

Design of Experiments (DOE) with R

Dieter Rasch, Jürgen Pilz and Petr Šimeček

Preface

Experimental design is the stepchild of applied and mathematical statistics. In hundreds of text books and monographs about basic and advanced statistics, nothing is said about planning a survey or a design - statistics is understood there as a collection of methods for analysing data only. As a consequence of this situation, experimenters seldom think about an optimal design and the necessary sample size needed for a precise answer for an experimental question. This situation consequently is reflected in statistical program packages – they mainly are packages for data analysis. This is also the case for the S- or R-packages and for books describing statistics by R as:

Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language*. Chapman & Hall, London, 1988.

W.J. Brown and D.J. Murdoch. *A First Course in Statistical Computing with R*. Cambridge University Press, Cambridge, 2008

John M. Chambers and Trevor J. Hastie. *Statistical Models in S*. Chapman & Hall, London, 1992.

John M. Chambers. *Programming with Data*. Springer, New York, 1998.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S. Fourth Edition*. Springer, New York, 2002. ISBN 0-387-95457-0.

William N. Venables and Brian D. Ripley. *S Programming*. Springer, New York, 2000.

Jose C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer, New York, 2000.

Peter Dalgaard. *Introductory Statistics with R*. Springer, 2002.

Stefano Iacus and Guido Masarotto. *Laboratorio di statistica con R*. McGraw-Hill, Milano, 2003.

John Maindonald and John Braun. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, 2003.

Julian J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL, 2004.

Richard M. Heiberger and Burt Holland. *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS*. Springer Texts in Statistics. Springer, New York, 2004.

Fionn Murtagh. *Correspondence Analysis and Data Coding with JAVA and R*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

Paul Murrell. *R Graphics*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

Michael J. Crawley. *Statistics: An Introduction using R*. Wiley, New York, 2005.

Brian S. Everitt. *An R and S-Plus Companion to Multivariate Analysis*. Springer, New York, 2005.

Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael Irizarry, and

- Sandrine Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. Springer, New York, 2005.
- Brian Everitt and Torsten Hothorn. *A Handbook of Statistical Analyses Using R*. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- Julian J. Faraway. *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- Jana Jureckova and Jan Picek. *Robust Statistical Methods with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006.
- Simon N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 1-584-88474-6.
- Bernhard Pfaff. *Analysis of Integrated and Cointegrated Time Series with R. Use R*. Springer, New York, 2006. ISBN 0-387-98784-3.
- Emmanuel Paradis. *Analysis of Phylogenetics and Evolution with R. Use R*. Springer, New York, 2006.
- Uwe Ligges. *Programmieren mit R*. Springer-Verlag, Heidelberg, 2nd edition, 2007.
- Dubravko Dolic. *Statistik mit R. Einführung für Wirtschafts- und Sozialwissenschaftler*. R. Oldenbourg, München, Wien, 2004.
- Andreas Behr. *Einführung in die Statistik mit R*. WiSo Kurzlehrbücher. Vahlen, München, 2005.
- Jim Albert. *Bayesian Computation with R*. Springer, New York, 2007.
- Lothar Sachs and Jürgen Hedderich. *Angewandte Statistik. Methodensammlung mit R*. Springer, Berlin, Heidelberg, 12th (completely revised) edition, 2006.
- Stefano M. Iacus. *Simulation and Inference for Stochastic Differential Equations: With R Examples*. Springer, New York, 2008.
- Maria L. Rizzo. *Statistical Computing with R*. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 1-584-88545-9.
- Robert Gentleman. *Bioinformatics with R*. Chapman & Hall/CRC, Boca Raton, FL, 2008. ISBN 1-420-06367-7.
- Deepayan Sarkar. *Lattice Multivariate Data Visualization with R*. Springer, New York, 2007.
- John M. Chambers. *Software for Data Analysis: Programming with R*. Springer, New York, 2007.
- W. John Braun and Duncan J. Murdoch. *A First Course in Statistical Programming with R*. Cambridge University Press, Cambridge, 2007.
- Julien Claude. *Morphometrics with R*. Springer, New York, 2008.
- Bernhard Pfaff. *Analysis of Integrated and Cointegrated Time Series with R, Second Edition*. Springer, New York, 2008.
- Phil Spector. *Data Manipulation with R*. Springer, New York, 2008.
- Jonathan D. Cryer and Kung-Sik Chan. *Time Series Analysis With Applications in R*. Springer, New York, 2008.
- Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, New York, 2006.
- Roger D. Peng and Francesca Dominici. *Statistical Methods for Environmental*

Epidemiology with R: A Case Study in Air Pollution and Health. Springer, New York, 2008.

Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R.* Springer, New York, 2008.

G. P. Nason. *Wavelet Methods in Statistics with R.* Springer, New York, 2008.

Christian Kleiber and Achim Zeileis. *Applied Econometrics with R.* Springer, New York, 2008. ISBN 978-0-387-77316-2.

Clemens Reimann, Peter Filzmoser, Robert Garrett, and Rudolf Dutter. *Statistical Data Analysis Explained: Applied Environmental Statistics with R.* Wiley, Chichester, UK, 2008.

Simon Sheather. *A Modern Approach to Regression with R.* Springer, New York, 2008. ISBN 978-0-387-09607-0.

Christian Ritz and Jens C. Streibig. *Nonlinear Regression with R.* Springer, New York, 2008. ISBN 978-0-387-09615-5.

Alain Zuur, Elena N. Ieno, Neil Walker, and Graham M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer, New York, 2009. ISBN 978-0-387-87457-9.

Vito Ricci. Rappresentazione analitica delle distribuzioni statistiche con R (prima parte). *Economia e Commercio*, 1/2:47-60, 2005.

By reading the title of most of the books above, we can already see that they deal with data analysis only.

The present book has the following goals:

1. to introduce experimenters into the philosophy of experimentation and the need for designing experiments and data collection,
2. to give experimenters and consulting statisticians an easy process for constructing optimum experimental designs and calculating the size needed in experimentation by using R programs.
3. to show by examples how the R programs should be used,
4. to give mathematicians interested in the theoretical background of experimental designs the theoretical background of the programs in a short theoretical last chapter 7.

Together with this book the R – program package DOE is developed by the third author. It follows the structure of the book, and it will be available as soon a the book is published.

Because the authors are not native English speakers, we are happy that we found help from Sandra Almgren, Kremmling, Colorado, USA.

..... We have to thank her for many helpful suggestions and corrections.

Wien, Klagenfurt and Prague, Spring 2009

Table of Contents

- 1 Introduction
- 1.1 Experimentation and empirical research
- 1.2 Designing experiments
- 1.3 Some basic definitions

1.4 Block designs

1.5 About the R programs

Part A Determining the minimal Size of Experiments for given Precision

2. Sample Size Determination in Completely Randomized Designs¹

2.1 Introduction

2.2 Confidence Estimation

2.2.1 Confidence intervals for expectations

2.2.1.1 One-sample case , σ^2 known:

2.2.1.2 One-sample case , σ^2 unknown:

2.2.1.3 Confidence Intervals for the Expectation of the Normal Distribution in the Presence of a Noise Factor

2.2.1.4 One-sample case , σ^2 unknown, paired observations:

2.2.1.5 Two-sample case , σ^2 unknown, independent samples – equal variances.:

2.2.1.6 Two-sample case , σ^2 unknown, independent samples – unequal variances

2.2.2 Confidence Intervals for Probabilities

2.2.3 Confidence Interval for the Variance of the Normal Distribution

2.3 Selection Procedures

2.4 Testing hypothesis

2.4.1 Testing Hypotheses about Means of Normal Distributions

2.4.1.1 One-sample problem, univariate

2.4.1.2 One-sample problem, bivariate

2.4.1.3 Two-sample problem, equal variances

2.4.1.4 Two-sample problem, unequal variances

2.4.1.5 Comparing more than two means, equal variances

2.4.2 Testing hypotheses about probabilities

2.4.2.1 One-sample problem

2.4.2.2 Two-sample problem

2.5 Summary of sample size formulae

3. Size of Experiments in Analysis of variance models

3.1 Introduction

3.2 One-way layout

3.3 Two-way layout

3.3.1 Two-way Analysis of Variance - Cross-classification

3.3.1.1 Two-way Analysis of Variance - Cross-classification - Model I

3.3.1.2 Two-way Analysis of Variance - Cross-classification - Mixed Model

3.3.2 Nested-classification ($A \succ B$):

3.3.2.1 Two-way Analysis of Variance - Nested Classification - Model I

3.3.2.1 Two-way Analysis of Variance - Nested Classification - Mixed Model, A Fixed and B Random

3.3.2.2 Two-way Analysis of Variance - Nested Classification - Mixed Model, B Fixed and A Random

3.4 Three-way layout

3.4.1 Three-way layout – cross classification

3.4.1.1 Three-way Analysis of Variance – classification – $A \times B \times C$ Model I

3.4.1.2 Three-way Analysis of Variance – classification $A \times B \times C$ – Model III

3.4.1.3 Three-way Analysis of Variance – classification $A \times B \times C$ -Model IV

3.4.2 Three-way Analysis of Variance – nested classification $A \succ B \succ C$

3.4.2.1 Three-way Analysis of Variance – nested classification -Model I

3.4.2.2 Three-way Analysis of Variance – nested classification -Model III

- 3..4.2.3 Three-way Analysis of Variance – nested classification -Model IV
- 3..4.2.4 Three-way Analysis of Variance – nested classification -Model V
- 3..4.2.5 Three-way Analysis of Variance – nested classification -Model VI
- 3..4.2.6 Three-way Analysis of Variance – nested classification -Model VII
- 3..4.2.7 Three-way Analysis of Variance – nested classification -Model VIII
- 3..4.3 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$
- 3..4.3.1 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$ Model I:
- 3..4.3.2 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$ Model III
- 3..4.3.3 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$ Model IV
- 3..4.3.4 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$ Model V
- 3..4.3.5 Three-way Analysis of Variance – mixed classification $(A \times B) \succ C$ Model VI
- 3..4.4 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$
- 3..4.4.1 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model I
- 3..4.4.2 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model III
- 3..4.4.3 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model IV
- 3..4.4.4 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model V
- 3..4.4.5 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model VI
- 3..4.4.6 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model VII
- 3..4.4.7 Three-way Analysis of Variance – mixed classification $(A \succ B) \times C$ Model VIII

4 Sample Size Determination in model II of regression analysis

- 4..1 Introduction
- 4.2 Confidence intervals
 - 4.2.1 A confidence interval for the correlation coefficient
 - 4.2.2 Confidence intervals for partial correlation coefficients.
 - 4.2.3 A confidence interval for $E(y|x) = \beta_0 + \beta_1 x$
- 4.3 Hypothesis testing
 - 4.3.1 Comparing the correlation coefficient with a constant
 - 4.3.2 Comparing two correlation coefficients.
 - 4.3.3 Comparing the slope with a constant
- 5. Sequential Designs
 - 5.1 Introduction
 - 5.2 Wald's sequential Likelihood Ratio Test (SLRT)

Part B Constructing optimal designs

- 6. Constructing Balanced Incomplete Block Designs
 - 6.6.1 Introduction
- 6.2 Testing the absence of interactions
- 6.3 Basic Definitions
- 6.4 Construction of BIBD
 - 6.4.1 Specific methods
 - 6.4.2 A general Method
- 6.5 Appendix

7 Constructing Fractional Factorial Designs

- 7.1 Introduction and basic notations
- 7.2 Factorial designs – basic definitions
- 7.3 Fractional factorials designs with two level of each factor (2^{p-k} -designs)
- 7.4 Fractional factorials designs with three level of each factor (3^p -designs)
- 7.5 Fractional factorials designs of mixed type (with two or three level per factor) ($2^m 3^{p-m}$ -designs)
- 8 Exact Optimal Designs and Sample Sizes in Model I of Regression Analysis and Mixture Designs
 - 8.1 Introduction
 - 8.2 The multiple linear regression model I
 - 8.3 Simple Polynomial Regression
 - 8.4 Intrinsically Non-linear Regression
 - 8.5 *Exact Φ* - optimal Designs
 - 8.5.1 Simple Linear regression.
 - 8.5.2 Polynomial regression.
 - 8.5.3 Intrinsically non-linear regression
 - 8.5.4 Replication free designs
 - 8.6 Determining the Size of an Experiment
 - 8.7 Central Composite Designs
 - 8.8 Mixture Designs
 - 8.8.1. Introduction
 - 8.8.2 The Simplex Lattice Designs
 - 8.8.3 Simplex Centroid Designs
 - 8.8.4 Augmented Lattice Designs
 - 8.8.5. Extreme Vertice Designs
 - 8.8.6 Constructing mixture designs with R

Part C Special designs and mathematical background

- 9. Weighing Designs
 - 9.1. Introduction
 - 9.2 The general approach
- 10. Theoretical background
 - 10.1 Groups, fields and finite geometries
 - 10.2 Difference sets
 - 10.3 Hadamard matrices
 - 10.4 Existence and Non-existence of BIBD

Why is designing experiments so important?

We demonstrate this by a negative example:

The following text in italics was taken from an information of the Ministry of Health in Austria entitled:

**SCIENTIFIC ARGUMENTS FOR AN IMPORT BAN OF GENETICALLY
MODIFIED MAIZE MON 863 (*Zea mays* L., line MON 863) OF MONSANTO
(NOTIFICATION C/DE/02/9)**

On 8th August 2005 the Decision (2005/608/EC) concerning the placing on the market, in accordance with Directive 2001/18/EC of genetically modified maize MON 863 was adopted by the Commission. The product may be placed on the market and put to the same uses as any other maize, with the exception of cultivation and uses as or in food.

On 13th January 2006 the placing on the market of foods and food ingredients derived from genetically modified maize line MON 863 as novel foods or novel food ingredients under Regulation (EC) No 258/97 was authorised.

*With regard to the **studies on nutritional equivalence assessment in farm animals**, which are quoted in HAMMOND et al. (2006) as scientific proof for the safety of maize MON 863, a lot of shortcomings have been detected:*

*In this document, the scientific arguments, which are justifying the Austrian import ban of this GMO, are described. They focus particularly on the toxicological safety assessment and the antibiotic resistance marker (ARM) gene *nptII*, which is contained in maize MON 863, but also on the given risk management measures to prevent accidental spillage.*

***Summarizing** the evaluation of the **toxicological safety assessment** of the dossier, it can be stated that a lot of deficiencies are obvious:*

*With regard to the **studies on nutritional equivalence assessment in farm animals**, which are quoted in HAMMOND et al. (2006) as scientific proof for the safety of maize MON 863, a lot of shortcomings have been detected:*

Concerning the experimental design it has to be criticised that reference groups are often contributing 60-80% of the sample size. Statistically significant differences between test and control groups are therefore often masked because group differences between iso- and transgenic diets fall into the broad range of reference groups.

An important factor is also the sensitivity of the animal model: HAMMOND et al. (2006) described the use of an outbred rat model. The study compared a high number of different lines of maize, among them MON 863. The data vary considerably in and between the groups. That would allow the assumption that only effects with great deviations from the control would have been detectable with the chosen trial setup.

Sandra do not correct the black text above, it is a citation from an EU document

So far a part of the official text of the ministry. Let us consider the design of Hammond et al. (2006) without the reference groups (the role say play is not quite clear to us). There were two control groups of rats fed with grain free of MON 863 and two groups fed with MON 863 one of them with 11% and one with 33% MON 863. The size of each group was $n = 20$. Now let us assume we like to test the null hypothesis that there is no decrease in fecundity of rats in the treatment groups compared with the control groups. That means that there is no argument (with regard to fecundity) not to use MON 863. Let us further assume that we use a risk of the first kind of 0.1 or alternatively 0.05 (the probability that we reject the null hypothesis if it is true, the producer risk of Monsanto) and at first a risk of the second kind of 0.01 (the probability that there is a decrease of fecundity even it is present) if the decrease is 2% or larger. Then we will find out by the method described in 2.4.2.2. assuming that the usual fecundity is around 0.8 that in each of two groups to be compared the group size should be 10893. An overview about the association between risk of the first and second kind, group

size and minimum detectable decrease (from 0.8) as calculated by R is given below (one-sided alternative hypothesis):

Risk β	$0.8 - \delta$	Size of each group; $\alpha = 0.1$	Size of each group; $\alpha = 0.05$
0.01	0.78	10893	13177
	0.70	504	607
	0.60	142	171
0.05	0.78	7201	9074
	0.70	338	423
	0.60	97	121
0.1	0.78	5547	7201
	0.7	264	339
	0.60	77	98
0.2	0.78	3338	5277
	0.70	188	250
	0.60	56	74

With 20 animals and with a high error rate of $100*\beta = 20\%$ the null hypothesis of no negative effect is accepted as long the real fecundity in the group treated with MON 863 lies above 0.4375. for $100*\beta = 1\%$ as long the real fecundity in the group treated with MON 863 lies above 0.1625. Of course such risks are totally unacceptable.