Unbalanced Analysis of Variance, Design, and Regression:
Applied Statistical Methods

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico
To Mark, Karl, and John
It was great fun.
Contents

Preface xv

1 Introduction 1
   1.1 Probability 1
   1.2 Random variables and expectations 4
      1.2.1 Expected values and variances 6
      1.2.2 Chebyshev’s inequality 9
      1.2.3 Covariances and correlations 10
      1.2.4 Rules for expected values and variances 12
   1.3 Continuous distributions 13
   1.4 The binomial distribution 17
      1.4.1 Poisson sampling 21
   1.5 The multinomial distribution 21
      1.5.1 Independent Poissons and Multinomials 23
   1.6 Exercises 24

2 One sample 27
   2.1 Example and introduction 27
   2.2 Parametric Inference about $\mu$ 31
      2.2.1 Significance tests 34
      2.2.2 Confidence intervals 37
      2.2.3 $P$ values 38
   2.3 Prediction intervals 39
   2.4 Model Testing 42
   2.5 Checking normality 43
   2.6 Transformations 48
   2.7 Inference about $\sigma^2$ 52
      2.7.1 Theory 54
   2.8 Exercises 55

3 General Statistical Inference 59
   3.1 Model Based Testing 60
      3.1.1 An Alternative $F$ Test 66
   3.2 Inference on Single Parameters: Assumptions 66
   3.3 Parametric Tests 68
   3.4 Confidence Intervals 71
   3.5 $P$ values 74
   3.6 Validity of tests and confidence intervals 76
   3.7 Theory of prediction intervals 77
   3.8 Sample size determination and power 80
   3.9 The Shape of Things to Come 82
   3.10 Exercises 86
4 Two sample problems
  4.1 Two correlated samples: paired comparisons 89
  4.2 Two independent samples with equal variances 92
    4.2.1 Model testing 97
  4.3 Two independent samples with unequal variances 98
  4.4 Testing equality of the variances 103
  4.5 Exercises 106

5 Contingency Tables 111
  5.1 One binomial sample 111
    5.1.1 The sign test 114
  5.2 Two independent binomial samples 114
  5.3 One multinomial sample 117
  5.4 Two independent multinomial samples 119
  5.5 Several independent multinomial samples 123
  5.6 Lancaster–Irwin partitioning 125
  5.7 Exercises 132

6 Simple linear regression 135
  6.1 An Example 135
  6.2 The Simple Linear Regression Model 141
  6.3 The Analysis of Variance Table 143
  6.4 Model Based Inference 144
  6.5 Parametric Inferential Procedures 145
  6.6 An Alternative Model 148
  6.7 Correlation 148
  6.8 Two-Sample Problems 149
  6.9 A Multiple Regression. 150
  6.10 Estimation Formulae for Simple Linear Regression 151
  6.11 Exercises 156

7 Model Checking 159
  7.1 Recognizing randomness: simulated data with zero correlation 159
  7.2 Checking assumptions: residual analysis 161
    7.2.1 Another example 165
    7.2.2 Outliers 166
    7.2.3 Effects of high leverage 170
  7.3 Transformations 171
    7.3.1 Circle of Transformations 171
    7.3.2 Box–Cox transformations 173
    7.3.3 Constructed Variables 177
  7.4 Exercises 180

8 Testing Lack of Fit 181
  8.1 Polynomial regression 182
    8.1.1 Picking a polynomial 183
    8.1.2 Exploring the chosen model 184
  8.2 Polynomial Regression and Leverages 187
  8.3 Other Basis Functions 191
    8.3.1 High Order Models 193
  8.4 Partitioning Methods 193
    8.4.1 Utts’ Method 197
CONTENTS

8.5 Fisher’s Lack-of-fit Test 199
8.6 Exercises 199

9 Multiple regression: introduction 203
9.1 Example of inferential procedures 203
9.2 Regression surfaces and prediction 209
9.3 Comparing regression models 211
9.4 Sequential fitting 214
9.5 Reduced models and prediction 216
9.6 Partial correlation coefficients and added variable plots 217
9.7 Collinearity 218
9.8 More on model testing 220
9.9 Additive Effects and Interaction 224
9.10 Generalized Additive Models 226
9.11 Final Comment 228
9.12 Exercises 228

10 Diagnostics and Variable Selection 233
10.1 Diagnostics 233
10.2 Best Subset Model Selection 239
   10.2.1 $R^2$ statistic 240
   10.2.2 Adjusted $R^2$ statistic 241
   10.2.3 Mallows’s $C_p$ statistic 242
   10.2.4 A combined subset selection table 244
10.3 Stepwise Model Selection 244
   10.3.1 Backwards elimination 245
   10.3.2 Forward selection 246
   10.3.3 Stepwise methods 246
10.4 Model Selection and Case Deletion 247
10.5 Exercises 249

11 Multiple regression: matrix formulation 253
11.1 Random vectors 253
11.2 Matrix formulation of regression models 254
11.3 Least squares estimation of regression parameters 260
11.4 Inferential procedures 264
11.5 Residuals, standardized residuals, and leverage 267
11.6 Principal components regression 268
11.7 Exercises 273

12 One-way analysis of variance 275
12.1 Example 275
   12.1.1 Inferences on a Single Group Mean 279
   12.1.2 Inference on Pairs of Means 279
   12.1.3 Inference on Linear Functions of Means 281
   12.1.4 Testing $\mu_1 = \mu_2 = \mu_3$ 282
   12.1.5 Computing 282
12.2 Theory 283
   12.2.1 Analysis of Variance Tables 288
12.3 Regression Analysis of ANOVA Data 289
   12.3.1 Testing a Pair of Means 290
   12.3.2 Model Testing 291
12.3.3 Another Choice 294

12.4 Modeling Contrasts 295
   12.4.1 A Hierarchical Approach 296
   12.4.2 Evaluating the Hierarchy 297
   12.4.3 Regression Analysis 301
   12.4.4 Relation to orthogonal contrasts 301
   12.4.5 Difficulties in General Unbalanced Analyses 301

12.5 Polynomial regression and one-way ANOVA 302
   12.5.1 Fisher’s Lack-of-Fit Test 308

12.6 Weighted least squares 310

12.7 Exercises 314

13 Multiple comparison methods 321
   13.1 “Fisher’s” least significant difference method 322
   13.2 Bonferroni adjustments 324
   13.3 Scheffé’s method 326
   13.4 Studentized range methods 328
      13.4.1 Tukey’s honest significant difference 328
   13.5 Summary of multiple comparison procedures 329
   13.6 Exercises 330

14 Two-Way Anova 333
   14.1 Unbalanced Two-Way Analysis of Variance 333
      14.1.1 Initial Analysis 333
      14.1.2 Hierarchy of Models 337
      14.1.3 Computing Issues 338
      14.1.4 Discussion of Model Fitting 339
      14.1.5 Diagnostics 340
      14.1.6 Outlier Deleted Analysis 340
   14.2 Modeling Contrasts 344
      14.2.1 Nonequivalence of tests 345
   14.3 Regression Modeling 347

14.4 Homologous Factors 350
   14.4.1 Symmetric Additive Effects 350
   14.4.2 Skew Symmetric Additive Effects 352
   14.4.3 Symmetry 353
   14.4.4 Hierarchy of Models 355

15 ACOVA and Interactions 357
   15.1 One Covariate Example 357
      15.1.1 Additive Regression Effects 358
      15.1.2 Interaction Models 360
      15.1.3 Multiple Covariates 364
   15.2 Regression modeling 365
      15.2.1 Using Overparameterized Models 367
   15.3 ACOVA and Two-Way ANOVA 368
      15.3.1 Additive Effects 368
   15.4 Near Replicate Lack of Fit Tests 372
   15.5 Exercises 373
## CONTENTS

### 16 Multifactor structures
- 16.1 Unbalanced multifactor analysis of variance
  - 16.1.1 Computing: Anova
  - 16.1.2 Regression Fitting
- 16.2 Balanced Three Factors
- 16.3 Higher order structures
- 16.4 Exercises

### 17 Basic experimental designs
- 17.1 Experiments and Causation
- 17.2 Technical Design Considerations
- 17.3 Completely randomized designs
- 17.4 Randomized complete block designs
  - 17.4.1 Paired comparisons
- 17.5 Latin square designs
- 17.6 Balanced incomplete block designs
  - 17.6.1 Special cases
- 17.7 Youden squares
- 17.8 Analysis of covariance in designed experiments
- 17.9 Discussion of experimental design
- 17.10 Analytic and enumerative studies
- 17.11 Exercises

### 18 Factorial Treatments
- 18.1 Factorial Treatment Structures
- 18.2 Analysis
- 18.3 Modeling factorials
- 18.4 Interaction in a Latin Square
- 18.5 A Balanced Incomplete Block Design
- 18.6 Extensions of Latin squares
- 18.7 Exercises

### 19 Dependent Data
- 19.1 The analysis of split plot designs
  - 19.1.1 Modeling with Interaction
- 19.2 A Four Factor Example
  - 19.2.1 Unbalanced Subplot Analysis
  - 19.2.2 Whole Plot Analysis
  - 19.2.3 Fixing Effect Levels
  - 19.2.4 Final Models and Estimates
- 19.3 Multivariate analysis of variance
- 19.4 Random effects models
  - 19.4.1 Subsampling
  - 19.4.2 Random effects
- 19.5 Exercises

### 20 Confounding and fractional replication in $2^n$ factorial systems
- 20.1 Confounding
- 20.2 Fractional replication
- 20.3 Analysis of unreplicated experiments
- 20.4 More on graphical analysis
- 20.5 Augmenting designs for factors at two levels

---

**Page Numbers:**
- Multifactor structures: 377
- Basic experimental designs: 395
- Factorial Treatments: 419
- Dependent Data: 439
- Confounding and fractional replication in $2^n$ factorial systems: 475
CONTENTS

24.3 Statistical inference 579
24.4 Linearizable models 589
24.5 Exercises 589

Appendix A: Matrices and Vectors 591
A.1 Matrix addition and subtraction 592
A.2 Scalar multiplication 592
A.3 Matrix multiplication 592
A.4 Special matrices 594
A.5 Linear dependence and rank 595
A.6 Inverse matrices 596
A.7 A list of useful properties 598
A.8 Eigenvalues and eigenvectors 598

Appendix B: Tables 601
B.1 Tables of the $t$ distribution 602
B.2 Tables of the $\chi^2$ distribution 604
B.3 Tables of the $W'$ statistic 608
B.4 Tables of orthogonal polynomials 609
B.5 Tables of the Studentized range 610
B.6 The Greek alphabet 614
B.7 Tables of the $F$ distribution 615

References 627
In 1996 I published a book *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Within four years, I knew that the book was not what I thought needed to be taught in the 21st century, cf. Christensen (2000). This book *Unbalanced Analysis of Variance, Design, and Regression: Applied Statistical Methods* shares with the earlier book most of the title, much of the data, and even some of the text, but the book is radically different. The original book focused greatly on balanced analysis of variance. By making the effort to treat unbalanced analysis of variance, one can easily handle a wide range of models for nonnormal data, because the same fundamental methods apply.

After the effort to complete this book, the word added to the title may refer to me just as much as it does to the analyses.
Preface to “First Edition”

This book examines the application of basic statistical methods: primarily analysis of variance and regression but with some discussion of count data. It is directed primarily towards Masters degree students in statistics studying analysis of variance, design of experiments, and regression analysis. I have found that the Masters level regression course is often popular with students outside of statistics. These students are often weaker mathematically and the book caters to that fact while continuing to give a complete matrix formulation of regression.

The book is complete enough to be used as a second course for upper division and beginning graduate students in statistics and for graduate students in other disciplines. To do this, one must be selective in the material covered, but the more theoretical material appropriate only for Statistics Masters students is generally isolated in separate subsections and, less often, in separate sections.

For a Masters level course in analysis of variance and design, I have the students review Chapter 2, I present Chapter 3 while simultaneously presenting the examples of Section 4.2, I present Chapters 5 and 6, very briefly review the first five sections of Chapter 7, present Sections 7.11 and 7.12 in detail and then I cover Chapters 9, 10, 11, 12, and 17. Depending on time constraints, I will delete material or add material from Chapter 16.

For a Masters level course in regression analysis, I again have the students review Chapter 2 and I review Chapter 3 with examples from Section 4.2. I then present Chapters 7, 13, and 14, Appendix A, Chapter 15, Sections 16.1.2, 16.3, 16.5 (along with analysis of covariance), Section 8.7 and finally Chapter 18. All of this is done in complete detail. If any time remains I like to supplement the course with discussion of response surface methods.

As a second course for upper division and beginning graduate students in statistics and graduate students in other disciplines, I cover the first eight chapters with omission of the more technical material. A follow up course covers the less technical aspects of Chapters 9 through 15 and Appendix A.

I think the book is reasonably encyclopedic. It really contains everything I would like my students to know about applied statistics prior to them taking courses in linear model theory or log-linear models.

I believe that beginning students (even Statistics Masters students) often find statistical procedures to be a morass of vaguely related special techniques. As a result, this book focuses on four connecting themes.

1. Most inferential procedures are based on identifying a (scalar) parameter of interest, estimating that parameter, obtaining the standard error of the estimate, and identifying the appropriate reference distribution. Given these items, the inferential procedures are identical for various parameters.

2. Balanced one-way analysis of variance has a simple, intuitive interpretation in terms of comparing the sample variance of the group means with the mean of the sample variances for each group. All balanced analysis of variance problems are considered in terms of computing sample variances for various group means.

3. Comparing different models provides a structure for examining both balanced and unbalanced analysis of variance problems and for examining regression problems. In some problems the most reasonable analysis is simply to find a succinct model that fits the data well.

4. Checking assumptions is a crucial part of every statistical analysis.
The object of statistical data analysis is to reveal useful structure within the data. In a model-based setting, I know of two ways to do this. One way is to find a succinct model for the data. In such a case, the structure revealed is simply the model. The model selection approach is particularly appropriate when the ultimate goal of the analysis is making predictions. This book uses the model selection approach for multiple regression and for general unbalanced multifactor analysis of variance. The other approach to revealing structure is to start with a general model, identify interesting one-dimensional parameters, and perform statistical inferences on these parameters. This parametric approach requires that the general model involve parameters that are easily interpretable. We use the parametric approach for one-way analysis of variance, balanced multifactor analysis of variance, and simple linear regression. In particular, the parametric approach to analysis of variance presented here involves a strong emphasis on examining contrasts, including interaction contrasts. In analyzing two-way tables of counts, we use a partitioning method that is analogous to looking at contrasts.

All statistical models involve assumptions. Checking the validity of these assumptions is crucial because the models we use are never correct. We hope that our models are good approximations to the true condition of the data and experience indicates that our models often work very well. Nonetheless, to have faith in our analyses, we need to check the modeling assumptions as best we can. Some assumptions are very difficult to evaluate, e.g., the assumption that observations are statistically independent. For checking other assumptions, a variety of standard tools has been developed. Using these tools is as integral to a proper statistical analysis as is performing an appropriate confidence interval or test. For the most part, using model-checking tools without the aid of a computer is more trouble than most people are willing to tolerate.

My experience indicates that students gain a great deal of insight into balanced analysis of variance by actually doing the computations. The computation of the mean square for treatments in a balanced one-way analysis of variance is trivial on any hand calculator with a variance or standard deviation key. More importantly, the calculation reinforces the fundamental and intuitive idea behind the balanced analysis of variance test, i.e., that a mean square for treatments is just a multiple of the sample variance of the corresponding treatment means. I believe that as long as students find the balanced analysis of variance computations challenging, they should continue to do them by hand (calculator). I think that automated computation should be motivated by boredom rather than bafflement.

In addition to the four primary themes discussed above, there are several other characteristics that I have tried to incorporate into this book.

I have tried to use examples to motivate theory rather than to illustrate theory. Most chapters begin with data and an initial analysis of that data. After illustrating results for the particular data, we go back and examine general models and procedures. I have done this to make the book more palatable to two groups of people: those who only care about theory after seeing that it is useful and those unfortunates who can never bring themselves to care about theory. (The older I get, the more I identify with the first group. As for the other group, I find myself agreeing with W. Edwards Deming that experience without theory teaches nothing.) As mentioned earlier, the theoretical material is generally confined to separate subsections or, less often, separate sections, so it is easy to ignore.

I believe that the ultimate goal of all statistical analysis is prediction of observable quantities. I have incorporated predictive inferential procedures where they seemed natural.

The object of most statistics books is to illustrate techniques rather than to analyze data; this book is no exception. Nonetheless, I think we do students a disservice by not showing them a substantial portion of the work necessary to analyze even ‘nice’ data. To this end, I have tried to consistently examine residual plots, to present alternative analyses using different transformations and case deletions, and to give some final answers in plain English. I have also tried to introduce such material as early as possible. I have included reasonably detailed examinations of a three-factor analysis of variance and of a split plot design with four factors. I have included some examples in which, like real life, the final answers are not ‘neat.’ While I have tried to introduce statistical ideas as soon as possible, I have tried to keep the mathematics as simple as possible for as long as possible.
For example, matrix formulations are postponed to the last chapter on multiple regression and the last section on unbalanced analysis of variance.

I never use side conditions or normal equations in analysis of variance.

In multiple comparison methods, (weakly) controlling the experimentwise error rate is discussed in terms of first performing an omnibus test for no treatment effects and then choosing a criterion for evaluating individual hypotheses. Most methods considered divide into those that use the omnibus $F$ test, those that use the Studentized range test, and the Bonferroni method, which does not use any omnibus test.

I have tried to be very clear about the fact that experimental designs are set up for arbitrary groups of treatments and that factorial treatment structures are simply an efficient way of defining the treatments in some problems. Thus, the nature of a randomized complete block design does not depend on how the treatments happen to be defined. The analysis always begins with a breakdown of the sum of squares into treatments, blocks, and error. Further analysis of the treatments then focuses on whatever structure happens to be present.

The analysis of covariance chapter includes an extensive discussion of how the covariates must be chosen to maintain a valid experiment. Tukey’s one degree of freedom test for nonadditivity is presented as an analysis of covariance test for the need to perform a power transformation rather than as a test for a particular type of interaction.

The chapter on confounding and fractional replication has more discussion of analyzing such data than many other books contain.

Minitab commands are presented for most analyses. Minitab was chosen because I find it the easiest of the common packages to use. However, the real point of including computer commands is to illustrate the kinds of things that one needs to specify for any computer program and the various auxiliary computations that may be necessary for the analysis. The other statistical packages used in creating the book were BMDP, GLIM, and MSUSTAT.

Acknowledgements

Many people provided comments that helped in writing this book. My colleagues Ed Bedrick, Aparna Huzurbazar, Wes Johnson, Bert Koopmans, Frank Martin, Tim O’Brien, and Cliff Qualls helped a lot. I got numerous valuable comments from my students at the University of New Mexico. Marjorie Bond, Matt Cooney, Jeff S. Davis, Barbara Evans, Mike Fugate, Jan Mines, and Jim Shields stand out in this regard. The book had several anonymous reviewers, some of whom made excellent suggestions.

I would like to thank Martin Gilchrist and Springer-Verlag for permission to reproduce Example 7.6.1 from Plane Answers to Complex Questions: The Theory of Linear Models. I also thank the Biometrika Trustees for permission to use the tables in Appendix B.5. Professor John Deely and the University of Canterbury in New Zealand were kind enough to support completion of the book during my sabbatical there.

Now my only question is what to do with the chapters on quality control, $p^n$ factorials, and response surfaces that ended up on the cutting room floor.

Ronald Christensen
Albuquerque, New Mexico
February 1996
BMDP Statistical Software is located at 1440 Sepulveda Boulevard, Los Angeles, CA 90025, telephone: (213) 479-7799

MINITAB is a registered trademark of Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, telephone: (814) 238-3280, telex: 881612.

MSUSTAT is marketed by the Research and Development Institute Inc., Montana State University, Bozeman, MT 59717-0002, Attn: R.E. Lund.