`" effect. Explain how to interpret the ANOVA output and what the results mean if the effect of `seq` or `trial` is significant.

   - Area under curve
   - Maximum peak value of curve
   - Time to maximum peak.

3. Incorporate "`time`" into the analysis. Note that `time` can be treated as a factor or as a covariate.

9.5 The Tall Redhead

A Discrimination Case

At the request of the President of Erewhon Industries, we have been asked to review the statistical analysis presented in a report prepared on behalf of the plaintiff with regard to the case below.

Analysis of the statistical relationship between the decision to grant pay raises and hair color/height

in the matter of

Ilias Offen vs. Erewhon Industries

Summary of the REPORT

In the REPORT, three factors were analyzed with regard to employees at Erewhon seeking pay raises. These were:

RAISE	Granted or Not Granted
COLOR	Redhead or Nonredhead
HEIGHT	Height of Employee (Short or Tall)

The HEIGHTs of employees were categorized as "Short" if their actual height was less than 180 cm; otherwise they were categorized as "Tall." The statistical analysis of these factors was divided into two parts and in both cases, Fisher's exact test was employed.

(A) RAISE versus COLOR

(B) RAISE versus HEIGHT.

The Results

The following table in the REPORT summarizes the results obtained from the statistical analyses above. For each analysis, a statistically significant result was obtained.

(A) Percent and number **not** granted RAISE by COLOR

Nonredhead		Redhead		*P*-value	SD Units
18.2%	(4 of 22)	66.7%	(4 of 6)	0.038	2.08

(B) Percent and number **not** granted RAISE by HEIGHT

Short		Tall		*P*-value	SD Units
0.0%	(0 of 9)	42.1%	(8 of 19)	0.029	2.18

The Conclusion of the REPORT

Based on the above results, the author of the REPORT states:

> "... the data provided to me show a statistically significant probability that redheads were treated differently than nonredheads, and taller employees were treated differently than shorter employees with regard to the decision to grant pay raises by the President of Erewhon Industries."

Assignment

Your review should be written in the form of a detailed report: *"Response to the analysis presented in the REPORT,"* which addresses the concerns you should (!) have regarding the analysis and conclusions presented in the plaintiff's REPORT. Although the situation is clearly fictitious, there are certainly some very "real" factors that can be associated with the decision to grant pay raises to employees. These factors should also be considered in your review.

9.6 Bentley's Revenge

Case Study 6.3 Continued ...

In Case Study 6.3, we investigated a manufacturer's claim that placing its device on an engine would result in a steady decrease in hydrocarbon and carbon monoxide emissions, and an increase in carbon dioxide emissions. In this case study exercise, we present a second set of experiments that was conducted using one of the Bentleys owned by StatCon Enterprizes.

The Bentley Experiment

- The car was operated under four different road conditions on each of 14 different days. The road conditions were:

  Condition 1: Moderate engine load at 55 mph.
  Condition 2: High engine load at 35 mph.
  Condition 3: Low engine load at 55 mph.
  Condition 4: Hill climb at 145 mph.

- On each test day, the experimental conditions were repeated three times.

- The device was placed on the engine after test day 6.

- Hydrocarbon (ppm), carbon dioxide, and carbon monoxide (% of volume) emissions were measured using an instrument that was recalibrated each day measurements were taken.

The Data

There were some measurements that could not be made due to road construction that are denoted by `"NA"` in the dataset `c96.dat` which contains the results from this experiment. The entries in each column are indicated by the column headers. The test days were not contiguous and the car was driven for approximately 600 to 4000 miles between tests. Tests began in December and were completed in July of the following year. We should also note that the magnitude of the measured emissions are, for the most part, less than the accuracy of the measuring instrument.

Assignment

1. Based on data gathered from tests using the manufacturer's car and your analysis of the above experiment, is there evidence to support the manufacturer's claims?

2. The variability in the emissions measurements could well be due to several other factors that were not controlled in these experiments. How would you have designed the experiment?

9.7 Wear What You Like?

Description

The main focus of this study was to examine whether the interaction between student and teacher differed according to the type of clothing worn by a student. The observational experiment consisted of the client observing several classes taught by one teacher and counting the number of interactions that each student had with the teacher. An interaction was classified as either "Positive" or "Negative" according to the client's judgment of the teacher's response or reaction to the student. The interaction `Counts` were crossclassified by the student's `Gender` and `Clothing` type.

The Data

The dataset `c97.dat` contains the results from this observational experiment. In `c97.dat`, two lines are employed for each student with the number of positive interactions listed first. Missing values are denoted by a period `"."` which can be interpreted as a zero count. The levels of each variable

are defined in the table below, in the order in which the variables appear in the dataset `c97.dat`.

Variable and Level Definitions

`Clothing`	Type of Clothing Worn by the Student
`Unisex`	Unisex Clothing
`Std`	Standard Gender-Specific Clothing
`Other`	Unusual Clothing
`Gender`	Gender of Student (Male/Female)
`Count`	Number of Interactions Observed
`Interaction`	Type of Interaction (from Teacher's Perspective)
`PosInt`	Positive Interaction Between Student and Teacher
`NegInt`	Negative Interaction Between Student and Teacher

Assignment

How should we analyze these "count" data? A simple approach would be to consider the difference between the number of positive and negative interactions and see if there is any effect due to `Gender` or `Clothing`. Alternatively, a multivariate approach could be employed where the positive and negative interactions represent two separate responses.

9.8 An AIDS Study

Description

Measuring the "cell count" of particular cells provides an effective means of monitoring patients who are affected by the AIDS virus, or have diseases such as cancer or hepatitis. For someone who is HIV-positive, two important diagnostics are their CD4 and CD8 cell counts. CD4 are white blood cells that the AIDS virus uses as a host to reproduce itself. Hence CD4 cell counts provide a key indicator of a person's immune system status:

Below 200	Full-blown AIDS
200 to 500	Intermediate stage
Above 500	Sound functioning immune system.

CD8 cells help suppress the infectiousness of the virus by killing cells the body decides are foreign. Unfortunately, the AIDS virus itself evades detection by residing within the CD4 cell. However, the CD8 cell count provides a measure of the person's ability to fight off other infections. Another measure of the viral "load" carried by a person is their RNA count. This is a single strand of DNA which the AIDS virus uses to reproduce itself.

The Study

The purpose of this study was to see if the three measures: CD4, CD8, and RNA counts, provided discrimination between two groups of couples classified as *Discordinant* (DP: only one partner HIV-positive) and *Concordinant* (CP: both HIV-positive). Only one partner from each couple was included in the study, with the infected partner being measured in the DP group. This provided a more homogeneous cohort and to eliminate confounding effects, drug users and nonmonogamous couples were excluded.

Assignment

The results from this study are provided in the dataset `c98.dat` which includes self-explanatory column headers. In this analysis, a transformation of the predictor variables may help. Interpret the results of your analysis for a client who is not familiar with the methods you employed.

Appendix A
Resources

An important skill that any statistical consultant needs to develop is the ability to locate information and resources. This will often be a reference to some particular statistical procedure with which the consultant is unfamiliar (or has never used before). However, it may also be totally unrelated to any statistical issue. What's the best way to set up an accounting system? A client's database is in a format or application that the consultant doesn't have access to, or the file(s) containing the data appear to be corrupted. So who does the consultant consult for help?

A.1 References

For the statistician, the Internet has become an important resource. For the statistical consultant, it has become essential. However, the information provided through the Internet is not complete; the consultant will still need to rely on hardcopy resources such as books. Hence the following list of journals remains a useful resource for a consultant.

Journals

JRSS The Journal of the Royal Statistical Society publishes four series: A, B, C and D. Series A has good book reviews and is concerned with statistics in society issues. Series B publishes articles on statistical

methodology. Series C is actually called *"Applied Statistics"* so the title is self-explanatory. It often has case studies articles. Series D (*"The Statistician"*) contains articles of general interest.

ASA The American Statistical Association publishes several journals including *JASA, Journal of the American Statistical Association,* which has a mixture of theory and applications, as well as book reviews. Like most journals, the quality and diversity of the articles depend on the editor. *The American Statistician* publishes articles for a general audience, as well as software (and book) reviews. *Technometrics,* published jointly with the American Society for Quality Assurance, focuses on the use and application in science and engineering fields. Two other notable journals that contain articles of interest to the statistical consultant are: *Journal of Business and Economic Statistics* and the *Journal of Computational and Statistical Graphics* (published jointly with the Institute of Mathematical Statistics (IMS) and the Interface Foundation).

IMS Publishes mainly theoretical articles in *The Annals of Statistics* and *The Annals of Probability.* Conferences, employment opportunities, and other news is provided in the *IMS Bulletin.* Articles of interest to the statistical consultant are more likely to be found in *The Statistician.*

Other Journals This is a rather cheap way out to say that there are many other journals worth reading, some of which are not within the traditional confines of statistics. For example, we mention: *Biometrika, Biometrics, Statistics in Medicine, Journal of Marketing, International Journal of Forecasting, IEEE Journal of Automatic Control.* ... Time to go to the Internet.

Internet The Internet has become an important resource for everybody. Originally developed for the U.S. military (Arpanet) and then passed onto academia, the Internet has now become an important public resource. Of the many sites we could mention for statistics, one of the best maintained is StatLib. This site has a wealth of information.

`lib.stat.cmu.edu` This is StatLib which is maintained by Carnegie Mellon University and fully or partially mirrored at certain international sites. As indicated above, this has a wealth of information which will be of interest to a statistical consultant. Here is a brief list of some of the resources included on StatLib.

- Other Sites: Contains international Website addresses.
- DESIGNS: A collection of designs and programs and algorithms for creating designs for statistical experiments.

- Directory of People: Lists of addresses and email addresses of statisticians.
- DOS/Windows: DOS and Windows software, available only via FTP and WWW.
- General Archive: Contains a variety of software written in FORTAN, C, and Lisp, some complete statistical systems, and other odds and ends.
- Genstat: Software and macros for the Genstat language.
- Glim: Macros and software relating to the GLIM package.
- JASA Software: This collection contains software related to articles published in the *Journal of the American Statistical Association.*
- JCGS: The jcgs archive contains contributed datasets and software and abstracts from articles published in the Journal of Computational Graphics and Statistics.
- MacAnova: The MacAnova statistical system, for Mac, Windows, and UNIX.
- Maps: A world map in a compressed tar file of over 4 Mb. Available only via FTP (use *binary* mode).
- Meetings: Calendars and programs for various domestic and international meetings.
- Minitab: Minitab Industrial Statistical macros and macros from the Minitab Users Group.
- Multivariate: Various routines for multivariate analysis and classification. Look in the general collection for other multivariate techniques.
- PoliSci Data: Data from political science journals and authors.
- R: 'GNU S' — A language and environment for statistical computing and graphics.
- S Archive: Software and extensions for the S (S-PLUS) language. Over 130 separate packages including many novel statistical ideas.
- Xlispstat: Luke Tierney's XlispStat system for UNIX systems.

`stat.fsu.edu` World Wide Web Virtual Library: Statistics. Maintained by the University of Florida Department of Statistics.

Government Sites The U.S. and other countries maintain Websites that provide "governmental" statistics on many aspects of society. For information about U.S. agencies go to: `www.fedstats.gov`. Here is a list of a few international government sites.

www.abs.gov.au	Australian Bureau of Statistics
www.info.gov.hk	Statistics Department — Hong Kong
www.bps.go.id	Bureau of Statistics — Indonesia
www.jbs.agrsci.dk	Danish Institute of Agricultural Science
www.dos.gov.jo	Department of Statistics — Jordan
www.statistics.gov.my	Department of Statistics — Malaysia
www.isical.ac.in	Indian Statistical Institute
www.ine.es	Instituto Nacional de Estadistica — Spain
www.cso.ie	Irish Central Statistics Office
www.istat.it	Italian National Institute of Statistics
www.stat.gov.jp	Japanese Statistics Bureau
www.ine.gov.bo	National Statisics Institute — Bolivia
www.nso.gov.kr	National Statistics Office — South Korea
www.nectec.or.th	National Statistical Office — Thailand
www.gks.ru	Russian Federation on Statistics
www.statcan.ca	Statistics Canada
www.dst.dk	Statistics Denmark
www.std.lt	Statistics Department — Lithuania
www.statgreen.gl	Statistics Greenland
www.cbs.nl	Statistics Netherlands
www.singstat.gov.sg	Statistics Singapore

Software Statistical consultants need statistical software. Where better to find it but on the Internet. Here is a list of some sites for statistical software:

www.cytel.com	CyTEL Software Corporation StatXact, LogXact, EaSt
www.datadesk.com	Data Description, Inc. Data Desk, Vizion, ActivStats
www.statgraphics.com	Statgraphics
www.mathsoft.com	Mathsoft, Inc. — **S-PLUS**
www.minitab.com	Minitab, Inc.
www.nag.co.uk	Numerical Algorithms Group Genstat, GLIM, NAG Libraries
www.sas.com	The **SAS** Institute
www.spss.com	SPSS

A.2 Datasets for Case Studies in Part II

The datasets used in the case studies of Part II can be accessed via the Springer-Verlag or the authors' Websites:

`www.springer-ny.com`

`www.rci.rutgers.edu/~cabrera`

`www.csam.montclair.edu/~mcdougal`

The datasets are named according to the subsection in which they appear (e.g., `"c64.dat"` is the dataset from Case Study 6.4). Additional information on SAS and S-PLUS code is provided.

The authors have also provided additional information such as specific ESL advice which was not included in this book.

A.3 Statistical Consulting Course

A.3.1 Course Description

OBJECTIVE The creation of a new graduate course on statistical consulting. Students will be exposed to realistic statistical and scientific problems that appear in typical interactions between statisticians and scientists. The lectures will be centered around case studies presented by invited speakers.

AUDIENCE Second-Year Graduate Students. The student should be familiar with computing and applied statistical methodology.

PREREQUISITES This depends on the level of sophistication possessed by the students and/or desired by the instructor. A possible list of prerequisites are:

Level 1 For first-year graduate students, the emphasis of the course can be placed on the communication aspects of statistical consulting. Analysis of case studies would necessarily be restricted to the level of statistical knowledge provided at the institution's undergraduate (senior) level.

- Exploratory data analysis.
- A computing course, where they would get the computing and data analysis base.
- Simple ANOVA, Regression.

Level 2 For a course designed for second year students, the statistical methodology can be extended to include:

- Design of experiments.

- Regression and GLM.
- Methodology or theory course. Theory of statistical computing basics.
- Statistical computing: SAS, S-PLUS (or equivalent) basics.

Level 3 For research-level doctoral students, this type of course can be used to ensure that these students are provided with the type of skills that will make them more attractive propositions for employment after graduation. Specifically, good communication skills are important.

- Exploratory data analysis (A review ... perhaps?).
- GLM, multivariate, time series, and categorical models.
- Statistical computing with SAS and S-PLUS.

CLASS SIZE Maximum of 12 to 15 students per class, divided into three to five consulting teams. We have observed that students often prefer to work in pairs but this clearly defeats the purpose of interacting as part of a team. We would strongly encourage instructors to maintain "team" sizes of at least three.

ASSIGNMENTS Consulting teams are given an assignment for every case study, with the full dataset, and they must produce a report and in most cases, give a presentation. For every case study the teams have a leader who will give the presentations and write the report. The position of team leader will rotate for different case studies.

CASE STUDIES It is of great importance to identify good case studies since they are the centerpiece of the course. We have decided to consider case studies from the following groups.

GROUP I Simple case studies where the answer is well known, but with some interesting statistical and scientific issues to be discussed.

Case studies of Group I will be presented in a one-hour lecture followed by a discussion. As a result of the discussion each team is assigned a project. The data for all case studies must be available to the student.

The following week the students will present the team reports.

GROUP II More complicated case studies where the statistical problem is generally well defined, but broader in scope than the Group I case studies. Several solutions may need to be evaluated. The complexity may also be partly due to the size or format of the dataset.

Case studies of Group II will be presented in similar format to Group I above.

Preliminary results to be presented in the following week. Discussion and questions addressed. Final presentation in the second week.

GROUP III Research-oriented case studies where it is hard to pick up the statistical problem but where the students have to do a lot of thinking and the answer may or may not be well known. Several stages of analysis may be needed to obtain suitable results. There may not necessarily be an "answer" to the statistical problem.

The following week the students will present the preliminary reports and continue the discussion of the principal issues and come up with the final objectives.

In the second or third week they will do the final presentations.

A.3.2 List of Topics by Week

Week 1 Introduction and organization of the course. The students will form consulting teams that will work together in the preparation and presentation of several reports involving the analysis of case studies.

History of science and the role of statistics. Introduction to the scientific method. Statistical consulting environments. The role of the statistician within a scientific environment.

Text section: *Chapter 1.*

Lecturer: Instructor.

Week 2 Communicating with researchers from other areas. Report writing.

Text section: *Chapter 2.*

Lecturer: Instructor.

Week 3 Methodological aspects. A review of statistical methods that will be used in the course.

An overview of computational tools and statistical software such as SAS and S-PLUS that are available to the students.

Text section: *Chapter 3, Appendix B.*

Lecturer: Instructor.

Week 4 Case Study 1 (from Group I).

Introduction to case studies. Description of the procedure that will be followed for case studies. Format for report writing and presentations by students.

Presentation of Case Study 1 by invited speaker. Discussion.

Assignment of projects to the student teams.

Text section: *Chapter 4.*

Lecturers: Instructor and invited speaker for Case Study 1.

Week 5 Case Study 2 (from Group I).

Presentation by consulting teams of reports from Case Study 1 with videotaping.

Presentation of Case Study 2 by invited speaker. Discussion.

Review of videotape. Discussion.

Assignment of projects to the student teams.

Text section: Review *Chapters 2,3,4.*

Lecturer: Invited speaker for Case Study 2.

Week 6 Case Study 3 (from Group I).

Presentations by consulting teams of the reports from Case Study 2. Discussion.

Assignment of Case Study 3. Projects to the student teams.

Review of reports submitted and presentations. Discussion of communication skills.

Text section: *Chapter 6.*

Lecturer: Instructor.

Week 7 Case Study 4 (from Group II).

Presentations by consulting teams of the reports from Case Study 3.

Presentation of Case Study 4 by invited speaker. Discussion.

Assignment of projects to the student teams.

Research articles and reports from other case studies assigned to students. Discussion on reading about other studies.

Text section: *Appendix A.*

Lecturer: Invited speaker for Case Study 4.

Week 8 Case Study 4 (from Group II).

Progress reports by consulting teams. Preliminary presentations of their initial analyses for Case Study 4.

Discussion, progress evaluation, and revision of the objectives of the projects. Collect research article reports.

Lecturer: Instructor

Week 9 Case Study 5 (from Group II).

Presentations by consulting teams of the reports from Case Study 4.

Presentation of Case Study 5. Discussion of Case Study 5. Review of article reports.

Text section: *Chapter 7.*

Lecturer: Instructor.

Week 10 Progress reports by consulting teams. Preliminary presentations of their initial analyses for Case Study 5.

Discussion, progress evaluation, and revision of the objectives of the projects.

Text section: *Chapter 7.*

Lecturer: Instructor.

Week 11 Case Study 6 (from Group III).

Presentations by consulting teams of the reports from Case Study 5. Second videotaping.

Invited speaker presentation of Case Study 6. Discussion.

Review of videotape. Self-critique of presentations. Assignment of projects to the student teams.

Text section: *Chapter 8.*

Lecturer: Invited speaker for Case Study 6.

Week 12 Case Study 6 (from Group III).

Progress reports by consulting teams. Preliminary presentations of their initial analyses for Case Study 6.

Discussion, progress evaluation, and revision of the objectives of the projects.

Text section: *Chapter 8.*

Lecturer: Instructor.

Week 13 Case Study 7 (unknown difficulty).

Presentations by consulting teams of the reports from Case Study 6.

Presentation of Case Study 7. Deadline: one-week turnaround. Discussion to be led by students.

Assignment of projects.

Text section: *Chapter 9.*

Lecturer: Instructor.

Week 14 Case Study 7.

Presentations of final reports for Case Study 7.

Turn in reports for final evaluations. Order pizza.

Open discussion.

Lecturer: Instructor.

Week 15 After Week 14 (optional). Real life experiences. Limited internships, or consulting projects.

A.3.3 Reference List

GENERAL: Additional texts that could used to supplement this book.

1 J. Derr (2000), *Statistical Consulting: A Guide to Effective Communication,* Duxbury Press, Pacific Grove, CA.

2 C. Chatfield (1995), *Problem Solving. A Statistician's Guide,* (2nd ed.) Chapman & Hall, London.

3 D. J. Hand and B. S. Everitt (Eds.) (1987), *The Statistical Consultant in Action,* Cambridge University Press, Cambridge.

4 J. Tanur (Ed.) (1988), *Statistics: A Guide to the Unknown,* (3rd ed.), Duxbury Press, Belmont, CA.

5 S. Conrad (Ed.) (1989), *Assignments in Applied Statistics,* John Wiley, New York.

6 D.R. Cox and E.J. Snell (1981), *Applied Statistics, Principles and Examples.* Chapman & Hall, London.

BIOSTATISTICS: Texts specifically dealing with biostatistical applications.

7 R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses (Eds.) (1980), *Biostatistics Casebook,* John Wiley, New York.

PHARMACOSTATISTICS: Texts specifically dealing with pharmacostatistical applications.

8 K. E. Peace (Ed.) (1988), *Biopharmaceutical Statistics for Drug Development,* Marcel Dekker, New York.

DATA: Some sources of data are available from:

9 D. F. Andrews and A. M. Herzberg (1985), *Data: A Collection of Problems from Many Fields for the Student and Research Worker,* Springer-Verlag, New York.

Appendix B
Statistical Software

B.1 SAS

SAS is a powerful statistical analysis package which is used extensively by many large (and not so large) companies, particularly pharmaceuticals. However, SAS does have a rather specialized language structure that can make data manipulations something of an adventure not recommended for the faint-hearted!

B.1.1 The SAS Setup

In the execution of a SAS session several files are involved. If we are running SAS interactively, then each of the following files will appear in a separate window.

prog.sas: The "program" file created by the user that contains SAS statements which, if successfully processed, will produce the output file `prog.lst`. The extension "`.sas`" is needed on most systems, but the name "`prog`" is our choice.

prog.lst: The "output" file which may **not** contain all the results we wanted because an error occurred at some point. To locate the error examine the `prog.log` file.

prog.log: The "error" file. It is really a log of what happened as SAS executed each statement in the program file `prog.sas`. It should **always**

be checked during the data input stage since SAS will continue to read a data file even though it may have encountered errors.

SAS Program Example

The following example illustrates some basic SAS statements that would all be contained in the `prog.sas` file.

```
options ls=78 ;
filename in1 'example.dat';
data a;
      infile in1;
      input x y ;
* proc print ;
data b ;
      set a ;
      if y > 10 ;
proc means ;
run ;
endsas ;
```

Note that the structure of a SAS program is actually quite simple. It consists of two basic steps which are employed as often as required by the user to perform the desired analysis:

DATA Step: Process and perform manipulations on the dataset.

PROC Step: Apply a specified procedure to the dataset.

These "steps" must conform to certain rules and each statement must follow the syntax required by SAS. As in the above example, the following features are generic to any SAS program.

syntax: All statements must end in a semicolon ; .

comments: Text (or statements) that appear between * <text> ; are treated as comments and will be ignored by SAS during execution.

options: Options may be specified to control the output to `prog.lst`. Here `ls=78` restricts the output to 78 characters per line.

filename: Tells SAS where to find an external data file.

endsas: Tells SAS to stop executing. If we are running SAS interactively, do **not** use this statement or the entire SAS session will terminate. (Use a final `run ;` statement instead.)

Now we can describe the particular analysis performed by this program.

1. `example.dat` is an external data file that contains two columns of numeric data separated by at least one white space (i.e., blank space).

2. This data file is assigned a *fileref* called `"in1."` The actual name is up to the user and hence more than one external file may accessed by using a different *fileref* in successive `filename` statements.

3. `data a ;` creates a "working" dataset called `"a"` (user's choice) which is how SAS will process the information in `example.dat`.

4. `infile in1 ;` sends SAS off to try to access the external file. If successful, then

5. `input x y ;` tells SAS how to read `example.dat` and what variable names to assign to each column.

   Note: If a column contains character data, the symbol $ must be used immediately after the variable name: e.g., `input x $ y ;` indicates that the `x` column contains character values.

6. `* proc print ;` will be ignored by SAS. If the * is removed this procedure will print the contents of the most recently created SAS dataset (in this case `"a"`) to the file `prog.lst`.

7. `data b ; set a ; if y > 10 ;` creates a new SAS dataset called `"b"` which starts with the contents of `"a"` (statement: `set a ;`) but only selects the rows of `"a"` for which `y > 10`. The values of `x` corresponding to these values of `y` will also be contained in the SAS dataset `"b"`.

8. `proc means ;` applies the MEANS procedure to the most recently created SAS dataset, in this case `"b"`. To apply this procedure to `"a"` the statement `proc means data=a ;` is required. There are additional statements and options that can be specified for this procedure. In the default form as given here, SAS will compute the mean, min, max, number of observations, and standard deviation of all numeric variables contained in `"b"` and print the results to `prog.lst`. SAS takes care of presenting the results in a suitable format.

B.1.2 Details on the DATA Step

In the above example, reading `example.dat` into SAS was made easy because it was in column format with one or more white spaces separating the columns. Clearly then, the amount of time and effort we need to spend on the `input ;` statement will be minimal if the external data file is already in column format. This applies even if the columns contain special character

values, dates, or times, since SAS provides a wide range of modifiers to deal with these formats. Hence:

Always try to prepare external data files in column format.

Of course, we will not always be this fortunate and now the time and effort that we need to spend can increase dramatically. However, provided the data file has some type of regular format, we should be able to avoid having to reformat the data file directly. The following examples illustrate some situations where a more complex input statement is required.

Example B.1 *Trailing @@*

```
input trt $ x1 x2 x3 y @@ ;
```

Read more than one observation (sequence) from a line. Here, `trt` is a treatment group (character), `x1 x2 x3` are three design variables, and `y` is the response variable. Data file format would be of the form:

```
placebo 0 1 0  23.5 drugA 0 1 1 13.4 drugB 1 0 0 11.4
placebo 1 0 1  33.1 drugA 1 0 0 12.7 drugB 1 1 1 16.9
...
```

Example B.2 *Fixed format*

```
input x 4-7 @15 y  ;
```

Read the X value from columns 4 to 7 then skip to column 15 to read the corresponding Y value.

Example B.3 *Multiple inputs*

```
if(mod(_N_,11) = 1) then
   do ;
      input group $ ;
      retain group ;
   end ;
else
   do ;
      input  qlabel :$15. answer ;
      output ;
   end ;
```

Use the internal observation counter `_N_` (provided by SAS) to read survey data consisting of 10 questions where the (numeric) responses `answer` are preceded by a (question) label which can be up to 15 characters long `qlabel :$15.` Although a respondent is classified by a `group` variable, this entry only appears at the beginning of each set of responses and hence the need for the `retain group ;` statement.

Example B.4 *Read anything*

```
input tmp $ @@ ;
```

Read every entry as a character value! The point here is that it is sometimes easier to read data entries as character variables and then apply the appropriate conversion.

SAS Errors

In most cases, syntax errors can be quickly resolved by examining the `prog.log` file since SAS will stop processing the DATA or PROC Step at the point where the error occurred. The `prog.log` file is also an important source for locating where a data file is not being read properly. The difficult errors to resolve are usually where the user wants to perform some type of data manipulation, but SAS keeps producing unexpected results. This can be particularly troublesome when the user wants to access or operate on certain row components of a variable. In SAS there is no direct access to the "ith element," $X[i]$ say, of a variable X. Dexterous use of several DATA Steps may be required to subset rows of a variable.

B.1.3 SAS Procedures

The PROC step essentially consists of applying a presupplied SAS procedure to a SAS dataset. That is, we must use the options available within a SAS procedure to perform a particular analysis. Fortunately, the range of options is extensive and there exist general procedures such as GLM and CATMOD which duplicate the analyses of several specialized procedures. For example, GLM and ANOVA can both be employed to analyze balanced designs, but GLM is needed when the design is unbalanced. The following provides a summary of some useful SAS procedures and options.

EDA

`FREQ` Multiway contingency tables and associated statistics.

```
proc freq  ;
tables a a*b / exact norow nocol nopercent ;
```

Produce a frequency table of variable A and the contingency table of $A \times B$ with Fisher's exact test computed. Suppress printing the row, column, and cell percentages. The default is to print these percentages.

`MEANS` Means, standard deviations, and other summary statistics for numeric variables.

```
proc means n mean std stderr t prt ;
var x1-x5 ;
output out=mdata mean=m1-m5 ;
```

Compute specified summary statistics and perform the t-test $H_o : \mu_i = 0$ for the five numeric variables X_i, $i = 1, 2, 3, 4, 5$. Note that `x1-x5` can be used to represent the sequence `x1 x2 x3 x4 x5`. The last statement outputs the five sample means to the SAS dataset `mdata`.

Other: `CHART, PLOT, PRINT, TIMEPLOT`.

Simple Statistics

`UNIVARIATE` A more detailed summary of numeric variables including tests of normality, other distributional properties, and diagnostic plots. The plots tend to be of limited value.

```
proc univariate ;
vars x1-x5 ;
by group ;
```

Univariate summary of $X_1, \ldots, X_5$ at each level of `group`.

`TTEST` Two-sample t-test. A two-level `class` variable is required. To perform a paired t-test, we first need to create the differences (in a DATA step) and then use `proc means` above.

```
proc ttest ;
class group ;
var x1-x5 ;
```

`CORR` Cronbach's alpha (Chapter 3, *Correlation*) can be obtained from `proc corr` as shown below. The `with <variable list>;` statement is optional. If omitted, all pairwise correlations are produced by the `var <variable list> ;` statement.

```
proc corr data=a alpha ;
var x1-x5 ;
with y1-y4 ;
```

Other: `TABULATE, SUMMARY`.

Regression

`REG` Fits a linear regression model. All regressors to be included in the model must be predefined (e.g., $X_{12} = X_1 * X_2$).

```
proc reg ;
model y = x1 x2 x12 / all ;
output out=rdata p=yhat r=resid ;
```

The option `all` requests that additional statistics be printed. Output of specific statistics to a SAS dataset is available via `keyword = name`. Here, `rdata` is the dataset name, and `p= r=` are keywords that will output the predicted values and residuals from the model into the variables `yhat` and `resid`, respectively.

`RSREG` Fits a response surface regression and performs a canonical (eigenvector) analysis of the surface. The following model statement is equivalent to the regression model specified above.

```
proc rsreg ;
model y = x1 x2 ;
```

Other: `CALIS, NLIN, ORTHOREG, TRANSREG.`

Experimental Design

`ANOVA` For balanced designs, ANOVA is more efficient than GLM but we may need to use GLM if we require certain statistics to be output for further analysis. The user is also responsible for performing an appropriate analysis using ANOVA when *random* effects and nonstandard F-tests are involved.

```
proc anova ;
class group trt ;
model y = group trt group*trt ;
means group trt / t lines ;
```

Two-way ANOVA with interaction. Again, the user is responsible for interpreting the output correctly. Multiple comparisons using pairwise t-tests with significant differences displayed by the `lines` option.

`GLM` This procedure can be used to fit general linear models and therefore essentially duplicates the analyses of special cases such as regression (REG) and balanced designs (ANOVA). However, GLM has certain features that are not available in these specialized procedures and in particular, is designed to handle unbalanced data.

```
proc glm ;
class group trt ;
model y = group trt group*trt ;
lsmeans group trt ;
means group trt ;
random group group*trt / test ;
test h=group e=group*trt ;
output out=gdata r=resid ;
```

In this example `group` is declared as a random effect (GLM does not automatically declare interactions as random) and `lsmeans` provides the equivalent of `means`, had the design been balanced. The `test` statement (also available in ANOVA) enables the appropriate error term to be used for testing `group`. Note that the appropriate F-tests for a random effect can also be requested via the option `/ test` in the `random` statement. Lastly, the

output statement enables the residuals from this model to be examined further.

Other: MIXED, NESTED, NPAR1WAY, PLAN, VARCOMP GLMMOD, LATTICE, MULTTEST.

Categorical Data

LOGISTIC When the response variable is binary or ordinal, a linear regression model based on the *logit* function $g(p) = p/(1-p)$ where $p = P[Y = 1|X]$, may be appropriate. That is, a model of the form: $g(p) = \alpha + \beta' X$.

```
proc logistic outest=beta ;
model y = x1 x2 / ctable details ;
output out=ldata p=phat ;
```

The format of the model statement is similar to REG (meaning that all explanatory variables specified in the model statement must be predefined). The options ctable details are specific to the logistic model.

CATMOD This procedure provides the general-purpose equivalent of GLM for analyzing categorical data. A wide variety of categorical models may be fitted including logistic (above) and log-linear models. The following example fits a log-linear model to the contingency table determined by $A \times B$.

```
proc catmod ;
response marginals ;
model a*b = _response_  ;
```

Other: CORRESP, GENMOD, PROBIT.

Multivariate Methods

CLUSTER, TREE, ACECLUS, FASTCLUS, MODECLUS, DISCRIM, CANDISC, PRINCOMP, CANCORR, FACTOR, CALIS, MDS.

There is quite an extensive range of multivariate procedures provided by SAS and the SAS/STAT manual (SAS 1990) is clearly the best place to learn the specific details and syntax associated with these procedures. (In Case Study 8.4, the use of proc cluster is illustrated.) Some readers who are unfamiliar with SAS may have noticed that no MANOVA procedure was listed. This is because MANOVA models are specified within the GLM (or ANOVA and MIXED) procedure. The specification of MANOVA analysis is somewhat cryptic, so it is important to follow the syntax carefully. We should also point out that SAS tends produce "volumes" of output from multivariate procedures. Be prepared to do some extensive reading!

Time Series *Requires SAS/ETS software.*

`ARIMA` The Box–Jenkins approach to fitting ARIMA(p, d, q) models (including the seasonal version) is employed in the ARIMA procedure. The differencing, parameter estimation, and forecasting steps are specified separately so several passes are usually required to obtain an appropriate model.

```
proc arima ;
i var=y(1,12) ;
e p=(1)(12) q=2 plot;
f out=adata lead=12 id=t ;
```

The above example fits the multiplicative seasonal model, denoted by $(1,1,0)_{12} \times (1,1,2)$, and uses this model to forecast 12 time periods ahead. Note that `q=2` is shorthand for `q=(1,2)` and the option `plot` produces correlograms and portmanteau tests of the residuals.

Other: `SPECTRA, STATESPACE, X12`.

Additional Systems As indicated above, the procedures for time series analysis are **not** included in the SAS/STAT software module. To use `proc arima` therefore, the SAS/ETS module must be installed on the user's system. While there are many other specialized SAS products available, we confine our attention to the QC and GRAPH modules.

`QC` The SAS/QC software provides a variety of quality control procedures that would be useful to a statistical consultant working on projects where quality assurance plays an important role. The following provides an example of a standard Shewhart $\overline{X}$-R run chart on the dataset `spray`, where `batch` denotes the samples taken on the response measurement `ingred` and `tests=1 to 8` performs several run chart diagnostic tests.

```
proc shewhart history=spray graphics;
xrchart ingred*batch /
   tests=1 to 8
   tableall
   zonelabels;
```

`GRAPH` Presentation quality graphics are an important part of any respectable statistical software package and the SAS GRAPH system provides this utility. While it is certainly possible to obtain high resolution graphics via the "G" PROCs (e.g., GPLOT,

GCONTOUR, GCHART, G3D) these procedures, in our opinion, do not compare to the ease and flexibility provided by the S-PLUS software (discussed in the next section). With high-resolution graphics the particular graphics *driver* needs to be identified and in the following example, two possibilities are indicated by `dev=XCOLOR` (suitable for interactive use of SAS on a UNIX workstation) and `dev=PS` (creates a PostScript file). Only one `goptions` statement should be used in practice, of course. The remainder of the example creates a scatter plot with a simple linear regression line and 95% confidence limits overlaid on the plot. Titles, footnotes, different fonts, colors, and symbols may be used to enhance the plot.

```
goptions dev=XCOLOR ; * <-- for interactive use ;
                      *     on UNIX workstation ;
filename gout1 '~/gexample.ps' ;
goptions gsfname=gout1 gsfmode=append dev=PS ;

filename in1 '~/datafilename' ;

data a ;
    infile in1 ;
    input x1-x4 ;
    rename x4=y ;

title1 f=xswiss 'Example of GPLOT for SLR' ;
title2 'SAS Graphics' ;
footnote '95% Confidence Interval Overlaid' ;

proc gplot data=a ;
      plot y*x3 ;
symbol1 c=red i=rlclm v=diamond ;
run;
```

`ASSIST, INSIGHT` In addition to the numerous manuals and texts published by SAS, there is also an interactive help system that provides details on all SAS procedures. This user interface software is available in SAS/ASSIST module. If implemented, access to SAS ASSIST will be found in the "Options" pull-down menu. Finally, we should also mention that there is a menu-driven version of SAS called SAS INSIGHT.

B.1.4 Further Details of SAS

Specification of Effects

SAS uses the following notation for specifying crossed and nested effects in analysis of variance procedures.

Crossed `A*B` Each level of A is observed at each level of B.

Nested `B(A)` The levels of B are *nested* within each level of A.

Bar notation `A|B` is shorthand notation for: `A + B + AB`.

Permanent SAS Datasets

For small- to moderate-sized datasets, reading in an external datafile every time we execute the `prog.sas` program is unlikely to be much of a concern. For large datasets, however, this is rather inefficient and a *permanent* SAS dataset should be created via the `libname` statement. The following example illustrates this for SAS running under UNIX.

```
options ls=78 ;
libname scp '~/sasuser/job1' ;
filename in1 'example.dat' ;
data a;
   infile in1 ;
   input x y ;
data scp.xydat ;
   set a ;
run ;
endsas ;
```

The directory `~/sasuser` should already have been created by SAS and we may create subdirectories such as `/job1` to separate projects. The *libref* `scp` directs SAS to that particular subdirectory and thus, in the second DATA step, a *permanent* SAS dataset called `xydat.ssd01` is created there (the extension is provided by SAS). Since this permanent dataset contains all the information that SAS requires, it suffices to access that file for analysis purposes. That is, the following `prog2.sas` program file can now be used to analyze the data directly.

```
options ls=78 ;
libname scp '~/sasuser/job1' ;
data a ;
   set scp.xydat ;
proc means ;       * The catch is that we need to know ;
   var x y ;       * what variables exist in "xydat".  ;
endsas ;           * Use "proc contents" to find out.  ;
```

Transporting SAS Files

Transporting permanent SAS datasets between computers can be more efficient than starting from the raw data. Often, a certain amount of data cleaning was performed prior to creating a permanent SAS dataset and reprocessing the raw data may not be worthwhile or even feasible.

In some cases, it may be possible to transport SAS datasets directly via a communications network. In other cases, the SAS dataset must be rewritten in so-called *transport format* or Export form. Depending on the operating system, transport formats may be created via a DATA step or one of the procedures PROC COPY, XCOPY, or CPORT with appropriate options such as `export`. To read a SAS dataset in transport format, use one of the above with the `import` option, or PROC CIMPORT if necessary. With large SAS datasets, the Export version may need to be uncompressed first.

SAS Macros

SAS provides a MACRO facility which allows us to build programs that are specifically designed to carry out a procedure or analysis that we expect to use often. The key point here is that a SAS macro needs to be used frequently enough to make the effort worthwhile. This is often the case in business and research consulting environments that deal with similar datasets, but usually this will not be the case for one-off clients. We have not used any SAS macros in this book.

B.2 S-PLUS

S-PLUS evolved from the S programming language software created by Becker and Chambers (1984), Becker et al. (1988) at AT&T Labs. Thus, while S-PLUS provides more statistical functionality than S does, its underlying approach remains program-oriented. One of the main strengths of S-PLUS is that it can be used effectively in an interactive environment. In addition, S-PLUS provides the user with excellent graphics capabilities. However, the price of this flexibility is that the user may need to do a lot of S-PLUS "programming" in order to perform a complete statistical analysis.

B.2.1 S-PLUS Preliminaries

S-PLUS is a high-level object-oriented application and once the user obtains a good understanding of some basic S-PLUS commands, building our own functions is straightforward and opens up a rich and extendable environment for data analysis. That is, the user is not constrained by the limitations of built-in S-PLUS functions, but since there is an extensive collection of these functions, the user is not subjected to dealing with the

intrinsic algorithms (C code) underlying these procedures. Before considering the art of creating our own functions, we first need to know two important commands.

Starting and Ending S-PLUS

To invoke S-PLUS from a UNIX[1] prompt type:

```
Splus
```

If this is our first time, the special directory `"/.Data"` will be created and a message saying as much. Then the prompt: `>` appears. We are now in interactive mode and can happily type in S-PLUS commands. To quit from S-PLUS type `q()` beside the > prompt:

```
> q()
```

Of course, we would probably like to do a little more than just start and end an S-PLUS session! Thus, we begin by covering some basic S-PLUS commands with reference to the following dataset which is assumed to exist as the file `"ex.dat"` on our UNIX home directory.

The `ex.dat` *Dataset*

The file `ex.dat` contains the entries (including column labels):

name	class	height	weight	size
clm	1	8	10	104
gag	2	1	3	11
gam	2	9	11	120
grc	1	2	4	12
mra	1	6	11	127
rmm	3	6	6	33
spm	3	6	7	52
tcs	3	10	12	150

Data Input

After invoking S-PLUS the first task is to read the dataset into S-PLUS and assign a variable to it. Remember that the following commands are typed *beside* the S-PLUS prompt > (don't type the prompt itself).

```
> x <- read.table("ex.dat", header=T, row.names=NULL)

    #   the assignment symbol "<-" consists of the
    #  "less than" key followed by the "dash" key
```

[1] UNIX is case-sensitive so `Splus` is *not* the same as `splus`.

```
> print(x)
     name  class  height  weight  size
      clm      1       8      10   104
      gag      2       1       3    11
      gam      2       9      11   120
      grc      1       2       4    12
      mra      1       6      11   127
      rmm      3       6       6    33
      spm      3       6       7    52
      tcs      3      10      12   150
> q()
```

Although the above S-PLUS session is very brief (!), it does illustrate some of the generic properties of an S-PLUS session.

Assignment There are several ways to assign one value to another.

- `<-` Is used for assigning a variable to a quantity.
- `=` Is **only** used within functions to set options.

Comments In S-PLUS, the `#` symbol precedes a comment.

Strings Character strings such as a filename `"ex.dat"` need to be enclosed in double quotes.

Logicals The logical `T` is short for `TRUE` and has the value 1. Similarly, `F = FALSE` and has the value 0.

Intrinsics `NULL` is a special value that essentially assigns no memory storage to a variable or sets an argument such as `row.names` to a default setting.

Functions Many S-PLUS functions provide default settings for options that can be reset by the user: `read.table("ex.dat", header=T)`. Some functions have no options, but (almost) all require a user-supplied argument to be passed to it: `print(x)`. An exception is `q()`.

/.Data `x` will be permanently stored in the `/.Data` directory.

B.2.2 The S-PLUS Setup

S-PLUS is perhaps better thought of as an environment rather than a software package. A typical S-PLUS user would work interactively in this environment, reading and writing data, creating variables and functions, as well as obtaining the results from a statistical procedure. All of these are referred to as *objects* and would be permanently stored in the `/.Data`

under the same name. Thus, in the above example, x will exist as a "file" in /.Data with the contents stored in S-PLUS format. Since /.Data can become rapidly cluttered with all the objects we create, the first step is separate our projects.

In UNIX, we can easily create a subdirectory via the mkdir command so all we need to do is type: mkdir .Data (the period "." is required and remember that UNIX is case-sensitive). This will create /.Data in the current directory, /project1 say, and whenever S-PLUS is invoked in /project1 it will store S-PLUS objects in /project1/.Data. If no /.Data subdirectory exists, S-PLUS will use the /.Data in our top-level directory ($HOME). In PC S-PLUS, objects are stored in the _DATA subdirectory.

The built-in S-PLUS functions (objects) will be located in a system directory, for example, /usr/local/splus, which would be set up by the system operator. The system directory for S-PLUS contains several subdirectories that become "attached" whenever we invoke S-PLUS from our current directory. To see what these are, start S-PLUS and type: search(). On our system we get:

S-PLUS Directories

```
> search()
[1] "/home/faculty/mcdougal/.Data"
[2] "/opt/local/ss/splus/splus/.Functions"
[3] "/opt/local/ss/splus/stat/.Functions"
[4] "/opt/local/ss/splus/s/.Functions"
[5] "/opt/local/ss/splus/s/.Datasets"
[6] "/opt/local/ss/splus/stat/.Datasets"
[7] "/opt/local/ss/splus/splus/.Datasets"
[8] "/opt/local/ss/splus/library/trellis/.Data"
```

The "positions" [1] [2] ... indicate the order in which S-PLUS will search for an object. Thus, our /.Data directory is therefore the first place S-PLUS would look for an object. Note that we "could" create our own print() function — not a good idea[2] — which will not overwrite S-PLUS' built-in print() function. To see what function objects are contained in the S-PLUS Functions directory, type: ls(pos=2). (Be prepared — there are quite a few!) A brief glossary of S-PLUS functions is provided at the end of this section. For details on the use and various options available, use the online help() function in S-PLUS. Note that there is an extensive collection of S and S-PLUS documentation available from StatLib: http://lib.stat.cmu.edu.

[2]The problem is that "our" version of print() is stored in /.Data and therefore the built-in S-PLUS version can no longer be used. (That is, until we rename our version.)

S-PLUS Terminal Window (UNIX)

In UNIX, S-PLUS is normally invoked from a terminal window so the lines we type eventually disappear as output from S-PLUS gets printed to this screen. Like the `prog.sas` file in SAS, we can also create a program file of S-PLUS commands which can be submitted by using the `source("prog.spl")` function (the extension "`.spl`" is used for convenience here and is not required). Everything in the file `prog.spl` will be executed by S-PLUS just as if we had typed the lines directly, errors included. Hence a nice way to run an S-PLUS session is from within an Emacs shell window with another Emacs window containing the "source" file `prog.spl`.

S-PLUS Graphics Window (UNIX)

To see high-resolution graphics generated by S-PLUS we need to specify a graphics device. There are several that are available for interactive use, as well as the `postscript()` device for saving graphics to a file. The device `printer()`, combined with the command `show()`, can be used to produce text graphics (which are usually worthless). Some standard interactive graphics device types are:

```
X11() motif() openlook() hp2468() tek4105()
```

All high-resolution graphics will now appear in the graphics window which will have controls for printing and other facilities.

B.2.3 Basic S-PLUS Commands

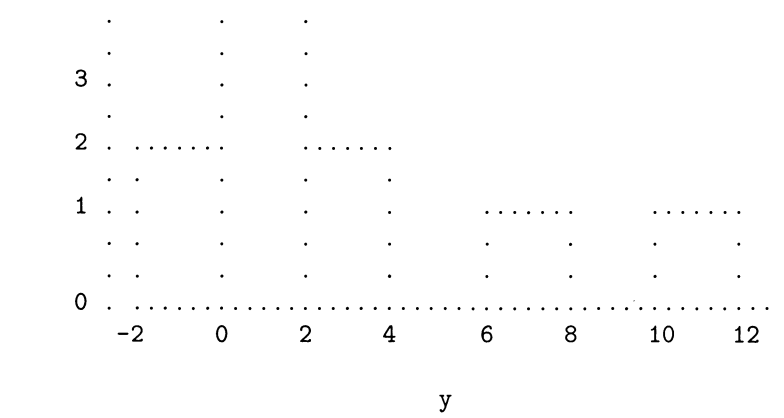
The following S-PLUS session illustrates some of the basic commands for manipulating a simple numeric data file called `mat.dat` which consists of two columns of five numbers. All the S-PLUS commands typed here are actually standard S commands. The data file `mat.dat` to be read into S-PLUS is shown below. Note that the two columns of numbers in this file are not in a tidy format, but there is at least one blank space between the column entries in each row. The file should **always** end with a final RETURN key press (the invisible character `^M`).

```
unix.prompt% more mat.dat
   2 4
   -1  7
   0 4
    11 2
   1 1
```

Now invoke S-PLUS and try to read `"mat.dat"` using the `scan()` function.

```
unix.prompt% S-PLUS
> scan("mat.dat")
 [1]   2   4  -1   7   0   4  11   2   1   1
> # not quite right ... hence
>
> y_matrix(scan("mat.dat"), ncol=2,byrow=T)
> # data is now in variable "y"
> # typing y just outputs its contents.
> y
     [,1] [,2]
[1,]    2    4
[2,]   -1    7
[3,]    0    4
[4,]   11    2
[5,]    1    1
>                    # Now we can try some simple manipulations
> y[,1]                       # subsetting 1st column
[1]   2  -1   0  11   1
> y[1,]                       # subsetting 1st row
[1] 2 4
> a_rev(y[-2,1])     # assign variable "a" to the 1st column,
                     # but exclude the 2nd element, and then
                     # reverse the order of those 4 elements

> a                  # show contents of "a" ... it worked
[1]   1  11   0  2

> y[,1]+y[1,]        # add 1st col and 1st row ?? (the lengths
[1]  4  3  2 15  3   # are different!) S-PLUS does it, but
                     # provides a warning message ...
Warning messages:
  Length of longer object is not a multiple of the length of
  the shorter object in: y[, 1] + y[1,  ]

 # S-PLUS does the above vector addition operation by its own
 # special rules which were partly violated. Hence a warning.

> solve(y)                  # S-PLUS won't do illegal operations
Error in solve.qr(a): matrix inverse only for square matrices
Dumped

> stem(y)  # do a stemplot of matrix y. Although y is matrix,
           # the S-PLUS function "stem" will treat y as a
           # vector here. Not all functions will do this.
```

```
N = 10   Median = 2
Quartiles = 1, 4

Decimal point is at the colon

  -1 : 0
  -0 :
   0 : z          # S-PLUS uses "z" for 0 on the 0 stem only
   1 : 00
   2 : 00
   3 :
   4 : 00
   5 :
   6 :
   7 : 0

High:  11          # S-PLUS puts outliers on a separate line

> args(stem)             # get the arguments of "stem" These
                         # can be used to modify the default
Stem-and-Leaf Display    # display. For full documentation
                         # use help(stem)
USAGE:
       stem(x, nl=0, scale=1000,
            twodig=FALSE, fence=2, head=TRUE, depth=FALSE)

> hist(y)                #  now try a histogram ... error?

Error in barplot(counts, width = breaks, histo = T..: Device
not active
Dumped
Error in barplot  # hist() actually uses barplot() to draw
                  # a histogram. However, barplot() requires
                  # a graphics device to be specified first.

                     # we will use the printer() device, but
> printer(60,20)     # make the plotting region smaller than
                     # the default settings: 80,64
> hist(y)            # hist can now be done, however device
> show()             # printer() needs show() to display

                     # result is a rather low quality display


     4 .       .......
```




```
> postscript("g.hist.ps")    # to get high quality printout,
> hist(y)                    # use postscript() device which
> q()                        # creates the PostScript file
                             # "g.hist.ps" on exit from S-PLUS
```

B.2.4 Efficient Use of S-PLUS

When we quit from S-PLUS only the objects we created during the S-PLUS session are retained and will be stored in the `/.Data` directory. At some point we should clean up the `/.Data` directory by deleting the variables we don't need such as `tmp, i, j,...`. Note that each of these variables is also a "file" and hence requires a certain minimum amount of space (such as 2048 bytes) which can rapidly consume our allowable account quota.

The `/.Data` Directory

Any object, function, or variable that we create (excluding variables created inside a function) is automatically stored as a special file in `/.Data`. After a period of time, the number of objects in `/.Data` can become quite large. Also the `.Audit` file, which essentially records every keystroke we made during an S-PLUS session, can get **very** large. Since nobody we know ever uses the `.Audit` file you should delete it via `rm .Data/.Audit` or by setting the appropriate environment variable to have size zero.

S-PLUS saves all the objects we create, but allows us to overwrite any object in `/.Data` — without warning! Hence, we should try to use appropriate names for objects that we **don't** want to be overwritten. The advantage of this is that once we have read in a dataset and assigned it to a well named variable: `a1.org.dat` say, then we don't need to read in the data again. Simply assign `y <- a1.org.dat` and proceed with the analysis using `y` (which is easier to type).

Adopting a good naming convention can be **very** helpful. One approach is to prefix (or postfix) objects that we want to keep in a systematic manner. For example, if **aname** denotes an object that we want to keep, use:

d.aname.dat	data frame or data matrix object
d.aname.fun	data function for processing data frame object
f.aname.out	function to output a result
f.aname.fun	generic function
f.aname.run	function for running a procedure
g.aname.fun	graphics function to produce a plot.

S-PLUS Program File

There are three main parts involved in creating an S-PLUS Program:

- Data entry/output
- Creating functions for data analysis
- Creating presentation graphics.

S-PLUS Program This is a file, `prog.spl`[3] say, created using Emacs for example, which contains a list of S-PLUS statements dictated by the user's objective. By using the S-PLUS function

```
> source("prog.spl")
```

all the S-PLUS statements in the file `prog.spl` will be executed in exactly the same way as if we had typed each line individually. Obviously, if a mistake is made, S-PLUS will say so and will usually stop processing the remainder of the statements in this file.

Debugging Simple debugging consists of running the S-PLUS program, then using Emacs to edit where the error occurred. Repeating this several times eventually debugs the entire S-PLUS program. A better method is to break the S-PLUS program into a collection of S-PLUS functions that can be debugged in isolation. Functions should be written in *generic* form so that the output from one function can be easily input into another.

Errors Errors from `> source("prog.spl")` are easy to fix, but can be hard to find since the actual syntax error may have occurred long before S-PLUS gives up executing. Examples of things to check are:

---

[3]Don't name your S-PLUS program as `"prog.s"`. Emacs may treat `"*.s"` as an assembly language file and can adopt weird behavior.

`ym = mean(y)`	direct assignment is done by: _ or <-
`sqrt(mean(ym)`	missing ")"
`{{ [[ ( ) ] }}`	make sure all the {} [] () balance
`# ym <- mean(y)`	comments: everything after # is ignored
`"missing }"`	S-PLUS won't source ?
	⇒ put a RETURN after the last "}"

Program errors (after source errors have been eliminated) are harder to fix since they usually imply an illegal operation or an improper computation.

Generalizing Adding options to the functions we create will make these functions more flexible to use. Similarly, we should avoid so-called "hard coding" in our functions. That is, `x[-length(x)]` is preferred over `x[-65]` since `x` may not always be of length 65.

Analysis Most of the program functions we create are likely to be for data manipulation and graphical purposes. The S-PLUS functions available for data analysis and modeling are generally more suited to direct interactive use since only minor modifications are usually made at each step of an analysis. Creating a function to run a complete analysis can be built up as we proceed with each stage of the analysis. This can then be run at the end to obtain all the necessary output for the report.

Output Once all sections of the S-PLUS program have been debugged, the output part is examined. Unfortunately, the output from S-PLUS procedures tends to be minimally formatted and we will likely need to create an output file for editing later. In some cases it may be worthwhile creating functions that perform certain formatting tasks, but this tends to be time consuming and it may be easier to simply cut and paste S-PLUS results into the output file and edit it.

Graphics For presentation quality graphics, creating a separate graphics function in another S-PLUS program file will make it easier to correct mistakes and modify the plot. The same debugging steps as above would be used to produce the final plot, save that changes would be based on the display shown in the graphics window. The final version should also be saved in PostScript so it can be reprinted at a later time if necessary.

Creating S-PLUS Functions

So far we have only employed the built-in functions of S-PLUS. However, the most powerful feature of S-PLUS is that we can create our own S-PLUS functions. Two examples are presented.

Example B.5 *A Simple Read Function*

Our first S-PLUS function is called `read.dat()` and the following S-PLUS code is all contained in our `prog.spl` file. The function could be created in S-PLUS directly, but any mistake we make means retyping the whole function over again.

```
read.dat <- function(file="",nc=1)
{                          # first check whether a file
   if(!missing(file)) {    #     has been specified
      if(nc == 1)          # if so, default is to assume
          y_scan(file)     #     input as a vector
      else
          y_matrix(scan(file), ncol=nc, byrow=T)
      return(y)            # return the result. Here
   }                       # "y" is a temporary variable
   else {                  # and is NOT saved in /.Data
      cat("No input file specified\n")
   }
}
```

Note that { } are not needed when only one statement is processed in an `if()` and/or `else` statement. The function `cat()` enables us to add helpful messages to be printed on the screen. Here is the result of running this function in S-PLUS.

```
> source("prog.spl")
> read.dat()
No input file specified
> read.dat("mat.dat")
 [1]   2   4  -1   7   0   4  11   2   1   1
> read.dat("mat.dat",nc=2)
     [,1] [,2]
[1,]    2    4
[2,]   -1    7
[3,]    0    4
[4,]   11    2
[5,]    1    1
> read.dat("mat.dat",nc=5)
     [,1] [,2] [,3] [,4] [,5]
[1,]    2    4   -1    7    0
[2,]    4   11    2    1    1
> y_read.dat("mat.dat",nc=2)
```

Remember that we must assign a variable to actually use the input data in S-PLUS and this y will be stored in `/.Data`. The `read.dat()` function is not perfect (try `nc = -1` or 3), and is somewhat redundant in view of the

`read.table()` S-PLUS command. (Don't reinvent the wheel.) However, it does illustrate the basic features (and problems) associated with creating our own functions.

Example B.6 *A Simple Graphics Function*

One problem with graphics is that every plot we want is invariably different from the last one, making it impractical to write a general plotting function (there are too many options). However, a simple skeleton that may be useful is the following.

```
gxy.fun_function(x,y,tk=1,skip=0,psfile="g.slr.ps",msg="")
{
    switch(tk,              # the switch variable is tk
       ,                    # default is to specify no
       printer(70,25),      # graphics device (option tk=1)
       tek4105(),           # since we may have already
       X11(),               # specified one.
       postscript(psfile) )
#-------------------------------------------------------------
# Any graphical display of interest to the user can go here.

    plot(x,y,pch="*")           # do a scatter plot
    if(skip == 1)               # if skip=1, add SLR fit
        abline(lsfit(x,y)$coef)
    if(!missing(msg))
        title(msg)       # put a title on plot if supplied
}
```

The main idea of this function is that it can be run using different graphics devices. For example, setting `tk=5` creates a PostScript file by either using the default filename `"g.slr.ps"` or setting `psfile="new_name"` in the options line. Even in this simple case it can be seen that there are an enormous number of options that could be added such as `onefile=T`. This could be incorporated in the options line of `gxy.fun` to allow the user to create a separate PostScript file per plot (assuming more plots were added to this function).

Printing S-PLUS Output

S-PLUS graphics output can be put into PostScript files. Data output is a slightly more complicated issue since we don't want to just dump an entire S-PLUS session into a file. That is, only the relevant parts of the analysis should be retained. The simplest method is to run S-PLUS in Emacs and cut and paste the relevant parts of the S-PLUS session into an output file `"a1.out"` say, and then print `a1.out`. With large datasets, however, this may be tedious and we should direct output to external

files. This can be done via the `write.table(file="dat.out", x)` or `write(file="dat.out", x)` S-PLUS functions. The slowest method is to use the `cat()` function:

```
for(i in 1:nrow(y)) {
   cat(file="dat.out", y[i,], append=T)
   cat(file="dat.out","\n",append=T)
}
```

B.2.5 S-PLUS Statistical Procedures

Data Frames

S-PLUS data frames were discussed in Case Study 6.4 and allow the user to employ statistical modeling procedures in a more natural way. We first illustrate some of the basic feastures of a data frame object with reference to the `"mat.dat"` file introduced previously.

```
> a_read.table("mat.dat")
> a
   V1 V2   # read.table() works out the number of columns
1   2  4   # for itself and labels them as V1 V2
2  -1  7
3   0  4
4  11  2
5   1  1
> a[1,]    # subsetting is as before
  V1 V2
1  2  4
> mode(a)    # However "a" is now a dataframe object ...
[1] "list"   # hence V1 is  a "list" component of "a"
> a$V1                    # ==> a$V1 = a[,1]
[1]   2  -1   0  11   1
> a[,1]
[1]   2  -1   0  11   1

> summary(a)  # provides summary statistics for
              # dataframe objects

       V1              V2
 Min.   : -1.0   Min.   :1.0
 1st Qu.:  0.0   1st Qu.:2.0
 Median :  1.0   Median :4.0
 Mean   :  2.6   Mean   :3.6
 3rd Qu.:  2.0   3rd Qu.:4.0
 Max.   : 11.0   Max.   :7.0
```

To illustrate the use of S-PLUS to fit logistic regression, consider the following dataset.

```
> d <- example.dat
> d  # print the dataframe object "example.dat"
     # read-in via "read.table()". To do this
     # using the "scan()" is more complicated.

Age Number Start    Y
 2    3      5    abs
12    3     25    pres
22    5      4    abs
 4    4     23    pres
15    5     31    abs
61    4     45    pres

> attach(d)   # a key feature of dataframes is that
              # the variables in "d" can now be
              # accessed directly. Thus, we can use
              # the glm() function to fit the logistic
              # regression model. Model statements in
              # S-PLUS are defined by the "~" character.

d.model <- glm(Y ~ Age + Start + Number,
               family = binomial, data = d )
```

The following functions can now extract various components from the `glm` object: `"d.model"`.

```
residuals()  # residuals
fitted()     # fitted values
predict()    # predicted values
coef()       # regression coefficents
deviance()   # the model deviance
```

For a more detailed description of `d.model` use:

```
summary(d.model)
```

To compare different models we can add or drop variables or observations from a model:

```
d2.model <- update(d.model, ~ . - Age)
       # drops the Age variable

d3.model <- update(d.model, subset= -79)
       # drops the 79th observation
```

Analysis of deviance tables:

```
anova(d.model, d2.model)  # produce an ANOVA Table
                          # for comparing models

drop1()       # produces similar tables by dropping
add1()        # or adding 1 variable at a time

step()        # stepwise selection of model variables
```

B.2.6 S-PLUS Glossary

We list some of the main S-PLUS commands and functions. For further information, use the S-PLUS `help()` command or see the S and S-PLUS references.

Note: All S-PLUS function commands require `"()"` or `(arg1,arg2,...)` to invoke a command: `length(x)` gives the length of the vector `x`. If we just type: `length` then S-PLUS returns the *source code* for the function `length` (which may or may not be useful).

```
> length
function(x)
.Internal(length(x), "S_extract", T, 6)
```

Online help is available via the S-PLUS `help()` or `args()` command:

```
> help(cmd)   full documentation on {\tt cmd}
> args(cmd)   brief documentation on {\tt cmd}
```

In some cases, double quotes may be required (e.g., `help("[[")`).

Basic S-PLUS Functions

Operators

`+ - / * ^ ** %% %/`	arithmetic
`< <= > >= == !=`	comparison
`_ -> <- assign`	assignment: `x_4` or `x <- 4` (both set $x = 4$)
`& | !`	logical: and, or, not
`[] : seq`	subset data: first four elements: `x[1:4]`

Data Attributes

`col row`	matrix column and row number vectors
`length mode`	length and mode of vector
`ncol nrow dim`	dimensions of a matrix
`missing`	returns logical of function argument
`attr attributes`	return attribute of S-PLUS object
`mode`	returns mode type of S-PLUS object
`names dimnames`	returns names of object, matrix

Data Directories

`attach detach`	attach or remove directory

`ls object`	list S-PLUS files (variables) in directory
`rm`	remove variables
`options`	returns current S-PLUS control settings
`search`	list the S-PLUS directories
`library`	attach an S-PLUS library
Data Manipulation	
`c`	create vector: `x <- c(1,0,-1)`
`cbind rbind`	create matrix by combining columns or rows
`order rep rev sort`	order, replicate, reverse, or sort vector
`paste`	paste character strings together
`split`	returns list object of vector split by group
Data Structures	
`matrix`	create matrix
`array`	create multidimensional array
`list`	create a list of S-PLUS objects
`ts`	create time series
`data.frame`	create a data frame object
Documentation	
`help args`	full or brief help documentation
Graphical displays	
`stem`	stem and leaf plot
`printer`	device type (low-resolution)
`show`	show graphics (for `printer()` device only)
`X11 postscript`	device types (high-resolution) — needed for:
`boxplot hist`	boxplot histogram
`plot`	plot(x,y, [many options])
`tsplot`	time series plot
`qqplot`	Q–Q plot
`contour`	contour plot
Graphics Add-Ons	
`par`	extensive list of graphics options
`abline axis box`	:
`lines points`	add lines, points, titles, and so on to plots
`arrows segments`	:
`legend title`	:
`text mtext symbols`	:
I/O Files	
`print`	print a variable to terminal screen
`scan`	read an external data file
`read.table`	read data file as `data.frame`
`write write.table`	write object or data frame to an external file
`source`	source external S-PLUS command file
`cat`	simple output to screen or external file
`unix`	execute a UNIX command: `unix("ls")`
`!`	UNIX shell escape: `"! ls"`

Linear Algebra

`%*% t`	matrix multiplication, transpose
`apply`	apply function to rows or columns of matrix
`diag`	diagonal elements of matrix
`eigen`	eigenvalues and eigenvectors of a matrix

Logical Operators

`all any`	used for conditional statements
`!= < <= == > >=`	`if((a&&b)||c){...}`
`if ifelse else`	"if (`a` *and* `b`)"
`&& || !`	logic operators used in if-else statements

Looping and Iteration

`apply sapply tapply`	apply function to S-PLUS object
`for(i in start:finish) {... S-PLUS commands ... }`	loops in S-PLUS can be **very** slow — use vectorized operations if possible.

Mathematical Operations

`abs ceiling floor trunc cos sin diff exp log log10`
`max min mean median order range rank round sort`
`sqrt sum prod var cor`

Matrices

`cbind rbind col row diag dim dimnames ncol nrow`

`data.matrix`	convert data frame into numeric matrix
`dist`	pairwise distance matrix
`scale`	centers and scales columns of matrix

Probability Distributions and Random Numbers

`*norm: p q r`	generate normal quantities:
`pnorm`	`pnorm(2) = P[Z < 2]`
`qnorm`	`qnorm(0.5)` $\Rightarrow$ `z st. P[Z<z]=0.5`
`rnorm`	`rnorm(10)` $\Rightarrow$ 10 quantiles from N(0,1)
`*chisq *t`	chi-square, t-distribution
`*exp *unif`	exponential, uniform
`*cauchy *f`	cauchy, F-distribution
`qqplot qqnorm`	Q–Q plots
`sample`	get sample from a vector

Robust/Resistant Techniques

`mad mean median sabl smooth twoway`

Statistical Operations

`hist min max mean median prod sum range stem var cor`

Time Series

`diff lag sabl sablplot smooth ts tsplot`

S-PLUS Statistical Procedures

Numerical functions

`mean median`	sample (trimmed) mean and median
`quantile`	quantiles. Default is 5 Number Summary
`var`	s^2 or $\widehat{\Sigma}$. To obtain s use: `sqrt(var(x))`
`summary`	provides all the above

Classical Tests

`t.test`	t-test. One-sample, two-sample, and paired
`cor.test`	correlation test ($\rho = 0$)
`chisq.test`	chi-square test: GOF or contingency tables
`fisher.test`	Fisher's exact test
`var.test`	F-test for variances
`wilcox.test`	Wilcoxon signed rank test
`kruskal.test`	Kruskal-Wallis test
`binom.test`	sign test
`prop.test`	z-test for proportions

Regression

`lm`	fit a linear model
`lsfit`	Least squares regression
`ls.summary`	LS regression summary
`l1fit`	$\mathcal{L}_1$ regression
`rbiwt rreg`	robust regression
`step`	stepwise model selection

ANOVA

`aov raov`	compute fixed or random effects ANOVA

GLM

`glm`	generalized linear model
`gam`	generalized additive model
`lm lme`	simple or complicated linear model
`varcomp`	variance components analysis

Graphics

`trellis`	generalized plot functions

Multivariate

`cancor`	canonical correlation analysis
`hclust plclust`	hierarchical clustering, dendogram plot
`kmeans`	k-means clustering
`pairs`	pairwise scatter plot
`prcomp princomp`	principal components analysis
`tree`	creates a tree object

Robust/Resistant Techniques

`loess lowess`	spline curves

Statistical Operations

`summary`	summary of S-PLUS object

Time Series

`acf`	correlograms
`ar arima.mle`	fit ar or arima model
`fft spectrum`	spectrum estimate

S and S-PLUS References

1 Ripley, B.D. (1994) *Introductory Guide to S-Plus.* PostScript file available from StatLib: `/S/sguide2.ps`

```
FTP:   ftp  lib.stat.cmu.edu  [ then cd to S ]
Web:   http://lib.stat.cmu.edu/S/
```

2 Venebles, W.N. and Ripley, B.D. (1994) *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

3 Chambers, J.M. and Hastie, T.J. (Editors) (1993) *Statistical Models in S.* Chapman & Hall, New York.

4 Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988) *The New S Language.* Wiley, New York.

Appendix C
Statistical Addendum

Ideally, this type of appendix would contain all the statistical details that a consultant needs. Sorry ... not in this book! In reality, this is too impractical and somewhat redundant since there are certainly other comprehensive sources available (e.g., Kotz and Johnson 1982). Instead, we have included several tables that contain the sort of information of which we (the authors) can never quite remember the precise details. Now, at least, *we* have the information in one place! Here is a short version of the tables presented in this appendix.

C.1 Univariate Distributions

TABLE C.1. Discrete Distributions

Discrete Distribution	Probability Function		Mean μ	Variance σ^2
Binomial	$\binom{n}{x}p^x q^{n-x}$	$0<p<1$ $x=0,1,\ldots,n$	np	npq $q=1-p$
Poisson	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\lambda>0$ $x=0,1,\ldots$	λ	λ
Geometric [1]	pq^x	$x=0,1,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$
	pq^{y-1}	$y=1,2,\ldots$	$\frac{1}{p}$	$\frac{q}{p^2}$
Pascal [1] *or*	$\binom{r+x-1}{x}p^r q^x$	$r>0$ $x=0,1,\ldots$	$\frac{rq}{p}$	$\frac{rq}{p^2}$
Negative Binomial	$\binom{y-1}{r-1}p^r q^{y-r}$	$y=r,r+1,\ldots$	$\frac{r}{p}$	$\frac{rq}{p^2}$
Hypergeometric	$\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$	$p^*=\frac{K}{N}$	np^*	$np^*q^*(1-f)$
	$x=\max\{0,n-N+K\},\ldots,\min\{K,n\}$			$q^*=1-p^*$ $f=\frac{n-1}{N-1}$
Multinomial [2]	$\frac{n!}{\prod_{i=1}^k n_i!}\prod_{i=1}^k p_i^{n_i}$		$n\boldsymbol{p}$	$n(\mathrm{diag}(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}')$
	$n=\sum n_i$ $\sum p_i=1$		$\boldsymbol{p}=(p_1,\ldots,p_k)'$	

[1] The geometric and Pascal distributions may be defined in terms of X = number of *failures*, or Y = number of *trials*.

[2] The multinomial is a multivariate distribution.

TABLE C.2. Continuous Distributions I

Distributions Defined on $\mathcal{R}$ $(-\infty < X < \infty)$

Distribution	Probability Function §		Mean	Variance
Normal [1]	$X \sim N(\mu, \sigma^2)$	$\sigma>0$	μ	σ^2
	$Z \sim N(0,1)$		0	1
t-Distribution [2]	$T \sim t_\nu$	$\nu>0$	If $\nu>1$	If $\nu>2$
	$T = \frac{Z}{\sqrt{W/\nu}}$	$Z \sim N(0,1)$ $W \sim \chi^2_\nu$	0	$\frac{\nu}{\nu-2}$
Cauchy	t_1	When $\theta=0$ $\beta=1$		
	$\frac{1}{\pi\beta\left[1+(\frac{x-\theta}{\beta})\right]^2}$	$\beta>0$	Does not exist	Does not exist
Laplace *or* *Double Exponential*	$\frac{\lambda}{2}e^{-\lambda\lvert x-\theta\rvert}$	$\lambda>0$	θ	$\frac{2}{\lambda^2}$
§ Logistic	$\left[1+e^{-(\frac{x-\theta}{\beta})}\right]^{-1}$	$\beta>0$	θ	$\frac{\beta^2\pi^2}{3}$
§ Gumbel *or* *Extreme Value*	$e^{-e^{-(x-\theta)/\beta}}$	$\beta>0$ $\gamma\approx0.577216$	$\theta+\beta\gamma$ [Euler's Constant]	$\frac{\beta^2\pi^2}{6}$

§ denotes cdf: $P[X \le x]$

[1] Normal (*Gaussian*) pdf: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/2\sigma^2}$.

[2] t-distribution pdf: $f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}\left[1+\frac{x^2}{\nu}\right]^{-(\nu+1)/2}$.

TABLE C.3. Continuous Distributions II

Distributions Defined on $\mathcal{R}^+$ $(X > 0)$

Distribution	Probability Function		Mean	Variance †
Chi-Square [3]	$W \sim \chi^2_\nu$ $\chi^2_n = \sum^n Z_i^2$	$\nu>0$ $Z_i \sim N(0,1)$	ν	2ν
F-Distribution [4]	$F \sim \mathcal{F}_{\nu_1,\nu_2}$ $F = \frac{W_1/\nu_1}{W_2/\nu_2}$	$\nu_i>0$ $W_i \sim \chi^2_{\nu_i}$	If $\nu_2>2$ $\frac{\nu_2}{\nu_2-2}$	If $\nu_2>4$ $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$
Exponential	$\lambda e^{-\lambda x}$	$\lambda>0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma	$\frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}$	$\lambda>0$ $a>0$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$
Weibull	$a\lambda x^{a-1} e^{-\lambda x^a}$	$\lambda>0$ $a>0$	$\frac{\Gamma(1+a^{-1})}{\lambda^{1/a}}$	$\mu_2 = \frac{\Gamma(1+2a^{-1})}{\lambda^{2/a}}$
Lognormal [5]	$Y = e^X$	$X \sim N(\mu,\sigma^2)$	$e^{\mu+(\sigma^2/2)}$	$\mu_2 = e^{2\mu+2\sigma^2}$

† For Weibull and lognormal: $\sigma^2 = \mu_2 - \mu^2$.

[3] $W \sim$ Gamma$(a = \frac{\nu}{2}, \lambda = \frac{1}{2})$.

[4] $f(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1-2)/2} \left[1 + x\left(\frac{\nu_1}{\nu_2}\right)\right]^{-(\nu_1+\nu_2)/2}$.

[5] $f(x) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-(\log_e y - \mu)^2/2\sigma^2}$.

TABLE C.4. Continuous Distributions III

Distributions Defined on $\mathcal{X} \subset \mathcal{R}$

Distribution	Probability Function		Mean	Variance
Uniform	$\frac{1}{b-a}$	$x \in [a,b]$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
Beta	$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$	$x \in [0,1]$ $a>0, b>0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b+1)(a+b)^2}$
Pareto	$\frac{\theta x_o^\theta}{x^{\theta+1}}$	$x>x_o>0$ $\theta>0$	If $\theta>1$ $\frac{\theta x_o}{\theta-1}$	If $\theta>2$ $\frac{\theta x_o^2}{(\theta-1)^2(\theta-2)}$

C.2 Multivariate Distributions

Multivariate Normal (MVN)

Let Y be a d-dimensional random vector with joint probability density function:

$$f(y) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left[-\tfrac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right],$$

where $\Sigma > 0$. Then Y has a (nonsingular) MVN distribution denoted by $Y \sim N_d(\mu, \Sigma)$. It follows that $E[Y] = \mu$ and $Var[Y] = \Sigma$. An alternative definition is:

Y is MVN if $\ell'Y$ is univariate normal for all $\ell \in \mathcal{R}^d$.

Wishart

Let $Y_1, Y_2, \ldots, Y_m$ be independently distributed as $N_d(0, \Sigma)$. Then $W = \sum_{i=1}^{M} Y_i Y_i'$ has a Wishart distribution with m degrees of freedom and $d \times d$ variance matrix Σ. This is denoted by $W \sim W_d(m, \Sigma)$. It follows that $\ell'W\ell \sim \ell'\Sigma\ell\chi^2_m$ for every $\ell \in \mathcal{R}^d$. However, this quadratic version of the MVN linear property does not uniquely determine the joint density function of the $d(d+1)/2$ distinct elements of W. For completeness:

$$f(w) = \frac{|W|^{(m-d-1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}W)\right]}{2^{md/2}|\Sigma|^{m/2}\pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(\frac{1}{2}(m+1-j)\right)} .$$

Hotelling's T^2

This is a multivariate version of the t-statistic. Let $Y \sim N_d(\mu, \Sigma)$ and $W \sim W_d(m, \Sigma)$ be independent. Then:

$$T^2 = m(y-\mu)'W^{-1}(y-\mu) \sim T^2_{d,m} ,$$

where $T^2_{d,m}(m-d+1)/md \equiv \mathcal{F}_{d,m-d+1}$. When $d = 1$, $T^2 \sim \mathcal{F}_{1,m} \equiv (t_m)^2$, where t_m denotes the t-distribution with m degrees of freedom.

MANOVA

The F-tests in the analysis of variance procedure are based on ratios of the form $T = H/E$ and $V = H/(H+E)$ where $H \sim \sigma^2\chi^2_{m_H}$ represents the "hypothesis" sum of squares and is independent of $E \sim \sigma^2\chi^2_{m_E}$, which represents the "error" sum of squares. The densities of these ratios are $T \sim (m_H/m_E)\mathcal{F}_{m_H,m_E}$ and $V \sim Beta(m_H/2, m_E/2)$ which have the distributional relationship: $1 - Beta(a,b) \sim Beta(b,a) \sim (1+(a/b)\mathcal{F}_{2a,2b})^{-1}$.

These results can be generalized to the multivariate case by taking $E \sim W_d(m_E, \Sigma)$ and $H \sim W_d(m_H, \Sigma)$ as independent Wishart matrices with $m_E, m_H \geq d$. The following eigenvalue decompositions are relevant.

$$|HE^{-1}| = \prod_{j=1}^{k} \psi_j = \prod_{j=1}^{k} \theta_j/(1-\theta_j)$$
$$|H(E+H)^{-1}| = \prod_{j=1}^{k} \theta_j \ ,$$

where $k = \min(d, m_H)$ is the number of nonzero eigenvalues. In the descriptions of the statistics[1] presented below, ν_1 and ν_2 refer to the following quantities.

$$\nu_1 = \tfrac{1}{2}(|m_H - d| - 1) \qquad \nu_2 = \tfrac{1}{2}(m_E - d - 1)$$

***U*-Statistic** This is also called Wilk's lambda.

$U \equiv \Lambda = |E|/|E+H| = \prod_{j=1}^{k}(1-\theta_j) \sim U_{d,m_H,m_E}$.

For $k = 1, 2$, the exact distribution of the U-statistic is given by

$(k=1)$ $\frac{1-U}{U}\frac{\nu_2+1}{\nu_1+1} \sim \mathcal{F}_{2\nu_1+2,2\nu_2+1}$

$(k=2)$ $\frac{1-U^{1/2}}{U^{1/2}}\frac{2\nu_2+2}{2\nu_1+3} \sim \mathcal{F}_{4\nu_1+6,4\nu_2+4}$.

For $k > 2$, the following approximations can be used.

For large m_E:

$-f \log U \sim \chi^2_{dm_H}$, where $f = m_E - (d - m_H + 1)/2$.

$\frac{1-U^{1/t}}{U^{1/t}}\frac{ft-g}{dm_H} \sim \mathcal{F}_{dm_H,ft-g}$, where $g = dm_H - 2/2$, and
$t = \max\left(1, \sqrt{(d^2m_H^2 - 4)/(d^2 + m_H^2 - 5)}\right)$.

Pillai's Trace $V = \mathrm{tr}[H(E+H)^{-1}]$. Approximation:

$\frac{2\nu_2+k+1}{2\nu_1+k+1}\frac{V}{k-V} \sim \mathcal{F}_{k(2\nu_1+k+1),k(2\nu_2+k+1)}$.

Lawley–Hotelling Trace Also called Hotelling's generalized T^2 statistic.

$T_g^2 = m_E \mathrm{tr}[HE^{-1}]$.

Approximation: $(1/c)\mathrm{tr}[HE^{-1}] \sim \mathcal{F}_{a,b}$, where

$a = dm_H$, $b = 4 + \frac{(a+2)(m_E-d-3)(m_E-d)}{(m_E+m_H-d-1)(m_E-1)}$ and $c = \frac{a(b-2)}{b(m_E-d-1)}$.

Roy's Maximum Root Tables for Roy's maximum root exist for the following, $\theta_{\max} = \lambda_{\max}(H(E+H)^{-1})$ and $\psi_{\max} = \lambda_{\max}(HE^{-1})$.

When $k = 1$, $\theta_{\max} \sim Beta(\nu_1 + 1, \nu_2 + 1)$. Hence,

$\frac{2(\nu_2+1)}{2(\nu_1+1)}\frac{\theta_{\max}}{1-\theta_{\max}} = \frac{2(\nu_2+1)}{2(\nu_1+1)}\psi_{\max} \sim \mathcal{F}_{2(\nu_1+1),2(\nu_2+1)}$.

Approximation: $\psi_{\max}(m_E - r + m_H)/r$, where $r = \max(d, m_H)$ is an upper bound on $\mathcal{F}_{r,m_E-r+m_H}$ that provides a lower bound on the significance level.

---

[1] For a more detailed discussion on these statistics and tables, see Seber (1984).

The case $m_H = 1$ When $m_H = 1$, the above statistics satisfy:

$$U = 1 - \theta_{\max} \qquad V = \theta_{\max} \qquad T_g^2 = m_E \frac{\theta_{\max}}{1-\theta_{\max}} \sim T^2_{d,m_E} \ .$$

Thus,

$$\frac{m_E - d + 1}{d} \frac{\theta_{\max}}{1-\theta_{\max}} \sim \mathcal{F}_{d,m_E-d+1} \ .$$

Sphericity Test Let $X_1, X_2, \ldots, X_n$ be iid $N_d(\mu, \Sigma)$. The sphericity test is $H_o : \Sigma = \lambda I_d$ versus $H_A : \Sigma \neq \lambda I_d$. The sphericity test is equivalent to testing that the eigenvalues of Σ are all equal:

$$1 = \frac{\text{geometric mean of the } \lambda_j}{\text{arithmetic mean of the } \lambda_j} = \frac{\left(\prod_j \lambda_j\right)^{1/d}}{\frac{1}{d}\sum_j \lambda_j} = \frac{|\Sigma|^{1/d}}{\text{tr}[\Sigma/d]} \ .$$

Replacing Σ by its maximum likelihood estimate gives the following likelihood ratio test statistic.

$$V = |\text{SS}_X| / (\tfrac{1}{d}\text{tr}[\text{SS}_X])^d \ ,$$

where $\text{SS}_X = \sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})'$. The test rejects H_o for small values of V. See Seber (1984) and Muirhead (1982) for details on large sample approximations for this test.

C.3 Statistical Tests

TABLE C.5. Standard One-Sample Tests

C.5-1	t-Test: $H_o : \mu = \mu_o$	
Test Statistic	H_o Distribution	Assumptions and Conditions
$t_o = \dfrac{\overline{x} - \mu_o}{s/\sqrt{n}}$	t_{n-1} $t_o \approx N(0,1)$	Normal Process $n > 50$

C.5-2	z-Test: $H_o : \mu = \mu_o$, σ known	
$z_o = \dfrac{\overline{x} - \mu_o}{\sigma/\sqrt{n}}$	$N(0,1)$ $z_o \xrightarrow{d} N(0,1)$	Normal Process $n > 30$

C.5-3	Paired t-Test: $H_o : \mu_A - \mu_B = \delta_o$	
$t_o = \dfrac{\overline{d} - \delta_o}{s_d/\sqrt{n}}$	t_{n-1}	*Before and After* Experiment Normal Differences $d_i = x_{i_{\text{after}}} - x_{i_{\text{before}}}$

C.5-4	Proportion: $H_o : p = p_o$	
$v_o = \sum_{i=1}^{n} x_i$	$Bin(n, p_o)$	Binomial Experiment $x_i \sim Bin(1, p_o)$
$z_o = \dfrac{v_o - np_o}{\sqrt{np_oq_o}}$	$z_o \xrightarrow{d} N(0,1)$	$np_o, nq_o \geq 5$ $(q_o = 1 - p_o)$
$w_o = z_o^2$	$w_o \sim \chi_1^2$	$np_o, nq_o \geq 5$

C.5-5	Variance: $H_o : \sigma^2 = \sigma_o^2$	
$w_o = \dfrac{(n-1)s^2}{\sigma_o^2}$	χ_{n-1}^2	Normal Process

C.5-6	Wilk–Shapiro Test of Normality $H_o : X \sim N(\mu, \sigma^2)$	
W §	*Tabulated* [1]	Departures from Normality Detected by Small W KS Test used for $n > 50$ See [C.7-3]

§ $W = \left\{ \sum_{i=1}^{[n/2]} a_{n-i+1}^{(n)} (x_{(n-i+1)} - x_{(i)}) \right\}^2 / \sum_i^n (x_{(i)} - \overline{x})^2$.

[1] See Seber (1984; Appendices D7, D8).

TABLE C.6. Standard Two-Sample Tests

C.6-1	t-Test: $H_o : \mu_x = \mu_y$ with ($\sigma_x = \sigma_y$)	
Test Statistic	H_o Distribution	Assumptions and Conditions
$t_o = \dfrac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$	$t_{n_x+n_y-2}$	Independent Normals $s_p^2 = \frac{(n_x-1)\sigma_x^2+(n_y-1)\sigma_y^2}{n_x+n_y-2}$

C.6-2	Smith–Satterhwaite Test: $H_o : \mu_x = \mu_y$ and $\sigma_x \neq \sigma_y$	
$t_o = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$	$t_o \approx t_\nu$	Independent Normals $\nu \geq \min(n_x, n_y) - 1$ §
	$t_o \approx N(0,1)$	$n_x, n_y > 50$

C.6-3	z-Test: $H_o : \mu_x = \mu_y$ and σ_x^2, σ_y^2 Known	
$z_o = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$	$N(0,1)$ $z_o \xrightarrow{d} N(0,1)$	Independent Normals $n_x, n_y > 30$

C.6-4	Proportions: $H_o : p_x = p_y \Leftrightarrow$ Fisher's Exact Test (2×2)	
$v_o = \sum_{i=1}^{n_x} x_i$	*Hypergeometric*	Independent Binomials $x_i, y_j \sim Bin(1,p)$ *see Table C.1*
$z_o = \dfrac{\hat{p}_x - \hat{p}_y}{\sqrt{\overline{p}\,\overline{q}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$	$z_o \xrightarrow{d} N(0,1)$	n_x, n_y large (see [C.5-4]) $\hat{p}_{x,y} = \sum x_i/n_x \ , \ \sum y_j/n_y$ $\overline{p} = \frac{n_x\hat{p}_x+n_y\hat{p}_y}{n_x+n_y}$
$z_o' = \dfrac{\psi(\hat{p}_y) - \psi(\hat{p}_x)}{\sqrt{1/4n}}$	$z_o' \approx N(0,1)$	$\psi(p) = \sin^{-1}(\sqrt{p})$ (arcsine)
$w_o = z_o^2$	$w_o \sim \chi_1^2$	*see Table C.3*

C.6-5	Variances: $H_o : \sigma_x^2 = \sigma_y^2$	
$f_o = s_x^2/s_y^2$	$\mathcal{F}_{n_x-1,n_y-1}$	Independent Normals

C.6-6	Correlation: $H_o : \rho = \rho_o$	
$z_o = \dfrac{V(r) - V(\rho_o)}{s_V}$	$z_o \xrightarrow{d} N(0,1)$	Bivariate Normal $V(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$ $s_V = 1/\sqrt{n-3}$
$t_o = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	t_{n-2}	$\rho_o = 0$

§ Data Approximation: $\nu = \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2 \Big/ \left[\frac{(s_x^2/n_x)^2}{n_x-1} + \frac{(s_y^2/n_y)^2}{n_y-1}\right]$.

TABLE C.7. Nonparametric Tests I

C.7-1	Sign Test: $H_o : \eta = \eta_o \Leftrightarrow H_o : p = p_o = P[X < \eta_o]$ Median Test: $H_o : M = \eta_o$ (*Median*) $\Leftrightarrow H_o : p = p_o = 0.5$	
Test Statistic	H_o Distribution	Assumptions and Conditions
$v_o = \sum_{i=1}^{n} v_i^+$	$Bin(n, p_o)$	$v_i^+ = \mathcal{I}[x_i > \eta_o]$ § For $n\{p_o, q_o\} \geq 5$ see [C.5-4]

C.7-2	Wilcoxon Signed-Rank Test: $H_o : M = \eta_o$	
$S_o = \sum_{i=1}^{n} v_i^+ R_i$	$S_o \sim \mathcal{S}_n$	Symmetric Distribution $R_i = Rank\big[\,\lvert x_i - \eta_o \rvert\,\big]$ v_i^+ as above Tables Exist for $\mathcal{S}_{n,\alpha} (n \leq 20)$
$z_o = \dfrac{S_o - S_m}{\sqrt{S_v}}$	$z_o \xrightarrow{d} N(0,1)$	$n > 20$ $S_m = n(n+1)/2$ $S_v = n(n+1)(2n+1)/24$

C.7-3	Kolmogorov–Smirnov (KS) Test: $H_o : X \sim F_o$	
$D_o = \sup_x \lvert \Delta(x) \rvert$	$D_o \sim \mathcal{D}_n$	$\Delta(x) = F_n(x) - F_o(x)$ $F_n(x) = n^{-1} \sum \mathcal{I}[x_i \leq x]$ Tables Exist for $\mathcal{D}_{n,\alpha}$
	$D_o \xrightarrow{d} H_1$ †	$n > 5$ $x_1 = D_o\big[\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\big]$ $P[D > D_o] \approx 1 - H_1(x_1)$
$V_o = D_o^+ + D_o^-$	$V_o \xrightarrow{d} H_2$ ‡	*Kuiper's Statistic*[1] $D_o^{\pm} = \sup_x \pm\Delta(x)$ $x_2 = V_o\big[\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\big]$ $P[V > V_o] \approx 1 - H_2(x_2)$

§ Indicator Function: $\mathcal{I}[A] = 1$ if A occurs, 0 otherwise.

† $H_1(x) = 1 - 2\sum_{j=1}^{\infty} (-1)^j e^{-2j^2x^2}$.

‡ $H_2(x) = 1 - 2\sum_{j=1}^{\infty} (4j^2x^2 - 1) e^{-2j^2x^2}$.

[1] See Press et al. (1992; §14.3).

TABLE C.8. Nonparametric Tests II

C.8-1	KS Test: $H_o : F = G$ ($X \sim F$ and $Y \sim G$)	
Test Statistic	H_o Distribution	Assumptions and Conditions
$D_o = \sup_x \lvert\Delta^*(x)\rvert$	$D_o \xrightarrow{d} H_1$	$\Delta^*(x) = F_n(x) - G_n(x)$ $n_e = n_x n_y/(n_x + n_y) > 5$
$V_o = D_o^+ + D_o^-$	$V_o \xrightarrow{d} H_2$	Use [C.7-3] with $n = n_e$

C.8-2	Runs Test: $H_o : F = G$ or $H_o : \{X_i\} = \textit{Random Sample}$	
$A_o = \#\{\text{runs}\}$	$A_o \sim \mathcal{A}_{n_x,n_y}$	Randomness Test §
$z_o = \dfrac{A_o - A_m}{\sqrt{A_v}}$	$z_o \xrightarrow{d} N(0,1)$	$n_x, n_y > 10$ $A_m = \frac{2n_x n_y}{n_x+n_y} + 1$ $A_v = \frac{(A_m-1)(A_m-2)}{n_x+n_y-1}$

C.8-3	Median Test: $H_o : F = G \Rightarrow H_o : \eta_X = \eta_Y$ (*Nedians*)	
$M_o = \#[x_i < M]$	*Hypergeometric*	$M = \text{med}\{x_i, y_j\}$ †

C.8-4	Mann–Whitney U-test: $H_o : F = G \Rightarrow H_o : \eta_X = \eta_Y$ Wilcoxon Rank-Sum Test (S): $S = U + n_x(n_x+1)/2$	
$U_o = \sum_i^{n_x} \sum_j^{n_y} u_{ij}$	$U_o \sim \mathcal{U}_{n_x,n_y}$	$u_{ij} = \mathcal{I}[x_i < y_j]$ Tables Exist for $\mathcal{U}_{n_x,n_y\alpha}$ $\mathcal{S}_{n_x,n_y\alpha}$
$S_o = \sum_i^{n_x} R_i^X$	$S_o \sim \mathcal{S}_{n_x,n_y}$	R_i^X = Rank of x_i in $\{x_i, y_j\}$
$z_o = \dfrac{U_o - U_m}{\sqrt{U_v}}$	$z_o \xrightarrow{d} N(0,1)$	$n_x, n_y > 10$ $U_m = n_x n_y/2$ $U_v = n_x n_y(n_x + n_y + 1)/12$

C.8-5	Kruskal-Wallis Test: H_o : *No Treatment Effect*	
$K_o = \sum_i^a g_{n_i}(R_{ij})$	$K_o \approx \chi^2_{a-1}$ ‡	Fixed Effects Model : $y_{ij} = \mu_i + \epsilon_{ij} \quad \epsilon_{ij} \sim (0, \sigma^2)$ $R_{ij} = \text{Rank}(y_{ij})$ $n_i > 5,\ i = 1, \ldots, a$

C.8-6	Spearman Correlation: $H_o : \rho = 0$	
$t_o = \dfrac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$	$t_o \approx t_{n-2}$	$r_s = 1 - 6\sum_{i=1}^n d_i^2/n(n^2-1)$ $d_i = \text{Rank}(x_i) - \text{Rank}(y_i)$

§ $P[A = a] = \binom{n_x-1}{[a/2]}\binom{n_y-1}{[(a+1)/2]} + \binom{n_x-1}{[(a+1)/2]}\binom{n_y-1}{[a/2]} \Big/ \binom{n_x+n_y}{n_x}$.

† $P[M = a] = \binom{n_x}{a}\binom{n_y}{[(n_x+n_y)/2]-a} \Big/ \binom{n_x+n_y}{[(n_x+n_y)/2]}$.

‡ $K_o = R_v^{-1}\left[\sum_{i=1}^a n_i^{-1}\left(\sum_{j=1}^{n_i} R_{ij}\right)^2 - n(n+1)^2/4\right]$,

where $n = \sum_{i=1}^a n_i$

and $R_v = (n-1)^{-1}\left[\sum_{i=1}^a \sum_{j=1}^{n_i} R_{ij}^2 - n(n+1)^2/4\right]$.

C.4 Sample Size

For the statistical tests and procedures listed in Table C.9, the sample sizes will provide 80% power at the specified significance level and effect size. For example, in order for an investigator to be 80% confident they will detect a nonzero correlation of at least 0.30 using the t-test (see [C.6-6]) with $\alpha = 0.05$, the X and Y samples need to have a minimum of 85 observations. Other points to note are:

- For the one-way ANOVA test, the sample size refers to the minimum number of observations required **per** treatment level. The total sample size required is therefore $N = I \times n$ where n is the entry given in the table.
- In general, the effect sizes given in the table would be regarded as "Medium" or "Large." (The actual value of the effect size is determined by the index quantity d.) These sample sizes can therefore be thought of as representing the *least* number of observations the investigator can employ, and still have a meaningful study from a statistical perspective.

Table C.10 contains the factorial, axial, and center point allocations for central composite designs up to seven factors. For a given number of factors, the key differences among the uniform precision, orthogonal, and orthogonal block designs are the number of center points and the axial scaling required to make the effects orthogonal. The notation `[8,8,0]` refers to a particular design component (factoral, axial, or center) and is the number of runs (for that design component) allocated to each block. This is abbreviated as $B_{1-8} = 8$ (meaning blocks 1 through 8 each have 8 observations) for block designs with $k = 6$ and 7 factors. The necessary and sufficient conditions for constructing a central composite design in b orthogonal blocks is that each block forms an orthogonal first-order design, X_j say, and $\text{trace}(X_j'X_j)/\sum_j \text{trace}(X_j'X_j) = n_j/n$ where n_j is the number of runs in the jth block.

TABLE C.9. Sample Sizes Required for 80% Power

		Significance:	α-Level	
Test	Index for ES	Effect Size (ES)	5%	1%
t-Test:	$d = \lvert\mu_1 - \mu_2\rvert/\sigma$	0.50	64	95
$H_o : \mu_1 = \mu_2$		0.80	26	38
Correlation:	$d = r$	0.30	85	125
$H_o : \rho = 0$		0.50	28	41
Proportion:	$d = p - 0.5$	0.15	85	127
$H_o : p = 0.5$		0.25	30	44
$H_o : p_1 = p_2$	$d = \psi(p_1) - \psi(p_2)$	0.50	63	93
z-Test	ψ =arcsine (see [C.6-4])	0.80	25	36

Chi-Square: $H_o : p_i = p_{oi}\ i = 1, \ldots, k$			α-Level	
Index		ES	5%	1%
$d = \sqrt{\sum_{i=1}^{k} \frac{(p_{ai} - p_{0i})^2}{p_{0i}}}$		0.50	26	38
		2	39	56
		3	44	62
		4	48	67
		5	51	71
		6	54	75

Regression: $H_o : \beta_0 = \cdots = \beta_k = 0$						
Index	α-Level	ES	k=2	k=3	k=4	k=5
$d = \frac{R^2}{1 - R^2}$	5%	0.15	67	76	84	91
		0.35	30	34	38	42
	1%	0.15	97	108	118	126
		0.35	45	50	55	59

ANOVA (One-Way) : $H_o : \mu_1 = \cdots = \mu_I$						
Index	α-Level	ES	I=3	I=4	I=5	I=6
$d = \frac{\sigma_{TRT}}{\sigma_E}$	5%	0.25	52	45	39	35
		0.40	26	21	18	16
	1%	0.25	76	63	55	49
		0.40	30	25	22	20

TABLE C.10. Uniform Precision (U.P.) and Orthogonal (OR.) Block Central Composite Designs

Factorial: $n_f = 2^k$ or 2^{k-1}
Axial Points: $n_a = 2k$
Center Points: n_o

Factors		Design Type			Axial Scaling: Orthogonal Effects: α_{or} Rotatable: $\alpha_{rot} = \sqrt[4]{n_f}$		
k	$n_\bullet$	U.P.	OR.	Block			
2	n_f	4	4	[4, 0]		1.267103	(up)
	n_a	4	4	[0, 4]	α_{or}	1.414214	(or)
	n_o	5	8	[3, 3]		1.414214	(bk)
	n_{runs}	**13**	**16**	**14**	α_{rot}	**1.414214**	
3	n_f	8	8	[4, 4, 0]		1.524649	(up)
	n_a	6	6	[0, 0, 6]	α_{or}	1.668032	(or)
	n_o	7	9	[2, 2, 2]		1.632993	(bk)
	n_{runs}	**21**	**23**	**20**	α_{rot}	**1.681793**	
4	n_f	16	16	[8, 8, 0]		1.770742	(up)
	n_a	8	8	[0, 0, 8]	α_{or}	2.0	(or)
	n_o	7	12	[2, 2, 2]		2.0	(bk)
	n_{runs}	**31**	**36**	**30**	α_{rot}	**2.0**	
5 ($\frac{1}{2}$rep)	n_f	16	16	[16, 0]		1.820359	(up)
	n_a	10	10	[0,10]	α_{or}	2.0	(or)
	n_o	6	12	[6, 1]		2.0	(bk)
	n_{runs}	**32**	**36**	**33**	α_{rot}	**2.0**	
5	n_f	32	32	[8, 8, 8, 8, 0]		2.096683	(up)
	n_a	10	10	[0, 0, 0, 0,10]	α_{or}	2.353860	(or)
	n_o	10	16	[2, 2, 2, 2, 4]		2.366432	(bk)
	n_{runs}	**52**	**58**	**54**	α_{rot}	**2.378414**	
6 ($\frac{1}{2}$rep)	n_f	32	32	[16,16, 0]		2.142723	(up)
	n_a	12	12	[0, 0,10]	α_{or}	2.353860	(or)
	n_o	9	14	[4, 4, 2]		2.366432	(bk)
	n_{runs}	**53**	**58**	**54**	α_{rot}	**2.378414**	
6	n_f	64	64	$B_{1-8} = 8$		2.481445	(up)
	n_a	12	12	$B_9 = 12$	α_{or}	2.828427	(or)
	n_o	15	24	B_{1-8} =1 B_9 =2		2.828427	(bk)
	n_{runs}	**91**	**100**	**90**	α_{rot}	**2.828427**	
7 ($\frac{1}{2}$rep)	n_f	64	64	$B_{1-8} = 8$		2.523223	(up)
	n_a	14	14	$B_9 = 14$	α_{or}	2.828427	(or)
	n_o	14	22	B_{1-8} =1 B_9 =4		2.828427	(bk)
	n_{runs}	**92**	**100**	**90**	α_{rot}	**2.828427**	

References

[1] Adcock, C. (1997). Sample size determination: a review. *The Statistician*, **46**(2), 261–283.

[2] Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

[3] Akaike, H. (1974). Stochastic theory of minimal realization. *IEEE Trans. Automatic Control*, **AC-19**, 667–674.

[4] Andersen, B. (1990). *Methodological Errors in Medical Research.* Blackwell Scientific, Oxford.

[5] Anderson, T. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

[6] Anderson, V. and McLean, R. (1974). *Design of Experiments: A Realistic Approach.* Marcel Dekker, New York.

[7] Andrews, D. and Herzberg, A. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer-Verlag, New York.

[8] Barnett, V. (1982). *Comparative Statistical Inference.* 2nd edition. Wiley, Chichester.

[9] Becker, R. and Chambers, J. (1984). *S: An Interactive Environment for Data Analysis and Graphics.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

[10] Becker, R., Chambers, J., and Wilks, A. (1988). *The New S Language — A Programming Environment for Data Analysis and Graphics.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

[11] Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

[12] Boen, J. and Zahn, D. (1982). *The Human Side of Statistical Consulting.* Lifetime Learning Publications, Belmont, CA.

[13] Box, G. (1957). Evolutionary operation: A method for increasing industrial productivity. *Applied Statistics*, **6**, 81–101.

[14] Box, G. and Cox, D. (1964). An analysis of transformations. *JRSS(B)*, **26**, 211–252.

[15] Box, G. and Draper, N. (1969). *Evolutionary Operation.* Wiley, New York.

[16] Box, G., Hunter, W., and Hunter, J. (1978). *Statistics for Experimenters.* Wiley, New York.

[17] Box, G. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco.

[18] Box, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control.* 3rd edition. Prentice Hall, Englewood Cliffs, NJ.

[19] Box, J. (1978). *R. A. Fisher, The Life of a Scientist.* Wiley, New York.

[20] Brieman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA.

[21] Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods.* 2nd edition. Springer-Verlag, New York.

[22] Carr, D. (1998). Presenting spatially-indexed summary statistics using linked micromap plots. Presented at the Data Visualization Conference, Drew University, New Jersey.

[23] Chambers, J. and Hastie, T., Editors (1993). *Statistical Models in S.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

[24] Chatfield, C. (1995). *Problem Solving: A Statistician's Guide.* 2nd edition. Chapman & Hall, London.

[25] Cleveland, W. (1985). *The Elements of Graphing Data.* Wadsworth, Inc., Belmont, CA.

[26] Cleveland, W. (1993). *Visualizing Data.* Hobart Press, Summit, NJ.

[27] Cochran, W. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, **13**, 261–281.

[28] Cochran, W. (1977). *Sampling Techniques.* 3rd edition. Wiley, New York. First published in 1953.

[29] Cochran, W. (1983). *Planning and Analysis of Observational Studies.* Wiley, New York.

[30] Cochran, W. and Cox, D. (1957). *Experimental Designs.* 2nd edition. Wiley, New York.

[31] Cohen, J. (1988). *Statistial Power Analysis for the Behavioral Sciences.* 2nd edition. Erbaum, Hillsdale, NJ.

[32] Cohen, J. (1992). A Power Primer. *Psych. Bulletin*, **112**(2), 155–159.

[33] Collett, D. (1991). *Modelling Binary Data.* Chapman & Hall, New York.

[34] Conover, W. (1980). *Practical Nonparametric Statistics.* 2nd edition. Wiley, New York.

[35] Conrad, S., Editor (1989). *Assignments in Applied Statistics.* Wiley, New York.

[36] Cox, D. (1958). *The Planning of Experiments.* Wiley, New York.

[37] Cox, D. and Snell, E. (1981). *Applied Statistics, Principles and Examples.* Chapman & Hall, London.

[38] Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**(3), 297–334.

[39] Crowder, M. and Hand, D. (1990). *Analysis of Repeated Measures.* Chapman & Hall, New York.

[40] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm. *JRSS(B)*, **39**, 1–22.

[41] Derr, J. (2000). *Statistical Consulting: A Guide to Effective Communication.* Duxbury Press, Pacific Grove, CA.

[42] Desu, M. and Raghavarao, D. (1990). *Sample Size Methodology.* Academic Press, Boston.

[43] Dillman, D. (1978). *Mail and Telephone Surveys: The Total Design Method.* Wiley, New York.

[44] Dixon, W. and Kronmal, R. (1965). The choice of origin and scale for graphs. *J. of the Assoc. for Computing Machinery*, **12**, 259–261.

[45] Dobson, A. (1990). *An Introduction to Generalized Linear Models.* 2nd edition. Chapman & Hall, London.

[46] Dodge, H. and Romig, H. (1959). *Sampling Inspection Tables: Single and Double Sampling.* Wiley, New York.

[47] Draper, N. and Smith, H. (1981). *Applied Regression Analysis.* Wiley, New York.

[48] Duncan, D. (1955). Multiple range and multiple F-tests. *Biometrics*, **11**, 1–42.

[49] Dunckel, J. and Parnham, E. (1993). *Effective Speaking for Business Success: Making Confident Presentations, Using Audio-Visuals, and More.* 2nd edition. Self-Counsel Press, Bellingham, WA.

[50] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, San Francisco.

[51] Everitt, B. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, **44**, 113–135.

[52] Fienberg, S. (1989). *The Evolving Role of Statistical Assessments as Evidence in the Courts.* Springer-Verlag, New York.

[53] Fink, A. (1995). *The Survey Handbook.* Sage, Thousand Oaks, CA.

[54] Finklestein, M. and Levin, B. (1990). *Statistics for Lawyers.* Springer-Verlag, New York.

[55] Finney, D. (1982). The questioning statistician. *Statistics in Medicine*, **1**, 5–13.

[56] Fisher, R. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh. 14th edition, 1970.

[57] Fleiss, J. (1986). *The Design and Analysis of Clinical Experiments.* Wiley, New York.

[58] Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

[59] Freedman, D. and Diaconis, P. (1981). On the maximum deviation between the histogram and the underlying density. *Z. Wahrsch. Verw. Gebiete*, **57**, 453–476.

[60] Friedman, L., Furberg, C., and Demets, D. (1985). *Fundamentals of Clinical Trials.* 2nd edition. PSG Publishing, Littleton, MA.

[61] Fuller, W. (1987). *Measurement Error Models.* Wiley, New York.

[62] Gail, M. and Gart, J. (1973). The determination of sample size for use with the exact conditional test in 2×2 comparative trials. *Biometrics*, **29**, 441–448.

[63] Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley, New York.

[64] Gopen, G. and Swan, J. (1990). The science of scientific writing. *American Scientist*, **78**, 550–558. Also available from the JCGS Website: `www.amstat.org`.

[65] Greenfield, A. (1979). Statisticians in industrial research: The role and training of the industrial consultant. *The Statistician*, **28**, 71–82.

[66] Hacker, D. (1998). *A Writer's Reference.* 4th edition. St. Martins Press, New York.

[67] Hahn, G. and Hoerl, R. (1998). Key challenges for statisticians in business and industry. *Technometrics*, **40**, 195–200.

[68] Hamilton, C. and Parker, C. (1993). *Communicating for Results: A Guide for Business and the Professions.* Wadsworth, Inc., Belmont, CA.

[69] Hand, D. (1994). Deconstructing statistical questions. *JRSS(A)*, **157**, 317–356.

[70] Hand, D. J. and Everitt, B. S., Editors (1987). *The Statistical Consultant in Action.* Cambridge University Press, Cambridge.

[71] Hastie, T., Tibshirani, R., and Buja, A. (1995). Flexible discriminant and mixture models. In Kay, J. and Titterington, D., Editors, *Neural Networks and Statistics.* Oxford University Press, Edinburgh. Web page: `http://www-stat.stanford.edu/~hastie/Papers/`.

[72] Hinkley, D. (1977). On quick choice of power transformation. *Applied Statistics*, **26**, 67–69.

[73] Hoaglin, D., Mosteller, F., and Tukey, J., Editors (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, New York.

[74] Hoaglin, D., Mosteller, F., and Tukey, J., Editors (1985). *Exploring Data Tables: Trends and Shapes.* Wiley, New York.

[75] Hoaglin, D., Mosteller, F., and Tukey, J., Editors (1991). *Fundamentals of Exploratory Analysis of Variance.* Wiley, New York.

[76] Hoerl, R., Hooper, J., Jacobs, P., and Lucas, J. (1993). Skills for industrial statisticians to survive and prosper in the emerging quality environment. *The American Statistician*, **47**, 280–292.

[77] Hurlbert, S. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

[78] Johnson, R. and Wichern, D. (1998). *Applied Multivariate Statistical Analysis.* 4th edition. Prentice Hall, Upper Saddle River, NJ.

[79] Joiner, B. (1982). Practicing statistics, or, what they forgot to say in the classroom. In Rustagi, J. and Wolfe, D., Editors, *Teaching of Statistics and Statistical Consulting.* Academic Press, New York.

[80] Kaye, D. and Zeisel, H. (1997). *Statistics for Lawyers.* Springer-Verlag, New York.

[81] Kettenring, J. (1995). What industry needs. *The American Statistician*, **49**, 2–4.

[82] Kettenring, J. (1997). Message to students: Will you get a job in industry? *AmStat News*, **240**, 9–10.

[83] Kimball, A. (1957). Errors of the third kind in statistical consulting. *JASA*, **52**, 133–142.

[84] Kirk, R. (1991). Statistical consulting in a university: Dealing with people and other challenges. *The American Statistician*, **45**, 28–34.

[85] Kotz, S. and Johnson, N., Editors (1982). *Encyclopedia of Statistical Sciences.* Wiley, New York. (in 9 volumes).

[86] Kraemer, H. and Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research.* Sage, Newbury Park, CA.

[87] Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–269.

[88] Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

[89] Lurie, W. (1958). The impertinent questioner: The scientist's guide to the statistician's mind. *American Scientist*, **46**, 57–61.

[90] Maindonald, J. (1992). Statistical design, analysis, and presentation issues. *New Zealand Journal of Agricultural Research*, **35**, 121–141.

[91] Manchester, L., Field, C., and McDougall, A. (1999). Regression for overdetermined systems. A fisheries example. *The Canadian Journal of Statistics*, **27**, 25–39.

[92] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis.* Academic Press, New York.

[93] Marquardt, D. (1979). Statistical consulting in industry. *The American Statistician*, **33**, 72–76.

[94] Mason, S. (1962). *A History of the Sciences.* MacMillian, New York.

[95] MathSoft (1997). *S-Plus Trellis Graphics User's Manual.* MathSoft Inc, Seattle, WA.

[96] MathSoft (2000). *S-PLUS 6.0 Guide to Statistics (for UNIX).* MathSoft Inc, Seattle, WA. Data Analysis Division.

[97] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* 2nd edition. Chapman & Hall, London.

[98] McDougall, A. J. and Cook, D. (1994a). *Exploring Time Series Using Interactive Graphics.* Video. ASA Statistical Graphics Video Library.

[99] McDougall, A. J. and Cook, D. (1994b). XQz: An X application for Interactive Exploratory Time Series Analysis. *Contributed software.* StatLib: `lib.stat.cmu.edu`.

[100] Meinert, C. (1986). *Clinical Trials: Design, Conduct and Analysis.* Oxford University Press, New York.

[101] Miller, R., Efron, B., Brown, B., and Moses, L., Editors (1980). *Biostatistics Casebook.* Wiley, New York.

[102] Montgomery, D. (1997). *Design and Analysis of Experiments.* 4th edition. Wiley, New York.

[103] Moore, D. and McCabe, G. (1998). *Introduction to the Practice of Statistics.* 3rd edition. Freeman, New York. (2nd ed., 1993).

[104] Morgan, B. (1992). *Analysis of Quantal Response Data.* Chapman & Hall, New York.

[105] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

[106] Myers, R. and Montgomery, D. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments.* Wiley, New York.

[107] Newton, J. and Schmiediche, H. (1993). Graphical analysis for time series. In *Proc. of the 27th Symp. on the Interface: Computing Science and Statistics*, Amsterdam. Elsevier.

[108] Ott, L. (1993). *An introduction to statistical methods and data analysis.* PWS-Kent, Boston.

[109] Peace, K., Editor (1988). *Biopharmaceutical Statistics for Drug Development.* Marcel Dekker, New York.

[110] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.*, **50**(Series 5), 157–175. Reprinted in *Karl Pearson's Early Statistical Papers*, ed. by E.S. Pearson (1948), Cambridge Universty Press.

[111] Plewis, I. (1985). *Analysing Change: Measurement and Explanation using Longitudinal Data.* Wiley, Chichester.

[112] Pocock, S. (1983). *Clinical Trials: A Practical Approach.* Wiley, Chichester.

[113] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C: The art of scientific computing.* 2nd edition. Cambridge University Press, New York.

[114] Priestley, M. (1981). *Spectral Analysis and Time Series.* Academic Press, London.

[115] Radelet, M. (1981). Racial characteristics and the imposition of the death penalty. *Amer. Sociolog. Review*, **46**, 918–927.

[116] Rencher, A. (1995). *Methods of Multivariate Analysis.* Wiley, New York.

[117] Ripley, B. (1994). Introductory guide to S-Plus. PostScript file available from StatLib: `http://lib.stat.cmu.edu/S/sguide2.ps`.

[118] Ross, N. (1995). What government needs. *The American Statistician*, **49**, 7–9.

[119] Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

[120] Rustagi, J. and Wolfe, D., Editors (1982). *Teaching of Statistics and Statistical Consulting.* Academic Press, New York.

[121] SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 6.* SAS Institute Inc., Cary, NC.

[122] SAS Institute Inc. (2000). *JMP, Version 4.* SAS Institute Inc., Cary, NC.

[123] Scheffé, H. (1959). *The Analysis of Variance.* Wiley, New York.

[124] Scott, D. (1979). On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

[125] Searle, S. (1987). *Linear Models for Unbalanced Data.* Wiley, New York.

[126] Seber, G. (1977). *Linear Regression Analysis.* Wiley, New York.

[127] Seber, G. (1984). *Multivariate Observations.* Wiley, New York.

[128] Seber, G. and Wild, C. (1989). *Nonlinear Regression.* Wiley, New York.

[129] Selvin, S. (1996). *Statistical Analysis of Epidemiologic Data.* Oxford University Press, Oxford.

[130] Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Wiley, New York.

[131] Shewhart, W. (1931). *Economic Control of Quality of Manufactured Products.* D. Van Nostrand, Princeton, NJ.

[132] Snedecor, G. (1937). *Statistical Methods.* Iowa State Press, Ames, IA.

[133] Sprent, P. (1993). *Applied Nonparametric Statistical Methods.* 2nd edition. Chapman & Hall, London.

[134] Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press, Cambridge, MA.

[135] Stoffer, D. S., Tyler, D. E., and McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, **80**, 611–622.

[136] Student (1908). The probable error of a mean. *Biometrika*, **6**, 1–25. Reprinted on pp. 11–34 in *"Student's" Collected Papers.* Edited by E.S. Pearson and John Wishart with a Foreword by Launce McMullen, Cambridge University Press for the Biometrika Trustees, 1942.

[137] Swayne, D., Cook, D., and Buja, A. (1991). XGobi: Interactive dynamic graphics in the X Window System with a link to S. In *ASA Proc. of the Section on Statistical Graphics*, pages 1–8.

[138] Tanur, J., Editor (1988). *Statistics: A Guide to the Unknown.* 3rd edition. Duxbury Press, Belmont, CA.

[139] Thisted, R. (1988). *Elements of Statistical Computing.* Chapman & Hall, New York.

[140] Thompson, S. (1992). *Sampling.* Wiley, New York.

[141] Tierney, L. (1991). *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics.* Wiley, New York.

[142] Tufte, E. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut.

[143] Tufte, E. (1997). *Envisioning Information.* Graphics Press, Cheshire, Connecticut.

[144] Tufte, E. (1999). *Visual Explanations: Images and Quantities, Evidence and Narrative.* Graphics Press, Cheshire, Connecticut.

[145] Tukey, J. (1953). The problem of multiple comparisons. Unpublished Notes, Princeton University.

[146] Tukey, J. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA. Originally published in 1970.

[147] Tweedie, R. (1998). Consulting: real problems, real interactions, real outcomes. *Statistical Science*, **13**, 1–29.

[148] Venables, W. and Ripley, B. (1994). *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

[149] Williford, W., Krol, W., Bingham, S., Collins, J., and Weiss, D. (1995). The multicenter clinical trials coordinating center statistician: More than just a consultant. *The American Statistician*, **49**, 221–225.

[150] Yeatts, D. and Hyten, C. (1998). *High-Performing Self-Managed Work Teams: A Comparison of Theory and Practice.* Sage, Thousand Oaks, CA.

Index