Simultaneous Inference for the Comparison of Overdispersed Multinomial Data

Von der Naturwissenschaftlichen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften

- Dr. rer. nat. -

genehmigte

Dissertation

von

Dipl.-Math. Katharina Charlotte Vogel
geboren am 06.06.1987 in Magdeburg

2018

Referent:	PD Dr. Frank Schaarschmidt
Korreferent:	Prof. Dr. Ludwig A. Hothorn
Tag der Promotion:	26.06.2018

Abstract

Multinomial data is present when the outcome of an experiment is a discrete choice of more than two mutually exclusive alternatives and a multinomial distribution is assumed as the underlying distribution. Usually, a multinomial regression model is considered for the analysis of multinomial count data. Frequently, such data exhibits overdispersion especially if the data is acquired in clusters. The collection of data in cell cultures, litters, members of a family or classroom will lead to observations that are more similar within clusters than observations from different clusters. Therefore, it has to be expected that some sources of variation may differ between clusters and overdispersion is present.

In addition, several treatments are often of interest for such data, causing a multiple testing problem. While for normally distributed data multiple comparison procedures proposed by Tukey and Dunnett are standard since decades, multiple treatment comparisons between several overdispersed multinomial samples have been rarely investigated. The primary objective of this thesis is to develop a method to obtain multiple hypothesis tests and simultaneous confidence intervals for the comparison of multiple polytomous vectors that lack independence among experimental units due to a collection of data in clusters. Building on previous work, overdispersion is considered and an asymptotic procedure is proposed for simultaneous inference of odds ratios between multiple multinomial samples by including an estimated dispersion parameter. To assess validity, a simulation approach utilizing the Dirichlet-multinomial distribution is applied to determine the simultaneous coverage probability of confidence intervals for different magnitudes of overdispersion.

As part of this thesis, the proposed test procedure is implemented in the statistical software environment `R` in a user-friendly way. The application of the novel method and corresponding `R`-functions is described comprehensively on two real data sets from toxicological research and one data set from a social study. Especially the first example is examined in detail and an alternative approach using multiple marginal models is presented. Possible problems are discussed and suggestions for future work are outlined.

Keywords: multinomial, polytomous, count data, overdispersion, multiple comparisons, simultaneous inference, multiple contrast test, simultaneous confidence intervals.

Zusammenfassung

Multinomiale Daten sind häufig Ergebnisse eines Experiments, die durch die Wahl einer Kategorie aus einem Set von mehr als zwei sich gegenseitig ausschließenden Kategorien entstehen. In der Regel wird für die Analyse von nominalen Zähldaten ein multinomiales Regressionsmodell verwendet. Multinomiale Daten weisen jedoch häufig eine Überdispersion auf, insbesondere wenn die Daten in Clustern beobachtet werden. Zum Beispiel führt die Erfassung von Daten in Zellkulturen, die Untersuchung mehrerer Jungtiere eines Muttertiers oder die Befragung von Familienmitgliedern zu Beobachtungswerten, die innerhalb ihres Clusters ähnlicher sind als Beobachtungen von verschiedenen Clustern. Durch einzelne Varianzquellen, die sich zwischen den Clustern unterscheiden, liegt als Folge dessen Überdispersion vor.

Darüber hinaus sind für solche Daten häufig mehrere Behandlungen von Interesse, die ein multiples Testproblem verursachen. Während für normalverteilte Daten die von Tukey und Dunnett vorgeschlagenen Vergleichsverfahren seit Jahrzehnten geläufig sind, wurden multiple Vergleiche zwischen mehreren überdispersen multinomialen Daten selten untersucht.

Das Hauptziel dieser Arbeit ist es, für den Vergleich mehrerer multinomialer überdisperser Vektoren ein multiples Testverfahren und simultane Konfidenzintervalle zu konstruieren. Aufbauend auf einer vorhergehenden Arbeit wird die Methode um Überdispersion erweitert und ein asymptotisches Verfahren zur simultanen Inferenz von Odds Ratios zwischen multiplen multinomialen Daten unter Berücksichtigung eines Dispersionsparameters vorgeschlagen. Die Validität der vorgeschlagenen Methode wird in einer extensiven Simulationsstudie auf Basis von überdispersen multinomialen Daten aus der multinomialen Dirichlet-Verteilung beurteilt. Dabei werden unterschiedliche Ausmaße von Überdispersion angenommen und der Fehler 1. Art sowie die simultane Überdeckungswahrscheinlichkeit von Konfidenzintervallen untersucht.

Im Rahmen dieser Arbeit wird das entwickelte Testverfahren in der statistischen Softwareumgebung `R` benutzerfreundlich implementiert. Die Anwendung der neuen Methode und der entsprechenden `R`-Funktionen wird umfassend anhand von zwei realen Datensätzen aus der toxikologischen Forschung und einer Sozialstudie beschrieben. Insbesondere wird das erste Beispiel im Detail untersucht und ein alternativer Ansatz unter Nutzung multipler marginaler Modelle vorgestellt. Mögliche Probleme und Vorschläge für zukünftige Arbeiten werden diskutiert.

Schlagworte: multinomiale, polytome, Zähldaten, Überdispersion, Mehrfachvergleiche, simultane Inferenz, multiple Hypothesentests, simultane Konfidenzintervalle

Contents

List of Figures

List of Tables

Abbreviations

CP	**c**overage **p**robability
FWER	**f**amily**w**ise **e**rror **r**ate
GLM	**g**eneralized **l**inear **m**odel
H_0	null **h**ypothesis
H_1	alternative **h**ypothesis
MCP	**m**ultiple **c**omparison **p**rocedure
MMM	**m**ultiple **m**arginal **m**odels
MMM-Poisson	**m**ultiple **m**arginal **Poisson m**odels
sCI	**s**imultaneous **c**onfidence **i**nterval
VGLM	**v**ector **g**eneralized **l**inear **m**odel

Symbols

$\boldsymbol{A}_{I\times C}$	Contrast matrix to define log odds (logits)
$\boldsymbol{B}_{J\times G}$	Contrast matrix to define group comparisons
b_g	Number of clusters in group g
C	Number of categories
df	Degrees of freedom
G	Number of (treatment) groups
I	Number of logit comparisons
$\boldsymbol{I}$	Identity matrix
J	Number of group comparisons
K	Number of hypotheses of interest
L	Number of marginal models
m_g	Sample size in group g
m_{gb}	Sample size in cluster b of group g
N	Number of clusters
p	p-value
P	Number of parameters in the model
q	Index for logits
r	Pearson residuals
$\boldsymbol{R}$	Correlation matrix
t	Critical value from a t distribution
$\boldsymbol{V}$	Variance-covariance matrix of $\boldsymbol{\theta}$
$\boldsymbol{x}$	Vector of covariates, i.e. group membership
$\boldsymbol{X}$	Design matrix
$\boldsymbol{y}_{gb} = (y_{gbc})$	Vector of observations in cluster b of group g with $c = 1, \ldots, C$
$\boldsymbol{Y}$	Random vector

α	Type-I-error rate
$\boldsymbol{\alpha}_g = (\alpha_{gc})$	Vector of Dirichlet parameters in group g with $c = 1, \ldots, C$
$\boldsymbol{\beta}$	Parameter vector in a multinomial logit model
$\boldsymbol{\delta}_g = (\delta_{gi})$	Vector of log odds in group g with $i = 1, \ldots, I$
$\boldsymbol{\delta} = (\boldsymbol{\delta}_g)$	Stacked vector of log odds with $g = 1, \ldots, G$
Δ	Simulated increase in (baseline) odds
$\boldsymbol{\eta}$	Linear predictor
$\boldsymbol{\pi}_g = (\pi_{gc})$	Vector of probabilities in group g with $c = 1, \ldots, C$
σ^2	Dispersion parameter
v	Element of $\boldsymbol{\Sigma}$, i.e. the variance of a parameter
$\boldsymbol{\Sigma}_g$	Variance-covariance matrix of $\boldsymbol{\delta}_g$
$\boldsymbol{\Sigma}$	Joint variance-covariance matrix of $\boldsymbol{\delta}$
$\boldsymbol{\theta} = (\theta_k)$	Vector of parameters of interest with $k = 1, \ldots, K$
X^2	Pearson's test statistic

Dedicated to my dear father.

Chapter 1

Introduction

In many clinical trials, toxicology or behavioural biology, categorical data are acquired. A categorical variable consists of a limited number of levels (categories), whereby an observation can only take one of these levels. For example in toxicology, the survival status of an organism can be measured as "alive", "malformed" and "dead" but only one manifestation is possible. Such variables may reflect an ordinal variable but often there is no clear order in the categories so that a nominal classification is most suitable. If experiments are carried out in a controlled manner, several treatments are usually examined. In the aforementioned example of toxicology, certain dosages of a chemical agent may be important. A comparison of two groups can be based on the ratio of two odds of a particular event in these groups. Yet, several detailed questions arise for the comparison of more than two groups or categories, i.e. which categories differ between which groups. Consequently, there exists a multiple testing problem.

Given that the number of subjects is fixed by experimental design in each group, a multinomial regression model can be considered for the analysis of such count data depending on a categorical explanatory variable with regard to multiplicity. However, in many scientific research studies, experimental units are not assigned individually to the treatment groups. Instead, units of observation are randomized in aggregates, e.g. in cell cultures or litters with dams passing on their genetic trait or a location comprises several elemental units or members of a family or classroom, region or population are allowed to interact with each other. Such data, in which individual observations are assembled in aggregates, are called clustered. It has to be expected that first, some sources of variation may differ between clusters and second, that observations from different clusters are likely to vary more than observations within the same cluster. Thus the data exhibit a larger variability than it is generally assumed in the multinomial model, which is described as overdispersion by McCullagh and Nelder (1989), among others. If overdispersion is present, a dispersion factor has to be taken into account.

Ignoring overdispersion and analysing data under the assumption of a multinomial distribution leads to underestimation of standard errors and inflated size of corresponding tests. Misleading inferences and false conclusions are the consequence.
An exemplary extract of such multinomial count data exhibiting overdispersion is provided in Table 1.1 (Hothorn, 2015) and will be explained shortly in the following (see more details in Section 2.2). Table 1.1 displays data from a study conducted by the National Toxicology Program on the maternal toxicity after exposure of timed-pregnant CD-1 mice to diethylene glycol dimethyl ether (DYME, in doses of 0, 62.5, 125, 250 or 500 mg/kg/day). On gestational day 17, the life status of the foetuses was classified into three categories: alive, malformed and dead. Since some foetuses belong to the same dam (DAM_ID) and each of the five treatment groups consists of several dams, overdispersion may be present and can be estimated.

Table 1.1: Multinomial responses measuring life status. Raw data snippet of the life status of the offspring of female mice.

DOSE	DAM_ID	alive	malformed	dead
0	51	10	0	0
0	60	14	0	0
0	61	11	0	1
...	...	...	...	...
500	175	6	18	7
500	176	7	37	21
500	185	9	2	2

In the high-dose group, it is clearly seen that the frequencies in clusters are different. Of course, this type of data is not restricted to the field of biological and medical sciences. Overdispersed multinomial data may be also found in epidemiology, public health, genetics, botany, behavioural sciences, sociology, econometrics, marketing and other areas. To all, the important question about the data is: In which categories does the probability increase or decrease between which groups? For instance, is there a statistically significant difference in probabilities of an event in one of the treatments groups compared to the control group and if so, what is the expected extent?
While for normally distributed data multiple comparison procedures proposed by Tukey (1953) and Dunnett (1955) are well explored in the statistical field, multiple comparisons between several multinomial samples have been rarely investigated. Goodman (1964) introduced a method for constructing simultaneous confidence intervals to compare a number of odds between single multinomial samples in contingency tables. Schaarschmidt et al. (2017) allow to observe multiple comparisons of odds ratios between multiple multinomial samples and improved his work by taking the correlation into account. Still, overdispersion is ignored due to the assumption of a basic multinomial regression model. Clearly, this will lead to incorrect results as in the case of Poisson or the binomial model (Cox, 1983, McCullagh and Nelder, 1989). For multinomial data, Yee (2015) defines

a dispersion parameter in a vector generalized linear model (VGLM) by full maximum likelihood. Conversely, this method is lacking a correction for multiple testing.
A novel approach combining multiple marginal models proposed by Pipper et al. (2012) may be considered as an alternative method. It offers a flexible option for analysing several separate category-specific models. As part of a univariate binomial or Poisson model a dispersion parameter can be easily incorporated. Furthermore, the correlation between the test statistics from multiple marginal models can be taken into consideration for the simultaneous inference.
In this thesis, we focus on an appropriate analysis of polytomous data without category ordering and a lack of independence among experimental units due to data collection in clusters. We extend the work of Schaarschmidt et al. (2017) and develop a method to obtain simultaneous confidence intervals between multinomial vectors which incorporate overdispersion as suggested by McCullagh and Nelder (1989). Our proposed method allows performing standard multiple comparisons to control and all pairwise comparisons as well as user-defined contrasts for multiple odds ratios in analogy to Dunnett (1955), Tukey (1953) and Bretz et al. (2001). Further, we assess the simultaneous coverage probability of confidence intervals in a simulation study with Dirichlet-multinomial distribution. The familywise error rate is enclosed by settings of true null hypotheses. Finally, we offer an easy and convenient implementation in `R`.
The outline of this thesis is organized as follows. Chapter 2 recites the toxicological example to describe the current status of the art for analysing multinomial data in `R` also explaining the reasons for overdispersion. Two more real data examples are given to further underline the demand for an adequate method. In Chapter 3 the statistical framework for approximate multiple comparisons of overdispersed multinomial data is provided. We develop an asymptotic procedure for simultaneous inference of odds ratios between multiple multinomial samples taking overdispersion into account. A simulation study is conducted and parameter settings underlying the simulations are given. Simulation results for the familywise error rate and coverage probability in a strong sense are examined in Chapter 4. Power simulations show the probability that the proposed test procedure will reject the null hypothesis when it is not true. Chapter 5 introduces the newly implemented functions corresponding to this thesis in `R` and illustrates their application to the introductory examples. In Chapter 6 an extension of the proposed method to the case of heterogeneous variances in treatment groups is shown. In addition, alternative approaches using multiple marginal models to analyse multinomial data are presented. To conclude this thesis, Chapter 7 discusses the strengths and limitations of the proposed method.

Chapter 2

Multinomial Responses, Overdispersion and Multiple Hypothesis Testing

This chapter includes sample data sets to show typical issues and questions concerning multinomial data. The specification of multinomial response data is described and underlined by a first example. We address the topic of overdispersion in multinomial samples and explain the need for an appropriate method to simultaneously compare odds of certain categories between several groups. Additionally, two more examples are presented to motivate the research. We will revisit the examples in Chapter 5 to show the application of the novel procedure developed in this thesis.

2.1 Multinomial Response Data

In many fields of application one encounters categorical variables. Observations which can take one manifestation out of a fixed set with three or more possible values are called *polytomous* variables. Depending on whether a natural order of the variables is present or not, these data are also referred to as *ordinal* or *nominal*, respectively. The present work is based on nominal endpoints where the order within the response category is not important. For instance, a toxicological study might analyse the status of an organism after exposure to a toxic substance as "alive", "malformed" or "dead". A distinct ordering of the categories is not possible here, because it is not clear whether "dead" or "malformed" should be rated better. On that account, a nominal classification is most suitable. Moreover, it is not possible that two categories are present at the same time for one observation. Each of the observational units must fit into exactly one discrete category.

2.2 Example 1: Developmental Toxicity

The administration of a toxin during pregnancy of a mouse can influence the life status of its offspring. In a study conducted by the National Toxicology Program (NTP) on the maternal toxicity after exposure of timed-pregnant CD-1 mice to diethylene glycol dimethyl ether (DYME, in doses of 0, 62.5, 125, 250 or 500 mg/kg/day), the life status of the offspring was classified into three categories "alive", "malformed" or "dead" on gestational day 17.

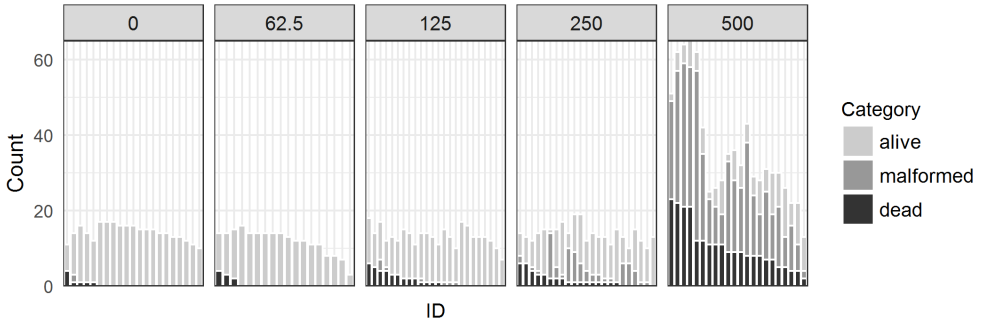

Figure 2.1: Bar chart of living, malformed and dead pups. Individual bars represent siblings from the same dam. Colours determine the counts of living, malformed and dead descendants in a cluster depending on treatment of the dam with a dose of DYME (0, 62.5, 125, 250 or 500 mg/kg/day).

Figure 2.1 shows a bar chart derived from the data of this study reported by Hothorn (2015). The data itself is available in the corresponding R-package "SiTuR" (Hothorn, 2014) and can be transformed to the format of Table 1.1 by reshaping (see Appendix B.2.1). In Figure 2.1 each dose panel displays the outcome of a brood per dam. In this experiment, a certain number of dams were randomized to the treatment doses: 21 mice received dose 0, 18 mice received dose 62.5, 24 mice received dose 125, 23 mice received dose 250 and 22 mice received a dose of 500 mg DYME/kg/day. These numbers of dams correspond to the number of bars in the diagram. Individual bars then represent siblings from the same dam where the height of the bars indicates the total number of siblings. The colours illustrate the counts of dead, malformed and living offspring. Black bars determine the observed number of dead offspring, dark grey bars determine the number of malformed offspring, and light grey bars determine the number of offspring alive. The clusters within each dose panel are arranged descending depending on the proportion of dead pups.

It can be assumed as a hypothesis that the counts of dead or malformed mice increase with higher doses and the proportion of mice alive strongly decreases in dose 500. But it should be noted that some clusters seem to generate higher counts of dead or malformed animals than other clusters, although they have received the same dose of DYME. Those ups and downs display a different variation in clusters which may indicate overdispersion.

2.3 Causes of Overdispersion

In practice, multinomial data often exhibit extra variation, which is also known as *overdispersion* in the literature (Agresti, 2013, McCullagh and Nelder, 1989, Tutz, 2011). The variation in the data is expected to be greater than the variance assumed under multinomial sampling if the statistical model does not sufficiently describe the data. In the example of the toxicology study above, we assumed that each young animal of a certain dose group has the same probability of being alive. Although the mice have been subjected to the same experimental conditions, the response probabilities differ depending on the cluster as it can be seen in both the extract of the data in Table 1.1 and Figure 2.1. Apparently, the survival status may also depend on genetic features, diseases or age of the dams during pregnancy. Because of missing explanatory factors, the proportions are not the same across the clusters. In this case, an omission of relevant explanatory components can lead to overdispersion and the data show more variation than induced by a multinomial model.

Even if the specified model contains all the detectable informative variables and their interactions, the underlying distribution may not be appropriate. Variables that are not observable may be responsible for more variation. Since it is not always possible to keep all study conditions constant, further conceivable explanations could be inhomogeneous conditions on the one hand or on the other, that the observational units have not been homogeneous. In both, variation between the response probabilities is the reason for overdispersion.

In addition to unobserved or unobservable variables, a correlation between observations may induce more variability. If the units are likely to be more similar within clusters than between clusters, multinomial responses are correlated and the variance in the data differs from the variance assumed under the multinomial distribution. Consider again the toxicology study. The existence of a malformed observation may increase the probability of further malformed observations in this cluster. Because the foetuses are descendants of the same dam, a positive correlation may exist between them.

Often, these explanations are interchangeable and lead to the same statistical model. It is clear that correlation between the observations causes more variation between the responses probabilities and therefore provokes overdispersion in the data. Conversely, a greater variation among the proportions can be explained by intercorrelation between the multinomial responses. However, there is no need to distinguish between these explanations for the estimation of overdispersion.

2.4 State of the art Using a Generalized Linear Model

Based on the introductory example, this section reports the current state of the art in the analysis of multinomial data. For this purpose, a multinomial logit model which belongs to the family of models called generalized linear model (GLM) is applied. The background to this method is explained in more detail in Section 3.1.3. Few parameters are anticipated now, but later clarified again.

To examine the effect of DYME on the survival status of young mice in example 2.2, we estimate a multinomial regression model using the dose group as an explanatory variable. As it is explained later, the multinomial logit model requires a baseline category, to which the probability of an event is compared. We choose "alive" as the baseline category and form a model as follows:

$$\log(\frac{\pi_q}{\pi_{\text{alive}}}) = \beta_{q0} + \boldsymbol{x}^T\boldsymbol{\beta}_q, \quad q \in \{1 = \text{malformed}, 2 = \text{dead}\}$$

where $\boldsymbol{x}^T = (\text{DOSE}[62.5], \text{DOSE}[125], \text{DOSE}[250], \text{DOSE}[500])$ is a dummy coded vector of group membership. That is, the vector $(0, 0, 0, 0)$ induces the scalar β_{q0}, which indicates the q-th log odds in the reference group, i.e. the log odds in group 0. $\boldsymbol{\beta}_q$ denotes the difference of log odds in the g-th group to reference. Both β_{q0} and $\boldsymbol{\beta}_q$ are unknown and need to be estimated in the multinomial model. In terms of a GLM, we assume that the random component is distributed multinomial and define a linear predictor produced by the explanatory variable "DOSE". A generalized logit is used as a link function.

In `R`, this can be achieved utilizing the `vglm()` function from the **VGAM** package. The `vglm()` function is used to fit a vector generalized linear model (VGLM), which offers a statistical framework for many parametric models and can also be used to fit a generalized linear model.

```
> head(bivar.re)
   DAM_ID DOSE alive malformed dead
1      51    0    10         0    0
8      60    0    14         0    0
9      61    0    11         0    1
15     70    0    17         0    0
16     71    0    15         0    0
23     79    0    17         0    0
> library(VGAM)
> multivgam <- vglm(cbind(alive,malformed,dead) ~ DOSE,
                    family = multinomial(refLevel=1),
                    data = bivar.re)
> coef(multivgam)
(Intercept):1 (Intercept):2    DOSE62.5:1    DOSE62.5:2     DOSE125:1
   -4.9698133    -3.5835189   -14.5687046     0.4577335     1.4361267
    DOSE125:2     DOSE250:1     DOSE250:2     DOSE500:1     DOSE500:2
    1.5257389     3.8473335     1.8378506     6.0338201     4.0268084
```

In the `vglm()` function, the first category "alive" is set as baseline. Hence there are two log odds, namely: $\log(\pi_{\text{malformed}}/\pi_{\text{alive}})$ and $\log(\pi_{\text{dead}}/\pi_{\text{alive}})$. For the predictor variable, the first dose group of 0 mg DYME/kg/day is automatically taken as a reference, because it was set as the first level when reading the factor variable into `R`. The `coef()`-function is used to extract the coefficients for the individual models. The coefficients are given in order of q within the dose levels. In particular, the estimates of β_{10} and β_{20} are indicated first. Then estimates of β_{11} and β_{21} follow and so forth, leading to $\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}} = (-14.57, 1.44, 3.85, 6.03)$ and $\widehat{\boldsymbol{\beta}}_2^{\mathrm{T}} = (0.46, 1.53, 1.84, 4.03)$. For purposes of illustration, the equation for the first log odds of malformed relative to alive is as follows:

$$\begin{aligned}\log(\widehat{\pi}_{\text{malformed}}/\widehat{\pi}_{\text{alive}}) = &-4.97 - 14.57 \cdot \text{DOSE}[62.5] + 1.44 \cdot \text{DOSE}[125] \\ &+ 3.85 \cdot \text{DOSE}[250] + 6.03 \cdot \text{DOSE}[500].\end{aligned}$$

In the model, dummy coding is used to determine the influence of group membership. If all dummy variables are equal to zero, the equation corresponds to the log odds for dose group 0. The estimated coefficient for dose group 125, for example, is 1.44, which gives the difference of log odds between dose 125 and dose 0. This means that animals in dose group 125 are more likely to be malformed than alive. For a unit increase in the predictor variable, the log odds of "malformed" relative to "alive" is expected to change by its respective parameter estimate while holding all other variables in the model constant. If a subject were treated with dose 125 instead of dose 0, the multinomial log odds of being malformed instead of alive would be expected to increase by 1.44. Note that this comparison is still on the log scale. The estimated odds ratio for comparing malformed versus alive is $\exp(1.44) = 4.22$. Given a change in dosage from 0 to 125 mg DYME/kg/day, the odds of being malformed instead of alive would be 4.22 times or 322% more likely.

2.5 Multiple Testing Problem in Multinomial Models and the Need for an Appropriate Method

These parameter estimates could now be tested for the null hypothesis whether a certain regression coefficient is equal to zero within a given model. For example, it may be interesting to see if the log odds ratio of 1.44 is found to be statistically different from zero and thus a significant difference between dose 0 and dose 125 for malformed relative to alive can be concluded. If all coefficients from both equations are tested with regard to an overall hypothesis, it gives rise to a multiple test problem.
Before such hypotheses can be investigated, it is important to be aware of any overdispersion in the data. Under the assumption of a generalized linear model (GLM) with

multinomial distributed errors, it is expected that the residual deviance is equal to the residual degrees of freedom. Thus, we will have a look at our goodness-of-fit statistic:

```
> summary(multivgam)
...
Residual deviance: 410.4707 on 206 degrees of freedom
...
```

The fact that the residual deviance is much greater than the residual degrees of freedom indicates that overdispersion of roughly 2 is present. The exact calculation of overdispersion and its handling is investigated in detail in Chapter 3. At the moment it is sufficient to say that before conclusions about significances are allowed to be drawn, overdispersion must not be ignored. So far there is the possibility to observe multiple comparisons of odds ratios between multiple multinomial samples (Schaarschmidt et al., 2017), but without incorporating overdispersion. Accordingly, there is a demand for an appropriate method of analysis, which allows comparing multiple treatments for multiple odds ratios while considering overdispersion in the data. To further illustrate the various occurrences of overdispersed multinomial data and the need for such a method, additional examples are given.

2.6 Example 2: Housing Satisfaction

In this example, a total of 35 neighbourhoods from two areas, including 17 from a rural area and 18 from an urban area, were surveyed according to their level of satisfaction with their homes. Five households from each neighbourhood were invited to submit a rating of their satisfaction in one of the categories as "unsatisfied" (us), "satisfied" (s) or "very satisfied" (vs). In contrast to example 2.2, the categories may be considered as ordered.

This survey is well-known under the topic of overdispersion and has previously been analysed by several authors, including Brier (1980), Koehler and Wilson (1986), Morel and Nagaraj (1993) and Morel and Neerchal (2012). We refer to the data as published by Wilson (1989), of which Figure 2.2 graphically illustrates the satisfaction of individual neighbourhoods per area. The data itself is included in the R-package "MM" separately for the metropolitan area and the non-metropolitan area under the name "wilson".

Note that in this example the total number of units per neighbourhood is fixed by design of the study. In Figure 2.2, each bar is assigned to a neighbourhood, with all bars having a uniform height of 5 households. Since the locals within a neighbourhood are generally in contact with each other and talk about their state of comfort, it is likely that the opinion of households in the same neighbourhood is more similar than the impression of another neighbourhood. Therefore, it can be assumed that overdispersion may be present in this example. This should then be included in the statistical analysis of

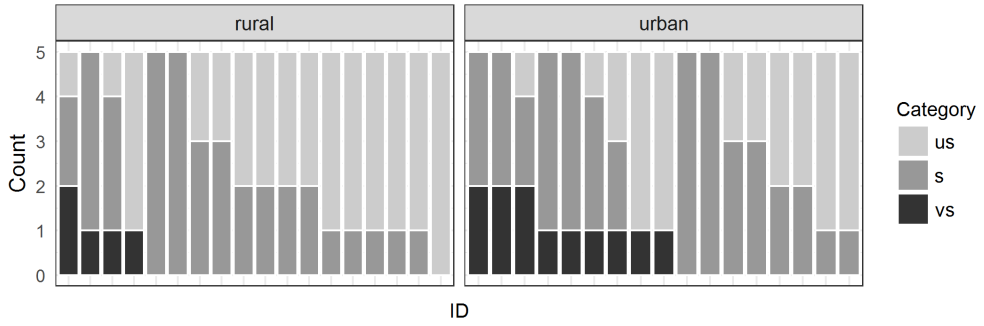

Figure 2.2: Bar chart of survey results in each neighbourhood. Each bar represents a cluster, in this case the households of equal size of 5, which are sorted by descending proportion of the events vs and s.

whether the regions differ in their satisfaction. Although the comparison concerns only two regions, it needs to be additionally adjusted for multiple comparisons, since several categories will be compared simultaneously.

2.7 Example 3: Differential Blood Count in Rats

For the third illustration, a toxicological example of white blood cell counts is considered. Table 2.1 shows an extract of raw count data from a toxicological study in rats (Hothorn, 2015). In this study, four treatment groups were investigated, consisting of one control group and three groups with different toxin dosages (low, mid and high). Rats of different gender were randomly assigned to these treatment groups. For each treatment group, ten male and ten female rats were examined, except for only eight male animals in the high dose group. A total of 200 leukocytes were counted per animal and classified into six categories: Eosinophils (Eos), Basophils (`Baso`), Neutrophilic Bands (`Stab`), Segmented Neutrophils (Seg), Monocytes (Mono) and Lymphocytes (Ly). The total number of counts per rat is fixed by experimental design. Further, there is obviously no clear order between the categories.

The data are presented separately according to gender and dose group in Figure 2.3 and are arranged according to the proportion of lymphocytes in descending order within

Table 2.1: Data snippet of the differential blood count in rats. Eosinophils (Eos), Basophils (Baso), Stab cells (Stab), Segmented Neutrophils (Seg), Monocytes (Mono) and Lymphocytes (Ly) were recorded per animal in four treatments groups.

sex	animal	Group	Eos	Baso	Stab	Seg	Mono	Ly
Males	1101	control	2	0	0	51	2	145
Males	1102	control	3	0	0	28	2	167
Males	1103	control	4	0	0	32	5	159
...	...	...	...	...	...	...	...	...
Females	2408	high dose	4	0	0	14	3	179
Females	2409	high dose	1	0	0	35	4	160
Females	2410	high dose	0	0	0	42	6	152

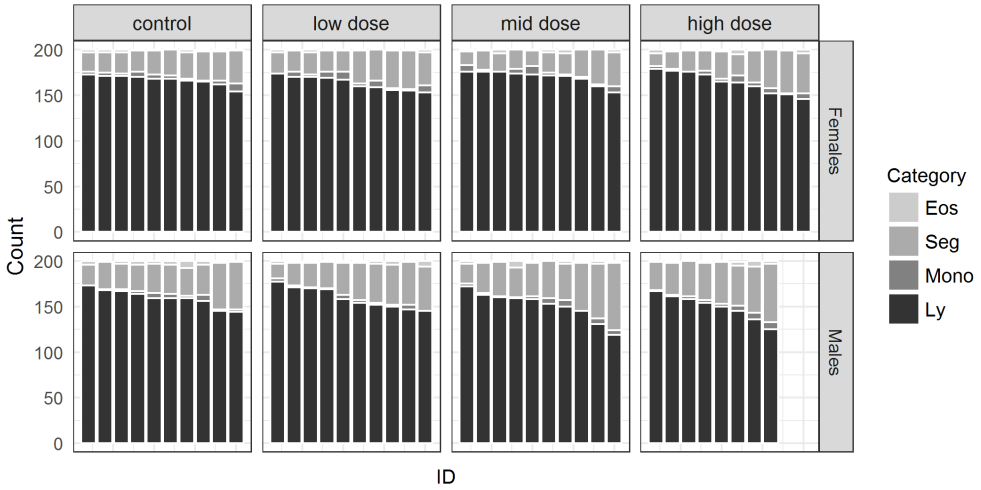

Figure 2.3: Bar chart of counts of the basic white blood cell types in male and female rats. Each bar determines the count of each type of white blood cell (Eosinophils, Basophils, Stab cells, Segmented Neutrophils, Monocytes and Lymphocytes) of a blood sample. Per animal 200 leukocytes were counted. The data are arranged in descending order of proportion of cell types, beginning with Lymphocytes and subsequently Monocytes, Segmented Neutrophils and Eosinophils.

the panels. The bars have a total height of 200, which corresponds to the number of total cells counted. Since no basophils or neutrophilic bands were detected among leukocytes, they are excluded from the graphical analysis. Although the conditions in this experiment have been attempted to be kept constant, the individual animals show variations in the counts of blood cell types. Because this might be an indication of overdispersion, a scale parameter should be estimated for this example. One is then interested in the question of whether the ratios of white blood cells change significantly depending on the dose of the toxin within sex.

Chapter 3

Methods

This chapter presents the statistical framework for modelling multinomial responses with overdispersion. First, principles of the multinomial distribution are revised in Section 3.1 and a general model for multinomial response variables is presented. Section 3.2 develops a method for the simultaneous analysis of overdispersed multinomial data. Herein, the mathematical structure of the underlying data is introduced. After defining the parameter of interest the calculation of overdispersion is presented. Under the condition that the formulated model satisfies the properties of a generalized linear model, multiple comparisons can be made by adapting the covariance matrix for overdispersion. In Section 3.3, the further course of a simulation study is outlined to validate the performance of the proposed method. Its results will be presented in Chapter 4.

3.1 Prerequisites

3.1.1 Multinomial Distribution

The *multinomial distribution* is first explained in an experiment with one group. Let m indicate the number of observational units that is examined in this group. Additionally there are C mutually exclusive categories, which occur with probabilities $\pi_1, \ldots, \pi_C \in [0, 1]$. All observations in a group can be clearly assigned to one of these categories depending on their outcome. Individual experiments thus each receive a vector of the form $(0, \ldots, 1, \ldots, 0)$ with 1 in the c-th place if the observation falls in category c. The sum of all outcomes yields a vector $\boldsymbol{y} = (y_1, \ldots, y_C)$, which specifies the cell counts in the categories $1, \ldots, C$. The probability mass function of the random counts Y_c is given by

$$Pr(Y_1 = y_1, \ldots, Y_C = y_C; m, \boldsymbol{\pi}) = \frac{m!}{y_1! \ldots y_C!} \pi_1^{y_1} \ldots \pi_C^{y_C} \tag{3.1}$$

for all $y_c \in \{0, \ldots, m\}$, where $\sum y_c = m$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)$ and $\sum \pi_c = 1$. The vector is said to follow a multinomial distribution with parameters m and $\boldsymbol{\pi}$, abbreviated with $\boldsymbol{Y} \sim Mult(m, \boldsymbol{\pi})$.

In the case of several groups, each random vector of a group g, which is labelled as $\boldsymbol{Y}_g = (Y_{g1}, \ldots, Y_{gC})$, follows its own multinomial distribution. Now let m_g indicate the sum of all observations in group g, i.e. $\sum_c y_{gc} = m_g$, adopting the convention that $\sum m_g = m$. The unknown probability that a unit of group g falls into the category c is denoted by the vector $\boldsymbol{\pi}_g = (\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC})$, provided that $\sum_c \pi_{gc} = 1$. The distribution of $\boldsymbol{Y}_g$ is then multinomial with parameters m_g and $\boldsymbol{\pi}_g$.

Important properties of a multinomial distributed random vector can be derived from the first derivatives of the moment generating function (Johnson et al., 1997). The first and second moments are

$$\mathrm{E}(Y_{gc}) = m\pi_{gc} \tag{3.2}$$

$$\mathrm{Var}(Y_{gc}) = m\pi_{gc}(1 - \pi_{gc}) \tag{3.3}$$

$$\mathrm{Cov}(Y_{gc}, Y_{gc'}) = -m\pi_{gc}\pi_{gc'} \tag{3.4}$$

and address the expected value of a random multinomial variable Y_{gc}, the variance and covariance, respectively.

3.1.2 Modelling Multinomial Data

A common model to link multinomial data to an explanatory variable is the *multinomial logit model*. Herein the logit function is used as a link function to calculate the odds for a category c to a reference category on the logarithmic scale, i.e. log odds also known as *logits*. If the first category is specified as a reference category, each baseline logit can be modelled assuming a linear model

$$\log \frac{\pi_{gc}(x)}{\pi_{g1}(x)} = \beta_{q0} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_q, \quad c = 2, \ldots, C,\ q = 1, \ldots, C-1. \tag{3.5}$$

In this model, the assignment of a given observation to a group $g = 1, \ldots, G$ is stored in a dummy coded vector $\boldsymbol{x}$ of dimension $G' = G - 1$. That is, for $g = 1$ the vector $\boldsymbol{x}$ consists of zeros only, otherwise, the vector $\boldsymbol{x}$ has the value 1 at position $g - 1$ if the observation belongs to group g. For a given logit q, β_{q0} and $\boldsymbol{\beta}_q = (\beta_{q1}, \ldots, \beta_{qG'})$ are unknown parameters, with β_{q0} denoting the q-th logit in the reference group and with $\boldsymbol{\beta}_q$ denoting the differences of log odds in groups $g = 2, \ldots, G$ to β_{q0}. The log odds vary according to the selected reference category and hence the coefficients β_{q0} and $\boldsymbol{\beta}_q$ depend on the chosen reference category. Typically, the first or the last category is chosen as a baseline. However, any other category, e.g. the most common one, can be selected as a reference to return the log odds. The log odds can be converted into each other so

that all paired contrasts are determined. For example, for three categories and given the first two logits with the first category serving as a reference, the logit of category 2 to 3 in group g can be obtained by $\log(\pi_{g2}/\pi_{g3}) = \log(\pi_{g2}/\pi_{g1}) - \log(\pi_{g3}/\pi_{g1})$. Therefore, given a certain choice of $C-1$ logits, the rest is redundant.
Of course, one can also directly express the probability of the responses by rewriting 3.5 to

$$\pi_{gc} = \frac{\exp(\beta_{q0} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_q)}{1 + \sum_{s=1}^{C-1} \exp(\beta_{s0} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_s)} \tag{3.6}$$

for all $c \in \{2, \ldots, C\}$ and $q = c - 1$. The reference category is calculated by $\pi_{g1} = 1 - \pi_{g2} - \ldots - \pi_{gC}$.

3.1.3 Multinomial Logit Model as a Generalized Linear Model

Generalized linear models (GLM) extend the class of classical linear models so that the response variable is linearly related to the explanatory factors via a link function (McCullagh and Nelder, 1989). In general linear models, random variables $\boldsymbol{Y}$ with realization $\boldsymbol{y}$ are assumed to have independent and normally distributed error terms with mean zero and constant variance. In a generalized linear model, the response is allowed to have an error distribution that is part of the exponential family. A GLM is characterized by three components which are the following in case of a multinomial assumption (Fahrmeir and Tutz, 2001):

(1) a *random component*: the random variable $\boldsymbol{Y}_g = (Y_{g1}, \ldots, Y_{gC})$ is assumed to follow an exponential family distribution, i.e. the multinomial distribution, with

$$\mathrm{E}(Y_{gc}) = m\pi_{gc}$$

and variance

$$\mathrm{Var}(Y_{gc}) = \sigma^2 m\pi_{gc}(1 - \pi_{gc})$$

(2) a *systematic component*: a linear predictor η_{gq} which is produced by a known explanatory vector $\boldsymbol{x}$ and unknown parameters β_{q0} and $\boldsymbol{\beta}_q$

$$\eta_{gq} = \beta_{q0} + \boldsymbol{x}\boldsymbol{\beta}_q$$

(3) a *link function*: a generalized logit link is defined between the random and the systematic component

$$\eta_{gq} = \log \frac{\pi_{gc}}{\pi_{g1}}, \quad c = 2, \ldots, C,\ q = 1, \ldots, C-1$$

If we assume the absence of overdispersion, $\sigma^2 = 1$. If the data exhibits variability greater than the assumed variance according to the multinomial model, a dispersion parameter of $\sigma^2 > 1$ can be estimated from the data (McCullagh and Nelder, 1989).
In more general models (not considered here) $\boldsymbol{x}$ may contain additional covariates for a given observation. Then $\boldsymbol{\beta}_q$ would contain additional parameters modelling their effects on the q-th logit.

3.2 Simultaneous Inference in a Multinomial Logit Model with Overdispersion

By comparing multiple log odds between several treatment groups, a multiple comparison problem arises. Assessing statistical inference for more than one hypothesis in case of the multinomial logit model is outlined in this section. In addition to the previous sections, we assume that the data was recorded in clusters, such as in Table 1.1, and the data exhibit overdispersion. Consequently, the dimension of the underlying data record increases at the cluster level and we introduce the mathematical notation for this kind of data structure in general first.

3.2.1 Data Structure

Consider a completely randomized design with $g = 1, \ldots, G,\ G \geq 2$ treatment groups each consisting of b_g clusters, $b = 1, \ldots, b_g$. Each observational unit is assigned to a group g and takes one of the mutually exclusive nominal categories $c = 1, \ldots, C$. Moreover, each unit belongs to a cluster b whose observations may not be independent. Let Y_{gbc} denote the number (count) of observations from cluster b of group g that fall into category c with observed value $y_{gbc} \in \mathbb{N}$. Furthermore, let m_g indicate the total number of units per group and m_{gb} the sample size in cluster b of group g. Note that the counts in the various clusters of one group add up to the total number of observations in that group, i.e. $\sum_{b,c} y_{gbc} = m_g$. By definition, let π_{gc} be the probability that an observation in treatment group g falls into category c, which is independent of the cluster and only depends on the group. In each group the probabilities always add up to 1, $\sum_{c=1}^{C} \pi_{gc} = 1$, regardless of the number of possible outcomes. We therefore assume that each multinomial response vector of counts in category c of cluster b of group g, $\boldsymbol{y}_{gb} = (y_{gb1}, y_{gb2}, \ldots, y_{gbC})$, could be distributed multinomial

$$(y_{gb1}, y_{gb2}, \ldots, y_{gbC}) \sim multinomial(m_{gb}, (\pi_{g1}, \pi_{g2}, \ldots, \pi_{gC}))$$

but may show higher variance than expected under multinomial distribution.

3.2.2 Parameters of Interest

For the analysis of such a data set, one might be interested in comparing the baseline odds between groups (Agresti, 2013). On the logarithmic scale, such parameters can be expressed as differences or linear combinations of log odds / logits, namely log odds ratios. According to Schaarschmidt et al. (2017) all possible log odds in group g can be estimated by

$$\boldsymbol{\delta}_g = \boldsymbol{A}_{(I\times C)} \log(\boldsymbol{\pi}_g^{\mathrm{T}}) \tag{3.7}$$

where $\boldsymbol{A} = (a_1^{\mathrm{T}}, \ldots, a_{\mathrm{I}}^{\mathrm{T}})^{\mathrm{T}}$ is a $I \times C$ matrix of rowwise stacked contrast vectors and contains all the comparisons of interest. For example, the first logit of group g for the ratio of the probability to fall in category $c = 2$ and the probability to fall in category $c = 1$ (first baseline logit) can be described by $\delta_{g1} = (-1, 1, 0, \ldots, 0) \cdot \log(\boldsymbol{\pi}_g^{\mathrm{T}}) = \log(\pi_{g2}/\pi_{g1})$. Further, these odds can be compared across groups. In the simple case that for all odds of interest the same set of baseline ratios is of interest, the parameter vector $\boldsymbol{\theta}$ is calculated as

$$\boldsymbol{\theta} = (\boldsymbol{B} \otimes \boldsymbol{A}) \begin{pmatrix} \log \boldsymbol{\pi}_1^{\mathrm{T}} \\ \log \boldsymbol{\pi}_2^{\mathrm{T}} \\ \vdots \\ \log \boldsymbol{\pi}_G^{\mathrm{T}} \end{pmatrix} \tag{3.8}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{B}$ is another contrast matrix of dimension $J \times G$. In a similar way to matrix $\boldsymbol{A}$, the matrix $\boldsymbol{B}$ defines all J comparisons between the G groups. The column vector $\boldsymbol{\theta}$ of dimension $K = I \cdot J$ then consists of the parameter values of the log odds ratios primarily ordered by the group comparisons and within them ordered according to the definition of odds contrasts in matrix $\boldsymbol{A}$.
The same can be achieved by stacking the column vectors $\boldsymbol{\delta}_g$ into a single column vector $\boldsymbol{\delta}$, i.e. $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \boldsymbol{\delta}_2^T, \ldots, \boldsymbol{\delta}_G^T)^T$. Then $\boldsymbol{\theta}$ can be written as

$$\boldsymbol{\theta} = (\boldsymbol{B} \otimes \boldsymbol{I}_I)\boldsymbol{\delta} \tag{3.9}$$

where $\boldsymbol{I}_I$ denotes the identity matrix of size I (Schaarschmidt et al., 2017). Asymptotic estimates for $\boldsymbol{\delta}_g$ and $\boldsymbol{\theta}$ may be obtained by using the empirical equivalent $\widehat{\boldsymbol{\pi}}$ in the formulas above and are indicated by $\widehat{\boldsymbol{\delta}}_g$ and $\widehat{\boldsymbol{\theta}}$, respectively.
Note that the set of parameters grows extremely fast, which needs to be considered later when adjusting for multiple testing. One should, therefore, be aware of the number of parameters and select a subset conscientiously.

3.2.3 Estimating Overdispersion

In our data structure, we have assumed that the probability π_{gc} is the same for each vector of observations in group g regardless of its belonging to a cluster. This is rather unlikely since observations in one cluster may behave similarly than observations from two different clusters of the same treatment group. Thus although the units were observed under the same conditions their response probabilities may vary between clusters. If this deviation to each other is due to unobserved or unobservable effects between clusters, it leads to a higher variance of y_{gc} than expected under the multinomial distribution. Assuming that out of $\sum b_g = N$ clusters the response probability for the gb-th observation of form $\boldsymbol{y}_{gb} = (y_{gb1}, y_{gb2}, \ldots, y_{gbC})$, which depends on a set of explanatory variables $\boldsymbol{X}$, is a random variable it can be shown that the unconditional mean and covariance of $\boldsymbol{Y}_{gb} \sim mult(m_{gb}, \boldsymbol{\pi}_g)$ are

$$\mathrm{E}(\boldsymbol{Y}_{gb}) = m_{gb}\boldsymbol{\pi}_g, \tag{3.10}$$

$$\mathrm{Cov}(\boldsymbol{Y}_{gb}) = \sigma^2 \boldsymbol{\Sigma}_Y \tag{3.11}$$

where $\boldsymbol{\pi}_g$ is the probability vector of success, m_{gb} is the total sample size per cluster with $m_{gb} = \sum_{c=1}^{C} y_{gbc}$, σ^2 is an unknown scale parameter, $\sigma^2 > 0$ (McCullagh and Nelder, 1989, p. 174) and $\boldsymbol{\Sigma}_Y$ is the multinomial variance-covariance matrix $\boldsymbol{\Sigma}_{Y_{gb}} = m_{gb}\{\mathrm{Diag}(\boldsymbol{\pi}_g) - \boldsymbol{\pi}_g\boldsymbol{\pi}_g^T\}$ (McCullagh and Nelder, 1989, p. 168), where $\mathrm{Diag}(\boldsymbol{\pi}_g)$ is a diagonal matrix with elements of the vector $\boldsymbol{\pi}_g$ on the main diagonal. An approximately unbiased estimator of σ^2 as an overall dispersion parameter is the value of the Pearson's X^2-statistic divided by its degrees of freedom for the full model

$$\widetilde{\sigma}^2 = X^2/\text{residual d.f.} \tag{3.12}$$

$$= X^2/N(C-1) - P, \tag{3.13}$$

with N as the number of clusters, C the number of categories of the nominal response and P the number of non-redundant parameters (McCullagh and Nelder, 1989, p. 174). When comparing G groups and there are no further covariates in the model, P is the number of groups multiplied by the number of categories minus 1, $P = G(C-1)$. Pearson's statistic is known as the sum of the squared Pearson residuals

$$X^2 = \sum_{g=1}^{G}\sum_{b=1}^{b_g}\sum_{c=1}^{C} r_{gbc}^2 \tag{3.14}$$

with Pearson residuals defined by

$$r_{gbc} = \frac{y_{gbc} - m_{gb}\widehat{\pi}_{gc}}{\sqrt{m_{gb}\widehat{\pi}_{gc}}} \tag{3.15}$$

where $m_{gb}\widehat{\pi}_{gc}$ are the estimated expected counts according to the fitted model (Agresti, 2013, p. 18).

In summary, to account for overdispersion the variance-covariance matrix of the parameter estimates is inflated by dispersion factor σ^2 (McCullagh and Nelder, 1989). This leads to maximum-likelihood estimates, which are still consistent but standard errors are multiplied by $\sqrt{\sigma^2} = \sigma$ (Agresti, 2013, p. 149).

3.2.4 Multiple Comparison Adjustment

Based on the parameter vector $\boldsymbol{\theta}$, simultaneous comparisons can now be formulated between the g independent treatment groups using the log odds ratios for $C > 2$ categories. One may wish to test for the alternative that at least one element of $\boldsymbol{\theta}$ differs from null or any other pre-specified value. Then p-values and confidence intervals must be adjusted for multiple testing.

Suppose, that we want to test the intersection of the k null hypotheses $\theta_k = 0$ against the union of k alternative hypotheses $\theta_k \neq 0$, that is,

$$H_0 : \bigcap_{k=1}^{K} \theta_k = 0 \quad \text{vs.} \quad H_1 : \bigcup_{k=1}^{K} \theta_k \neq 0. \tag{3.16}$$

Considering that our postulated multinomial logit model is a subform of a multivariate generalized linear model (McCullagh and Nelder, 1989) which is a special case of general parametric models (Hothorn et al., 2008), the linear function $\widehat{\boldsymbol{\theta}} = (\boldsymbol{B} \otimes \boldsymbol{I}_I)\widehat{\boldsymbol{\delta}}$ asymptotically follows a multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance-covariance matrix $\boldsymbol{V}$ (Agresti, 2013). The variance-covariance matrix $\boldsymbol{V}$ of the differences of the log odds is obtained via

$$\boldsymbol{V} = (\boldsymbol{B} \otimes \boldsymbol{I}_I)\boldsymbol{\Sigma}(\boldsymbol{B} \otimes \boldsymbol{I}_I)^T \tag{3.17}$$

where $\boldsymbol{\Sigma}$ is the corresponding variance-covariance matrix of the parameter vector $\boldsymbol{\delta}$. This matrix $\boldsymbol{\Sigma}$ consists of the individual covariance matrices of the $\boldsymbol{\delta}_g$ and is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2\boldsymbol{\Sigma}_1 & & \dots & \boldsymbol{0} \\ & \sigma^2\boldsymbol{\Sigma}_2 & & \vdots \\ \vdots & & \ddots & \\ \boldsymbol{0} & \dots & & \sigma^2\boldsymbol{\Sigma}_g \end{pmatrix} \tag{3.18}$$

which is the joint covariance matrix of $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \dots, \boldsymbol{\delta}_G$ with zero matrices in the off-diagonal blocks due to the fact that the groups $g = 1, \dots, G$ are assumed to be independent and hence $\text{cov}(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{i'}) = \boldsymbol{0}, i \neq i'$. Each vector $\boldsymbol{\delta}_g$ has the asymptotic covariance

matrix

$$\boldsymbol{\Sigma}_g = m_g^{-1}(\boldsymbol{A}\text{Diag}(\boldsymbol{\pi}_g)^{-1}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{A}\mathbf{1}\mathbf{1}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}) \tag{3.19}$$

where $\text{Diag}(\boldsymbol{\pi}_g)^{-1}$ is the inverse of the diagonal matrix with elements of $\boldsymbol{\pi}$ (Agresti, 2013, p. 591). Estimators for $\boldsymbol{V}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_g$, which are respectively indicated by $\widehat{\boldsymbol{V}}$, $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_g$, can be approximated by using $\widehat{\boldsymbol{\pi}}_g$ as a point estimator for $\boldsymbol{\pi}_g$ and $\widetilde{\sigma}^2$ as a point estimator for σ^2.

The initial null hypotheses can be tested using the test statistics

$$t_k = \frac{\widehat{\theta}_k}{\sqrt{\widehat{v}_k}} \tag{3.20}$$

where $\widehat{v}_k$ is the variance of the k-th log odds ratio, i.e. the k-th diagonal element of the variance-covariance matrix $\widehat{\boldsymbol{V}}$. Since the true covariance matrix is unknown and its estimate $\widehat{\boldsymbol{V}}$ depends on both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\sigma}^2$, we thus work with a multivariate t-distribution rather than a multivariate normal distribution for the distribution of the test statistics under the null hypothesis. p-values can then be calculated from this multivariate t-distribution (Genz and Bretz, 2009). Alternatively, two-sided simultaneous confidence intervals based on t integrals can be constructed by

$$\exp[\widehat{\theta}_k \pm t_{1-\alpha,df=N(C-1)-P,\widehat{\boldsymbol{R}}}\sqrt{\widehat{v}_k}], \quad k = 1, \ldots, K \tag{3.21}$$

with k parameters of interest in $\widehat{\boldsymbol{\theta}}$. The critical value $t_{1-\alpha,df=N(C-1)-P,\widehat{\boldsymbol{R}}}$ (two-sided) can be obtained from the K-variate t distribution with correlation matrix $\widehat{\boldsymbol{R}}$, which is the standardised variance-covariance matrix $\widehat{\boldsymbol{V}}$ (Genz and Bretz, 2009, Hothorn et al., 2008).

Hence, we extended the multiple comparison procedure of Schaarschmidt et al. (2017) by taking overdispersion into account. Adjusted p-values and simultaneous confidence intervals can be calculated for standard multiple comparisons such as multiple comparisons to control (Dunnett, 1955) and all pairwise comparisons (Tukey, 1953) but also user-defined contrasts for multiple odds ratios in analogy to Bretz et al. (2001).

3.2.5 Multiple Contrast Test for Categories and Groups

Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ define the comparisons of interest between the categories and between the groups, respectively. In a multinomial logit model, baseline logits in $\boldsymbol{A}$ are often of interest. Depending on the importance of the different categories, each level can be selected as a reference category. Furthermore, other contrasts than comparisons to baseline may be important. All pairwise logits that are possible can be written as

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}_{I \times C}$$

from which any line can be deleted depending on the comparisons of interests. Note that $\boldsymbol{A}$ is a matrix of dimension $I \times C$, where the number of columns corresponds to the number of categories. The number of rows is set according to scientific interest.
Likewise, different comparisons may be of interest within groups. Widely used contrast matrices are the contrasts of Dunnett (1955) and Tukey (1953). Multiple comparisons to a control group ("Dunnett-type"), e.g. the first treatment group, can be defined as

$$\boldsymbol{B} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{(G-1) \times G}$$

The number of columns is equal to the number of groups. Since each group is compared to a control group, the number of rows is equal to $G - 1$. All pairwise comparisons between groups ("Tukey-type") are determined via

$$\boldsymbol{B} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & \cdots & 0 & 1 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}_{J \times G}$$

The number of rows expands to the total number of group comparisons of immediate interest. Depending on the experimental question, the contrasts in $\boldsymbol{B}$ can also be specified as requested. A variety of contrasts have been investigated (Bretz et al., 2010). Thus, a contrast matrix can also be set up for comparisons to the grand mean, which is

the mean of all means. For a balanced setup that is

$$\boldsymbol{B} = \begin{pmatrix} {}^{G-1}/_{G} & {}^{1}/_{G} & \cdots & {}^{1}/_{G} \\ {}^{1}/_{G} & {}^{G-1}/_{G} & \cdots & {}^{1}/_{G} \\ \vdots & \vdots & \ddots & \vdots \\ {}^{1}/_{G} & {}^{1}/_{G} & \cdots & {}^{G-1}/_{G} \end{pmatrix}_{G \times G}$$

A trend test suggested by Williams (1971) can also be carried out. The contrast coefficients for a balanced design are given by

$$\boldsymbol{B} = \begin{pmatrix} -1 & 0 & \cdots & 0 & 0 & 1 \\ -1 & 0 & \cdots & 0 & {}^{1}/_{2} & {}^{1}/_{2} \\ -1 & 0 & \cdots & {}^{1}/_{3} & {}^{1}/_{3} & {}^{1}/_{3} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ -1 & {}^{1}/_{G-1} & \cdots & {}^{1}/_{G-1} & {}^{1}/_{G-1} & {}^{1}/_{G-1} \end{pmatrix}_{(G-1) \times G}$$

where each individual contrast test consists of comparisons of the weighted average of the last groups to control. In case of unequal sample sizes, contrast coefficients can be defined according to Bretz (2006), who describes an extension of Williams' test to general unbalanced linear models.
As mentioned earlier, the choice of contrast should be well considered. The more hypotheses are simultaneously defined in a contrast matrix, the more adjustment is needed to control the type-I-error rate.

3.3 Simulation Study

3.3.1 Dirichlet-Multinomial Distribution and Overdispersion

Overdispersed multinomial data can be simulated by a Dirichlet mixture of multinomials. To generate multinomially distributed random counts, non-negative vectors π_{gc} are drawn from a Dirichlet distribution, such that

$$(y_{gb1}, y_{gb2}, \ldots, y_{gbC}) \sim multinomial(m_{gb}, (\pi_{gb1}, \pi_{gb2}, \ldots, \pi_{gbC}))$$
$$(\pi_{gb1}, \pi_{gb2}, \ldots, \pi_{gbC}) \sim Dirichlet(\boldsymbol{\alpha}_g)$$

specifying the probability for C classes and where $\boldsymbol{\alpha}_g = (\alpha_{g1}, \alpha_{g2}, \ldots, \alpha_{gC})$ is the Dirichlet parameter vector for the g-th treatment group. Let $\alpha_{g.} = \sum_c \alpha_{gc}$, meaning that $\alpha_{g.}$ is the sum of the parameter vector across categories $c = 1, ..., C$ in group g. In the

Dirichlet-multinomial distribution the variance of $\boldsymbol{Y}_{gb}$ is given by

$$\text{var}(\boldsymbol{Y}_{gb}) = m_{gb}\{\text{Diag}(\boldsymbol{\pi}_g) - \boldsymbol{\pi}_g\boldsymbol{\pi}_g^{\text{T}}\}\left(\frac{m_{gb} + \alpha_{g.}}{1 + \alpha_{g.}}\right) \tag{3.22}$$

(Johnson et al., 1997, p. 81), whereas the corresponding moment of the multinomial distribution (Johnson et al., 1997, p. 34) is defined by

$$\text{var}(\boldsymbol{Y}_{gb}) = m_{gb}\{\text{Diag}(\boldsymbol{\pi}_g) - \boldsymbol{\pi}_g\boldsymbol{\pi}_g^{\text{T}}\}. \tag{3.23}$$

Thus a common overdispersion factor σ^2, given equal sizes in the clusters, can be modelled by

$$\sigma^2 = \frac{m_{gb} + \alpha_{g.}}{1 + \alpha_{g.}} \tag{3.24}$$

leading to

$$\alpha_{g.} = \frac{\sigma^2 - m_{gb}}{1 - \sigma^2} \tag{3.25}$$

for each multinomial vector.

Accordingly multinomial data with overdispersion can be generated by converting a given overdispersion parameter σ^2 and cluster size m_{gb} to a scaling parameter $\alpha_{g.}$ and first generating a random vector of probabilities $\boldsymbol{\pi}_{gb}$ for each cluster b in group g from the Dirichlet distribution with parameters $\boldsymbol{\alpha}_g = \alpha_{g.}\boldsymbol{\pi}_g$. Subsequently each vector of probabilities is used to draw a vector of observations $\boldsymbol{y}_{gb}$ from the multinomial distribution with probability $\boldsymbol{\pi}_g$ and equal number of trials m_{gb}.

3.3.2 Simulation Settings

In a simulation study, the performance of the method proposed was investigated for a variety of characteristics. For this purpose, simulations of the type-I-error were performed and the coverage probability of simultaneous confidence intervals was studied. Furthermore, the power for selected parameters was examined.

The simulation basis for both the type-I-error and coverage probability are certain sets of probability vectors, which can be found in the Appendix A.2. Depending on whether the true proportions of the categories were constant or varied over treatment groups, the sets can be divided into scenarios under the null hypothesis and scenarios under the alternative hypothesis. In the constant case (settings under H_0), 21 scenarios were considered, for which the categories had the same probability across groups. The first scenario additionally covers the case that all categories had the same probability ($\pi_{g1} = 1/3, \pi_{g2} = 1/3, \pi_{g3} = 1/3$). The next scenarios varied from high probabilities in the first

category and low probabilities in the last, e.g. ($\pi_{g1} = 0.90, \pi_{g2} = 0.09, \pi_{g3} = 0.01$) to low probabilities in the first category and high probability in the last category, e.g. ($\pi_{g1} = 0.01, \pi_{g2} = 0.09, \pi_{g3} = 0.90$). In the non-constant case, 38 scenarios were arranged, in which the probabilities for the categories in the first group were similar to those of the constant case. However, some log odds showed differences between some treatment groups leading to settings under the alternative.

Type-I-Error (α)

In order to assess the impact of different sampling schemes on the familywise error rate (FWER), a simulation with a reasonable selection of parameter settings was conducted for the 21 scenarios where all parameters were under the true null hypothesis. In these, only Dunnett-type many-to-one contrasts on the basis of $G = 4$ treatment groups were considered for the comparison of baseline log odds. If one of the comparisons within a scenario became statistically significant at the 5% level, a type-I-error was committed. The sample size in clusters m_{gb} was set to $m_{gb} = \{10, 20, 50\}$, with $b_g = \{5, 10, 20\}$ clusters per treatment group. Both m_{gb} and b_g were assumed to be equal among the groups. For each setting, $10,000$ random samples of multinomial count data with $C = 3$ unordered categories were drawn from a Dirichlet-multinomial distribution with different degrees of overdispersion (see 3.3.1). The degree of overdispersion was chosen with $\sigma^2 = \{1.01, 1.5, 2, 5\}$, in the interest of a broad range from no overdispersion to strong overdispersion. To illustrate the effect of ignoring present overdispersion, the proposed procedure as described in this Chapter 3 was simulated along with a procedure that ignores overdispersion. That is, on the one hand, overdispersion was estimated and included in the analysis using the enhanced covariance-matrix $\widetilde{\sigma}^2\widehat{\Sigma}$ whereas on the other hand, despite the fact that the data may exhibit overdispersion, $\widetilde{\sigma}^2$ was set to 1.

Simultaneous Confidence Intervals

Again the test setup assumed $C = 3$ outcome categories and $G = 4$ treatment groups but equally acquired $b_g = \{5, 10, 15, 20, 50, 100, 500, 1000\}$ clusters per treatment group, resulting in $N = \{20, \ldots, 4000\}$ clusters in total. The sample size in clusters m_{gb} was assumed to be equal among clusters and among groups with a total number of subjects of $m_{gb} = \{10, 50, 100, 500, 1000\}$ per cluster. A dispersion factor was chosen in a way as to yield an overdispersion of $\sigma^2 = \{1.01, 2, 5, 8\}$ in the data.

Baseline log odds with the first category selected as a reference were compared between groups initially with the first group set as a control group (Dunnett within categories and treatment groups). Additionally, all pairwise comparisons between treatment groups for baseline logits were studied in a second simulation (Dunnett within categories and Tukey within groups).

Separately for each of the 160 parameter combinations for the two types of multiple

comparisons, a number of 10,000 datasets were generated from a Dirichlet-multinomial model and simultaneous 95%-confidence intervals were examined whether they all contain the corresponding true odds ratios. As soon as one of the intervals from a confidence set did not contain the corresponding true value, the complete confidence set was recorded as "not covering the true parameter vector". The entire proportion of confidence sets covering the true parameter vector was calculated, which we will refer to as simultaneous coverage probability.

Power

A simulation study investigating the power was conducted on a selection of three probability vectors, which specifically cover the cases of low probability for one of the $C = 3$ categories. Namely these vectors are $(\pi_{g1} = 0.95, \pi_{g2} = 0.04, \pi_{g3} = 0.01)$, $(\pi_{g1} = 0.90, \pi_{g2} = 0.05, \pi_{g3} = 0.05)$ and $(\pi_{g1} = 0.80, \pi_{g2} = 0.10, \pi_{g3} = 0.10)$. The same proportions were assumed for the first three treatment groups. Power was simulated for an increase Δ in the second baseline odds, $\frac{\pi_{g3}}{\pi_{g1}}$, in the fourth group only. The effect Δ was chosen as the factor by which the second baseline odds in the fourth group increases compared to the first group, i.e. $\frac{\pi_{13}}{\pi_{11}} \cdot \Delta = \frac{\pi_{43}}{\pi_{41}}$. Therefore, a Delta of 1 does not alter the probabilities and random samples are drawn under the assumption of a true null hypothesis. Overall, an effect of $\Delta = \{1, \ldots, 10\}$ increasing in constants of 0.5 was realized. By increasing the effect Δ, the percent of datasets wherein at least one null hypothesis is rejected determines the global power for that specific setting. The number of clusters per treatment group was reduced to $b_g = \{5, 10, 20\}$ and the size of a cluster was set to $m_{gb} = \{10, 20, 50\}$, again both considered equally in all groups. Further, only Dunnett-type many-to-one comparisons were studied and the degree of overdispersion was set at $\sigma^2 = \{1.01, 1.5, 2, 5\}$. A number of 5,000 simulation runs were obtained.
Note, that Δ is defined on the scale of the odds but not on the logarithmic scale. An increase of $\Delta = 1.5$ for the first vector, for example, leads to an expression of $(\pi_{41} = 0.9453, \pi_{42} = 0.0398, \pi_{43} = 0.0149)$ in the categories of the fourth group. Although the odds ratio between the second category and the first category remains the same at $\delta_{41} = \log(0.0421)$, the ratio of the third category to the first category is increased by formerly $\delta_{12} = \log(0.0105)$ in the first group to $\delta_{42} = \log(0.0158)$ in the last group, i.e. by $\log(1.5) = 0.4054$.

All simulations were performed in `R`, version 3.3.3 (R Core Team, 2015), using package `MCMCpack` for Dirichlet random numbers (Martin et al., 2011). The proposed method of this Chapter 3 was implemented in `R` and used for evaluation of all simulated data. For more details on the implementation and use of `multcomp`, `mvtnorm` and `VGAM`, see Chapter 5.

3.3.3 Sampling Zeros

Difficulties in calculating the FWER, the simultaneous coverage probability and the power exist in the observation of zero event counts. Datasets with groups containing only zeros in a category were allowed unless all groups were affected. Then this dataset was discarded and redrawn. Nevertheless note that there are settings where data records are sparse for individual categories. This occurs when the probability for a category π_{gc} is very small and there are only a few observations m_{gb} in a few clusters b. In such cases, the confidence limits of parameters become very large and approach infinity within limits. A type-I-error can then no longer occur under the null hypothesis. The same applies to the power under the alternative.

Chapter 4

Results

A simulation study was performed to investigate the characteristics of the proposed simultaneous test procedure under relevant conditions. First, the familywise error rate (FWER) is assessed for scenarios under the null hypothesis. The selected scenarios cover a small sample size in total as well as small sample sizes per cluster. Thereafter, simultaneous confidence intervals (sCI) of the linear contrasts analogous to Tukey and Dunnett are examined. In order to evaluate the intervals, the associated coverage probabilities (CP) for different settings under the null and alternative hypothesis are used as criteria. Finally, the power is considered for certain vectors of probability to further characterize the performance of the test procedure.

4.1 Violation of the Type-I-error when Ignoring Overdispersion

For illustration purposes, Figure 4.1 shows the serious inflation of the FWER in case of a two-sided Dunnett contrast for four groups compared to control, if overdispersion in the data is ignored. On assumption of $\sigma^2 = 1.01$, which describes nearly no overdispersion in the data, the estimated type I error remains below the chosen alpha level of 0.05 for all combinations of b_g number of clusters and m_{gb} units per cluster. As the minimal expected event count in clusters increases the estimated type-I-error levels off at $\alpha = 0.05$. If overdispersion comes into play, the FWER increases rapidly and exceeds the nominal alpha level. In the case of slight overdispersion of $\sigma^2 = 1.5$, the Dunnett test procedure leads to type-I-errors that are nearly two times higher than $\alpha = 0.05$ for an expected number of events per cluster starting from 20. For $\sigma^2 = 2$ the FWER further increases and reaches values up to a maximum of $\alpha = 0.74$ at $\sigma^2 = 5$.
In comparison, Figure 4.2 shows the simulated FWER additionally for the same set of scenarios but taking overdispersion into account. As expected, the mean FWER remains below the predetermined level of 0.05 in all cases of overdispersion. The sample size in

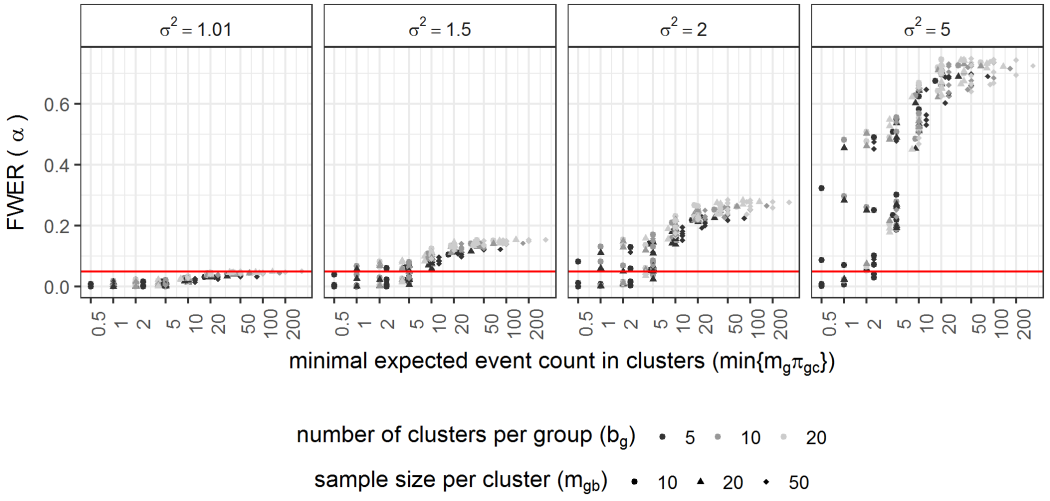

Figure 4.1: Simulated familywise error rate for multiple comparisons to a control without incorporating overdispersion. Estimated familywise error rate for a multiple Dunnett-type contrast test with different underlying levels of overdispersion but no incorporation of overdipersion. The rates are calculated on different numbers of clusters and different numbers of units per cluster. In addition, different proportions were simulated in the categories. The a priori chosen alpha level of 0.05 is represented by the horizontal line.

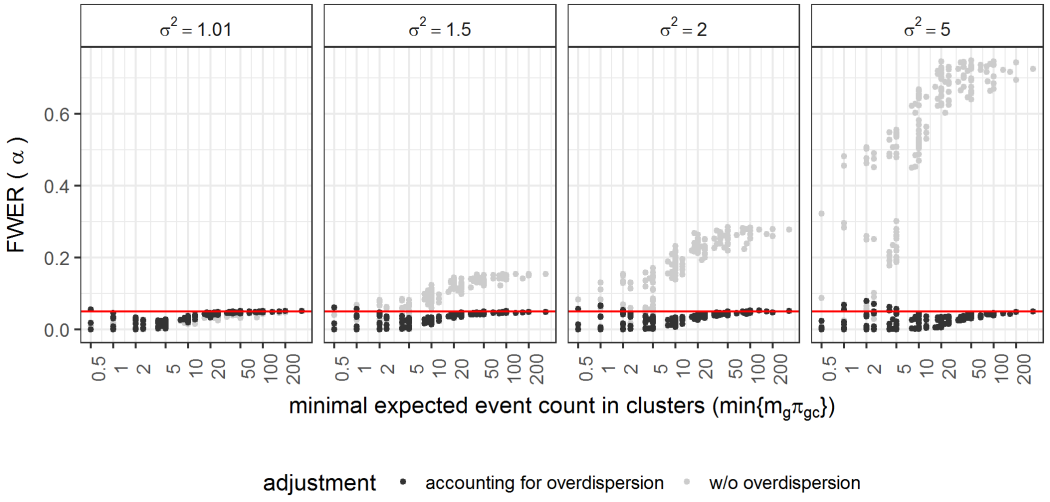

Figure 4.2: Simulated familywise error rate for multiple comparisons to a control with and without incorporating overdispersion. Comparison of the estimated familywise error rate including overdispersion and without inclusion of overdispersion depending on the minimal expected event count in clusters. The a priori chosen alpha level of 0.05 is represented by the horizontal line.

total or the sample size per cluster does not affect the FWER, as long as overdispersion is estimated and incorporated in the analysis.
A more detailed analysis will be given in the next subsection. The assessment of the coverage probability of sCI includes the FWER in a strong sense. In addition, Tukey-type contrasts are considered next to Dunnett-type comparisons.

4.2 Simultaneous Coverage Probability when Accounting for Overdispersion

The simultaneous coverage probability is examined for all scenarios of the null and alternative hypotheses. The points in Figure 4.3 show the probability that the 95% sCI contains the true value for four groups compared to control and all pairwise comparisons, depending on the minimal expected event count ($\min\{m_g\pi_{gc}\}$). That is the minimum of the expected counts per category of all clusters among groups. The cases considered had minimum expected event counts ranging from 0.5 to nearly $200,000$ with the smallest sample size per cluster at $m_{gb} = 10$ and $b_g = 5$ clusters at least. The plot is divided row-wise according to settings under the global null hypothesis and settings with at least one true alternative. The columns correspond to the preselected overdispersion specifications. To better illustrate the behaviour of the proposed method in settings with small sample sizes and a few clusters, Figure 4.4 shows the CP for small event counts of 1 to 50 only. It is assumed that these are also the more relevant settings in praxis.
According to the minimal expected event count the pattern is similar for all cases of overdispersion in settings under the null hypothesis for comparisons with a control in Figure 4.3 in the top two rows. The investigated method shows a conservative behaviour for small expected counts per cluster or rare events. With increasing sample size the CP asymptotically approaches the 95% level for all settings. In case of no overdispersion described by $\sigma^2 = 1.01$, CP is close to 97.5% if the lowest expected event count equals 10 and near the nominal level of 95% from 20 event counts onwards. At minor overdispersion ($\sigma^2 = 2$) a minimal expected event count of at least 35 is needed to reach the nominal level. In case of very strong overdispersion ($\sigma^2 \geq 5$), the method shows its more conservative behaviour up to a minimum expected count of 100.
Under the alternative, a broader scattering is present for smaller event counts as overdispersion increases. The probability of a category containing only zeros in one group becomes more likely as the minimum event count drops lower. Settings with sparse data show a high coverage probability of the 95%-sCI (see Appendix A.1). For event counts less than 5 the probability that at least one of the groups just contains zeros is more than 20% when $\sigma^2 = 2$ and the average CP is 97.24%. For the same range of event

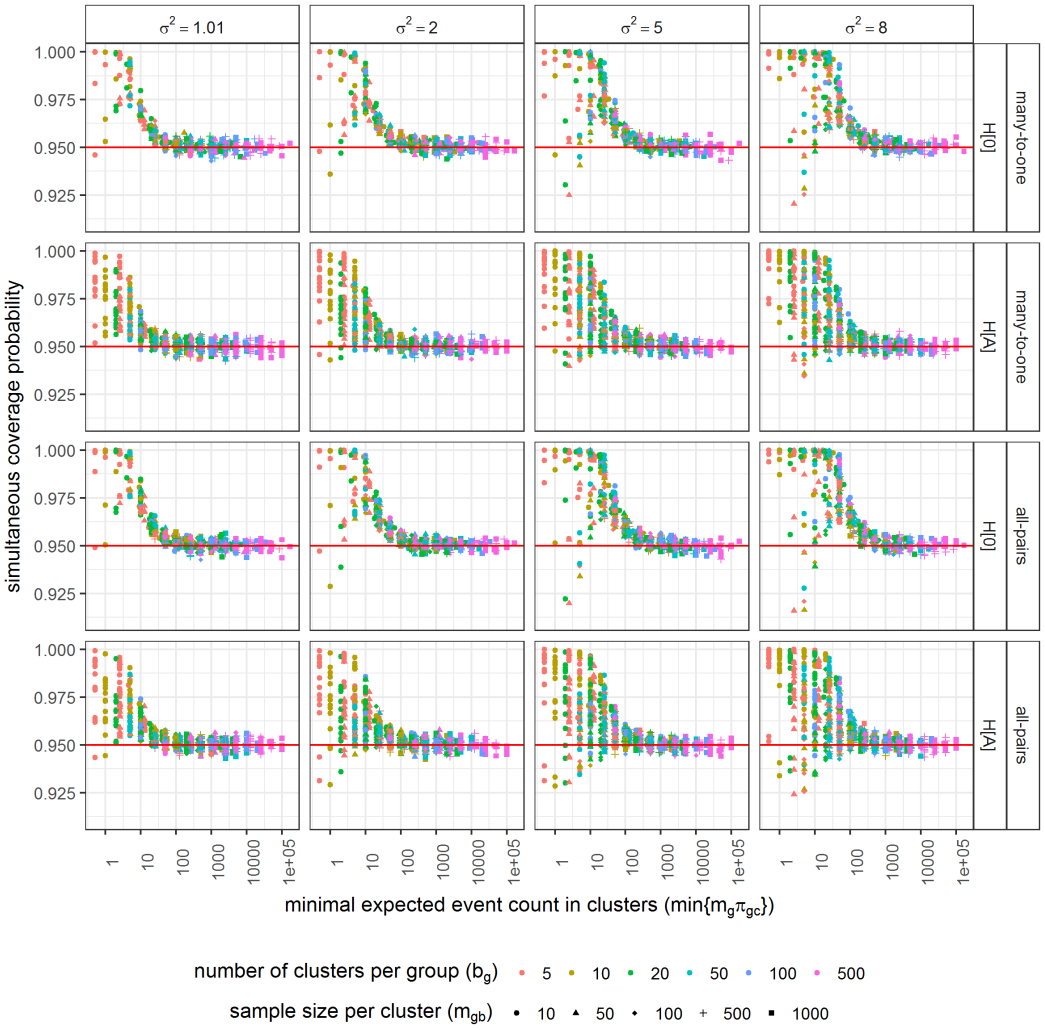

Figure 4.3: Simultaneous coverage probability for multiple comparisons to a control and all pairwise comparisons. Estimated coverage probability for two-sided simultaneous confidence intervals depending on the minimum expected event count. First, for 4 groups comparing to control in the top two rows (many-to-one), and second, for all pairwise comparisons with 4 groups in the bottom two rows (all-pairs). Scenarios are separated line by line according to settings in which all logits of interest are equal in all treatments groups (H[0]) and settings in which at least one logit is different between treatments groups (H[A]). The colour differs according to the number of clusters per treatment group and the symbols indicate the number of observations per cluster. The clusters are of equal sizes and the number of clusters remains the same in each group. Each point is evaluated by 10,000 simulation runs. A nominal level of 0.95 coverage is represented by the horizontal red line.

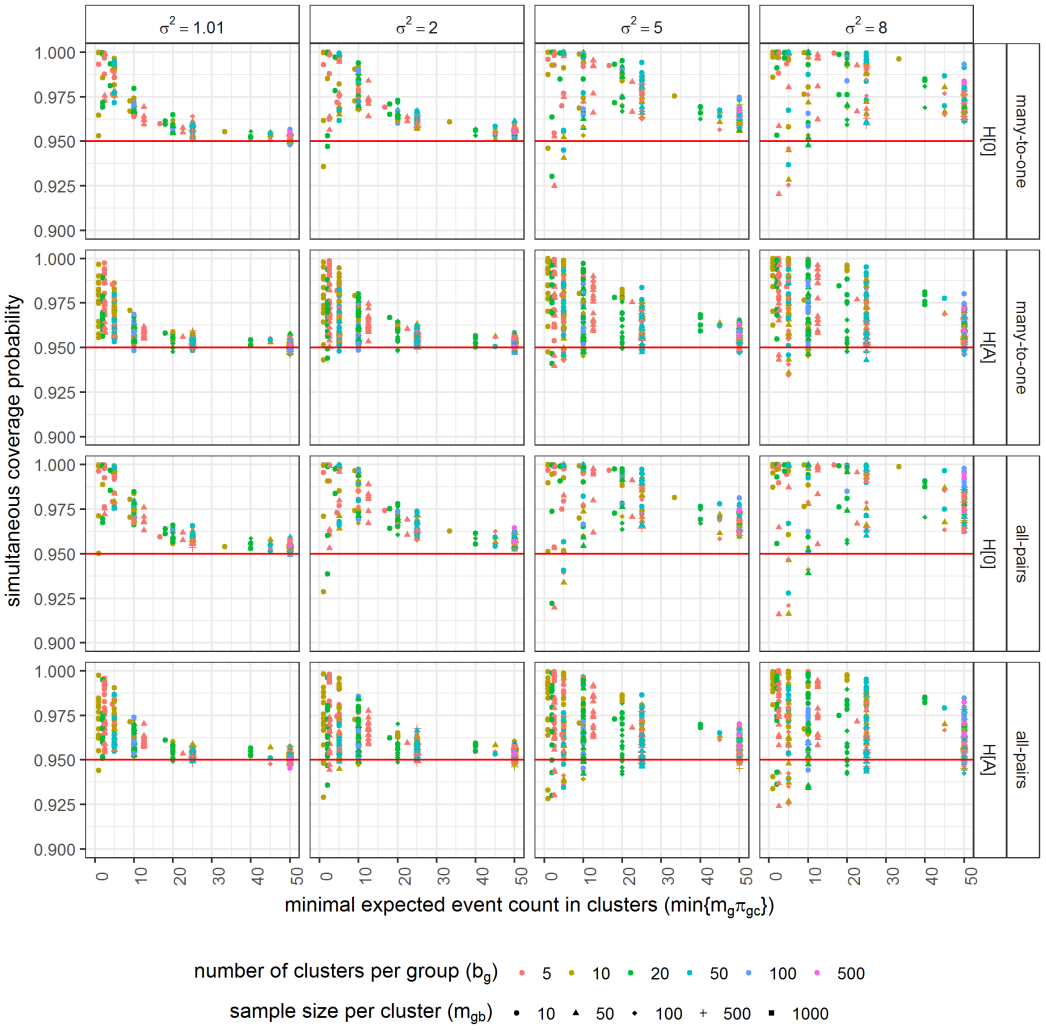

Figure 4.4: Detail of the simultaneous coverage probability for multiple comparisons to a control and all pairwise comparisons. Estimated coverage probability for two-sided simultaneous confidence intervals restricted to a set of scenarios with minimal expected event counts of 1 to 50. First, for 4 groups comparing to control in the top two rows (many-to-one), and second, for all pairwise comparisons with 4 groups in the bottom two rows (all-pairs). Scenarios are separated line by line according to settings in which all logits of interest are equal in all treatments groups (H[0]) and settings in which at least one logit is different between treatments groups (H[A]). The colour differs according to the number of clusters per treatment group and the symbols indicate the number of observations per cluster. The clusters are of equal sizes and the number of clusters remains the same in each group. Each point is evaluated by 10,000 simulation runs. A nominal level of 0.95 coverage is represented by the horizontal red line.

counts but an overdispersion of $\sigma^2 = 8$, the mean CP is 98.30%. With increasing event counts, the method asymptotically approaches the 95% level for all settings.
A similar figure occurs for the coverage probability of two-sided sCI of all pairwise comparisons with four groups in Figure 4.3 in the bottom two rows. The pattern of the all-pairs comparisons is generally similar to the previous simulation of many-to-one comparisons: If the minimum expected event count is small, the 95%-sCI cover the true parameter too often; this applies to settings with a minimum expected event count below 15 for $\sigma^2 \leq 2$ or 100 for $\sigma^2 = 8$. From then on, the procedure shows a coverage close to 95%, e.g. the average CP of all settings with $\sigma^2 = 5$ and minimal expected count of 20 under the null hypothesis is 95.73% where the lowest point estimate is at 94.29%. Under the alternative the mean CP for the nominal 95%-sCI is 95.35%; the minimum at 94.21%.
The simultaneous coverage probability depending on the number of clusters per group, divided according to the sample size in clusters, is shown in Figure 4.5. The figure contains multiple comparisons to a control and all pairwise comparisons as well as scenarios under the null hypothesis and scenarios under the alternative hypothesis. sCI cover the parameters too often if the number of clusters is small and the sample size in clusters (m_{gb}) is small and/or overdispersion is high. If the sample size per cluster is large, e.g. $m_{gb} = 500$ or $m_{gb} = 1000$, CP is equal to the nominal confidence level even for a small number of clusters. If the sample size in clusters is moderate, e.g. $m_{gb} = 50$, a CP close to 95% is reached from $b_g = 50$ clusters per group at an overdispersion of $\sigma^2 = 2$. At $m_{gb} = 50$ and $\sigma^2 = 5$, a CP close to 95% is achieved at $b_g = 100$.

4.3 Power-Simulations

Concerning the power, the lines in Figure 4.6 display the probability that the test procedure will reject at least one null hypothesis regardless of whether the contrast is truely under the alternative (global power). Power curves are shown for 5, 10, and 20 clusters in each treatment group when varying the underlying parameters. By way of example, the simulations were carried out for three vectors with extreme expectations of probabilities of individuals being in three different categories. It is expected that less extreme proportions provide better results of power.
All over the power increases as the effect Δ increases, depending on the number of clusters and their size as well as the degree of overdispersion. The results demonstrate that if the effect is sufficiently large, a power of at least 80% can be achieved even at substantial overdispersion of $\sigma^2 = 5$ for sparse categories of $\pi_{gc} = 0.1$ (third framework of Figure 4.6). If the sample size in clusters decreases, the power eventually reaches an acceptable value later. For a number of clusters of $b_g = 20$ and a cluster size of $m_{gb} = 20$ sufficient power is attained from a delta of $\Delta = 4.5$. The power to detect an effect of

$\Delta = 4.5$ at a cluster size of $m_{gb} = 10$ is 40.56%. The situation is similar for the impact of the number of clusters. As b_g decreases, the power diminishes. Thus, with a slight overdispersion of $\sigma^2 = 1.5$ and a cluster size of $m_{gb} = 10$, sufficient power is reached from $\Delta = 3.5$ for a number of $b_g = 20$ clusters. For a number of $b_g = 10$ or $b_g = 5$ clusters, the detectable effect shifts to 5.5 or 10, respectively.

The power is low for proportions with very small proportions for individual categories, e.g. $\pi_{gc} = 0.01$. The first framework of Figure 4.6 shows a power of less than 80% for all settings with a low cluster size. The power can be increased by increasing the number of clusters or the number of units per cluster. Even with $b_g = 20$ clusters and $m_{gb} = 20$ units per cluster, an adequate power is achieved at $\sigma^2 = 1.5$ from $\Delta = 8$.

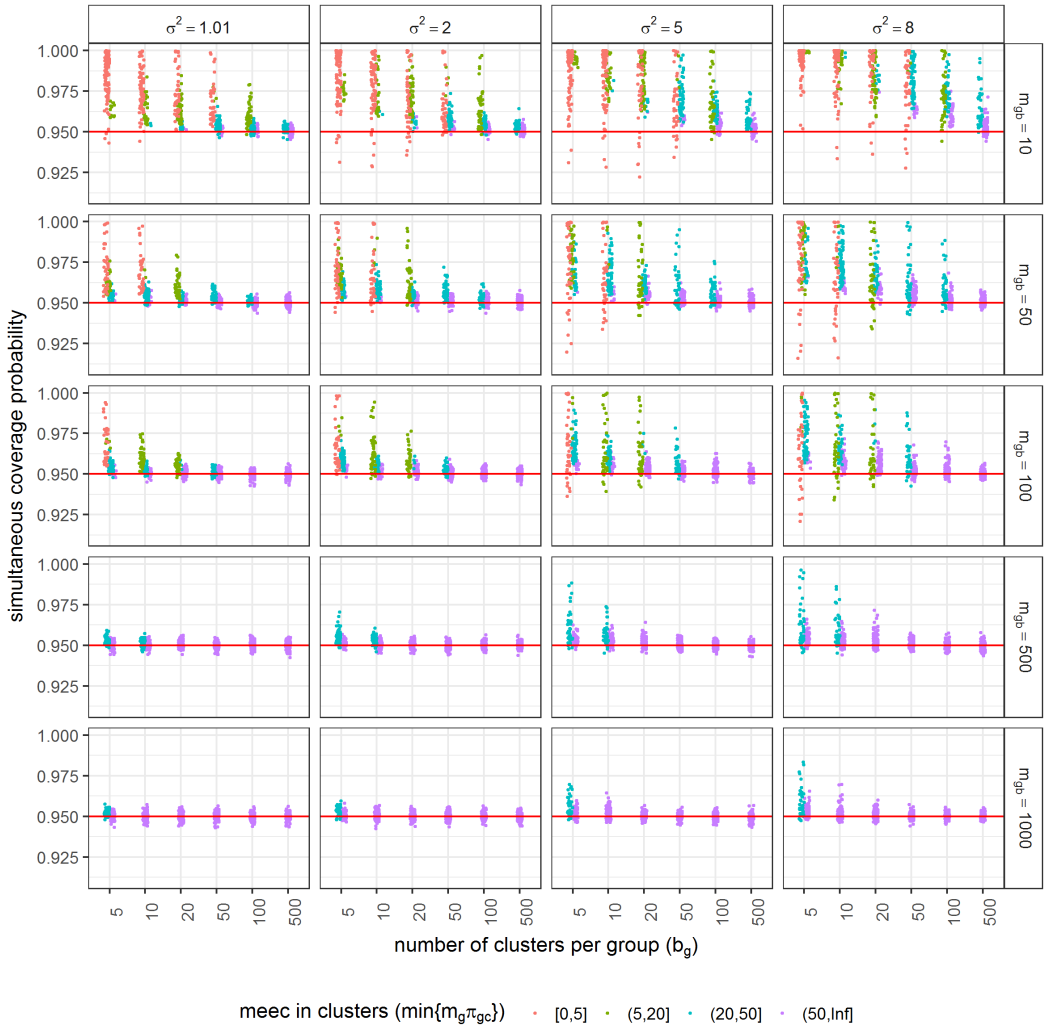

Figure 4.5: Simultaneous coverage probability of nominal 95% simultaneous confidence intervals for multiple comparisons to a control and all pairwise comparisons, in dependence of the number of clusters per group. Estimated coverage probability is presented depending on the number of clusters per group. Rows show different sample size within clusters (m_{gb}). The colour distinguishes the minimal expected event count into intervals: leftmost interval corresponds to low meec and rightmost intervals corresponds to high meec. Clusters are of equal sizes and number of clusters remains the same in each group. Each point is evaluated by 10,000 simulation runs. A nominal level of 0.95 coverage is represented by the horizontal red line.

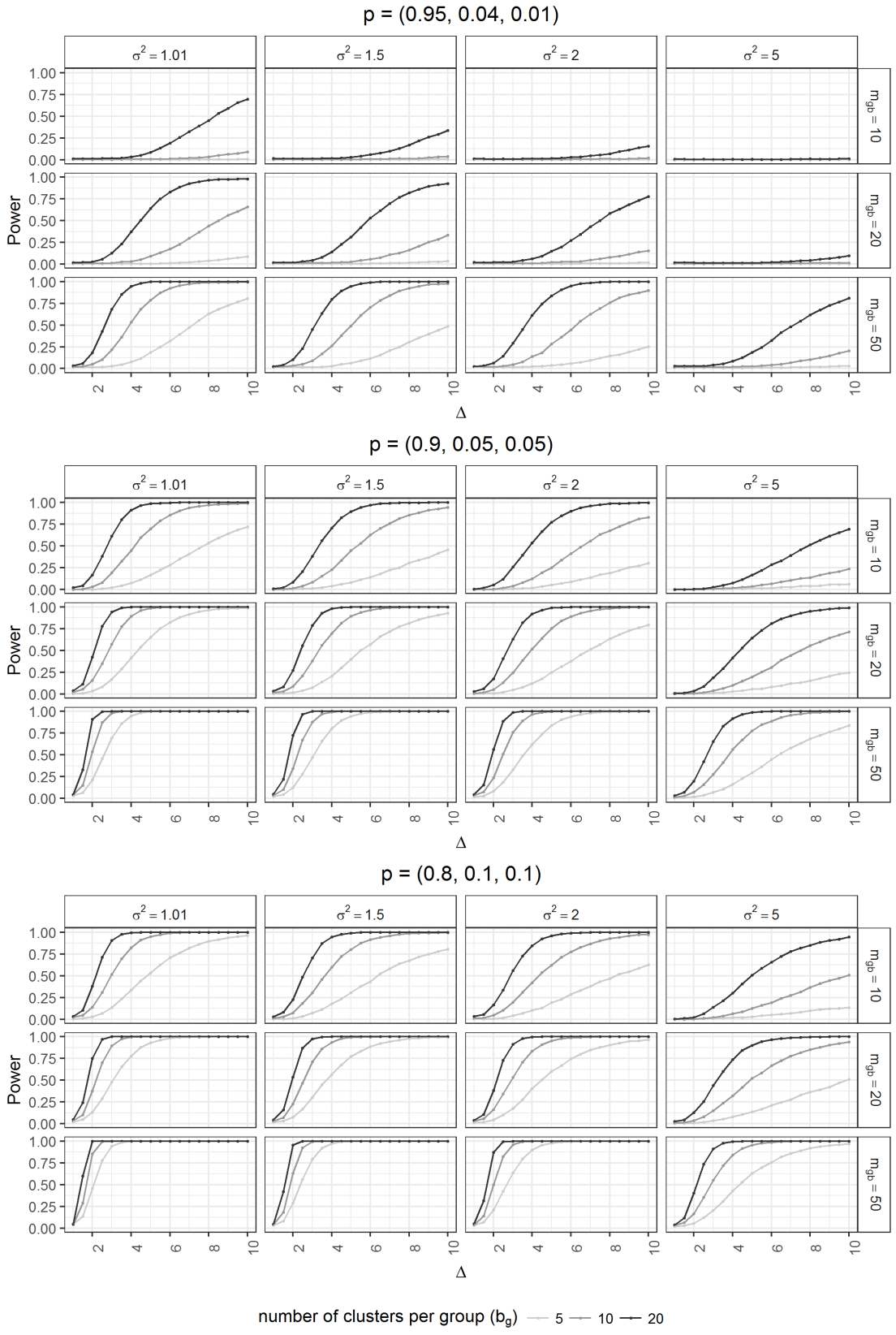

Figure 4.6: Simulated power for multiple comparisons to a control. Estimated power curves for a multiple Dunnett-type contrast on three vectors with sparse event probability. The values are calculated on different numbers of clusters and different numbers of units per cluster with different underlying levels of overdispersion.

Chapter 5

Evaluation of the Examples by Specially Implemented Software

The availability of an adequate software for the user-friendly evaluation of multiple multinomial test problems is essential. This chapter deals with the processing of multiple comparisons of multinomial vectors in the statistical software environment `R` and the implementation of the proposed asymptotic method. Presented functions can be applied to models of the class "vglm". The sample data sets from Chapter 2 are evaluated and exemplary `R` code is provided.

5.1 Computational Issues and Software Implementation

In `R` already a variety of options for modelling multinomial data exist. For example, the function `multinom()` of the `nnet` package (Venables and Ripley, 2002), or `mlogit()` from the `mlogit` package (Croissant, 2013) can be used to estimate a multinomial logit model. However, these options do not include an estimation of a dispersion parameter to blend into the model. By calling `summary(..., dispersion = 0)` of a `VGAM` object a dispersion parameter is estimated which is used to adjust the underlying variance-covariance structure of the model (Yee, 2008). Nevertheless, all functions mentioned are missing the possibility of multiple testing.

An implementation of multiple comparisons for multinomial regression models according to the methodology of this thesis is available by sourcing the code from the website powered by the Institute for Biostatistics (`https://www.biostat.uni-hannover.de/fileadmin/institut/r-code/methods.R`). The code accesses available functions of the package `multcomp` (Hothorn et al., 2008) for multiple contrast tests in general parametric models and extends it to handle vector generalized linear models from the `VGAM` package (Yee, 2015). In addition, the `mvtnorm` package (Genz and Bretz (2009)) is used in the

Table 5.1: Overview of additional functions. List of additionally implemented functions, where * indicates the mandatory arguments.

function	arguments	description
overdispersion()	model* data strata	calculates overdispersion for objects inheriting from class "vglm"
multin2mcp()	object* dispersion*=c("none", "overall", "stratified") data strata	returns a list of parameters (coef, vcov, df)
mcp2matrix.vglm()	model* linfct*	sets up a list that includes a specific contrast matrix K
model.frame.vglm()	model*	returns a data.frame with the variables of a "vglm"-model
modelparm.vglm()	model*	enables `multcomp` to extract model parameters from a "vglm"-model
vcov.disp()	model* linfct*	multiplies the model-vcov with an estimated dispersion parameter

background for computing multivariate normal probabilities and multivariate t probabilities and quantiles. An overview of additionally implemented functions is summarized in Table 5.1.

In particular, the framework of `multcomp` has been extended to objects of class "vglm" as part of this thesis. By implementation of `model.frame.vglm()`, `modelparm.vglm()` and `mcp2matrix.vglm()`, general linear hypotheses and multiple comparisons according to Hothorn et al. (2008) are now available for vector generalized linear models from the `VGAM` package. In addition, the functions `overdispersion()` and `vcov.disp()` have been implemented. This provides the following functionality: Given an object of class "vglm", overdispersion can be calculated using the function `overdispersion()`. In particular, overdispersion can be estimated overall or group-specific for each individual group level. For the former, only the model is needed to compute the dispersion parameter over all observations. For group-specific overdispersion (see Section 6.1), the data frame and a group variable to stratify for must also be specified. Certainly, overdispersion can directly be incorporated in the linear comparisons of interest. Using the function `multin2mcp()` on an object of class "vglm" and a contrast matrix with the same number of columns as numbers of parameters in the model, the function `glht()` of the `multcomp` package can be called as usual for multiple comparisons with or without (group-specific) overdispersion.

Associated p-values are calculated using the `summary()` function of `multcomp`, which calculates the quantiles from the multivariate t-distribution utilizing the `mvtnorm` package. Simultaneous confidence intervals can be constructed by applying the internal `confint()` function to the `glht` object.

5.2 Evaluation of the Examples

5.2.1 Example 1: Developmental Toxicity

The study on the toxicity of DYME, which was introduced in Section 2.2, recorded the survival status of pups of toxin-treated mice. Obviously, the primary endpoint measured as "alive", "malformed" or "dead" is nominal and can be described by a multinomial logit model using the dose group as an explanatory variable. Already in Section 2.5 overdispersion was suspected behind the data, which we now want to estimate exactly.

```
> str(bivar.re)
'data.frame':   108 obs. of  5 variables:
 $ DAM_ID    : int  51 60 61 70 71 79 80 88 95 104 ...
 $ DOSE      : Factor w/ 5 levels "0","62.5","125",..: 1 1 1 1 1 1 1 1 1 1
    ...
 $ alive     : num  10 14 11 17 15 17 11 13 16 12 ...
 $ malformed: num  0 0 0 0 0 0 2 0 0 0 ...
 $ dead      : num  0 0 1 0 0 0 1 0 0 0 ...
> multivgam <- vglm(cbind(alive,malformed,dead) ~ DOSE,
+                   family=multinomial(refLevel = "alive"),
+                   data=bivar.re)
> sum(residuals(multivgam, type = "pearson")^2)
[1] 485.7833
> overdispersion(multivgam)
[1] 2.358171
```

At first, the probability of observing a live, malformed or dead pup at each dose is estimated in a vector generalized linear models (VGLM), i.e. a multinomial logit model, with setting "alive" as the reference category. Subsequently, the sum of Pearson residuals of $X^2 = 485.78$ can be calculated. The number of degrees of freedom is $N \times (C-1) - P$, where N is the number of clusters (equal to the number of rows) and C is the number of levels in the multinomial response. $P = 2 \cdot 5$ is the number of parameters in the model corresponding to 2 baseline logits in each of the 5 dose groups. The ratio of the Pearson statistic to the degrees of freedom is then $485.78/(108 \cdot (3-1)-10) = 2.358$, which is much larger than 1 and means that substantial overdispersion exists in this data. To simplify this calculation step, a function called `overdispersion()` was implemented. Evidently, standard errors of parameter estimates have to be multiplied by $\sqrt{\sigma^2} = 1.536$.

The calculations of the logits and log odds ratio estimates are as follows: The vectors of probabilities for each group according to the model are presented in Table 5.2, e.g. the parameter vector of dose group 0 is $\widehat{\boldsymbol{\pi}}_{\mathrm{d0}}^{\mathrm{T}} = (\widehat{\pi}_{11}, \widehat{\pi}_{12}, \widehat{\pi}_{13}) = (0.9664, 0.0067, 0.0268)$.

Table 5.2: Estimated probabilities of the developmental toxicity study. Estimated probabilities of each life status, sample size and minimal event count per dose group.

dose	$\widehat{\pi}_{\text{alive}}$	$\widehat{\pi}_{\text{malformed}}$	$\widehat{\pi}_{\text{dead}}$	m_g	$\min_c\{m_g\widehat{\pi}_{gc}\}$
0	0.9664	0.0067	0.0268	298	2
62.5	0.9579	< 0.0001	0.0421	214	< 1
125	0.8644	0.0252	0.1104	317	8
250	0.6667	0.2170	0.1164	318	37
500	0.1833	0.5312	0.2855	802	147

On the right, the number of observations in group $g \in \{0, 62.5, 125, 250, 500\}$ and the minimum expected event count is given.

Due to the main focus on baseline logits comparing the probabilities of categories "malformed" and "dead" to the first category "alive" the contrast matrix $\boldsymbol{A}$ is set to

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

and log odds can be estimated for each treatment group, e.g. in the first dose group by means of

$$\widehat{\boldsymbol{\delta}}_{d0} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \log \begin{pmatrix} 0.9664 \\ 0.0067 \\ 0.0268 \end{pmatrix} = \begin{pmatrix} -4.97 \\ -3.58 \end{pmatrix},$$

leading separately for all dose groups to: $\widehat{\boldsymbol{\delta}}_{d0} = \left(\begin{smallmatrix} -4.97 \\ -3.58 \end{smallmatrix}\right)$, $\widehat{\boldsymbol{\delta}}_{d62.5} = \left(\begin{smallmatrix} -19.54 \\ -3.13 \end{smallmatrix}\right)$, $\widehat{\boldsymbol{\delta}}_{d125} = \left(\begin{smallmatrix} -3.53 \\ -2.06 \end{smallmatrix}\right)$, $\widehat{\boldsymbol{\delta}}_{d250} = \left(\begin{smallmatrix} -1.12 \\ -1.75 \end{smallmatrix}\right)$, $\widehat{\boldsymbol{\delta}}_{d500} = \left(\begin{smallmatrix} 1.06 \\ 0.44 \end{smallmatrix}\right)$.

Furthermore, multiple comparisons to control are of interest in this experiment now as recommended in the OECD Guideline for statistical analysis of ecotoxicity data (OECD, 2014). Therefore, contrast matrix $\boldsymbol{B}$ is set up with comparisons to dose group 0,

$$\boldsymbol{B} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Multiplying the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ by the vector of probabilities returns the parameter vector $\widehat{\boldsymbol{\theta}}$.

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{B} \otimes \boldsymbol{A}) \log \begin{pmatrix} \boldsymbol{\pi}_{d0}^{\mathrm{T}} \\ \boldsymbol{\pi}_{d62.5}^{\mathrm{T}} \\ \boldsymbol{\pi}_{d125}^{\mathrm{T}} \\ \boldsymbol{\pi}_{d250}^{\mathrm{T}} \\ \boldsymbol{\pi}_{d500}^{\mathrm{T}} \end{pmatrix}$$

The vector $\widehat{\boldsymbol{\theta}}$ consists of all estimated log odds ratios, which are evaluated in Table 5.3

Table 5.3: Comparison of adjusted and unadjusted simultaneous test results of the developmental toxicity example with / without accounting for overdispersion. Simultaneous test results for the $K = 8$ comparisons to control for baseline odds ratios malformed/alive and dead/alive. The first column shows the results for comparison without adjustment for multiplicity and without accounting for overdispersion; the second column shows results for comparison with multiplicity adjustment including overdispersion. MCP = multiple comparison procedure.

		w/o MCP, $\sigma^2 = 1$		MCP, $\sigma^2 = 2.358$	
Linear Hypotheses	$\widehat{\boldsymbol{\theta}}_k$	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value
malformed/alive: 62.5 - 0	-14.5687	740.8176	0.9843	1137.6246	1.0000
malformed/alive: 125 - 0	1.4361	0.7951	0.0709	1.2209	0.7343
malformed/alive: 250 - 0	3.8473	0.7230	0.0000	1.1102	0.0039
malformed/alive: 500 - 0	6.0338	0.7160	0.0000	1.0995	0.0000
dead/alive: 62.5 - 0	0.4577	0.4944	0.3546	0.7593	0.9821
dead/alive: 125 - 0	1.5257	0.4009	0.0001	0.6156	0.0730
dead/alive: 250 - 0	1.8379	0.4003	0.0000	0.6147	0.0176
dead/alive: 500 - 0	4.0268	0.3737	0.0000	0.5738	0.0000

along with corresponding standard errors and p-values. The results are presented for the case of neither adjustment for multiple testing nor for overdispersion and for the case of simultaneous comparisons with adjustment for overdispersion.

Both analyses can be reproduced in R. For instance, the analysis adapting to overdispersion is as follows:

```
> mcp <- glht(model = multin2mcp(multivgam, dispersion="overall"),
+             linfct = mcp2matrix(multivgam,
+                                 linfct = mcp(DOSE = "Dunnett"))$K)
> summary(mcp)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Linear Hypotheses:
                                 Estimate Std. Error t value Pr(>|t|)
malformed/alive: 62.5 - 0 == 0  -14.5687  1137.6246  -0.013  1.00000
malformed/alive: 125 - 0 == 0     1.4361     1.2209   1.176  0.73426
malformed/alive: 250 - 0 == 0     3.8473     1.1102   3.465  0.00398 **
malformed/alive: 500 - 0 == 0     6.0338     1.0995   5.488  < 0.001 ***
dead/alive: 62.5 - 0 == 0         0.4577     0.7593   0.603  0.98214
dead/alive: 125 - 0 == 0          1.5257     0.6156   2.479  0.07282 .
dead/alive: 250 - 0 == 0          1.8379     0.6147   2.990  0.01758 *
dead/alive: 500 - 0 == 0          4.0268     0.5738   7.017  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

After fitting the multinomial logit model, two functions facilitate the multiple comparisons of odds ratios. The first function `multin2mcp()` modifies the variance-covariance

matrix under consideration of overdispersion if desired. The choice of modification can be set by using one of the expressions "none", "overall" or "stratified" in the `dispersion` argument. For a single dispersion parameter, the argument is set to "overall" in this case. Since the `glht()` function still requires a contrast matrix, the second function helps with a little workaround. The `mcp2matrix()` function specifies multiple comparisons similar to the `multcomp` package but especially for `VGAM` objects. The contrast matrix is extracted by accessing $K. The outcome is stored in the object `mcp` so that one can continue working with it. Finally, p-values of a corresponding hypotheses test are computed by the `summary()` function.

As one can see, the summary of our `glht` object matches the right-hand column of Table 5.3. If the user is interested in multiple comparisons without modification of the variance-covariance matrix, i.e. without adaptation of overdispersion, the `dispersion` argument has to be set to "none". Alternatively, an analysis with several dispersion parameters might be considered. The bar chart in 2.1 suggests heterogeneous variances in treatment groups and the variability seems to be much higher for the animals at the highest dose level. As a choice of the `dispersion` argument, "stratified" can be selected to estimate separate dispersion parameters for each group. In contrast, the analysis without consideration of the multiplicity problem and without accounting for overdispersion (left column of Table 5.3) can be accomplished via the simple output of summary results of the model: `summary(multivgam)`.

Note that the first line of the output and Table 5.3 respectively highlights a result of small event counts. Because the observations in the category "malformed" of group 62.5 were all zero, the estimated parameters become very large. The calculation of the estimates and standard error is based on a model in `VGAM`, which terminates numerically after a certain number of iterations whereas the actual estimates approach infinity.

The p-values of the right-hand column of Table 5.3, which corresponds to the recommended evaluation of this dataset, indicate that there is a significant difference of log odds of "malformed" relative to "alive" between dose 250 and dose 0 and between dose 500 and the control group. Also, there is a significant difference in the log odds of "dead" relative to "alive" between dose 250 and dose 0 and between dose 500 and dose 0. Thus, the analyses only differ when comparing dose 125 to dose 0 on the basis of the log odds of "dead" relative to "alive".

Simultaneous confidence intervals (on the logit scale) can be calculated by using the `confint()` statement on our `glht` object named `mcp`:

```
> confint(mcp)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts
```

```
Fit: NULL

Quantile = 2.6244
95% family-wise confidence level


Linear Hypotheses:
                                Estimate   lwr         upr
malformed/alive: 62.5 - 0 == 0 -1.457e+01 -3.000e+03  2.971e+03
malformed/alive: 125 - 0 == 0   1.436e+00 -1.768e+00  4.640e+00
malformed/alive: 250 - 0 == 0   3.847e+00  9.337e-01  6.761e+00
malformed/alive: 500 - 0 == 0   6.034e+00  3.148e+00  8.919e+00
dead/alive: 62.5 - 0 == 0       4.577e-01 -1.535e+00  2.450e+00
dead/alive: 125 - 0 == 0        1.526e+00 -8.979e-02  3.141e+00
dead/alive: 250 - 0 == 0        1.838e+00  2.247e-01  3.451e+00
dead/alive: 500 - 0 == 0        4.027e+00  2.521e+00  5.533e+00
```

By default, 95% sCIs are reported, which can be changed by request via the `level` argument, e.g. adding `level=0.90` as an argument in `confint()` for 90% sCIs. The sCIs indicate the precision of the estimated log odds ratios and can be converted to the original scale using the natural exponential function. The respective 95% sCIs after back-transformation to the response scale are shown in Table 5.4. According to the reported sCI in the second line, there is a 95% probability that the true difference of log odds of "malformed" relative to "alive" between dose 125 and dose 0 lies between the lower confidence limit of 0.17 and the upper confidence limit of 103.43. Since the interval crosses the null value, i.e. 1 on the original scale and 0 on the log scale, this implies that there is no significant difference between dose 125 and dose 0. In contrast, the estimate in line 3 is the odds of being "malformed" rather than "alive" in dose group 250 which is increased by factor 46.87 compared to the control group. The true odds ratio is between 2.55 and 862.39 assuming there is no bias or confounding. This result is statistically significant as the sCI does not overlap 1 on the original scale. In

Table 5.4: Simultaneous 95% confidence intervals including overdispersion of the developmental toxicity example on the original scale. Simultaneous confidence intervals for the $K = 8$ comparisons to control for baseline odds ratios malformed/alive and dead/alive. Simultaneous confidence intervals are given for a significance level of $\alpha = 5\%$. MCP = multiple comparison procedure.

		MCP, $\sigma^2 = 2.358$	
	$\widehat{\boldsymbol{\theta}}_k$	2.5%	97.5%
malformed/alive: 62.5 - 0	0.0000	0.0000	∞
malformed/alive: 125 - 0	4.2044	0.1709	103.4325
malformed/alive: 250 - 0	46.8679	2.5471	862.3903
malformed/alive: 500 - 0	417.3061	23.3269	7465.3861
dead/alive: 62.5 - 0	1.5805	0.2157	11.5823
dead/alive: 125 - 0	4.5985	0.9148	23.1171
dead/alive: 250 - 0	6.2830	1.2528	31.5093
dead/alive: 500 - 0	56.0816	12.4468	252.6875

fact, this is revealed in Table 5.3 too, which shows a p-value of less than 0.01 for this comparison.

5.2.2 Example 2: Housing Satisfaction

For the second example, whose study design was explained in Section 2.6, we will have a closer look at the data structure first.

```
> head(housing)
  us s vs  type
1  3 2  0 rural
2  3 2  0 rural
3  0 5  0 rural
4  3 2  0 rural
5  0 5  0 rural
6  4 1  0 rural
> str(housing)
'data.frame':   35 obs. of  4 variables:
 $ us  : num  3 3 0 3 0 4 3 2 4 0 ...
 $ s   : num  2 2 5 2 5 1 2 3 0 4 ...
 $ vs  : num  0 0 0 0 0 0 0 0 1 1 ...
 $ type: Factor w/ 2 levels "rural","urban": 1 1 1 1 1 1 1 1 1 1 ...
```

In the dataset, each row represents a cluster of 5 households. This is crucial for the correct counting of N in the function `overdispersion()`. The belonging of a cluster to an area is stored in the factor variable "type". Next, a multinomial logit model with levels of satisfaction as a response and type of area as an independent predictor is set up.

```
> multivgam <- vglm(cbind(us, s, vs) ~ type,
                    family=multinomial(refLevel="us"),
                    data=housing)
> sum(residuals(multivgam, type = "pearson")^2)
[1] 107.2768
> overdispersion(multivgam)
[1] 1.625406
```

The sum of Pearson residuals is $X^2 = 107.28$. With $N = 35$ clusters, $C = 3$ categories and $P = 4$ parameters in the model, the overdispersion is estimated to be $\widetilde{\sigma}^2 = 107.28/(35(3-1)-4) = 1.625$. Therefore it is recommended to account for overdispersion.

Beforehand, the parameters of interest are calculated. In Table 5.5 the three-dimensional vectors of parameter estimates for the proportions $\widehat{\boldsymbol{\pi}}_g$ are presented for each group.

Table 5.5: Estimated probabilities of the housing data. Estimated probabilities of satisfaction, sample size and minimal event count per area

g	$\widehat{\pi}_{\text{us}}$	$\widehat{\pi}_{\text{s}}$	$\widehat{\pi}_{\text{vs}}$	m_g	$\min_c\{m_g\widehat{\pi}_{gc}\}$
rural	0.5222	0.4222	0.0556	90	5
urban	0.3529	0.5059	0.1412	85	12

To compute the log odds based on the satisfaction level, i.e. "s" versus "us" and "vs" versus "us", Matrix A is set to

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

and log odds in each group are $\widehat{\boldsymbol{\delta}}_{\text{rural}} = \left(\begin{smallmatrix} -0.21 \\ -2.24 \end{smallmatrix}\right)$, $\widehat{\boldsymbol{\delta}}_{\text{urban}} = \left(\begin{smallmatrix} 0.36 \\ -0.92 \end{smallmatrix}\right)$. In order to compare the baseline logits for "urban" versus "rural", contrast matrix $\boldsymbol{B}$ is defined by $\boldsymbol{B} = (-1 \quad 1)$. Log odds ratios $\widehat{\boldsymbol{\theta}}$ can be obtained by processing the vector $\log(\widehat{\boldsymbol{\pi}}_{\text{rural}}, \widehat{\boldsymbol{\pi}}_{\text{rural}})^{\text{T}}$ or by using the stacked vector of log odds $\widehat{\boldsymbol{\delta}}$. The latter is calculated by means of

$$\widehat{\boldsymbol{\theta}} = \left(\begin{pmatrix} -1 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} -0.21 \\ -2.24 \\ 0.36 \\ -0.92 \end{pmatrix} = \begin{pmatrix} 0.57 \\ 1.32 \end{pmatrix}.$$

A complete analysis of the log odds ratios for the different levels of satisfaction considering multiple comparisons and overdispersion is provided in Table 5.6 in the last column. In addition, the results are presented on the one hand for an analysis without adjustment for multiple comparisons and on the other hand for an analysis without adaptation to overdispersion and neither. In all ways of evaluation, the estimates are the same, but standard errors have been multiplied by $\sqrt{\widetilde{\sigma}^2} = 1.275$ in the third and forth analysis when overdispersion is taken into account. The variance-covariance matrix is multiplied by $\widetilde{\sigma}^2$ in this case,

$$\widehat{\boldsymbol{V}} = 1.625 \begin{pmatrix} 0.104 & 0.055 \\ 0.055 & 0.338 \end{pmatrix},$$

and adjusted p-values are computed from the bivariate t distribution with 66 degrees of freedom, whereby the vector of test statistics $\boldsymbol{t} = (1.391, 1.787)^{\text{T}}$ is associated with the

Table 5.6: Comparison of adjusted and unadjusted simultaneous test results of the housing example each with and without accounting for overdispersion. Simultaneous test results for the $K = 2$ comparisons of "urban" vs. "rural" for I. the log odds of s/us and II. the log odds of vs/us. The first column shows the results for comparison without adjustment for multiplicity and without accounting for overdispersion (w/o MCP, $\sigma = 1$); the second column shows results for comparison with multiplicity adjustment but without adjustment for overdispersion (MCP, $\sigma^2 = 1$); the third column shows results for comparison with adjustment for overdispersion but without accounting for multiplicity (w/o MCP, $\widetilde{\sigma}^2 = 1.625$); the fourth column shows results for comparison with multiplicity adjustment and with adjustment for overdispersion (MCP, $\widetilde{\sigma}^2 = 1.625$). MCP = multiple comparison procedure.

		w/o MCP, $\sigma^2 = 1$		MCP, $\sigma^2 = 1$		w/o MCP, $\widetilde{\sigma}^2 = 1.625$		MCP, $\widetilde{\sigma}^2 = 1.625$	
k	$\widehat{\boldsymbol{\theta}}_k$	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value
I	0.5726	0.3228	0.0761	0.3228	0.1503	0.4115	0.1688	0.4115	0.3001
II	1.3244	0.5813	0.0227	0.5813	0.0501	0.7411	0.0785	0.7411	0.1465

correlation matrix

$$\widehat{\boldsymbol{R}} = \begin{pmatrix} 1.000 & 0.291 \\ 0.291 & 1.000 \end{pmatrix}.$$

Table 5.6 clearly shows the differences in the standard error and thus in the evaluation of the underlying hypotheses. Compensating for overdispersion, the p-values have increased and a difference of log odds of "very satisfied" relative to "unsatisfied" by 1.32 when changing from rural to urban can no longer be rated as significant.
The full analysis of the example by taking overdispersion into account and adjusting for multiple testing can be achieved in `R` by means of the following code, formulated in one commando:

```
> summary(glht(model = multin2mcp(multivgam, dispersion="overall"),
+              linfct = mcp2matrix(model = multivgam,
+                                  linfct = mcp(type = "Dunnett"))$K))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
s/us: urban - rural == 0    0.5726     0.4115   1.391    0.300
vs/us: urban - rural == 0   1.3244     0.7411   1.787    0.146
(Adjusted p values reported -- single-step method)
```

Two-sided simultaneous 95% confidence intervals can, in turn, be calculated by the `confint()` statement and are given on the original scale in Table 5.7. Only sCIs for the multiplicity-adjusted analysis adapting to overdispersion are reported since these correspond to the recommended evaluation method.

Table 5.7: Simultaneous 95% confidence intervals including overdispersion of the housing satisfaction in two areas on the original scale. Simultaneous confidence intervals for the $K = 2$ comparisons for baseline odds ratios. Simultaneous confidence intervals are given for a significance level of $\alpha = 5\%$. MCP = multiple comparison procedure.

		MCP, $\sigma^2 = 1.625$	
	$\widehat{\boldsymbol{\theta}}_k$	2.5%	97.5%
s/us: urban - rural == 0	1.7728	0.6941	4.5280
vs/us: urban - rural == 0	3.7600	0.6946	20.3542

5.2.3 Example 3: Differential Blood Count in Rats

For the third illustration, we consider the toxicological example of white blood cells of Section 2.7. The study question is whether gender and dose group affect the number of white blood cells. There seems to be a downward trend in lymphocytes with increasing dose (see Figure 2.3), but is this trend significant? The data structure is as follows:

```
> head(dbb)
    sex animal   Group Eos Baso Stab Seg Mono  Ly    factorcomb
1 Males   1101 control   2    0    0  51    2 145 Males:control
2 Males   1102 control   3    0    0  28    2 167 Males:control
3 Males   1103 control   4    0    0  32    5 159 Males:control
4 Males   1104 control   3    0    0  32    6 159 Males:control
5 Males   1105 control   8    0    0  30    3 159 Males:control
6 Males   1106 control   1    0    0  52    3 144 Males:control
> str(dbb)
'data.frame':   78 obs. of  10 variables:
 $ sex       : Factor w/ 2 levels "Females","Males": 2 2 2 2 2 2 2 2 2 2
   ...
 $ animal    : num  1101 1102 1103 1104 1105 ...
 $ Group     : Factor w/ 4 levels "control","low dose",..: 1 1 1 1 1 1 1
   1 1 1 ...
 $ Eos       : num  2 3 4 3 8 1 4 4 4 1 ...
 $ Baso      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Stab      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seg       : num  51 28 32 32 30 52 29 23 33 30 ...
 $ Mono      : num  2 2 5 6 3 3 3 0 7 1 ...
 $ Ly        : num  145 167 159 159 159 144 164 173 156 168 ...
 $ factorcomb: Factor w/ 8 levels "Females:control",..: 5 5 5 5 5 5 5 5 5
    5 ...
```

First, we will build a vector generalized linear model for the counts of Eosinophils (Eos), Segmented Neutrophils (Seg), Monocytes (Mono) and Lymphocytes (Ly) depending on gender and dose group. As stated initially, Basophils (`Baso`) and Neutrophilic bands (`Stab`) are excluded from the analysis because their counts were all zero. Since the currently implemented software can only evaluate models with a single explanatory variable in terms of multiple comparisons with overdispersion, we defined a combination of gender and dose group (`factorcomb`), which is included in the model as a categorical predictor variable.

```
> multivgam <- vglm(cbind(Eos, Seg, Mono, Ly) ~ factorcomb,
                    family = multinomial, data = dbb)
> sum(residuals(multivgam , type = "pearson")^2)
[1] 455.7761
> overdispersion(multivgam)
[1] 2.170362
```

The analysis of residuals shows that there is an overdispersion of $\widetilde{\sigma}^2 = 455.78/(78 \cdot (4-1) - 24) = 2.1704$. This estimate indicates that the variance is about twice what we would expect for the multinomial model. Evidently, our analysis will be adjusted for overdispersion later.

In this example, the comparison of some baseline logits between female and male animals is not of interest, therefore a typical Dunnett contrast matrix contains too few comparisons and a Tukey contrast matrix contains too many comparisons. For the calculation of individual log odds ratios, we want to equip matrix $\boldsymbol{A}$ with comparisons to one reference group and matrix $\boldsymbol{B}$ should contain comparisons separated by gender. That is, within

categories logits are compared to Lymphocytes (`Ly`) serving as the baseline category and these are again compared between treatments ($g = \text{low}, \text{mid}, \text{high}$) and control group, separately for males and females. Of course, any other category or group may be chosen as a reference. In our case, $\boldsymbol{A}$ and $\boldsymbol{B}$ are as follows:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

Thus we obtain a vector $\widehat{\boldsymbol{\theta}}$ which is similar to the estimates of the first column in Table 5.8. Again, different standard errors and p-values are shown for the evaluation without overdispersion and without adjustment for multiple tests as opposed to the calculation taking overdispersion and multiple testing into account. Without adjustments, the log odds of Segmented Neutrophils against Lymphocytes appear to be increased in males

Table 5.8: Comparison of adjusted and unadjusted simultaneous test results of the differential blood count in rats with / without accounting for overdispersion. Simultaneous test results for the $K = 9$ comparisons to control for baseline odds ratios each in female and male. The first column shows the results for comparison without adjustment for multiplicity and without accounting for overdispersion; the second column shows results for comparison with multiplicity adjustment and with overdispersion.

		w/o MCP, $\sigma^2 = 1$		MCP, $\sigma^2 = 2.1704$	
Linear Hypotheses	$\widehat{\boldsymbol{\theta}}_k$	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value	SE($\widehat{\boldsymbol{\theta}}_k$)	p-value
Eos/Ly: fem.:low - fem.:con	-0.2152	0.3471	0.5353	0.5114	1.0000
Eos/Ly: fem.:mid - fem.:con	-0.0725	0.3307	0.8265	0.4872	1.0000
Eos/Ly: fem.:high - fem.:con	0.1152	0.3185	0.7176	0.4693	1.0000
Seg/Ly: fem.:low - fem.:con	0.1357	0.0904	0.1332	0.1331	0.9963
Seg/Ly: fem.:mid - fem.:con	-0.1270	0.0947	0.1796	0.1395	0.9990
Seg/Ly: fem.:high - fem.:con	0.1097	0.0907	0.2264	0.1336	0.9997
Mono/Ly: fem.:low - fem.:con	0.1578	0.2165	0.4662	0.3190	1.0000
Mono/Ly: fem.:mid - fem.:con	-0.0684	0.2263	0.7624	0.3334	1.0000
Mono/Ly: fem.:high - fem.:con	-0.0876	0.2294	0.7027	0.3380	1.0000
Eos/Ly: male:low - male:con	-0.2676	0.2629	0.3087	0.3873	1.0000
Eos/Ly: male:mid - male:con	-0.2533	0.2659	0.3407	0.3917	1.0000
Eos/Ly: male:high - male:con	-0.2947	0.2890	0.3079	0.4257	1.0000
Seg/Ly: male:low - male:con	0.0466	0.0837	0.5776	0.1233	1.0000
Seg/Ly: male:mid - male:con	0.2913	0.0809	0.0003	0.1192	0.2218
Seg/Ly: male:high - male:con	0.3191	0.0852	0.0002	0.1255	0.1747
Mono/Ly: male:low - male:con	-0.2462	0.2693	0.3605	0.3967	1.0000
Mono/Ly: male:mid - male:con	0.1148	0.2489	0.6447	0.3667	1.0000
Mono/Ly: male:high - male:con	0.3180	0.2510	0.2052	0.3698	0.9995

Table 5.9: Simultaneous 95% confidence intervals including overdispersion of the differential blood count in rats on the original scale. Simultaneous confidence intervals for the $K = 9$ comparisons to control for baseline odds ratios each in female and male. Simultaneous confidence intervals are given for a significance level of $\alpha = 5\%$. MCP = multiple comparison procedure.

		MCP, $\sigma^2 = 2.1704$	
	$\widehat{\boldsymbol{\theta}}_k$	2.5%	97.5%
Eos/Ly: fem.:low - fem.:con	0.8064	0.1739	3.7391
Eos/Ly: fem.:mid - fem.:con	0.9301	0.2157	4.0107
Eos/Ly: fem.:high - fem.:con	1.1221	0.2746	4.5848
Seg/Ly: fem.:low - fem.:con	1.1454	0.7683	1.7076
Seg/Ly: fem.:mid - fem.:con	0.8807	0.5796	1.3382
Seg/Ly: fem.:high - fem.:con	1.1160	0.7474	1.6663
Mono/Ly: fem.:low - fem.:con	1.1709	0.4498	3.0482
Mono/Ly: fem.:mid - fem.:con	0.9339	0.3435	2.5387
Mono/Ly: fem.:high - fem.:con	0.9162	0.3324	2.5248
Eos/Ly: male:low - male:con	0.7652	0.2394	2.4454
Eos/Ly: male:mid - male:con	0.7762	0.2397	2.5134
Eos/Ly: male:high - male:con	0.7448	0.2077	2.6706
Seg/Ly: male:low - male:con	1.0477	0.7238	1.5166
Seg/Ly: male:mid - male:con	1.3382	0.9358	1.9136
Seg/Ly: male:high - male:con	1.3759	0.9444	2.0046
Mono/Ly: male:low - male:con	0.7817	0.2378	2.5693
Mono/Ly: male:mid - male:con	1.1216	0.3734	3.3691
Mono/Ly: male:high - male:con	1.3744	0.4533	4.1676

receiving high or mid dose compared to males in the control group. A similar interpretation is drawn when adjusting for multiplicity but not for overdispersion (not explicitly calculated here). However, after adjusting for multiplicity and overdispersion, the null hypothesis of equal odds "Segmented/Lymphocytes" between males of high, mid and control group cannot be rejected.

In Table 5.9 the respective two-sided simultaneous 95% confidence intervals on the original scale are given only for the analysis with multiplicity adjustment and adapting to overdispersion. In this analysis, none of the sCIs crosses the null value of 1, so that none of the comparisons is statistically significant. This corresponds to the interpretation of the p-values of Table 5.8.

The equivalent analysis in `R` is partially confusing since an individual contrast matrix has to be defined. On the one hand, we use the `multin2mcp()` function to access the estimates from the `VGAM`-model, so that the contrast matrix must be determined via $\boldsymbol{B} \otimes \boldsymbol{I}_I$. On the other hand, the order of estimates in `VGAM` is different from the arrangement initially assumed for our parameter vector $\boldsymbol{\delta}$. This has not been noticed so far since `mcp2matrix()` inherits this rearrangement. Now we order the estimates manually using the `order()` function. Additionally, the first three columns must be set to zero, because in the beginning the model was fitted with intercept. The entire

calculation of this contrast matrix therefore is:

```
> I <- diag(3)
> library(Matrix)
> B <- matrix(c(-1,1,0,0,
+               -1,0,1,0,
+               -1,0,0,1), byrow=TRUE, nrow=3)
> B <- as.matrix(bdiag(B,B))
> K <- kronecker(B,I)
> Kstar <- K[do.call(order, as.data.frame(K)),]
> Kstar[ ,c(1:3)] <- 0
```

`Kstar` is a contrast matrix of dimension 18×24, which is not explicitly stated in the output. Optionally, the contrast matrix can be provided with names for the hypotheses as row names. In summary, the simultaneous comparison of the linear hypotheses can be carried out as follows:

```
> rownames(Kstar) <- c("Eos/Ly: fem.:low - fem.:con",
+                      "Eos/Ly: fem.:mid - fem.:con",
+                      "Eos/Ly: fem.:high - fem.:con",
+                      "Seg/Ly: fem.:low - fem.:con",
+                      "Seg/Ly: fem.:mid - fem.:con",
+                      "Seg/Ly: fem.:high - fem.:con",
+                      "Mono/Ly: fem.:low - fem.:con",
+                      "Mono/Ly: fem.:mid - fem.:con",
+                      "Mono/Ly: fem.:high - fem.:con",
+                      "Eos/Ly: male:low - male:con",
+                      "Eos/Ly: male:mid - male:con",
+                      "Eos/Ly: male:high - male:con",
+                      "Seg/Ly: male:low - male:con",
+                      "Seg/Ly: male:mid - male:con",
+                      "Seg/Ly: male:high - male:con",
+                      "Mono/Ly: male:low - male:con",
+                      "Mono/Ly: male:mid - male:con",
+                      "Mono/Ly: male:high - male:con")
>
> mcp <- glht(model = multin2mcp(multivgam, dispersion="overall"),
+             linfct = Kstar)
> summary(mcp)

         Simultaneous Tests for General Linear Hypotheses

Linear Hypotheses:
                                    Estimate Std. Error t value Pr(>|t|)
Eos/Ly: fem.:low - fem.:con == 0     0.11518    0.46926   0.245    1.000
Eos/Ly: fem.:mid - fem.:con == 0    -0.07248    0.48722  -0.149    1.000
Eos/Ly: fem.:high - fem.:con == 0   -0.21518    0.51142  -0.421    1.000
Seg/Ly: fem.:low - fem.:con == 0     0.10974    0.13364   0.821    1.000
Seg/Ly: fem.:mid - fem.:con == 0    -0.12705    0.13947  -0.911    0.999
Seg/Ly: fem.:high - fem.:con == 0    0.13572    0.13314   1.019    0.996
Mono/Ly: fem.:low - fem.:con == 0   -0.08755    0.33796  -0.259    1.000
Mono/Ly: fem.:mid - fem.:con == 0   -0.06842    0.33341  -0.205    1.000
Mono/Ly: fem.:high - fem.:con == 0   0.15778    0.31897   0.495    1.000
Eos/Ly: male:low - male:con == 0    -0.29466    0.42572  -0.692    1.000
Eos/Ly: male:mid - male:con == 0    -0.25335    0.39172  -0.647    1.000
Eos/Ly: male:high - male:con == 0   -0.26764    0.38734  -0.691    1.000
```

```
Seg/Ly: male:low - male:con == 0    0.31910    0.12547   2.543    0.175
Seg/Ly: male:mid - male:con == 0    0.29130    0.11924   2.443    0.222
Seg/Ly: male:high - male:con == 0   0.04661    0.12330   0.378    1.000
Mono/Ly: male:low - male:con == 0   0.31804    0.36982   0.860    0.999
Mono/Ly: male:mid - male:con == 0   0.11476    0.36668   0.313    1.000
Mono/Ly: male:high - male:con == 0 -0.24623    0.39669  -0.621    1.000
(Adjusted p values reported -- single-step method)
```

By utilizing the `confint()` function, simultaneous 95% confidence intervals are obtained on the logarithmic scale. As before, facility for converting the sCIs to the original scale exists by de-logarithmizing the obtained confidence limits using the inverse function `exp()`.

Chapter 6

Extensions and Alternative Approaches

In this chapter, we pay attention to possible extensions of the proposed method and examine alternative ways analysing multinomial data. The procedure developed in this thesis can be extended to the case of heterogeneous variances in treatment groups. In Section 6.1, it is shown how overdispersion can be estimated separately for each group and modifications in multiple testing are addressed. As possible alternatives, Section 6.2 discusses evaluation approaches using multiple marginal models. Arguments are based on the model structure and model assumptions as well as parameter estimates and interpretation.

6.1 Estimating Group-Specific Overdispersion

In the previous chapters, homogeneous dispersion was assumed across groups and one overdispersion parameter overall has been estimated. It may be desirable for various applications to estimate heterogeneous dispersion across groups. For instance, Figure 2.1 hints at different variances in treatment groups. For that reason, an individual dispersion parameter is suggested for each group.

A *group specific estimator for overdispersion* can be defined as the sum of residuals for a group g divided by the degrees of freedom from the related intercept-only model

$$\widetilde{\sigma}_g^2 = \frac{\sum_{b=1}^{b_g}\sum_{c=1}^{C} r_{gbc}^2}{b_g(C-1)-P_g} = \frac{\sum_{b=1}^{b_g}\sum_{c=1}^{C} r_{gbc}^2}{(b_g-1)(C-1)}, \tag{6.1}$$

with b_g as the number of clusters belonging to that group and P_g the number of non-redundant parameters needed for the intercept-only model which equals $(C-1)$.

Suppose the parameters of interest and the set of k null hypotheses remain the same as in 3.16. Then the underlying variance-covariance matrix can be estimated via

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} \widetilde{\sigma}_1^2 \widehat{\boldsymbol{\Sigma}}_1 & & \dots & \mathbf{0} \\ & \widetilde{\sigma}_2^2 \widehat{\boldsymbol{\Sigma}}_2 & & \vdots \\ \vdots & & \ddots & \\ \mathbf{0} & \dots & & \widetilde{\sigma}_g^2 \widehat{\boldsymbol{\Sigma}}_g \end{pmatrix} \tag{6.2}$$

Simultaneous inference can be drawn according to Section 3.2.4 using a normal distribution as the underlying distribution of test statistics. Since homogeneous variances are required for using a general degree of freedom in a t-distribution, a joint multivariate t-distribution is not available here. Hasler and Hothorn (2008) state that a single degree of freedom does not maintain the type-I-error and propose the use of comparison-specific degrees of freedom. Each test statistic is then compared with a separate quantile coming from a k-variate t-distribution with adjusted degrees of freedom and correlation matrix $\widehat{\boldsymbol{R}}$, which is the standardised variance-covariance matrix $\widehat{\Sigma}$.

The approach using a normal distribution as a reference distribution is realized as part of the additionally implemented functions. The function `overdispersion()` has the ability to specify a strata variable to estimate separate dispersion parameters for each strata level. Consider the developmental toxicity example of Section 5.2.1. An evaluation with group-specific overdispersion can be accomplished via:

```
> overdispersion(multivgam, strata="DOSE", data=bivar.re)
       0     62.5      125      250      500      all
2.488869 1.222012 1.735192 3.112491 3.045517 2.358171
> mcpSTRAT <- glht(model = multin2mcp(multivgam, dispersion="stratified",
+                                     data=bivar.re, strata="DOSE"),
+                  linfct = mcp2matrix(multivgam,
+                                      mcp(DOSE = "Dunnett"))$K)
Warning message:
In multin2mcp(multivgam, dispersion = "stratified", data = bivar.re,  :
  results will rely on normal approximation in case of group-specific
    overdispersion
> summary(mcpSTRAT)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Linear Hypotheses:
                              Estimate Std. Error t value Pr(>|t|)
malformed/alive: 62.5 - 0 == 0 -14.5687   818.9344  -0.018   1.0000
malformed/alive: 125 - 0 == 0    1.4361     1.0473   1.371   0.5978
malformed/alive: 250 - 0 == 0    3.8473     1.2755   3.016   0.0166 *
malformed/alive: 500 - 0 == 0    6.0338     1.2495   4.829   <0.001 ***
dead/alive: 62.5 - 0 == 0        0.4577     0.5466   0.837   0.9196
dead/alive: 125 - 0 == 0         1.5257     0.5280   2.889   0.0237 *
```

```
dead/alive: 250 - 0 == 0          1.8379     0.7062   2.603   0.0532 .
dead/alive: 500 - 0 == 0          4.0268     0.6521   6.175   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Again, this is an asymptotic approach and the performance for small sample sizes has not been reviewed yet. Therefore, further simulations with the same background as in Chapter 3 are recommended.

6.2 Alternative Approaches Using Multiple Marginal Models

An alternative approach to simultaneously evaluate count data can be based on the formulation of multiple marginal models. Of these, estimates are stacked into a joint vector of coefficients and a variance-covariance matrix is estimated to simultaneously compare a set of linear hypotheses. For this purpose, multivariate count data may be modelled as univariate linear models for binomial data or Poisson-distributed data. To incorporate overdispersion, quasi-binomial or quasi-Poisson models can be formulated. However, this implies that the model assumptions are different than previously considered in a multinomial model.

6.2.1 The Approach of Multiple Marginal Models

A flexible approach has been introduced by Pipper et al. (2012) for multiplicity adjustment when analysing treatment effects from *multiple marginal models* (MMM). These models can be evaluated simultaneously by estimating the correlation between the test statistics using a score decomposition.

Consider L marginal models and let $\boldsymbol{\beta}_l$ be the vector of parameters of the l-th marginal model. Each maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_l$ has an asymptotic representation based on standardized score functions for observations (van der Vaart, 1998). These asymptotic results still hold when stacking the parameters of interest into a single vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_L)$. According to Pipper et al. (2012), the asymptotic representation of the stacked version of parameter estimates converges in distribution to the l-variate normality $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ on behalf of the multivariate central limit theorem. A consistent estimator of the variance-covariance matrix, $\widehat{\boldsymbol{\Sigma}}$, is obtained by plugging in the parameter estimates from the different model fits. Hence no explicit formulation of the variance-covariance matrix of parameter estimates is required. p-values and quantiles for simultaneous confidence intervals will be adjusted using a reference distribution, based on the estimated simultaneous variance-covariance matrix (Hothorn et al., 2008).

In `R` the latter is implemented in the `R` package `multcomp` by means of the functions `mmm()` and `glht()` for the calculation of the correlation matrix and simultaneous testing of hypotheses, respectively. In the basic formulation of `glht()`, when no degrees of freedom are stated, p-values and quantiles for simultaneous confidence intervals rely on a multivariate normal distribution. Otherwise, degrees of freedom may be specified as an additional `df` argument to `glht()` and the multivariate t distribution is used for evaluation.

6.2.2 Model Choice in Case of Multivariate Count Data

In the last chapters, we assumed that each multinomial response vector of counts of a cluster is distributed multinomially. Multinomial data arise from counting the number of mutually exclusive categories out of an a priori fixed number of categorical distributed trials. Clearly, the number of 5 respondents per cluster was determined in advance in Example 2 of Section 2.6 and the satisfaction rating is exactly one of the levels "unsatisfied" or "satisfied" or "very satisfied". Example 3 of Section 2.7 lists exactly 200 cells per animal and the classification of a white blood cell is restricted to one type of leukocytes. Thus in both experiments, two important assumptions of a multinomial distribution hold: the total number of counts per cluster is fixed and every result of a trial can take exactly one of the possible categories.
But in statistics also cases with count data in several categories occur where these assumptions do not apply. Therefore, different choices of models for multivariate count data are possible, e.g. *binomial*, *multinomial* or *Poisson*. Understanding the mechanism that generated the data helps users identify a suitable model to describe the information. The following parts deal with the two important features of the multinomial distribution assumption, i.e. fixed cluster sizes and mutually exclusive categories. The distinction of multinomially distributed data to Poisson distributed data are discussed as well as the clear difference between the multinomial model and a multivariate binary/binomial model.

Fixed Cluster Size

In case of a multinomial distribution, it is assumed that the total number of counts per cluster has been fixed by the design of the experiment or the sampling process, e.g. in Example 2 of Section 2.6 and Example 3 of Section 2.7. After careful consideration of this assumption, the multinomial distribution can be sharply distinguished from the Poisson distribution which is used for count data whose sample size in clusters is random. The Poisson distribution applies when the number of events has been collected within a fixed time interval. This is often accompanied by an unknown upper bound for the maximum number of events or an upper limit that is very large and therefore assumed as irrelevant.

The outcome of Example 1 of Section 2.2 are counts, whose total number per cluster can be different in each trial. The sample size in clusters is determined by the number of foetuses one dam carries which is a random variable. Hence the counts in each category might also be modelled as Poisson. The events of being "alive", "malformed" or "dead" are considered to be independent of all other events. An alternative option to analyse such multivariate Poisson data is the approach of MMM. If we assume that the number of foetuses is fixed, the method developed in Chapter 3 can be applied to Example 1. Then the random counts no longer follow a Poisson distribution but the random vectors in each cluster become multinomial. The counts of this vector are no longer independent, but have to sum up to the total number per cluster. Therefore, the counts of the categories in a cluster are always negatively correlated in a multinomial model.

Mutually Exclusive Categories

Many examples, which appear to be multinomial at first glance, are often multivariate binary on closer inspection. Such applications measure several binary variables simultaneously in a sequence of a fixed number of trials. These classifications are not mutually exclusive categories, but rather a collection of binary responses. The number of events in each endpoint can be summarized per group and cluster. Separately, each count of a cluster in a group follows a binomial distribution.

Consider a toxicological dose-response study on the offspring of mice after exposure to a chemical substance to investigate its carcinogenicity. In each mouse pup, it is detected whether mutations in the lung (EP_1), liver (EP_2) or spleen (EP_3) have occurred. This type of study involves the collection of several binary outcomes, whereby each response variable can take one of just two possible values (cancer: yes/no). The data can be recorded as shown in Table 6.1. In this table, each row corresponds to one experimental

Table 6.1: Data example of ungrouped binary responses. To give an example, imagine that data was obtained in several clusters with different dose levels on three response variables (EP_1, EP_2, EP_3) and outcome was assembled with one subject per line. Each outcome is binary coded and can take one of two values (0/1). The simultaneous occurrence of an event in two or more response variables is possible.

DOSE	DAM_ID	EP_1	EP_2	EP_3
0	11	0	1	1
0	11	1	1	0
...	...	...	...	...
0	11	1	0	1
0	15	0	0	1
0	15	1	1	1
...	...	...	...	...
0	15	0	1	1
62.5	52	1	0	1
...	...	...	...	...

Table 6.2: Data example of grouped binary responses. This Table is generated by summarizing the data of Table 6.1. The outcome was observed as the number of "successes" in a given number of trials m_g. The sample size m_g corresponds to the number of subjects in each cluster, which is considered to be equal for reasons of simplification.

DOSE	DAM_ID	m_g	EP_1	EP_2	EP_3
0	11	5	1	3	3
0	15	5	2	2	5
0	52	5	5	1	1
...	...	...	...	...	...

unit with all response variables being binary coded. For instance, $(1, 0, 0)$ denotes a mutational effect in the lung, but not in the liver or spleen. However, in this example, it is also possible to obtain vectors of the form (1,1,0) which indicate mutations in both the lung and the liver.

Another representation may be obtained by grouping the data as shown in Table 6.2. Each row represents the summarized outcome of m_g trials, i.e. how many subjects out of m_g had a particular response. The latter illustration may look multinomial. Still, more than one organ can be affected by a mutation, which means that the categories are no longer mutually exclusive. Therefore, based on the scientific question in a study, several binary endpoints can be combined into a multivariate binary endpoint. When the binary outcomes are grouped, each count follows a binomial distribution with a certain cell probability in m_g trials.

The method developed in Chapter 3 is not suitable for the analysis of such data since the assumption of mutually exclusive categories is violated. Despite that, the response variables may be correlated which can be either positive or negative. An analysis of multivariate-binary data with mutually exclusive categories is shown in Section 6.2.5. Multivariate-binomial data with non-exclusive categories, i.e. multiple binary findings as in Table 6.2, can be evaluated similarly by use of MMM as suggested by Hothorn (2015).

6.2.3 Overdispersion in a Single Marginal Model

The approach of Pipper et al. (2012) allows to combine multiple marginal models, i.e. for each category c or response EP_c a single marginal model is fitted. Let c be the index for each outcome and later model. Similar to the assumption in multinomial data, both binomial and Poisson data may exhibit overdispersion under certain circumstances. In the presence of clusters, the ordinary model may not be appropriate since the variation in the data is expected to be greater than the variance assumed under the Poisson model or the binomial model, respectively. Instead, the variance is assumed to be inflated by an unknown factor σ^2. In principle as in the case of multinomial data, to account for extra variance in the data a dispersion component is introduced and the variance function is extended through a dispersion parameter $\sigma^2 \operatorname{var}(\cdot)$ (McCullagh and Nelder, 1989).

A Poisson random variable is expected to have the same mean and variance, i.e. $\mathrm{E}(Y_{gbc}) = \mathrm{var}(Y_{gbc}) = \mu_{gbc}$. If the observed variance is larger than the assumed variance, the quasi-poisson variance function $\mathrm{var}(Y_{gbc}) = \sigma^2\mu_{gbc}$ can be used to allow for overdispersion. For a binomial model with $\mathrm{E}(Y_{gbc}) = m_{gb}\pi_{gc}$ and $\mathrm{var}(Y_{gbc}) = m_{gb}\pi_{gc}(1-\pi_{gc})$, the quasi-binomial variance function is $\mathrm{var}(Y_{gbc}) = \sigma^2 m_{gb}\pi_{gc}(1-\pi_{gc})$.
Employing the standard approach of Equation 3.12

$$\widetilde{\sigma}^2 = X^2/\text{residual d.f.}$$

by use of Pearson's statistic for the c-th model

$$X_c^2 = \sum_{g=1}^{G}\sum_{b=1}^{b_g} r_{gbc}^2 \tag{6.3}$$

with Pearson residuals defined by

$$r_{gbc} = \frac{y_{gbc} - \widehat{\mu}_{gbc}}{\sqrt{\widehat{\mu}_{gbc}}} \tag{6.4}$$

in the case of Poisson and

$$r_{gbc} = \frac{y_{gbc} - m_{gb}\widehat{\pi}_{gc}}{\sqrt{m_{gb}\widehat{\pi}_{gc}(1-\widehat{\pi}_{gc})}} \tag{6.5}$$

in the binomial case, a consistent scale parameter for each marginal model can be estimated by

$$\widetilde{\sigma}_c^2 = X_c^2/\{N - P_c\} \tag{6.6}$$

with N as the number of clusters and P_c the number of non-redundant parameters (McCullagh and Nelder, 1989) in the c-th marginal model. Equation 6.6 holds for both the Poisson and the binomial model. Therefore, to deal with overdispersion, the enhanced covariance matrix for $\widehat{\boldsymbol{\theta}}$ is estimated by means of $\widehat{\boldsymbol{\Sigma}} = \sigma_c^2\boldsymbol{\Sigma}$ and the usual standard errors are multiplied by $\sqrt{\sigma_c^2} = \sigma_c$ (Agresti, 2013).

6.2.4 Evaluation of Poisson-Distributed Data Using Multiple Marginal Models

For illustration, Example 1 of Section 2.2 is to be evaluated in `R` through multiple marginal models. We now assume that the number of events in one category is the random variable to be modelled, following a Poisson distribution. For C categories, C univariate generalized linear models (GLM) are set up: one for each category. Strictly

speaking, the individual models are of the same structure, each following a Poisson distribution with its own dispersion parameter.

```
> m1 <- glm(alive ~ DOSE, family=quasipoisson(), data=bivar.re)
> m2 <- glm(malformed ~ DOSE, family=quasipoisson(), data=bivar.re)
> m3 <- glm(dead ~ DOSE, family=quasipoisson(), data=bivar.re)
```

The three models are combined using the function `mmm()` of `multcomp` to estimate the correlation of the parameters of interest via Pipper et al. (2012).

```
> library(multcomp)
> mcpPOI <- glht(mmm(alive = m1, malf = m2, dead = m3),
+                 mlf(mcp(DOSE = "Dunnett")))
> summary(mcpPOI)

         Simultaneous Tests for General Linear Hypotheses

Linear Hypotheses:
                       Estimate Std. Error z value Pr(>|z|)
alive: 62.5 - 0 == 0   -0.18580    0.08666  -2.144   0.2204
alive: 125 - 0 == 0    -0.18336    0.08003  -2.291   0.1588
alive: 250 - 0 == 0    -0.39735    0.08582  -4.630   <0.001 ***
alive: 500 - 0 == 0    -0.71905    0.09613  -7.480   <0.001 ***
malf: 62.5 - 0 == 0   -14.95121 1387.31605  -0.011   1.0000
malf: 125 - 0 == 0      1.25276    1.34181   0.934   0.9567
malf: 250 - 0 == 0      3.44999    1.21742   2.834   0.0386 *
malf: 500 - 0 == 0      5.31477    1.20296   4.418   <0.001 ***
dead: 62.5 - 0 == 0     0.27193    0.77585   0.350   1.0000
dead: 125 - 0 == 0      1.34238    0.62572   2.145   0.2195
dead: 250 - 0 == 0      1.44050    0.62256   2.314   0.1507
dead: 500 - 0 == 0      3.30776    0.57429   5.760   <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

In `mcpPOI`, the expected difference in log counts is estimated on a Dunnett-type contrast for each category. This means that the expected log count for each treatment group to control is calculated. As it is of interest whether the increase between the treatments in all categories is consistent, the parameters just calculated are once again put into a set of hypotheses. Accordingly, the matrix of linear functions is defined individually. Because these comparisons are no longer based on a model, the estimated model parameters and the corresponding covariance matrix must be passed to `glht()` via the `parm()` function. Note that the coefficients involved in the intermediate step of comparisons are different from those in the multinomial model.

```
> K <- matrix(c(#malf/alive
+               -1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
+               0, -1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
+               0, 0, -1, 0, 0, 0, 1, 0, 0, 0, 0, 0,
+               0, 0, 0, -1, 0, 0, 0, 1, 0, 0, 0, 0,
+               #dead/alive
+               -1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
```

```
+                0, -1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
+                0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
+                0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 1),
+                ncol = 12, byrow = TRUE)
> rownames(K) <- c("malf/alive: 62.5 - 0", "malf/alive: 125 - 0",
+                  "malf/alive: 250 - 0", "malf/alive: 500 - 0",
+                  "dead/alive: 62.5 - 0", "dead/alive: 125 - 0",
+                  "dead/alive: 250 - 0", "dead/alive: 500 - 0")
> summary(glht(model = parm(coef(mcpPOI), vcov(mcpPOI)),
+              linfct = K))

         Simultaneous Tests for General Linear Hypotheses

Linear Hypotheses:
                           Estimate Std. Error z value Pr(>|z|)
malf/alive: 62.5 - 0 == 0  -14.7654  1387.3267  -0.011  1.00000
malf/alive: 125 - 0 == 0     1.4361     1.3601   1.056  0.74408
malf/alive: 250 - 0 == 0     3.8473     1.2404   3.102  0.00994 **
malf/alive: 500 - 0 == 0     6.0338     1.2212   4.941  < 0.001 ***
dead/alive: 62.5 - 0 == 0    0.4577     0.8037   0.570  0.97333
dead/alive: 125 - 0 == 0     1.5257     0.6716   2.272  0.09932 .
dead/alive: 250 - 0 == 0     1.8379     0.6590   2.789  0.02570 *
dead/alive: 500 - 0 == 0     4.0268     0.6158   6.540  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

A comparison with the results from Table 5.3 shows that the final estimates are exactly the same, except the comparison of "malf/alive: 62.5 - 0" which is due to the low cell count in "malformed" of Dose 62.5. Still, the standard errors of the two analyses slightly vary because different assumptions for the variance have been considered. The analysis working with multiple marginal models uses separate dispersion estimates for each model, while the analysis in Table 5.3 incorporates an overall dispersion parameter for the whole dataset.

Figure 6.1 shows the estimated correlation matrix ($\widehat{\boldsymbol{R}}$) of the coefficients of the comparisons to control (the latter estimates). The correlation matrix is estimated in several steps in this type of evaluation via multiple Poisson models: First, the `mmm()` function estimates an empirical variance-covariance matrix of the stacked parameters according to Pipper et al. (2012). Next, `glht()` computes its variance-covariance matrix on the linear combination according to Dunnett and the just estimated covariance matrix of the coefficients of `mmm()`. Then again this covariance matrix is processed to calculate a variance-covariance matrix on the linear combination of estimated coefficients of the comparisons to control within a category as defined in K. Finally, the latter matrix is standardized resulting in the correlation matrix $\widehat{\boldsymbol{R}}$. This correlation matrix is used for simultaneous inference, calculating z-statistics and associated p-values. Parameters referring to the control group for the same log odds are positively correlated. That is within the set of hypotheses of "malf/alive" and within "dead/alive", whereas "malf/alive" shows a higher correlation than "dead/alive".

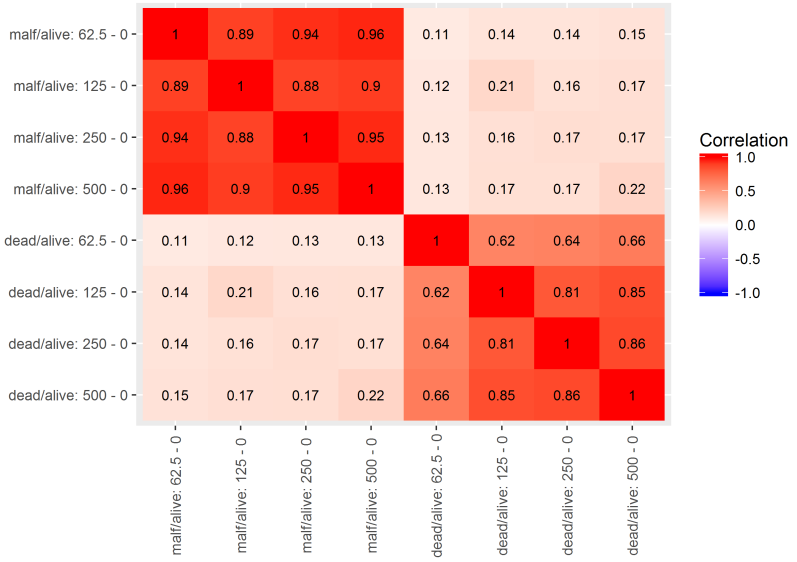

Figure 6.1: Estimated correlation matrix ($\widehat{R}$) for the Poisson analysis. Respective correlation matrix for the simultaneous analysis of Poisson distributed variables in the developmental toxicity example. Correlations are shown, with red indicating a positive correlation and blue negative correlation.

By way of comparison, Figure 6.2 displays the estimated correlation matrix used for the simultaneous inference of the coefficients of the multinomial model of Table 5.3. Since the correlation matrix is not estimated via `mmm()` but based directly on the model parameters of the multinomial model, of which the category "malformed" of group 62.5 has an estimator with a very high standard deviation, all correlations involving comparisons

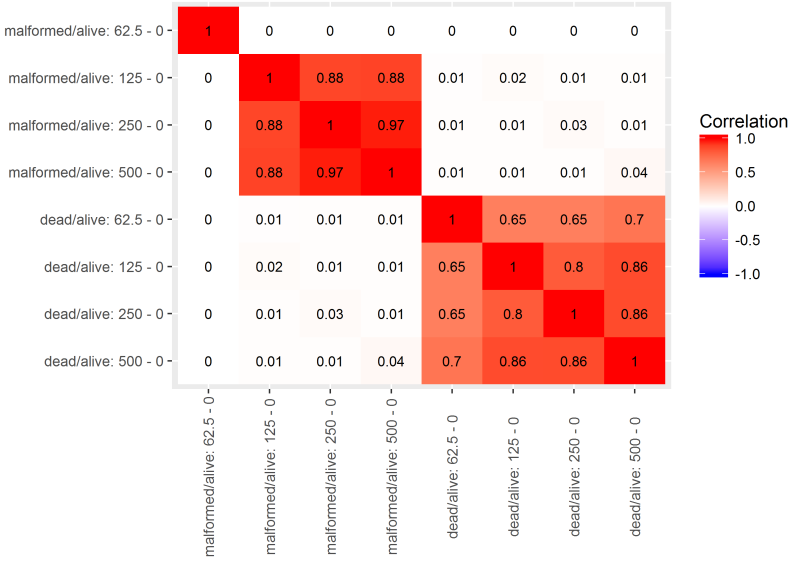

Figure 6.2: Estimated correlation matrix ($\widehat{R}$) for analysis under multinomial assumption. Respective correlation matrix for the simultaneous analysis of multinomial responses in the developmental toxicity example. Correlations are shown, with red indicating a positive correlation and blue negative correlation.

to that estimate are "0". The correlations within "dead/alive" are similar to those in Figure 6.1. The off-diagonal blocks show little correlation. More conclusions on comparing these different analysis methods are summarized in Section 6.2.6.

6.2.5 Evaluation of Multivariate-Binomial Data Using Multiple Marginal Models

Alternative approaches to evaluate multinomial data with C mutually exclusive categories as multivariate binomial exist in two ways. On the one hand, binary logit models can be separately fitted for pairs of response categories using only observations from two response categories, one of which is the baseline category (Agresti, 2003). For instance, the baseline-category logits in case of three mutually exclusive categories can be defined as

$$\log \frac{\pi_2}{\pi_1}, \quad \log \frac{\pi_3}{\pi_1}.$$

These logits can be assessed in univariate models and evaluated simultaneously by means of MMM, so that correlation is taken into account. However, estimates from separate logistic models containing only two categories out of C will be different to those from a simultaneous model and also tend to have larger standard errors, although the efficiency loss when choosing the most frequent category as a reference category is small (Agresti, 2003, p. 273).

On the other hand, separate binary logit models can be fitted with one category versus the sum of the remaining categories, which can also be assessed simultaneously using MMM. For example, the logits modelled in case of three categories are defined as

$$\log \frac{\pi_1}{\pi_2 + \pi_3}, \quad \log \frac{\pi_2}{\pi_1 + \pi_3}, \quad \log \frac{\pi_3}{\pi_1 + \pi_2}.$$

That means that each marginal model consists of a response variable which uses all C response categories. This approach appears to be similar to cumulative logits, except that one category c forms one outcome and all other categories $(1, \ldots, c-1, c+1, \ldots, C)$ form the second outcome. The first and last so defined logit is the same as in the model for cumulative logits regardless of the sign. We will use logits as just defined and evaluate them by using MMM in the next paragraph.

The data of Example 1 of Section 2.2 is used to demonstrate the aforementioned procedure of separate binary logit models as in the last definition. Initially, the dependent multinomial variable is split into multiple binomial responses to get the data format as needed. By always maintaining a reference category and combining the other categories to one, the sample size in clusters remains the same. Thus, the probability for the reference category stays the same and only the probabilities of the remaining categories add up.

```
> bivar.re$maldead <- bivar.re$malformed + bivar.re$dead
> bivar.re$alivedead <- bivar.re$alive + bivar.re$dead
> bivar.re$alivemal <- bivar.re$alive + bivar.re$malformed
> head(bivar.re)
   DAM_ID DOSE alive malformed dead maldead alivedead alivemal
1      51    0    10         0    0       0        10       10
8      60    0    14         0    0       0        14       14
9      61    0    11         0    1       1        12       11
15     70    0    17         0    0       0        17       17
16     71    0    15         0    0       0        15       15
23     79    0    17         0    0       0        17       17
```

Now we assume that the random vectors of (alive, maldead), (malformed, alivedead) and (dead, alivemal) arise from a binomial experiment. Accordingly, three univariate GLM with quasi-binomial error distribution are fitted, each having its own dispersion parameter. Afterwards, all marginal models can be assessed simultaneously by `glht()` via calling `mmm()`.

```
> m1 <- glm(cbind(alive, maldead) ~ DOSE,
+           family=quasibinomial(), data=bivar.re)
> m2 <- glm(cbind(malformed, alivedead) ~ DOSE,
+           family=quasibinomial(), data=bivar.re)
> m3 <- glm(cbind(dead, alivemal) ~ DOSE,
+           family=quasibinomial(), data=bivar.re)
> library(multcomp)
> mcp <- glht(mmm(MD.alive = m1, AD.malf = m2, AM.dead = m3),
+             mlf(mcp(DOSE = "Dunnett")))
> summary(mcp)

         Simultaneous Tests for General Linear Hypotheses

Linear Hypotheses:
                          Estimate Std. Error z value Pr(>|z|)
MD.alive: 62.5 - 0 == 0   -0.2346     0.8338  -0.281   0.9998
MD.alive: 125 - 0 == 0    -1.5084     0.6427  -2.347   0.0960 .
MD.alive: 250 - 0 == 0    -2.6672     0.6104  -4.369    <0.01 ***
MD.alive: 500 - 0 == 0    -4.8546     0.5951  -8.157    <0.01 ***
AD.malf: 62.5 - 0 == 0   -15.2516  1478.4375  -0.010   1.0000
AD.malf: 125 - 0 == 0      1.3433     1.1361   1.182   0.6859
AD.malf: 250 - 0 == 0      3.7139     1.0327   3.596    <0.01 **
AD.malf: 500 - 0 == 0      5.1221     1.0193   5.025    <0.01 ***
AM.dead: 62.5 - 0 == 0     0.4647     0.7344   0.633   0.9667
AM.dead: 125 - 0 == 0      1.5039     0.5952   2.527   0.0619 .
AM.dead: 250 - 0 == 0      1.5630     0.5924   2.639   0.0462 *
AM.dead: 500 - 0 == 0      2.6733     0.5449   4.906    <0.01 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Certainly, this is a completely different analysis and can not be compared with the previous evaluations. The data has been restructured and three binary logistic regression models have been undertaken. The parameters are therefore others than before and the interpretation varies. The output of `summary(mcp)` displays the coefficients and

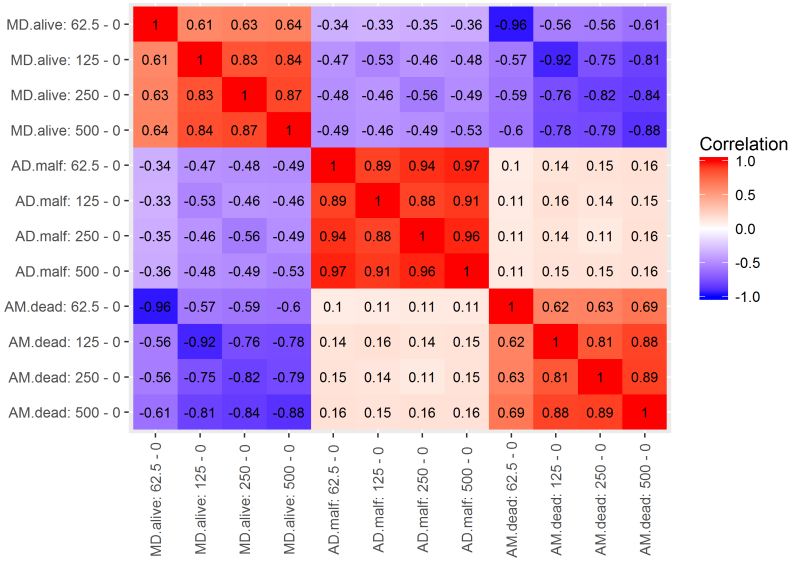

Figure 6.3: Estimated correlation matrix ($\widehat{\boldsymbol{R}}$) for the analysis of multivariate binary responses. Respective correlation matrix for the simultaneous analysis of binary response variables in the developmental toxicity example. Correlations are shown, with red indicating a positive correlation and blue negative correlation.

their standard errors of all logistic regression models together with the simultaneous z-statistics and the associated p-values. The logistic regression coefficients give the change in the log odds ratios of the outcome if the subject moves from the control group to another dose group.

The underlying estimated correlation matrix ($\widehat{\boldsymbol{R}}$) used for simultaneous inference is shown in Figure 6.3. The log odds ratios of being alive versus being malformed or dead are negatively correlated to the log odds ratios of being malformed versus being alive or dead and to the log odds ratios of being dead versus being alive or malformed. This is due to the fact that as the survival rate decreases, the rates of malformed and dead increase.

6.2.6 General Considerations

In general, the model choice has to be considered in order to show differences in the presented alternative approaches. Depending on the structure and associated assumptions, the models provide different model parameters, i.e. estimates, standard errors and variance-covariance matrix, which are used for simultaneous inference.

In comparison, estimates are nearly the same in case of the multinomial model and the multinomial marginal Poisson-models (MMM-Poisson). That is estimates are log odds ratios of one category compared to baseline between groups in the multinomial case and correspondingly ratios of expected counts compared to reference in the Poisson case. In the evaluation of multivariate binomial data with MMM, the estimates of interest are the differences of log odds ratios of an observational unit having a characteristic versus

not having the characteristic, which is the union of all complementary events (e.g. log odds ratios of being dead versus being alive or malformed).
Therefore, standard errors are estimated on different variance assumptions according to the model theory. In the multinomial model, one variance estimation with one dispersion parameter takes place, whereas in the approach of MMM several univariate Poisson models are used, all having their own category-specific dispersion parameter. The estimation of the variance-covariance matrix is also different in each approach. In case that the distribution is assumed to be multinomial, the variance-covariance matrix is derived on the basis of this assumption. In the approach of MMM, the variance-covariance matrix is empirically estimated. Together, this results in different test statistics with different correlations. In addition, the procedures differ in their reference distribution. The test statistics of the multinomial model parameters are compared with quantiles of a k-variate t-distribution. When evaluating the test statistics of MMM the reference distribution is multivariate normal in `R` since no corresponding degrees of freedom were specified. It can be expected that by specifying degrees of freedom, the approach of MMM can be improved.
Overall, we recommend using the multinomial method, if the multinomial assumption is ensured by the sampling process. That is that the total number of counts per cluster has been fixed by design of the experiment, as for example in highly controlled studies. If this number is rather a random variable, it is currently not clear which approach is preferable: MMM-Poisson accounts for random cluster sizes, but it is obscure whether and how this affects the precision of simultaneous confidence intervals or the type-I-error. In general, both the multinomial method and the approach of MMM are asymptotic. In case of small sample sizes, the former may provide more robust results. As opposed to this MMM could lead to better results if the distributional assumption is wrong because the variance-covariance matrix is estimated empirically. Of course, this guesswork should be verified in detailed simulation studies.

Chapter 7

Discussion

This thesis underlines that adjusting for overdispersion and multiple comparisons is important in statistical inference of multinomial clustered data. A novel method was proposed to adjust either p-values or confidence intervals appropriately. For multiple comparisons of odds ratios between multiple multinomial clustered response probabilities, this method allows for overdispersion in sense that a dispersion parameter is incorporated in the test distribution. Corresponding `R`-code of the procedure is available the Appendix B.1 and also ready for use as a source file from the website of the Institute (see details in Chapter 5 and Appendix B.2). It enables the R-package `multcomp` for multiple comparisons based on multinomial models fitted by the `VGAM`-package while accounting for overdispersion.

The detailed simulations indicate that the proposed method provides control of the familywise error rate in a strong sense. The familywise error rate is retained in case of high event counts for various magnitudes of overdispersion, i.e. a degree of overdispersion of $\sigma^2 = 1.01$ up to $\sigma^2 = 8$. Simultaneous confidence intervals were computed for multiple comparisons to control and all pairwise comparisons and their coverage probability was evaluated. It is shown that the coverage probability is close to the nominal level when the minimal expected event count is at least 35 at moderate overdispersion ($\sigma^2 = 2$). The higher the amount of overdispersion, the more clusters and higher sample sizes in clusters are needed to achieve a nominal coverage. If the minimum expected event count is small, simultaneous coverage is higher than the nominal level. Power simulations demonstrate that sufficient power is reached even in extreme expectations of probabilities. A power of at least 80% can be achieved even at substantial overdispersion for sparse categories. The power decreases with fewer clusters and fewer sample size in the clusters. Due to the small sample performance, we recommend the use of this methodology in trials with a sufficient number of observations or highly controlled biological experiments.

The simulations are restricted to settings with equal cluster sizes and equal number of clusters. Therefore, the simulations do not cover applications in which cluster sizes and

number of clusters can be expected to be very different like in observational studies. Yet, this method is versatile and may be considered in many other areas not addressed in this thesis.

All simulations, graphics and data analyses in this thesis were programmed in R (R Core Team, 2015). The use of implemented R functions was explained using three examples from different research areas. These functions provide simultaneous tests and confidence intervals for standard multiple comparisons such as multiple comparisons to control and all pairwise comparisons as well as user-defined contrasts for multiple odds ratios.

It is a limitation that our approach is based on asymptotic results. Nevertheless, it may also be used for small event counts, if one can accept that the coverage probability will usually be too high, but may rarely fall below the nominal level. Westfall and Wolfinger (1997) investigate an exact method that shows a close approximation to the nominal level for multinomial data without overdispersion. According to present knowledge no exact method exists in case of overdispersion.

Attention should be paid to the fact that dealing with groups mainly containing zeros need to be improved in our method. A very low probability of events in a group or category affects model fitting (Agresti, 2013). When counts of at least one category become zero for all clusters within one group, very extreme estimates occur and standard errors become very large. If such events occur frequently, the actual coverage probability is much larger than the nominal confidence level (see Figure A.1 in Appendix A.1). It can be assumed that with more than three categories, the proportion of at least some categories will continue to get low. Thereby, the definition of many categories also affects the minimal expected event count, which tends to smaller values in that case. These circumstances may lead to more rare categories and sparse data. A way of dealing with sparse data tables could be based on approaches that have already been proposed for two-way contingency tables. Anscombe (1956) suggests adding 0.5 or some other certain small constant to all cells if any cell is zero. Plackett (1962), on the other hand, recommends replacing the zero cell entries by 0.5 only for the affected contrasts. Fienberg (1969) discusses replacement values depending on the row and column of the cell entry. Some of these suggestions for improvement are addressed by Schaarschmidt et al. (2017) regarding multiple comparisons of odds ratios between multiple multinomial samples without overdispersion. In addition, they propose a MCMC technique based on sampling from Dirichlet posterior distributions with vague Dirichlet priors as a small sample approach. An analogue adaptation for the case with overdispersion remains to be investigated. In summary, various possibilities of continuity correction may lead to more adequate intervals and provide scope for further improvement.

There is little in the professional literature on accounting for overdispersion in multinomial data. References such as Bilder and Loughin (2014), Agresti (2013), Crawley

(2013), Tutz (2011) and Venables and Ripley (2002) comprehensively explain the handling of overdispersion in a binomial or Poison model. Up to now a naive approach to the analysis of clustered multinomial data is the aggregation of clusters into a single observational vector per group. The result is a $g \times c$ contingency table which is often evaluated using Pearson's test statistic (Agresti, 2013). This analysis tests the null hypothesis that the probabilities of the categories are consistent across groups and no statistically significant relationship between the categories and treatment groups exist. For a more detailed interpretation, Schaarschmidt et al. (2017) analyse the example of developmental toxicity in terms of group differences between odds ratios but also assume the absence of clusters, i.e. $\widetilde{\sigma}^2 = 1$. We strongly advise against collapsing over individual clusters and thereby ignoring possibly present overdispersion. Statistical inference will lead to erroneous decisions by underestimating the variability of the data.

The first to consider overdisersion in a multinomial model by inflation of the variance-covariance matrix are McCullagh and Nelder (1989). Studying vector generalized linear and additive models, Yee (2015) generalizes the estimation of a dispersion parameter by full maximum likelihood. Nevertheless, adjustment of statistical inference for multiple comparisons is missing. Obviously, resulting p-values and confidence intervals can be adjusted by the Bonferroni-method. However, this is known for being conservative with respect to familywise error rate control (Bretz et al., 2010). Less conservative methods to adjust against committing a type-I-error like the Bonferroni-Holm-method may be considered, but do not provide simultaneous confidence intervals with a straightforward interpretation (Strassburger and Bretz, 2008).

The alternative approach by using multiple marginal models (MMM) as proposed by Pipper et al. (2012) offers a flexible option for analysing several quasi-binomial or quasi-Poisson models at once by taking their correlation into account. By modelling univariate Poisson-models with dispersion parameter on each category, the example of developmental toxicity can be evaluated while assuming random cluster sizes. It is important to emphasize the different assumptions of this procedure and the difference in model parameters. Thus, for the developmental toxicity example estimates are identical to the multinomial analysis after defining an appropriate contrast matrix, but standard errors are higher than in the multinomial model. The dispersion parameter was taken to be 2.358 in the multinomial model overall, while for the individual Poisson models it is estimated at $\widetilde{\sigma}^2_{alive} = 0.899$, $\widetilde{\sigma}^2_{malf} = 2.881$ and $\widetilde{\sigma}^2_{dead} = 2.549$. Also this approach is asymptotic and even has an additional asymptotic element in comparison to the multinomial approach since the variance-covariance matrix is estimated on the observed information matrix. The control of the type-I-error in case of an analysis with MMM has not been studied separately in this thesis. As Pipper et al. (2012) admits that the performance in case of small sample sizes may break down it should be further investigated. Furthermore, a direct comparison of the two methods, i.e. our proposed method

and MMM, might be helpful in making a recommendation on the practical application. We have seen that the example of developmental toxicity can be analysed with both the multinomial method and MMM. Nevertheless, no statement can be made as to which method is more suitable. Additional simulations on the suitability of one or the other method and their robustness when assumptions are violated are recommended.

In this thesis only multinomial responses, i.e. assuming nominal categories, were considered. The proposed method can also be applied to ordered categorical response categories, but no benefit is taken from the additional information of order. Separate models for ordinal scales exist and are based on cumulative response probabilities rather than categorical probabilities. The ordered logit model only applies if the cumulative odds ratio is constant across categories. Morel and Neerchal (2012) discuss how to test this assumption in a multinomial model under overdispersion.

In the simulations of this thesis, we have assumed homogeneous dispersion across groups as estimated in one overdispersion parameter overall. As noted in Chapter 6, there is the possibility to develop group-specific overdispersion parameters and incorporate them individually into the variance-covariance matrix. If, for instance in the example of differential blood count in rats, the overdispersion is calculated individually for each group, one obtains $\widetilde{\sigma}^2 = (1.36, 2.52, 1.99, 2.92)$ as the vector of overdispersion parameters with single estimators in the order of control group, low dose, mid dose and high dose. This suggests heterogeneous variances in treatment groups, which may change statistical inference. Although estimates of group-specific dispersion parameters are ready for use, a multiple test procedure needs to be investigated. A joint multivariate t-distribution as in our proposed method is not accessible in that case. One can still use the multivariate normal distribution for approximation. Otherwise specifying a multivariate t-distribution with dispersion-specific degrees of freedom may improve the performance (Hasler and Hothorn, 2008). It would be of interest if this approach still controls for the type-I-error rate. One might also consider heterogeneous variances across categories in further investigations.

The focus of this thesis was on many-to-one and all-pairwise comparisons. As indicated in the last example, user-defined contrasts are feasible as well as other multiple contrasts, e.g. comparisons to the grand mean. The proposed method may also be extended to overdispersed multinomial regression models involving covariates. It is possible to include interaction terms in the model but the definition of a suitable contrast matrix for multiple comparisons might be very complex. Similar to the paper of Schaarschmidt et al. (2017), this is rather a computational effort than a methodological problem.

Appendix A

A.1 Simulation Results Dependent on the Probability of Zeros in Groups

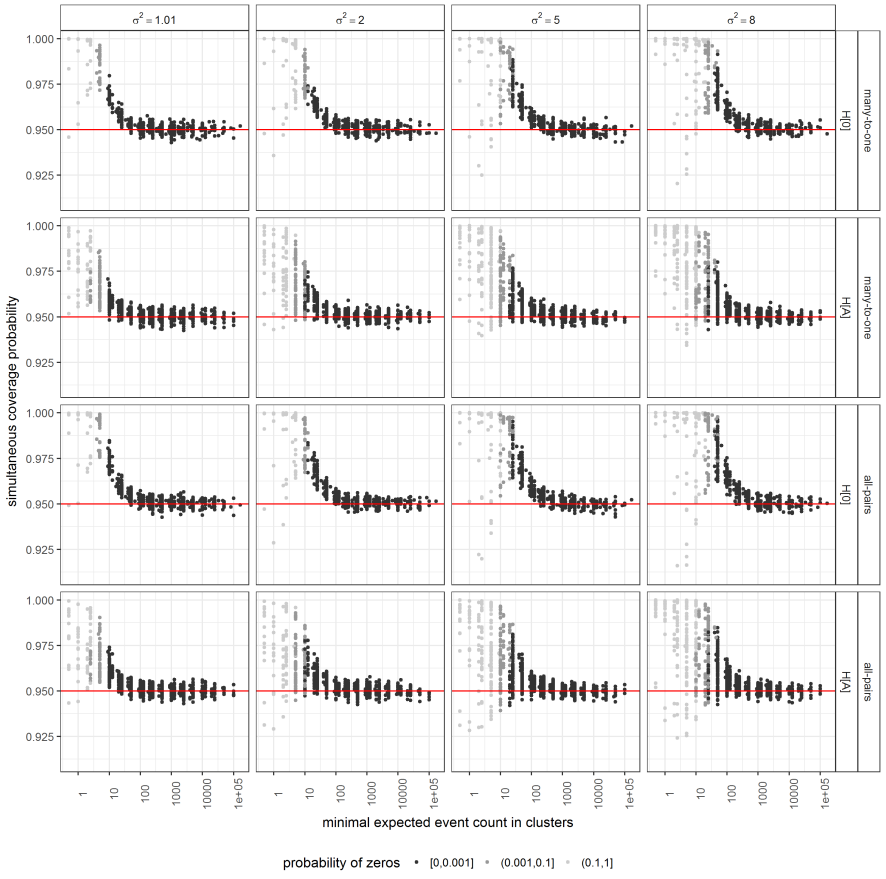

Figure A.1: Simultaneous coverage probability for multiple comparisons to a control and all pairwise comparisons. Estimated coverage probability for two-sided simultaneous confidence intervals depending on the minimum expected event count. First, for 4 groups comparing to control in the top two rows (many-to-one), and second, for all pairwise comparisons with 4 groups in the bottom two rows (all-pairs). Scenarios are separated line by line according to settings in which all logits of interest are equal in all treatments groups (H[0]) and settings in which at least one logit is different between treatments groups (H[A]). Grayscale is used to distinguish between the probabilities that at least one group contains only zeros, which may be in intervals of $[0, 0.001]$, $(0.001, 0.1]$ or $(0.1, 1]$. Each point is evaluated by 10,000 simulation runs. A nominal level of 0.95 coverage is represented by the horizontal red line.

A.2 Parameter Settings Used for Multinomial Proportions in Simulations

Table A.1: All parameters settings used for simulation. Parameters settings for π_{gc} used in simulations of $C = 3$ categories and $G = 4$ groups.

	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$	$c=1$	$c=2$	$c=3$
$g=1$	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10
$g=2$	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10
$g=3$	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10
$g=4$	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10
$g=1$	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05	0.90	0.08	0.02	0.90	0.09	0.01	0.20	0.30	0.50
$g=2$	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05	0.90	0.08	0.02	0.90	0.09	0.01	0.20	0.30	0.50
$g=3$	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05	0.90	0.08	0.02	0.90	0.09	0.01	0.20	0.30	0.50
$g=4$	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05	0.90	0.08	0.02	0.90	0.09	0.01	0.20	0.30	0.50
$g=1$	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80
$g=2$	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80
$g=3$	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80
$g=4$	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80
$g=1$	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.09	0.90	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10
$g=2$	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.09	0.90	0.33	0.33	0.33	0.50	0.30	0.20	0.50	0.40	0.10
$g=3$	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.09	0.90	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05
$g=4$	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.09	0.90	0.50	0.30	0.20	0.50	0.40	0.10	0.50	0.45	0.05
$g=1$	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05
$g=2$	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05
$g=3$	0.46	0.45	0.08	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01	0.80	0.15	0.05	0.90	0.08	0.02
$g=4$	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01	0.90	0.05	0.05	0.80	0.10	0.10	0.90	0.09	0.01
$g=1$	0.90	0.08	0.02	0.20	0.30	0.50	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80
$g=2$	0.90	0.09	0.01	0.20	0.30	0.50	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80
$g=3$	0.90	0.09	0.01	0.33	0.33	0.33	0.20	0.30	0.50	0.10	0.40	0.50	0.05	0.45	0.50	0.05	0.45	0.50
$g=4$	0.90	0.09	0.01	0.33	0.33	0.33	0.33	0.33	0.33	0.20	0.30	0.50	0.08	0.45	0.46	0.10	0.40	0.50
$g=1$	0.05	0.15	0.80	0.01	0.19	0.80	0.05	0.05	0.90	0.02	0.08	0.90	0.33	0.33	0.33	0.50	0.30	0.20
$g=2$	0.05	0.15	0.80	0.01	0.19	0.80	0.05	0.05	0.90	0.01	0.09	0.90	0.33	0.33	0.33	0.50	0.30	0.20
$g=3$	0.10	0.10	0.80	0.05	0.15	0.80	0.05	0.15	0.80	0.01	0.09	0.90	0.20	0.30	0.50	0.33	0.33	0.33
$g=4$	0.08	0.45	0.46	0.10	0.10	0.80	0.08	0.45	0.46	0.01	0.09	0.90	0.20	0.30	0.50	0.20	0.30	0.50
$g=1$	0.50	0.40	0.10	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01
$g=2$	0.50	0.40	0.10	0.50	0.40	0.10	0.46	0.45	0.08	0.80	0.10	0.10	0.80	0.15	0.05	0.80	0.19	0.01
$g=3$	0.50	0.30	0.20	0.50	0.45	0.05	0.50	0.45	0.05	0.46	0.45	0.08	0.80	0.10	0.10	0.46	0.45	0.08
$g=4$	0.33	0.33	0.33	0.20	0.30	0.50	0.20	0.30	0.50	0.10	0.40	0.50	0.46	0.45	0.08	0.05	0.45	0.50
$g=1$	0.90	0.05	0.05	0.90	0.08	0.02	0.20	0.30	0.50	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46
$g=2$	0.90	0.05	0.05	0.90	0.05	0.05	0.20	0.30	0.50	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46
$g=3$	0.80	0.19	0.01	0.80	0.19	0.01	0.10	0.40	0.50	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80
$g=4$	0.20	0.30	0.50	0.80	0.15	0.05	0.05	0.45	0.50	0.08	0.45	0.46	0.10	0.10	0.80	0.05	0.15	0.80
$g=1$	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80	0.05	0.05	0.90	0.02	0.08	0.90			
$g=2$	0.10	0.10	0.80	0.05	0.15	0.80	0.01	0.19	0.80	0.05	0.05	0.90	0.05	0.05	0.90			
$g=3$	0.05	0.15	0.80	0.01	0.19	0.80	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.19	0.80			
$g=4$	0.01	0.19	0.80	0.05	0.05	0.90	0.02	0.08	0.90	0.01	0.09	0.90	0.05	0.15	0.80			

Appendix B

B.1 Implementation in R

B.1.1 Updated multcomp Functions for vglm-Objects

Some parts of the R code is taken from the R package multcomp (Hothorn et al., 2008) to customize it for the use with objects of class "vglm" of the R package VGAM (Yee, 2015).

```
model.frame.vglm <- function(model){
  model.frame <- model.framevlm(model)
}

modelparm.vglm <- function(model, ...){
  df <- (dim(model@y)[2]-1)*model@misc$n - length(coef(model))
  multcomp:::modelparm.default(model, coef. = coef, vcov. = vcov, df = df)
}

mcp2matrix.vglm <- function(model, linfct){
  fc <- multcomp:::factor_contrasts(model)
  contrasts <- fc$contrasts
  factors <- fc$factors
  intercept <- fc$intercept
  mf <- fc$mf
  mm <- fc$mm
  alternative <- NULL
  if (!is.list(linfct) || is.null(names(linfct)))
    stop(sQuote("linfct"), "is not a named list")
  nhypo <- names(linfct)
  checknm <- nhypo %in% rownames(factors)
  if (!all(checknm))
    stop("Variable(s) ", sQuote(nhypo[!checknm]), " have been specified in ",
         sQuote("linfct"), " but cannot be found in ", sQuote("model"),
         "! ")
  if (any(checknm)) {
    checknm <- sapply(mf[nhypo[checknm]], is.factor)
    if (!all(checknm))
```

```
    stop("Variable(s) ", sQuote(paste(nhypo[!checknm],
                                      collapse = ", ")), " of class ",
         sQuote(paste(sapply(mf[nhypo[!checknm]], class), collapse = ", ")),
         " is/are not contained as a factor in ", sQuote("model"), ".")
}
m <- c()
ctype <- c()
for (nm in nhypo) {
  if (is.character(linfct[[nm]])) {
    Kchr <- function(kch) {
      types <- eval(formals(contrMat)$type)
      pm <- pmatch(kch, types)
      if (!is.na(pm)) {
        tmpK <- contrMat(table(mf[[nm]]), type = types[pm])
        ctype <<- c(ctype, types[pm])
      }
      else {
        tmp <- chrlinfct2matrix(kch, levels(mf[[nm]]))
        tmpK <- tmp$K
        m <<- c(m, tmp$m)
        if (is.null(alternative)) {
          alternative <<- tmp$alternative
        }
        else {
          if (tmp$alternative != alternative)
            stop("mix of alternatives currently not implemented")
        }
      }
      if (is.null(rownames(tmpK)))
        rownames(tmpK) <- paste(kch, 1:nrow(tmpK),
                                sep = "_")
      if (length(nhypo) > 1)
        rownames(tmpK) <- paste(nm, rownames(tmpK),
                                sep = ": ")
      list(K = tmpK)
    }
    tmp <- lapply(linfct[[nm]], Kchr)
    linfct[[nm]] <- do.call("rbind", lapply(tmp, function(x) x$K))
  }
}
for (nm in nhypo) {
  if (is.character(contrasts[[nm]])) {
    C <- do.call(contrasts[[nm]], list(n = nlevels(mf[[nm]])))
  }
```

```
    else {
      C <- contrasts[[nm]]
    }
    if (intercept || (!intercept && nm != colnames(factors)[1])) {
      Kstar <- linfct[[nm]] %*% C
    }
    else {
      Kstar <- linfct[[nm]]
    }
    pos <- factors[nm, ] == 1
    if (sum(pos) > 1)
      warning("covariate interactions found -- ",
              "default contrast might be inappropriate")
    attr(mm,"assign") <- unlist(attr(mm,"assign"))
    newlist <- attr(mm,"vassign")
    flist <- newlist[grep(nm, names(newlist))]
    level <- lapply(flist, function(x) attr(mm,"assign") %in% x)
    hypo <- vector(mode = "list", length = length(level))
    for(dp in seq_along(level)){
      hypo[[dp]] <- list(K = Kstar, where = level[[dp]])
    }
  }
  Ktotal <- matrix(0, nrow = sum(sapply(hypo, function(x) nrow(x$K))),
                   ncol = ncol(mm))
  colnames(Ktotal) <- colnames(mm)
  count <- 1
  for (h in hypo) {
    Ktotal[count:(count + nrow(h$K) - 1), h$where] <- h$K
    count <- count + nrow(h$K)
  }
  if (!is.matrix(Ktotal))
    Ktotal <- matrix(Ktotal, nrow = 1)
  nlist <- lapply(hypo, function(x) rownames(x$K))
  refkatnr <- model@extra$use.refLevel
  ratios <- paste(rep(dimnames(model@y)[[2]][-refkatnr],
                      each=length(nlist[[1]])),
                  dimnames(model@y)[[2]][refkatnr], sep="/")
  rnames <- paste(ratios,
                  unlist(nlist), sep=": ")
  rownames(Ktotal) <- rnames
  if (is.null(ctype))
    ctype <- "User-defined"
  ctype <- paste(unique(ctype), collapse = ", ")
  attr(Ktotal, "type") <- ctype
```

```
  if (length(m) == 0)
    m <- 0
  list(K = Ktotal, m = m, alternative = alternative, type = ctype)
}
```

B.1.2 Estimating Overdispersion

Function to estimate overdispersion overall or group-specific for each individual group level.

```
overdispersion <- function(model, data=NULL, strata=NULL){
  if(is.null(model)){
    warning("object ", sQuote(model), " not found", call. = TRUE)
  }

  osd <- sum(residuals(model, type = "pearson")^2)/(model@df.residual)

  if(!is.null(strata) & is.null(data)){
    warning("data.frame with strata variable needed, ",
            "please specify a 'data' argument", call. = TRUE)
  }
  else if (!is.null(strata)) {
    group.residuals <- as.data.frame(residuals(model, type = "pearson"))
    group.residuals$group <- data[,strata]
    rij.list <- split(group.residuals, group.residuals$group)
    rij.list$all <- group.residuals
    osd.strata <- sapply(rij.list, function(x) {
      npar <- (ncol(x)-1)*length(levels(factor(x$group)))
      x["group"] <- NULL
      sum(x ^ 2) / (nrow(x) * (ncol(x)) - npar)
    })
    return(osd.strata)
  }
  else return(osd)
}
```

B.1.3 Prepare Model Parameters for use in `glht()`

Construct a covariance matrix and extract degrees of freedom depending on the definition in the dispersion argument to perform asymptotic multiple comparisons.

```
multin2mcp <- function(object, dispersion=c("none","overall","stratified"),
                       data=NULL, strata=NULL){
  disptype <- match.arg(dispersion)
  if (disptype=="stratified" && (is.null(strata) | is.null(data)))
```

```
    warning("additional arguments needed", call. = TRUE)
  if (disptype=="stratified")
    warning("results will rely on normal approximation in case of
             group-specific overdispersion", call. = TRUE)
  switch(disptype,
         none = {
           vcov <- vcov(object)
           df = (dim(object@y)[2]-1)*object@misc$n - length(coef(object))
           },
         overall = {
           vcov <- vcov.disp(object)
           df = (dim(object@y)[2]-1)*object@misc$n - length(coef(object))
           },
         stratified = {
           vcov <- vcov.disp(object, data, strata)
           if (!isSymmetric(vcov, tol = sqrt(.Machine$double.eps)))
             {
             vcov <- as.matrix(forceSymmetric(vcov))
             df = 0
             }
           }
         )
  parm.object <- parm(coef(object), vcov = vcov, df=df)
  return(parm.object)
}

vcov.disp <- function(model, data=NULL, strata=NULL){
  xvar <- model@assign[names(model@assign) != "(Intercept)"]
  nrxvar <- length(xvar)
  if(nrxvar > 1){stop("no vcov computable: currently dispersion-adjustments of
                      vcov are implemented for one independent factor only",
                      call. = TRUE)}

  osd <- overdispersion(model)
  vcov.tilde <- osd * vcov(model)

  if(!is.null(strata) & is.null(data)){
    warning("data.frame with strata column needed", call. = TRUE)
  }
  else if (!is.null(strata)) {
    osdstrat <- overdispersion(model, data, strata)
    osdstrat <- osdstrat[names(osdstrat) != "all"]
    phi <- rep(osdstrat, each=ncol(model@y) - 1)
    phi.tilde <- diag(sqrt(phi))
```

```
    vcov.tilde <- phi.tilde%*%vcov(model)%*%phi.tilde
  }
  return(vcov.tilde)
}
```

B.2 R Code for Reproducing the Analysis of the Examples

The following code allows to reproduce the three examples from this thesis. The data is described in Chapter 2 and evaluated in Chapters 5 and 6. Make sure you have the relevant packages installed, i.e. `VGAM`, `multcomp` and `Matrix`.

B.2.1 Example 1: Developmental Toxicity Data from Hothorn (2015)

```
# Methods
source("https://www.biostat.uni-hannover.de/fileadmin/institut/r-code/methods.R")

# Data
library("devtools")
install_github(repo="lahothorn/SiTuR")
data("bivar", package="SiTuR")
levels(bivar$DEFECT_TYPE) <- c("alive","malformed","dead")
bivar$DOSE <- as.factor(bivar$DOSE)

# Reformat data
bivar$alive <- ifelse(bivar$DEFECT_TYPE == "alive", 1, 0)
bivar$malformed <- ifelse(bivar$DEFECT_TYPE == "malformed", 1, 0)
bivar$dead <- ifelse(bivar$DEFECT_TYPE == "dead", 1, 0)

library(plyr)
bivar.re <- ddply(bivar, .(DAM_ID, DOSE), summarize,
                  alive=sum(alive), malformed=sum(malformed), dead=sum(dead))
bivar.re <- bivar.re[order(bivar.re$DOSE),]

# Model
multivgam <- vglm(cbind(alive,malformed,dead) ~ DOSE,
                  family=multinomial(refLevel=1), data=bivar.re)
overdispersion(multivgam)

# Multiple comparison procedure with overdispersion and simultaneous CIs
summary(glht(model = multin2mcp(multivgam, dispersion="overall"),
             linfct = mcp2matrix(multivgam, linfct = mcp(DOSE = "Dunnett"))$K))
confint(glht(model = multin2mcp(multivgam, dispersion="overall"),
             linfct = mcp2matrix(multivgam, linfct = mcp(DOSE = "Dunnett"))$K))
```

```
# Multiple comparison procedure without overdispersion
summary(glht(model = multin2mcp(multivgam, dispersion="none"),
             linfct = mcp2matrix(multivgam, linfct = mcp(DOSE = "Dunnett"))$K))

# Multiple comparison procedure with group-specific overdispersion
summary(glht(model = multin2mcp(multivgam, dispersion="stratified",
                                observed.data=bivar.re, strata="DOSE"),
             linfct = mcp2matrix(multivgam, linfct = mcp(DOSE = "Dunnett"))$K))
```

B.2.2 Example 2: Housing Data from Wilson (1989)

```
# Methods
source("https://www.biostat.uni-hannover.de/fileadmin/institut/r-code/methods.R")

# Data
data("wilson", package = "MM")
non_met <- as.data.frame(non_met)
met_area <- as.data.frame(met_area)
non_met$type <- "rural"
met_area$type <- "urban"
housing <- rbind(non_met, met_area)
housing$type <- factor(housing$type, levels = c("rural", "urban"))
colnames(housing) <- c("us", "s", "vs", "type")

# Model
multivgam <- vglm(cbind(us, s, vs) ~ type,
                  family=multinomial(refLevel=1), data=housing)
overdispersion(multivgam)

# Multiple comparison procedure with overdispersion
summary(glht(model = multin2mcp(multivgam, dispersion="overall"),
             linfct = mcp2matrix(multivgam, linfct = mcp(type = "Tukey"))$K))

# Multiple comparison procedure without overdispersion
summary(glht(model = multin2mcp(multivgam, dispersion="none"),
             linfct = mcp2matrix(multivgam, linfct = mcp(type = "Tukey"))$K))
```

B.2.3 Example 3: Differential Blood Count Data from Hothorn (2015)

```
# Methods
source("https://www.biostat.uni-hannover.de/fileadmin/institut/r-code/methods.R")

# Data
library("devtools")
install_github(repo="lahothorn/SiTuR")
```

```
data("dif", package="SiTuR")
dbb <- cbind(dif[, c(1:3)], dif[,5:10]*2)
colnames(dbb) <- c("sex", "animal", "Group", "Eos", "Baso",
                   "Stab", "Seg", "Mono", "Ly")
dbb$Group <- factor(dbb$Group,
                    levels = c("control", "low dose", "mid dose", "high dose"))
dbb$factorcomb <- dbb$sex:dbb$Group

# Model
multivgam <- vglm(cbind(Eos, Seg, Mono, Ly) ~ factorcomb,
                  family = multinomial, data = dbb)
overdispersion(multivgam)

# Define contrast matrix
I <- diag(3)
library(Matrix)
B <- matrix(c(-1,1,0,0,
              -1,0,1,0,
              -1,0,0,1), byrow=TRUE, nrow=3)
B <- as.matrix(bdiag(B,B))
K <- kronecker(B,I)
# order K according to estimates from VGAM object
Kstar <- K[do.call(order, as.data.frame(K)),]
# set first columns to zero, because model was fitted with intercept
Kstar[ ,c(1:3)] <- 0

rownames(Kstar) <- c("Eos/Ly: fem.:low - fem.:con",
                     "Eos/Ly: fem.:mid - fem.:con",
                     "Eos/Ly: fem.:high - fem.:con",
                     "Seg/Ly: fem.:low - fem.:con",
                     "Seg/Ly: fem.:mid - fem.:con",
                     "Seg/Ly: fem.:high - fem.:con",
                     "Mono/Ly: fem.:low - fem.:con",
                     "Mono/Ly: fem.:mid - fem.:con",
                     "Mono/Ly: fem.:high - fem.:con",
                     "Eos/Ly: male:low - male:con",
                     "Eos/Ly: male:mid - male:con",
                     "Eos/Ly: male:high - male:con",
                     "Seg/Ly: male:low - male:con",
                     "Seg/Ly: male:mid - male:con",
                     "Seg/Ly: male:high - male:con",
                     "Mono/Ly: male:low - male:con",
                     "Mono/Ly: male:mid - male:con",
                     "Mono/Ly: male:high - male:con")
```

```
# Multiple comparison procedure with overdispersion
summary(glht(model = multin2mcp(multivgam, dispersion="overall"),
             linfct = Kstar))
```

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471249689.

A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9780470463635.

F. J. Anscombe. On estimating binomial response relations. *Biometrika*, 43(3/4):461–464, 1956. ISSN 00063444.

C. R. Bilder and T. M. Loughin. *Analysis of Categorical Data with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2014. ISBN 9781439855676.

F. Bretz. An extension of the williams trend test to general unbalanced linear models. *Computational Statistics & Data Analysis*, 50(7):1735 – 1748, 2006. doi: https://doi.org/10.1016/j.csda.2005.02.005.

F. Bretz, A. Genz, and L. A. Hothorn. On the numerical availability of multiple comparison procedures. *Biometrical Journal*, 43(5):645–656, 2001. ISSN 1521-4036. doi: 10.1002/1521-4036(200109)43:5⟨645::AID-BIMJ645⟩3.0.CO;2-F.

F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. Taylor & Francis, 2010. ISBN 9781584885740.

S. S. Brier. Analysis of contingency tables under cluster sampling. *Biometrika*, 67(3):591–596, 1980. doi: 10.1093/biomet/67.3.591.

D. R. Cox. Some remarks on overdispersion. *Biometrika*, 70(1):269, 1983. doi: 10.1093/biomet/70.1.269.

M. J. Crawley. *The R Book*. Wiley, 2013. ISBN 9780470973929.

Y. Croissant. *mlogit: multinomial logit model*, 2013. URL https://CRAN.R-project.org/package=mlogit. R package version 0.2-4.

C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955. doi: 10.1080/01621459.1955.10501294.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer New York, 2001. ISBN 9780387951874.

S. E. Fienberg. Preliminary graphical analysis and quasi-independence for two-way contingency tables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18(2):153–168, 1969. ISSN 00359254, 14679876.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer Berlin Heidelberg, 2009. ISBN 9783642016899.

L. A. Goodman. Simultaneous confidence limits for cross-product ratios in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(1):86–102, 1964. ISSN 00359246.

M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. *Biometrical Journal*, 50(5):793–800, 2008. doi: 10.1002/bimj.200710466.

L. A. Hothorn. *SiTuR: Data files for Statistics in Toxicology using R*, 2014. R package version 1.0.

L. A. Hothorn. *Statistics in Toxicology Using R*. Chapman & Hall/CRC the R series. Taylor & Francis, 2015. ISBN 9781498701273.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008. doi: 10.1002/bimj.200810425. URL http://CRAN.R-project.org/package=multcomp. R package version 1.4-5.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Discrete multivariate distributions*. A Wiley-Interscience publication. Wiley, 1997. ISBN 0471128449.

K. J. Koehler and J. R. Wilson. Chi-square tests for comparing vectors of proportions for several cluster samples. *Communications in Statistics—Theory and Methods*, 15: 2977–2990, 1986.

A. D. Martin, K. M. Quinn, and J. H. Park. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22, 2011. URL http://www.jstatsoft.org/v42/i09/.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606.

J. G. Morel and N. K. Nagaraj. A finite mixture distribution for modelling multinomial extra variation. *Biometrika*, 80(2):363–371, 1993. doi: 10.1093/biomet/80.2.363.

J. G. Morel and N. K. Neerchal. *Overdispersion Models in SAS*. SAS Publishing, 2012. ISBN 9781607648819.

OECD. *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document)*. OECD Publishing, Paris, 2014. doi: http://dx.doi.org/10.1787/9789264085275-en.

C. B. Pipper, C. Ritz, and H. Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(2):315–326, 2012. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2011.01005.x.

R. L. Plackett. A note on interactions in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(1):162–166, 1962. ISSN 00359246.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org.

F. Schaarschmidt, D. Gerhard, and C. Vogel. Simultaneous confidence intervals for comparisons of several multinomial samples. *Computational Statistics & Data Analysis*, 106:65–76, 2017. doi: 10.1016/j.csda.2016.09.004.

K. Strassburger and F. Bretz. Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. *Statistics in Medicine*, 27 (24):4914–4927, 2008. doi: 10.1002/sim.3338.

J. W. Tukey. The problem of multiple comparisons. Unpublished, but widely circulated as a manuscript, 1953.

G. Tutz. *Regression for Categorical Data*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2011. ISBN 9781139499583. doi: 10.1017/CBO9780511842061.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 978-0-387-95457-8.

P. H. Westfall and R. D. Wolfinger. Multiple tests with discrete distributions. *The American Statistician*, 51(1):3–8, 1997. ISSN 00031305.

D. A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971.

J. R. Wilson. Chi-square tests for overdispersion with multiparameter estimates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 38(3):441–453, 1989.

T. W. Yee. *Family Functions for generalized Linear and Additive Model*, 2008. URL https://www.stat.auckland.ac.nz/~yee/VGAM/doc/glmgam.pdf.

T. W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer-Verlag New York, 2015. ISBN 978-1-4939-2817-0.

Acknowledgements

First, I would like to thank my first advisor Prof. Dr. Ludwig A. Hothorn for the continuous support of my thesis and related research. The door to Prof. Hothorn office was always open and even after retirement, he continued to help with his patience, motivation, and immense knowledge. He guided me in the right direction whenever he thought I needed it.

With the same emphasis I would like to thank PD Dr. Frank Schaarschmidt who provided insight and expertise that greatly assisted my research. Without his passionate participation and input it would not be possible to conduct this research.

I would also like to thank my friends of the Electric Energy Storage Systems Section of the Institute of Electric Power Systems (EES-IfES) for always welcoming me to their cafeteria circle and making everyday work a little bit brighter.

Equally, I am grateful to Cornelia Frömke for inspiring me in my professional life and Clemens Buczilowski for providing technical support at the Institute of Biostatistics.

And finally, I would like to thank my forever interested and enthusiastic sister for always being keen to know what I was doing, although she already admitted herself that she has never grasped what it was all about (little hint: my research did not take place in a lab).

Thanks for all your encouragement!

Curriculum Vitae

Persönliche Angaben

Name	Katharina Charlotte Vogel
Geburtstag	6. Juni 1987
Geburtsort	Magdeburg, Deutschland

Schulbildung

08/1993–07/1997	**Grundschule**, *Grundschule Lindenhof*, Magdeburg, Deutschland.
08/1997–08/1999	**Weiterführende Schule**, *Sekundarschule August-Wilhelm-Francke*, Magdeburg, Deutschland.
09/1999–03/2006	**Weiterführende Schule**, *Hegel-Gymnasium*, Magdeburg, Deutschland.
06.07.2006	**Allgemeine Hochschulreife**, *Hegel-Gymnasium*, Magdeburg, Deutschland, *2,1*.

Studium

10/2006–09/2012	**Studium der Mathematik**, *Otto-von-Guericke-Universität (OvGU)*, Magdeburg, Deutschland.
01/2010–06/2010	**Auslandssemester**, *University of Glasgow*, Glasgow, Großbritannien.
11.09.2012	**Abschluss Diplom-Mathematikerin**, *OvGU*, Magdeburg, Deutschland, *1,7*.

Promotionsstudium

03/2015–09/2018	**Promotion**, *Institut für Biostatistik, Leibniz Universität Hannover (LUH)*, Hannover, Deutschland.
26.06.2018	**Disputation**, *Institut für Biostatistik, Leibniz Universität Hannover (LUH)*, Hannover, Deutschland, *magna cum laude*.

Berufstätigkeit

08/2010–10/2010	**Statistical Analyst**, *WeeWorld Ltd., Finance and Controlling*, Glasgow, Großbritannien.
09/2012–01/2015	**Wissenschaftliche Mitarbeiterin**, *Institut für Biometrie, Medizinische Hochschule (MHH)*, Hannover, Deutschland.
03/2015–02/2018	**Wissenschaftliche Mitarbeiterin**, *Institut für Biostatistik, LUH*, Hannover, Deutschland.
seit 03/2018	**Wissenschaftliche Mitarbeiterin**, *Institut für Biometrie, Epidemiologie und Informationsverarbeitung, Stiftung Tierärztliche Hochschule Hannover (TiHo)*, Hannover, Deutschland.

Veröffentlichungen

Methodische Publikationen

Schaarschmidt, F., Gerhard, D., **Vogel, C.**, *Simultaneous confidence intervals for comparisons of several multinomial samples.* Computational Statistics and Data Analysis, 2017, 106, S. 65-76

Medizinische Publikationen

Bauer, B. U., Rapp, C., Mülling, C. K.W., Meissner, J., **Vogel, C.**, Humann-Ziehank, E., *Influence of dietary zinc on the claw and interdigital skin of sheep.* Journal of Trace Elements in Medicine and Biology, 2018, 50, S. 368-376

Karch, A., Vogelmeier, C., Welte, T., Bals, R., Kauczor, H.-U., Biederer, J., Heinrich, J., Schulz, H., Gläser, S., Holle, R., Watz, H., Korn, S., Adaskina, N., Biertz, F., **Vogel, C.**, Vestbo, J., Wouters, E.F.M., Rabe, K.F., Söhler, S., Koch, A., Jörres, R.A., *The German {COPD} cohort COSYCONET: Aims, methods and descriptive analysis of the study population at baseline.* Respiratory Medicine, 2016, 114, S. 27-37

Doose, M., Ziegenbein, M., Hoos, O., Reim, D., Stengert, W., Hoffer, N., **Vogel, C.**, Ziert, Y., Sieberer, M., *Self-selected intensity exercise as augmentation in the treatment of major depression: a pragmatic RCT.* International Journal of Psychiatry in Clinical Practice, 2015, 19(4), S. 266-275

Kreuzer, M., Prüfe, J., Bethe, D., **Vogel, C.**, Großhennig, A., Koch, A., Oldhafer, M., Dierks, M. L., Albrecht, U. V., Müther, S., Brunkhorst, Pape, L. and Study group of the German Society for Pediatric Nephrology (Gesellschaft für Pädiatrische Nephrologie, GPN), *The TRANSNephro-study examining a new transition model for post-kidney transplant adolescents and an analysis of the present health care: study protocol for a randomized controlled trial.* Trials, 2014, 15(505)

Lenzen, H., Musmann, L. E., Ernst, S., **Vogel, C.**, Schönemeier, B., Köhnlein, T., Manns, M. P., Lankisch, T. O., *Gastrointestinale Blutung in der Notaufnahme: Mann, höheres Alter, Winter und die Nacht sind ein Risiko.* Zeitschrift für Gastroenterologie, 2014, 52

Schweitzer, N., Hoffmann, M., Papendorf, F., **Vogel, C.**, Scherer, R., Manns, M.P., Vogel, A., *Prognostic Factors in Patients with Cholangiocellular Carcinoma - Comprehensive Analysis of 570 Patients.* Journal of Hepatology, 2013, 58, S. 273

Unter Review

Vogel, C., Schaarschmidt, F., Hothorn, L. A., *A multiple comparison procedure for overdispersed multinomial data.* Computational Statistics and Data Analysis

Vorträge und Poster

Vorträge

Vogel, C., Schaarschmidt, F., Hothorn, L. A., *A multiple comparison procedure for overdispersed multinomial data.* 29. International Biometric Conference, Barcelona, 10. Juli 2018.

Vogel, C., Schaarschmidt, F., Hothorn, L. A., *A multiple comparison procedure for overdispersed multinomial data.* 64. Biometrisches Kolloquium, Frankfurt, 28. März 2018.

Vogel, C., Hothorn, L. A., *Simultaneous Confidence Intervals in Subgroup Analysis - A MMM Approach* MMM Workshop, Hannover, 14. Dezember 2016.