

Springer Series in Statistics

Edgar Brunner  
Arne C. Bathke  
Frank Konietzschke

# Rank and Pseudo- Rank Procedures for Independent Observations in Factorial Designs

Using R and SAS

 Springer

# Springer Series in Statistics

*Advisors:*

P. Diggle, U. Gather, S. Zeger

More information about this series at <http://www.springer.com/series/692>

Edgar Brunner • Arne C. Bathke • Frank  
Konietschke

# Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs

Using R and SAS

Edgar Brunner  
Department of Medical Statistics  
University of Göttingen  
University Medical Center  
Göttingen, Germany

Arne C. Bathke  
Department of Mathematics  
University of Salzburg  
Salzburg, Austria

Frank Konietschke  
Institute of Biometry and Clinical Epidemiology  
Charité – University Medical School  
Berlin, Germany

SAS<sup>®</sup> is a registered trademark of SAS Institute.

ISSN 0172-7397

ISSN 2197-568X (electronic)

Springer Series in Statistics

ISBN 978-3-030-02912-8

ISBN 978-3-030-02914-2 (eBook)

<https://doi.org/10.1007/978-3-030-02914-2>

Mathematics Subject Classification (2010): 62G10, 62G15, 62G20, 62P10, 62P15

© Springer Nature Switzerland AG 2018, corrected publication 2019

Partly based on a translation from the German language edition: Nichtparametrische Datenanalyse by Edgar Brunner and Ullrich Munzel. © Springer-Verlag Berlin Heidelberg 2013. All Rights Reserved

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

This is a book on modern nonparametric statistics for factorial designs, using ranks and pseudo-ranks. The field of nonparametric statistics may be most easily described by first introducing its counterpart, parametric statistics. Parametric statistics is concerned with modeling, representing, and analyzing data assumed to originate from known parameterized classes of distributions, for example from normal, exponential, or Poisson distributions. Typical effect sizes in parametric models are differences or ratios of parameters such as means and variances. Consequently, the validity of conclusions from parametric methods depends on whether the models and classes of distributions are appropriate, and whether the used effect sizes make sense.

This book describes a class of statistical methods for factorial designs that does not have to rely on specific parametric models, and the model distributions may be rather general. Indeed, response variables may be metric, ordinal or ordered categorical, or even binary. The approach presented here is unified; it is applicable for the analyses of discrete and continuous data. Thus, also corrections for tied values, as sometimes found in classical books, are obsolete. The underlying effect size is the nonparametric relative effect, which has a simple and intuitive probability interpretation. Its meaning, interpretation, and relation to other effect measures are explained in detail in Chap. 2 of the book, preceded by a chapter describing the designs covered in this book, distinguishing the variable scales, and introducing other terminology. Chapters 3–6 explain in detail statistical methodology, illustrative examples, and application using SAS and R, moving from the two-sample design to several samples, two factors, and finally three and more factors. The approach to data analysis presented here attempts to be as comprehensive as possible, including appropriate *descriptive* statistics which follow a nonparametric paradigm, as well as corresponding *inferential* methods using hypothesis tests and confidence intervals based on pseudo-ranks.

We generally recommend a unified nonparametric approach toward data analysis, as opposed to a cherry-picking use of nonparametric methods when data appear skewed or seem to exhibit outliers. The latter may lead to biased results and

to problems in comparing and interpreting different analyses, due to different invariance properties of the used methods (see Sect. 5.2.3 for more details).

With this book, we try to address a wide range of readers, from those who attempt to understand the methodologically underlying mathematical derivations and asymptotic results to those who only have a very basic statistical background and simply want to apply modern nonparametric techniques using R or SAS. In particular with the latter audience in mind, we have decided on a writing style that is generally as non-technical as possible, avoiding much of the theoretical terminology of probability and statistics, while still trying to be methodologically precise. The more technical details can, for the most part, be found in Chap. 7.

Finally, it should be mentioned that this book project originally started with the idea of updating, translating, and *slightly* extending the earlier textbook in German *Nichtparametrische Datenanalyse*, Springer, 2002 1st ed., 2013 2nd ed., that the first author of this volume coauthored with Ullrich Munzel. When we started this project, we didn't think it would keep on growing so much!

Göttingen, Germany  
Salzburg, Austria  
Berlin, Germany  
May 2018

Edgar Brunner  
Arne C. Bathke  
Frank Konietzschke

# Contents

<b>1</b>	<b>Types of Data and Designs</b>	1
1.1	Types of Data	1
1.1.1	Accuracy of a Scale	2
1.1.1.1	Continuous Scale	2
1.1.1.2	Discrete Scale	2
1.1.2	Distances on a Scale	2
1.1.2.1	Metric Scale	2
1.1.2.2	Ordinal Data	4
1.1.2.3	Binary (Dichotomous) Data	4
1.1.2.4	Nominal Data	5
1.2	Factors and Designs	5
1.2.1	Configuration of Factors	6
1.2.2	Designs	9
1.2.3	Use of Indices	11
1.2.4	Classification of Designs	11
<b>2</b>	<b>Distributions and Effects</b>	15
2.1	Distribution Functions	15
2.2	Relative Effects	16
2.2.1	Two Distributions	17
2.2.2	Application to Diagnostic Trials	25
2.2.3	How to Measure Effect Sizes?	29
2.2.3.1	Relative Effect	29
2.2.3.2	Standardized Mean Difference	29
2.2.3.3	Area Under the Receiver Operating Characteristic Curve (AUC of ROC Curve)	30
2.2.4	Several Distributions	30
2.2.4.1	Generalization of $p$	30
2.2.4.2	Relative Effects for Several Distributions, Efron's Paradoxical Dice	33
2.2.4.3	Independent Replications	37

2.2.5	Summary .....	42
2.2.5.1	Two Distributions .....	42
2.2.5.2	Several ( $a \geq 2$ ) Distributions: General Case ....	43
2.2.5.3	Several ( $a \geq 2$ ) Distributions: $n_i$ Independent Replications .....	44
2.3	Empirical Distributions and Ranks .....	45
2.3.1	Empirical Distribution Functions .....	45
2.3.2	Ranks and Pseudo-Ranks .....	50
2.3.3	Estimators of Relative Effects .....	61
2.3.4	Summary .....	63
2.4	Software for Computing Ranks and Pseudo-Ranks .....	67
2.4.1	Computing Ranks and Pseudo-Ranks Using SAS .....	67
2.4.2	Computing Ranks and Pseudo-Ranks Using R .....	70
2.5	Exercises and Problems .....	71
<b>3</b>	<b>Two Samples</b> .....	<b>75</b>
3.1	Introduction and Motivating Examples .....	75
3.1.1	Weight Gain .....	76
3.1.2	Number of Implantations .....	77
3.1.3	Irritation of the Nasal Mucosa .....	78
3.1.4	Leukocytes in the Urine .....	79
3.1.5	Features of the Examples .....	80
3.2	Models, Effects, and Hypotheses .....	80
3.2.1	Normal Distribution Model .....	80
3.2.2	Location Model .....	82
3.2.3	Lehmann Model .....	83
3.2.4	Nonparametric Model .....	85
3.3	Effect Estimators and Hypotheses .....	86
3.4	Wilcoxon–Mann–Whitney Test .....	88
3.4.1	Exact (Permutation) Distribution .....	89
3.4.1.1	Recursion Algorithm: No Ties .....	89
3.4.1.2	Shift Algorithm: No Ties .....	93
3.4.1.3	Recursion Algorithm: Ties Allowed .....	95
3.4.1.4	Shift Algorithm: Ties Allowed .....	96
3.4.2	Procedure for Large Sample Sizes .....	97
3.4.3	The So-Called Rank Transform .....	102
3.4.4	Application to Dichotomous Data .....	104
3.4.4.1	Fisher’s Exact Test .....	105
3.4.4.2	The Large Sample $\chi^2$ -Test .....	106
3.4.5	Analysis of the Examples .....	108
3.4.5.1	Analysis of Example 3.1.1 (Weight Gain) .....	111
3.4.5.2	Analysis of Example 3.1.2 (Number of Implantations) .....	112

	3.4.5.3	Analysis of Example 3.1.3 (Irritation of the Nasal Mucosa) .....	113
	3.4.5.4	Analysis of Example 3.1.4 (Leukocytes in the Urine) .....	114
	3.4.6	Summary .....	116
3.5		Nonparametric Behrens–Fisher Problem .....	117
	3.5.1	Large Sample Procedure .....	120
	3.5.2	Small Sample Approximation .....	124
	3.5.3	Separated Samples .....	126
	3.5.4	Example .....	127
	3.5.5	Software .....	129
	3.5.6	Summary .....	131
3.6		Consistency of Two-Sample Rank Tests .....	132
	3.6.1	Consistency of the WMW-Test .....	133
	3.6.2	Consistency of the Fligner–Policello and Brunner–Munzel Tests .....	136
3.7		Confidence Intervals .....	137
	3.7.1	Location Shift Effects .....	137
	3.7.1.1	Hodges–Lehmann Confidence Interval (No Ties) .....	138
	3.7.1.2	Hodges–Lehmann Confidence Interval (Ties Allowed) .....	139
	3.7.2	Relative Effects .....	141
	3.7.3	Summary .....	147
3.8		Power and Required Sample Size .....	149
	3.8.1	General Considerations and Notations .....	149
	3.8.2	Sample Size Planning for the General Case .....	153
	3.8.2.1	Case (1): No Prior Knowledge on $F_1$ and $F_2$ Available .....	154
	3.8.2.2	Case (2): $F_1$ and $F_2$ Known .....	157
	3.8.2.3	Brief Review of the Literature .....	160
	3.8.3	Software for Sample Size Planning .....	161
	3.8.4	Examples for Planning Sample Sizes .....	166
	3.8.5	Summary .....	168
3.9		Software .....	170
	3.9.1	General Remarks .....	170
	3.9.2	SAS: PROC NPAR1WAY .....	170
	3.9.3	Macro: NPTSD.SAS .....	171
	3.9.4	R-Package <i>rankFD</i> .....	172
	3.9.5	Application of the Software .....	173
	3.9.5.1	Analysis of the Two-Sample Design .....	173
3.10		Exercises and Problems .....	174

<b>4</b>	<b>Several Samples</b>	181
4.1	Introduction and Motivating Examples	181
4.2	Models, Effects, and Hypotheses	183
4.2.1	Normal Distribution and Location-Shift Model	184
4.2.2	Nonparametric Model	185
4.3	Effect Estimators and Test Statistics	189
4.3.1	Effect Estimators	190
4.3.2	Statistics	192
4.4	Kruskal–Wallis Test	198
4.4.1	Procedures for Large Sample Sizes	199
4.4.2	Consistency of the Kruskal–Wallis Test	200
4.4.3	Permutation Procedures for Small Samples	202
4.4.4	Discussion of the Rank Transform	203
4.4.5	Comparing Rank- and Pseudo-Rank Procedures	204
4.4.6	Application to Dichotomous (Binary) Data	207
4.4.7	Example and Software	209
4.4.8	Summary	212
4.5	Patterned Alternatives	214
4.5.1	Hettmansperger–Norton Test	216
4.5.2	Jonckheere–Terpstra Test	217
4.5.3	Comparison of Different Tests for Patterned Alternatives	218
4.5.4	Analysis of the Example	220
4.5.5	Software: SAS	220
4.5.6	Software: R	223
4.5.7	Summary	223
4.6	Confidence Intervals for Relative Effects	225
4.6.1	Direct Application of the Central Limit Theorem	226
4.6.2	Application of the $\delta$ -Method for Range Preserving Intervals	229
4.6.3	Summary	230
4.6.4	Application to an Example and Software	232
4.7	Multiple Comparisons	234
4.7.1	Basic Considerations: Global Versus Pairwise Rankings	237
4.7.1.1	Global Ranking	237
4.7.1.2	Pairwise Ranking	239
4.7.1.3	Conclusions	239
4.7.2	Multiple Testing Procedures	241
4.7.2.1	Bonferroni Adjustment	241
4.7.2.2	Holm’s Step-Down Procedure	242
4.7.2.3	Hochberg’s Step-Up Procedure	243
4.7.2.4	Closed Testing Principle	244

4.7.3	Multiple Contrast Tests and Simultaneous Confidence Intervals .....	246
4.7.3.1	Test Statistics for $H_0^F$ .....	247
4.7.3.2	Test Statistics for $H_0^P$ and Simultaneous Confidence Intervals .....	248
4.7.3.3	Test Statistics for All Pairwise Comparisons ....	250
4.7.3.4	Test Statistics for Particular Multiple Contrasts.....	251
4.7.4	Software and Example .....	252
4.7.5	Summary .....	254
4.8	Exercises and Problems .....	258
<b>5</b>	<b>Two-Factor Crossed Designs</b> .....	<b>263</b>
5.1	Introduction and Motivating Examples .....	263
5.2	Models, Effects, and Hypotheses .....	266
5.2.1	Linear Model .....	267
5.2.2	Nonparametric Model.....	269
5.2.3	Relative Effects .....	275
5.3	Effect Estimators .....	279
5.4	Test Statistics .....	281
5.4.1	General Results for Large Samples .....	281
5.4.2	Consistency of Tests Based on $C\hat{p}$ and $C\hat{\psi}$ .....	283
5.4.3	Wald-Type Statistic (WTS) .....	287
5.4.4	ANOVA-Type Statistic (ATS) .....	288
5.5	Computational Aspects and Software .....	293
5.5.1	General Computational Aspects.....	293
5.5.2	Computational Aspects Using SAS .....	294
5.5.3	Computational Aspects Using R .....	296
5.5.4	Application to an Example .....	297
5.5.5	Summary .....	299
5.6	Confidence Intervals and Patterned Alternatives .....	302
5.6.1	Confidence Intervals .....	302
5.6.2	Patterned Alternatives.....	303
5.6.3	Computational Aspects Using SAS .....	305
5.6.4	Computational Aspects Using R .....	306
5.6.5	Summary .....	306
5.7	Global vs. Stratified Ranking: $a \times 2$ Design .....	307
5.7.1	Procedures Using Stratified Ranking.....	308
5.7.2	Underlying Ideas of the Procedures Using Stratified Ranking.....	308
5.7.3	Procedures Using Global Ranking .....	310
5.7.4	Global vs. Stratified Ranking .....	311
5.8	Special Case: $2 \times 2$ Design .....	317
5.8.1	Special Models, Hypotheses, and Statistics.....	317
5.8.2	Application to an Example .....	323

5.9	Exercises and Problems .....	326
5.10	Alternative Procedures .....	329
<b>6</b>	<b>Designs with Three and More Factors</b> .....	<b>333</b>
6.1	Introduction and Motivating Examples .....	333
6.2	Models, Effects, and Hypotheses .....	334
6.3	Effect Estimators Based on Pseudo-Ranks .....	338
6.4	Test Statistics .....	339
6.4.1	Wald-Type Statistic .....	340
6.4.2	ANOVA-Type Statistic .....	340
6.5	Consistency of Statistics Based on $M\hat{\psi}$ .....	342
6.6	Software .....	342
6.6.1	Computations Using SAS .....	343
6.6.1.1	Analysis of Example 6.1 .....	343
6.6.1.2	SAS Procedures and Statements .....	344
6.6.2	Computations Using R .....	344
6.7	Confidence Intervals for Relative Effects .....	346
6.8	Summary .....	347
6.9	Generalization to Higher-Way Layouts .....	350
6.10	Software in General Factorial Designs .....	350
6.10.1	SAS Standard Procedures and IML Macros .....	350
6.10.2	R-Package <i>rankFD</i> .....	352
6.11	Exercises and Problems .....	353
6.12	Alternative Procedures .....	354
6.12.1	Some Historical Remarks .....	354
6.12.2	Hypotheses About Relative Effects .....	354
<b>7</b>	<b>Derivation of Main Results</b> .....	<b>357</b>
7.1	Models, Effects, and Hypotheses .....	357
7.1.1	General Nonparametric Model .....	357
7.1.2	Nonparametric Effects .....	358
7.1.3	Nonparametric Hypotheses .....	361
7.2	Estimators .....	361
7.2.1	Estimators for Relative Effects .....	362
7.2.2	Empirical Distribution Functions .....	362
7.2.3	Rank Estimators .....	368
7.3	Permutation Techniques .....	372
7.3.1	Exchangeable Random Variables .....	373
7.3.2	Limitations of Permutation Procedures .....	375
7.4	Asymptotic Results .....	377
7.4.1	Expectation and Covariance Matrix of the Rank Vector .....	377
7.4.2	Asymptotic Equivalence .....	382
7.4.2.1	General Case: Several Samples .....	382
7.4.2.2	Special Case: Two Samples .....	384

7.4.3	Asymptotic Normality Under $H_0^F$ .....	387
7.4.3.1	General Case: Several Samples .....	387
7.4.3.2	Special Case: Two Samples .....	392
7.5	Test Statistics .....	396
7.5.1	Quadratic Forms .....	397
7.5.1.1	Wald-Type Statistics .....	397
7.5.1.2	ANOVA-Type Statistics .....	398
7.5.1.3	Comparison of WTS and ATS .....	406
7.5.1.4	Discussion of the Rank Transform .....	408
7.5.2	Linear (Generalized) Rank Statistics .....	411
7.6	Asymptotic Normality Under Fixed Alternatives .....	413
7.6.1	Confidence Intervals for $\psi_i$ .....	414
7.7	Special Topics .....	418
7.7.1	One-Point Distributions .....	418
7.7.2	Score-Functions .....	421
7.8	Exercises and Problems .....	426
<b>8</b>	<b>Mathematical Techniques</b> .....	429
8.1	Particular Results from Matrix Algebra .....	429
8.1.1	Notations .....	429
8.1.2	Functions of Square Matrices .....	431
8.1.3	Partitioned Matrices .....	432
8.1.4	Direct Sum and Kronecker Product .....	432
8.1.5	Particular Results .....	433
8.1.6	Generalized Inverse .....	435
8.1.7	Matrix Techniques for Factorial Designs .....	436
8.2	Results from Analysis and Probability Theory .....	441
8.2.1	Inequalities .....	441
8.2.2	Asymptotic Equivalence .....	442
8.2.3	Central Limit Theorems .....	443
8.2.4	$\delta$ -Theorems .....	444
8.2.5	Distribution of Quadratic Forms .....	445
	<b>Correction to: Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs</b> .....	C1
<b>A</b>	<b>Software and Program Code</b> .....	447
A.1	SAS Macros and Standard Procedures .....	447
A.1.1	SAS Standard Procedures .....	447
A.1.1.1	PROC RANK .....	448
A.1.1.2	PROC TTEST .....	448
A.1.1.3	PROC NPAR1WAY .....	448
A.1.1.4	PROC POWER .....	450
A.1.1.5	PROC FREQ .....	450
A.1.1.6	PROC MIXED .....	450

A.1.2	SAS IML Macros .....	451
A.1.2.1	PSR.SAS .....	452
A.1.2.2	NPTSD.SAS .....	453
A.1.2.3	NOETHER.SAS .....	454
A.1.2.4	WMWSSP.SAS .....	455
A.1.2.5	OWL.SAS .....	456
A.2	R Code and the Packages rankFD, nparcomp, and coin .....	457
A.2.1	R Standard Procedures .....	458
A.2.1.1	The R-function <i>rank</i> (...) .....	458
A.2.1.2	The R-function <i>t.test</i> (...) .....	458
A.2.2	The Package rankFD .....	459
A.2.2.1	The Function <i>psr</i> .....	460
A.2.2.2	The Function <i>rank.two.samples</i> .....	460
A.2.2.3	The Function <i>noether</i> .....	462
A.2.2.4	The Function <i>wmwssp</i> .....	464
A.2.2.5	The Function <i>rankFD</i> .....	465
A.2.3	The Package nparcomp .....	467
A.2.3.1	The Function <i>Steel</i> .....	468
A.2.3.2	The Function <i>nparcomp</i> .....	469
A.2.4	The Package coin .....	471
A.2.4.1	The Function <i>wilcox_test</i> .....	472
A.2.4.2	The Function <i>kruskal_test</i> .....	473
<b>B</b>	<b>Data Sets and Descriptions</b> .....	475
B.1	Two-Sample Designs .....	475
B.1.1	Toxicity Trial .....	475
B.1.2	Organ Weights .....	476
B.1.3	$\gamma$ -GT Prior to Gall Bladder Surgery .....	477
B.1.4	Ferritin and IGF-1 .....	478
B.1.5	Number of Implantations/Data Set-1 .....	479
B.1.6	Number of Seizures in an Epilepsy Trial .....	480
B.1.7	Leukocytes in the Urine .....	481
B.2	One-Factorial Designs .....	482
B.2.1	Head-Coccyx Length .....	482
B.2.2	Closure Techniques of the Pericardium .....	483
B.2.3	Relative Liver Weights .....	484
B.2.4	Number of Corpora Lutea/Data Set-1 .....	485
B.3	Two-Way Layouts .....	486
B.3.1	Abdominal Pain Study .....	486
B.3.2	Irritation of the Nasal Mucosa .....	487
B.3.3	O <sub>2</sub> -Consumption of Leukocytes .....	488
B.3.4	Kidney Weights .....	489
B.3.5	Number of Corpora Lutea/Data Set-2 .....	490
B.3.6	Number of Implantations and Resorptions/Data Set-2 .....	491
B.3.7	Major Depression Trial .....	492

B.4	Three-Way Layouts .....	493
B.4.1	Number of Leukocytes .....	493
B.4.2	Luting Agents for Root Canal Dentin .....	494
<b>Acknowledgments</b>	.....	497
<b>References</b>	.....	501
<b>Index</b>	.....	511

# Glossary, Symbols, and Abbreviations

## General Symbols

$X_i$	Summation over all levels of the second index
$\overline{X}_i$	Arithmetic mean over all levels of the second index
$\sim$	Distributed as, distributed according to
$\dot{\sim}$	Approximately distributed as
$\doteq$	Asymptotically equivalent, see Remark 8.3 in Sect. 8.2.2, p. 443
$\oplus$	Direct sum, see Sect. 8.1.3
$\otimes$	Kronecker product, see Sect. 8.1.1
'	The symbol ' denotes a transposed vector or a transposed matrix
$\hat{\phantom{x}}$	The symbol $\hat{\phantom{x}}$ on a letter denotes an estimator of the respective quantity <i>Remark:</i> $\widehat{F}(x)$ denotes the empirical distribution function
$Cov(X)$	Covariance matrix of the random vector $X$
$E(X)$	Expected value (or expectation) of $X$
$H_0^F$	Nonparametric hypothesis (formulated in the distributions, see, e.g., p. 189 or p. 272ff)
$H_0^\mu$	Parametric hypothesis formulated in the parameters $\mu_1, \dots, \mu_a$ , see, e.g., p. 268
$H_0^P$	Nonparametric hypothesis formulated in the weighted relative effects $p_1, \dots, p_a$ , see, e.g., p. 361
$H_0^\psi$	Nonparametric hypothesis formulated in the unweighted relative effects $\psi_1, \dots, \psi_a$ , see, e.g., p. 361
$\log(x)$	Natural logarithm of $x$
$logit(x)$	$= \log\left(\frac{x}{1-x}\right)$ , logit of $x$
$\max(\cdot)$	Maximum of $(\cdot)$
$\min(\cdot)$	Minimum of $(\cdot)$
$\mu$	Constant parameter, e.g., expectation
$p_i$	Weighted relative treatment effect in a one-way layout, see, e.g., (4.2), p. 186

$\psi_i$	Unweighted relative treatment effect in a one-way layout, see, e.g., (4.4), p. 186
$\sigma^2$	Variance
$Var(X)$	Variance of $X$

## Vectors and Matrices

$C$	(General) contrast matrix, see Definition 4.1, p. 185 and, e.g., Sect. 5.2.1, Remark 5.1, p. 269
$C_A$	Contrast matrix for factor $A$ in a several-factorial design, see Sect. 4.2.2
$diag\{\dots\}$	Diagonal matrix of the elements within the parentheses
$\mathbf{1}_a$	$a$ -dimensional vector of 1's $(1, \dots, 1)'$ , understood as a column vector, see Sect. 8.1.1
$\mathbf{1}'_a$	$a$ -dimensional vector of 1's $(1, \dots, 1)$ , understood as a row vector, see Sect. 8.1.1
$I_a$	$a$ -dimensional unit matrix, see Sect. 8.1.1
$J_a$	$a \times a$ -dimensional matrix of 1's, $J_a = \mathbf{1}_a \mathbf{1}'_a$ , see Sect. 8.1.1
$r(M)$	Rank of matrix $M$
$M^-$	Generalized inverse ( $g$ -inverse) of matrix $M$
$M^+$	Moore–Penrose inverse of matrix $M$
$\mu_d$	Column vector of constants $\mu_1, \dots, \mu_d$ with $d$ components
$\mu'_d$	Row vector of constants $\mu_1, \dots, \mu_d$ with $d$ components
$P_a$	$= I_a - \frac{1}{a} J_a$ , centering matrix, see Sect. 8.1.1
$p$	Vector of the (weighted) relative effects $p_i$ , the dimension depends on the particular design
$\psi$	Vector of the (unweighted) relative effects $\psi_i$ , the dimension depends on the particular design
$ S $	Determinant of a quadratic matrix $S$ <i>Remark:</i> If $S$ denotes a $1 \times 1$ matrix (scalar), then $ S $ denotes the absolute value
$S^{-1}$	Inverse of a (non-singular) square matrix $S$
$w$	Vector of the weights $w_1, \dots, w_a$ for the pattern in patterned alternatives, see, e.g., Sect. 4.3.2
$L_N(w)$	Statistic for patterned alternatives, to be understood as a linear form in $w = (w_1, \dots, w_a)'$ , see, e.g., Sect. 4.3.2
$tr(M)$	Trace of a square matrix $M$

## Distributions, Functions, and Random Variables

$c(x)$	Normalized version of the count function (see Definition 2.12, p. 45)
$c^-(x)$	Left-continuous version of the count function (see Definition 2.12, p. 45)
$c^+(x)$	Right-continuous version of the count function (see Definition 2.12, p. 45)
$\chi_f^2$	Central Chi-square distribution with $f$ degrees of freedom
$\chi_{f;1-\alpha}^2$	Lower $(1 - \alpha)$ -quantile of $\chi_f^2$
$\chi_f^2/f$	Distribution function of the random variable $Z/f$ where $Z \sim \chi_f^2$ . Then, $\chi_f^2/f = F(f, \infty)$
$F(f_1, f_2)$	Central $F$ -distribution with $f_1$ and $f_2$ degrees of freedom
$F(f, \infty)$	See: $\chi_f^2/f$
$F_{1-\alpha}(f_1, f_2)$	Lower $(1 - \alpha)$ -quantile of $F(f_1, f_2)$
$F^+(x)$	Right-continuous version of the distribution function (see Definition 2.1, p. 16)
$F^-(x)$	Left-continuous version of the distribution function (see Definition 2.1, p. 16)
$F(x)$	Normalized version of the distribution function (see Definition 2.1, p. 16)
$H(x)$	Weighted mean of all distribution functions in a trial
$G(x)$	Unweighted mean of all distribution functions in a trial
$N(\mu, \sigma^2)$	Univariate normal distribution with expectation $\mu$ and variance $\sigma^2$
$N(0, 1)$	Standard normal distribution
$N(\boldsymbol{\mu}, \mathbf{S})$	Multivariate normal distribution with expectation $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}$
$R_{ik}$	Mid-rank of $X_{ik}$ among all observations—briefly called <i>rank of <math>X_{ik}</math></i> (see Definition 2.20, p. 55)
$R_{ik}^\psi$	Pseudo-rank of $X_{ik}$ among all observations (see (2.30) in Definition 2.20, p. 55)
$R_{ik}^{(i)}$	Mid-rank of $X_{ik}$ among all observations within sample $i$ —briefly called <i>internal rank of <math>X_{ik}</math></i> (see Definition 2.20, p. 55)
$R_{ik}^{(ir)}$	Pairwise rank of $X_{ik}$ among all $n_i + n_r$ observations within groups $i$ and $r$ for $i \neq r$ (see (2.29) in Definition 2.20, p. 55)
$R^W$	Wilcoxon rank sum (see (3.3), p. 90)
$t_f$	Central $t$ -distribution with $f$ degrees of freedom
$t_{f;1-\alpha}$	Lower $(1 - \alpha)$ -quantile of $t_f$

## Abbreviations

ANOVA	Analysis of variance
ART	Asymptotic rank transform, see Sect. 4.4.4, p. 203
ATS	ANOVA-type statistic, see Sect. 7.5.1.2
GART	Generalized asymptotic rank transform, see (7.31), p. 388ff
PRT	Pseudo-rank transform, see Remark 7.15, p. 410
RAA	Ranking after alignment, see Sect. 7.3.2
RT	Rank transform, see Remark 7.14, p. 409
WTS	Wald-type statistic, see Sect. 7.5.1.1

# Chapter 1

## Types of Data and Designs



**Abstract** This chapter provides an introduction into basic statistical terminology regarding different data types, measurement scales, variables, factors, and study designs, illustrated with several examples. Good scientific practice requires research reproducibility. This includes sound statistical modeling and informed choice of appropriate statistical methods for inference. Choosing valid statistical methods requires a firm understanding of the basic terminology and concepts presented in this chapter. Readers will be able to differentiate between the different data types encountered in practice, and understand why this is important. Further, readers will gain familiarity with concepts and notation of experimental design, so that they can choose appropriate designs and valid models for many typical situations themselves, or evaluate correct use by others.

### 1.1 Types of Data

The scientific method of acquiring knowledge requires empirical and measurable evidence. It involves formulating and testing hypotheses that are evaluated through methodic, carefully planned experiment, observation, and measurement. Measurements vary due to systematic effects whose detection and characterization is typically one of the scientific aims. However, they also vary due to random noise and factors that are too complex to be feasibly modeled or measured explicitly. These latter types of variation are subsumed into the random error. Separating systematic effects from random variation is an integral part of reproducible research, and it requires conducting experiments repeatedly and under the same conditions, and evaluating them as a whole.

Another important aspect of making research reproducible, and thus advancing science, is the choice of an appropriate statistical model and consequently, of an appropriate statistical method for inference. Among the first steps in choosing an adequate model is determining which variables are involved, along with their measurement scales. The following sections are dedicated to a description of the different scales of measurement. Differentiating between them is not actually

difficult, but it helps substantially in narrowing down the choice of possibly valid models and methods. There is a surprisingly large body of published research articles where inappropriate choices of statistical methods could have been avoided if the measurement scales of the variables involved had been evaluated more carefully.

### ***1.1.1 Accuracy of a Scale***

#### **1.1.1.1 Continuous Scale**

A variable is said to have a continuous measurement scale if, at least in theory, it could be measured with an arbitrarily detailed degree of precision (length, height, velocity, etc.). The intervals into which observations fall could be subdivided further and further. For example, instead of measuring length to the precision of meters (between 1 and 2) it could be measured in centimeters (between 112 and 113) or millimeters (between 1124 and 1125). Any specific value in the real numbers is only taken with probability zero. As a consequence, also ties (the same value observed more than once) can only occur with probability zero.

#### **1.1.1.2 Discrete Scale**

Contrary to continuous scales, subdividing the intervals of measurement precision is not sensible for discrete variables (number of children, day of week, sex, etc.). The possible outcomes occur with positive probabilities. Therefore, also ties occur with positive probability, in particular when the number of observations is large and the number of possible outcome values is small.

Based on the relations between different points on the measurement scale, we further differentiate between *metric* or *quantitative*, *ordinal*, *nominal*, and *binary* or *dichotomous* scales of measurement. Table 1.1 shows a summary of these different scales, along with some typical examples.

### ***1.1.2 Distances on a Scale***

#### **1.1.2.1 Metric Scale**

Key property of metric or quantitative data is that it makes sense to take differences and measure distances between observed values. Typically, it helps to think of quantitative data as data where numbers occur naturally in the measurement process, but this memory hook may fail in some situations, as for example passport numbers and postal codes are numbers, but in most instances, it would *not* make sense or

**Table 1.1** Different types of measurement scales and examples

Structure	Property	Type of Scale	Examples
metric (quantitative)	continuous	distance measure without ties	length, weight, volume
	discrete	distance measure with ties	counts; discretized: length, weight, volume
ordinal	continuous	ordered scale without ties	analog-scale, calibration scale
	discrete	ordered scale with ties	quality of life, pain score, damage score, rating scale
nominal	discrete	not ordered scale with ties	ethnic group, therapy, color
binary (dichotomous)	discrete	0-1-values with ties	indicators for success, morbidity, sex

be of interest to take their differences, thus they do not constitute metric data. For metric data, differences can be taken and interpreted, and it can be decided whether the difference between the points in one data pair is larger, smaller, or equal to the difference in another pair. From a mathematical point of view, a *metric* can be defined, thus the name. If, in addition, ratios between measured values can be taken and interpreted, the scale is often referred to as *ratio scale*, as opposed to the *interval scale*, which allows only for differences, but not for ratios. For example, something may be *twice as long* as something else (ratio scale), but when measuring temperature, the concept of *twice as hot* is typically not useful (interval scale). The relation between interval and ratio scales somewhat resembles that between groups and fields in algebra.

Metric data can be continuous (length, weight, and volume). In practice however, continuous scales are usually discretized because measurement instruments are not arbitrarily precise, or it would not make sense to measure at a higher precision. As an example, body height is typically measured in centimeters because smaller differences are irrelevant for most questions of interest. Thus, ties may occur for *discretized* continuous data. Their frequency depends on the precision of the measurement instrument (including possible rounding), and the number of observations.

Examples for metric data without ties are the ferritin values [ng/ml] in the toxicity trial in Data Set B.1.4, p. 478. The organ weights [g] in the toxicity trial in Data Set B.1.2, p. 476, as well as the  $\gamma$ -GT values [U/l] in the gall bladder study in Data Set B.1.3, p. 477, constitute metric data with ties.

If, independent of measurement precision, only certain, fixed values can be taken by a metric variable, then the data are *metric discrete*. This includes count data which are discrete because possible outcome values are only nonnegative integers. Also, count data are metric since differences between values can be calculated, interpreted, and compared. Finally, they are measured on a ratio scale because ratios

of counts make sense to be computed and interpreted (e.g., *twice as many*). This data type occurs in the fertility trial in Data Set B.1.5, p. 479 (Number of Implantations) and in Data Set B.2.4, p. 485 (Number of Corpora Lutea).

### 1.1.2.2 Ordinal Data

A measurement scale is called *ordinal* if the measured values can be sorted. That is, for any data pair, it can be determined which of the two values is larger (or better, greater, faster) according to some criterion. However, it is not sensible to add two values, take their difference, or calculate their distance. Therefore, it does not make sense either to calculate averages or standard deviations of ordinal data. Because of these limitations, ordinal measurement scales need to be carefully distinguished from metric scales which contain much more information regarding location and distance of the observed points.

There are examples of continuous ordinal scales such as the visual analog scales (VAS), where a subjective pain score is assigned by choosing a number between a minimum and a maximum possible value. Much more common are discrete ordinal scales. For example, the severity of a disease, or the rating of health or damage of an experimental unit, is often measured by classifying the observed unit into a particular category (e.g., very healthy) on an ordered scale of available categories. For this reason, discrete ordinal scales are also referred to as *ordered categorical scales*.

For convenience, the different categories on a discrete ordinal scale (grading scale, and rating scale) are usually encoded as integers, for example 0, 1, 2, . . . . However, the encoding is chosen arbitrarily and only reflects the order structure in the data, often in such a way that a worse category is assigned a smaller number. Instead of the selected numbers, a different encoding, for example using the letters  $A, B, C, \dots$  with the convention that  $A < B < C < \dots$  would contain the same information. Different choices of encoding do not change the amount of information contained in an ordinal (discrete or continuous) variable, as long as the relation between the encodings can be described by an order-preserving, thus strictly isotone (i.e., monotonically increasing,  $x < y$  implies  $m(x) < m(y)$ ) transformation. This has implications regarding the choice of adequate inference procedures. Indeed, when analyzing ordinal data, the results should not change when the data are transformed using any order-preserving function. In other words, appropriate statistical methods for ordinal data have to be invariant under strictly isotone transformations of the data. Examples for ordinal data are given by the nasal mucosa trial in Data Set B.3.2, p. 487 and the abdominal pain study in Data Set B.3.1, p. 486.

### 1.1.2.3 Binary (Dichotomous) Data

If there are only two possible values that can be taken by a variable, its scale is called *binary* or *dichotomous*. Examples are yes/no, good/bad, or healthy/diseased. Similar

to ordinal data, the outcomes are typically encoded as numbers, typically 0 and 1. Therefore, binary data are also sometimes referred to as  $(0,1)$ -data. Depending on the context, a binary measurement scale can be considered a special case of an ordinal scale if one of the two categories can be considered better than the other in some sense. Otherwise, they represent a special case of a nominal scale, which is defined next.

#### 1.1.2.4 Nominal Data

Data are referred to as *nominal* when it is not possible to order the individual categories in a natural way. A simple check whether data are nominal or ordinal can be done as follows: If for any three observations which are falling into three different categories, it can be determined which of the three naturally fits between the other two, then the respective variable has an order structure and its scale is (at least) ordinal. If not, it is nominal.

Typical nominal categories are, for example, left- vs. right-handedness, or localizations of heart attacks as front, back, septal, etc. Also, election surveys where possible choices are candidates  $A$ ,  $B$ ,  $C$ , or no opinion yield nominal data.

Sometimes, the term *qualitative* is used synonymously with nominal. However, this is not a universal convention, and sometimes qualitative is equated with being either nominal or ordinal.

This book does not present inference procedures for nominal response variables (apart from binary responses). However, nominal variables do appear as explanatory variables, that is, as variables describing the classification of experimental units into different treatment groups or strata. For an elaboration of the two terms explanatory variable and response variable, see Sect. 1.2. Rank-based procedures require that the response variables of interest are measured at least on an ordinal measurement scale. Regarding inference methods for nominal data, see, for example, the excellent textbook by Agresti (2013).

## 1.2 Factors and Designs

When choosing appropriate statistical methods for the analysis of data from experiments or observational studies, one needs to pay attention to the scales of the variables involved, and evaluate which distributional assumptions may be appropriate in modeling the values taken by the observations. Another important aspect is the underlying structure or design of the study. In this context, we distinguish between different types of variables, denoted *response variables* (dependent variables) and *explanatory variables* (independent variables). The terms in parentheses should only be used with caution because of possible confusion with the homonymous but different concepts of dependence and independence in statistical or probabilistic sense.

The response variable quantifies the success or effect of an experiment. In other words, it describes the response of an experimental unit to certain conditions the unit was subjected to. For example, the effect of a drug on the fertility of rats could be quantified by the number of implantations that each rat has. The number of implantations is thus the response variable, observed at each experimental unit (here: female rat in the experiment). In psychiatric research, the severity of depression is often measured on the Hamilton rating scale, an ordinal response variable. Then, the success of a psychotropic drug could be evaluated by assessing how much the drug lowers patients' Hamilton scale values. In clinical trials, response variables are often also referred to as *endpoints*.

An explanatory variable is conjectured to have some effect on the response variable. That is, under different given values of the explanatory variable, the distribution of the response variable may be different. Explanatory variables, in particular those measured on nominal or ordinal scales, are often called *factors* (see Sect. 1.2.1), and their values are called *factor levels*. For example, the distribution of Hamilton scale values may differ for male and female patients. Here, *sex* is a binary, nominal explanatory variable. The typical number of implantations may change with the dose level of the drug. In this case, the explanatory variable *dose level* may, for example, be measured on an ordinal scale (none, low, middle, and high). Or, it may be on a metric scale, by specifying the exact amounts of the drug that are administered.

The measurement scales of both, response and explanatory variable(s), are important in choosing appropriate statistical methods. For instance, if the explanatory variable has at least an ordinal structure, as in the last example, it may be of interest to check for certain shapes in the dose–response relationship (see Sect. 4.5 for details).

When there is more than one explanatory variable, different combinations of the levels of these variables will occur. The number and selection of explanatory variables and of relevant combinations of their levels constitutes the basic structure called the *design* of a study.

In the following sections, the main terminology of study design is introduced and illustrated using some examples. This overview will by no means be exhaustive. Some more complex designs will be described in later chapters, but for a detailed introduction into the design of experiments, and of other studies, see the excellent textbook by Kirk (1982, 2013).

### 1.2.1 Configuration of Factors

There are two different types of variables having an effect on the outcome that is quantified by the response variable. One type consists of those that are measured, observed, or even deliberately controlled in a study. These are the actual explanatory variables or *factors* whose effects are explicitly included in a statistical model. Their influence is then called *factor effect*. On the other hand, those variables that are not

recorded are subsumed into the *random error* term, together with possible random noise.

A major goal of designing a study is to minimize the random error by choosing relevant explanatory variables, appropriately modeling their (combined) factor effects, and statistically controlling the remaining variability through randomization.

Some of the factors are the subject of direct scientific interest and important research questions, and others are *confounding variables* (*covariates*, *covariables*) that are only recorded in order to minimize the random error. For example, among the former are dose levels of a drug, whereas the latter includes centers in a multi-center trial, or litters in an experiment involving rats.

The values taken by a factor are called *factor levels* or *levels*. For example, the factor *concentration* in the nasal mucosa trial (see Data Set B.3.2, p. 487) has the three levels 1 [ppm], 2 [ppm], and 5 [ppm].

Depending on the respective measurement scales, factors can be *metric*, *ordinal*, or *nominal*. Metric factors are also often called *regressors*, while ordinal or nominal factors are referred to as *categorical factors*. The factor *substance* in the nasal mucosa trial is clearly nominal, but for the factor *concentration*, it is not a priori clear, whether it should be assumed metric, ordinal, or nominal. Indeed, considering it to be metric also assumes a homogeneous influence of concentration on the irritation score, as in linear regression models. However, in this trial (see Table 1.2), the manufacturer was not interested in analyzing all possible concentrations between 1 and 5 ppm, or in estimating a dose–response curve. For that purpose, several different dose levels within this range would have been analyzed. Instead, a precise statement was desired for the three chosen dose levels. If these three concentrations

**Table 1.2** Irritation and damage of the nasal mucosa of 150 mice after inhalation of two different test substances, each at three different dose levels (see Data Set B.3.2, p. 487)

		Number of Animals with Damage Scores 0, 1, 2, 3, 4		
		Concentration		
Substance	Score	1 [ppm]	2 [ppm]	5 [ppm]
1	0	20	15	4
	1	4	7	6
	2	1	3	8
	3	0	0	5
	4	0	0	2
	Total Number	25	25	25
2	0	19	9	1
	1	5	9	6
	2	1	4	11
	3	0	2	5
	4	0	1	2
	Total Number	25	25	25

are regarded as levels of a nominal factor, the effect differences between 1 and 2 ppm and between 2 and 5 ppm, respectively, could be modeled independently. In this case, not even a possible monotonicity or similar dose–response relation would be considered. If a monotone relationship between concentration and response can be assumed, one may be able to improve the model, without having to specify a more precise functional relationship, by regarding the factor *concentration* as ordinal. However, the development of statistical inference for designs involving ordinal factors has been difficult and is still the subject of ongoing methodological research. In the following, we will in general assume that factors are nominally scaled. If other measurement scales are considered, this will be pointed out explicitly.

In addition to differentiating factors based on their measurement scale, another important aspect is their *reproducibility*, which leads to the distinction between *fixed* and *random* factors.

**Definition 1.1 (Fixed Factor)** A factor is called *fixed* if its levels are reproducible and if they are determined already at the outset of the study.

If an experiment was repeated, exactly the same levels of a fixed factor would be used. These are already known at the beginning of the experiment, and therefore can be reproduced. As a consequence, statements about levels of a fixed factor and their effects cannot per se be generalized to other possible levels of the respective factor. The design only allows to make statements about the levels actually used. In the nasal mucosa study, the factors *concentration* and *substance* are both fixed. In a potential replication of the experiment, the same substances would be used, with the same concentration levels.

Often, the goal is to make statements about a set of factor levels that extends beyond those explicitly used in the study. In that case, the levels actually used have to be chosen randomly from an appropriate population of factor levels.

**Definition 1.2 (Random Factor)** A factor is called *random* if its levels are selected randomly from an appropriate and comprehensive population of possible factor levels.

The levels of a random factor would be selected randomly again if an experiment was repeated. They are not known at the outset of the experiment but are determined during its course. The random selection allows to make statements about the whole population of factor levels based on the chosen sample. However, one needs to ensure that the sample of levels is representative of its underlying population. In the shoulder tip pain trial (Lumley 1996), *patient* can be modeled as a factor, in order to account for person-specific pain sensation, in particular when measurements are taken repeatedly over time on each subject. When repeating the experiment, it would

not be possible to observe the exact same patients, in the exact same condition. One would obtain a new sample. Therefore, *patient* is a random factor.

The main characteristics of fixed and random factors are summarized in the following rules:

### Replication Rule

If a study was repeated,

- a **fixed factor** would have exactly the same levels that have been determined at the outset of the study,
- a **random factor** would have a new random selection of levels, randomly chosen from an appropriate population of factor levels.

### Generalization Rule

In case of

- a **fixed factor**, statements about the factor levels and their effects are only valid for those levels which are involved in the trial and cannot be generalized to other possible levels of the fixed factor,
- a **random factor**, statements about the factor levels and their effects can be generalized to the population of factor levels from which they were randomly chosen.

A study design with just one factor is called *one-factor design* or *one-way layout*. An example is given by the fertility trial in Data Set B.1.5, p. 479 and in Data Set B.3.6, p. 491, analyzing the effect of the factor *substance* on the number of implantations. Another example is the toxicity trial in Data Set B.2.3, p. 484, where the influence of the *concentration* of the drug (dose level) on the relative liver weight is investigated. If there is more than one factor in the study design, it is called a *multifactor design* or *multi-way layout*. Examples for multifactor designs are the nasal mucosa trial in Table 1.2 with the factors *substance* and *concentration*, as well as the abdominal pain study in Table 1.3 with the factors *treatment* and *sex* (for details see Data Set B.3.1, p. 486).

## 1.2.2 Designs

In order to analyze data from a multifactor design appropriately, it is important to know how the levels of the different factors are combined with each other. If all theoretically possible combinations of the levels of two factors actually appear in

**Table 1.3** Pain scores at the morning of the third day after two different surgical interventions (techniques 1 and 2) for 25 patients (11 female and 14 male) with technique 1 and 28 patients with technique 2 (16 female and 12 male). The pain scores range from 0 (no pain) to 5 (severe pain)

Technique	Pain Score	
	Sex	
	Female	Male
1	0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 4	0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3
2	0, 0, 1, 2, 2, 2, 2, 3, 4, 4, 4, 4, 5, 5, 5	0, 1, 1, 2, 3, 3, 3, 3, 3, 4, 4, 5

the study design (Cartesian product), these factors are called *crossed*. Every level of the first factor is combined simultaneously with every level of the second factor.

When factors are crossed, one distinguishes between *main effects* and *interaction effects* (short: *interactions*). The main effect of factor *A* is the influence that this factor alone has on the response variable. On the other hand, the interaction between two factors *A* and *B* is an effect that cannot be explained by examining the factors individually, but only by looking at the combination of their levels. Interaction effects can be *synergistic* (reinforcing) or *antagonistic* (interfering).

If the main research interest concerns factor *A*, then an interaction effect with factor *B* is a confounding factor prohibiting a uniform analysis of factor *A* across all levels of factor *B*. For example, an interaction between the factors *substance* and *center* in a multi-center pharmaceutical trial means that the effect of the substance differs between the centers. Thus, further analysis of this effect has to be carried out differentiated for the individual centers.

However, often interaction effects are of research interest themselves. In the immune system trial in Data Set B.4.1, p. 493, the three factors *treatment*, *food*, and *stimulation* are all crossed with each other. A main effect of factor *treatment* could be interpreted as the influence of the treatment drug *versus* placebo, averaged over all the levels of the other two factors. An interaction effect between factors *treatment* and *food* would mean that the effect of treatment differed, depending on whether the animals received normal food or a reduced food diet.

If each of the different levels of factor *B* is only observed in combination with a specific level of factor *A*, then *B* is hierarchically nested (short: nested) within *A*. This relation between the factors is shown by using the notation  $B(A)$  instead of *B*. Designs with nested factors are called *hierarchical*.

*Remark 1.1* Hierarchical designs are mainly considered in the context of mixed models (see, e.g., Kirk 2013, Section 11.2, p. 492) which are beyond the scope of this book. Within the context of fixed linear models, Hocking (2003, Section 12.3) considers a hierarchical model with fixed effects. For details, we refer to this textbook.

The main practical problem in a hierarchical design with fixed effects is the verification of a random allocation of the subjects to the treatment levels. Therefore, such designs will not be considered in this book since all designs considered here are motivated by practical examples and real data sets.