Springer Series in Statistics

Edgar Brunner
Arne C. Bathke
Frank Konietschke

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

Springer

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

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

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 — Edgar Brunner
Salzburg, Austria — Arne C. Bathke
Berlin, Germany — Frank Konietschke
May 2018

The original version of the book was revised: Author affiliation in Copyright page was updated. The correction is available at https://doi.org/10.1007/978-3-030-02914-2_9

Contents

Glossary, Symbols, and Abbreviations

General Symbols

$X_{i\cdot}$	Summation over all levels of the second index
$\overline{X}_{i\cdot}$	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
$\widehat{\ }$	The symbol $\widehat{\ }$ on a letter denotes an estimator of the respective quantity *Remark*: $\widehat{F}(x)$ denotes the empirical distribution function
$Cov(\boldsymbol{X})$	Covariance matrix of the random vector $\boldsymbol{X}$
$E(X)$	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^μ	Parametric hypothesis formulated in the parameters $\mu_1, \ldots, \mu_a$, see, e.g., p. 268
H_0^p	Nonparametric hypothesis formulated in the weighted relative effects $p_1, \ldots, p_a$, see, e.g., p. 361
H_0^ψ	Nonparametric hypothesis formulated in the unweighted relative effects $\psi_1, \ldots, \psi_a$, see, e.g., p. 361
$\log(x)$	Natural logarithm of x
$logit(x)$	$= \log(\frac{x}{1-x})$, logit of x
$\max(\cdot)$	Maximum of $(\cdot)$
$\min(\cdot)$	Minimum of $(\cdot)$
μ	Constant parameter, e.g., expectation
p_i	Weighted relative treatment effect in a one-way layout, see, e.g., (4.2), p. 186

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

Vectors and Matrices

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

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)
χ^2_f	Central Chi-square distribution with f degrees of freedom
$\chi^2_{f;1-\alpha}$	Lower $(1-\alpha)$-quantile of χ^2_f
χ^2_f/f	Distribution function of the random variable Z/f where $Z \sim \chi^2_f$. Then, $\chi^2_f/f = F(f, \infty)$
$F(f_1, f_2)$	Central F-distribution with f_1 and f_2 degrees of freedom
$F(f, \infty)$	See: χ^2_f/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 μ and variance σ^2
$N(0, 1)$	Standard normal distribution
$N(\boldsymbol{\mu}, \boldsymbol{S})$	Multivariate normal distribution with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{S}$
R_{ik}	Mid-rank of X_{ik} among all observations—briefly called *rank of* X_{ik} (see Definition 2.20, p. 55)
R^{ψ}_{ik}	Pseudo-rank of X_{ik} among all observations (see (2.30) in Definition 2.20, p. 55)
$R^{(i)}_{ik}$	Mid-rank of X_{ik} among all observations within sample i—briefly called *internal rank of* X_{ik} (see Definition 2.20, p. 55)
$R^{(ir)}_{ik}$	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

E. Brunner et al., *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*, Springer Series in Statistics,
https://doi.org/10.1007/978-3-030-02914-2_1

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 γ-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 $0, 1, 2, \ldots$. 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 $A, B, C, \ldots$ with the convention that $A < B < C < \ldots$ 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 multicenter 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)

Pain Score		
	Sex	
Technique	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, 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.