Simultaneous Confidence Intervals for Biodiversity Indices with Application to Overdispersed Multinomial Count Data

von
Dipl. Ing. (FH) **Ralph Scherer**
aus Hannover

zur Erlangung des Grades eines

– Master of Science –

der Gartenbauwissenschaften
der Naturwissenschaftlichen Fakultät
der Leibniz Universität Hannover

vorgelegt von

Name: Scherer Vorname: Ralph

Email Ralph Scherer

Geb. am: 25.05.1976 in: Darmstadt

Erstprüfer: Prof. Dr. L. A. Hothorn
Zweitprüfer: Prof. Dr. B. Hau

Hannover, January 12, 2010

Für meine Familie

Abstract

In ecological trials with multiple treatments and many observed species, the point of interest may be comparing the species' diversity among treatments. This might be done in a proof of safety for a genetically modified organism as processed in this work or any other kind of diversity comparison. There exist several diversity indices which summarize the pattern of relative species abundances in treatment in a single number.
This work adresses statistical inference for two diversity indices, namely the Shannon and the Simpson index. Since the variances of these indices can be unequal although the indices are the same, standard multiple comparison procedures like the ANOVA are not adequate. There are asymptotic methods for multiple comparisons of diversity indices considering heterogeneous variances available in literature. These methods are constructed under the assumption of multinomial distributed counts. As this assumption may not hold true for ecological trials due to overdispersed species counts, this work addresses simultaneous bootstrap confidence intervals, which take overdispersion into account.
In this thesis, three different methods to construct simultaneous bootstrap confidence intervals are compared with the asymptotic methods in simulation studies and further applied to two real data sets. It is found, that the bootstrap methods perform well under overdispersed multinomial counts, whereas the asymptotic methods exhibit poor coverage probability.
Keywords: bootstrap, multiple comparisons, biodiversity, overdispersion, heteroscedasticity

Contents

Abbreviations

ANOVA	Analysis of variance
FWE	Familywise-Error-Rate
GMO	Genetically modified organism
iid	independent and identically distributed
isogene	identical organism without genetic transformation
IUT	Intersection-Union Testing
MCA	All-Pairwise Comparisons
MCC	Multiple Comparisons with Control
MCP	Multiple Comparison Procedures
SCIs	Simultaneous confidence intervals
STP	Simultaneous Test Procedure
UIT	Union-Intersection Testing

List of Symbols

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Introduction

In agricultural trials for genetically modified plants (GMOs) it can be of interest to prove safety for non-target species by comparing the effects of genetically modified plants on species with the effects of non-transgenic plants on species.

In such kind of agricultural trials one may compare GMO plants with several control treatments. These could be isogenic plants treated with standard pesticides, as well as untreated, isogenic plants. For proving safety of the GMOs one is interested in establishing equivalence effects of the different treatments on the non-target species, since a decrease in diversity may detect trouble in ecosystem. This could be for example a decrease of benefical organisms as well as a decrease of essential species in the associated food chain.

To detect effects on the biodiversity of habitats one may summarize the colonizing species in so-called biodiversity indices. This yields to a single number representing species richness and/ or evenness. There are several different indices available in literature for this application area, see Magurran [2004]. In this thesis I will focus on two of the most popular indices, the Shannon [Shannon and Weaver, 1949] and the Simpson [Simpson, 1949] index. Furthermore we will apply a recently modified Shannon index by Chao and Shen [2003], considering unseen species in sample.

Since there are more than two treatments to compare, applying multiple test procedures (MCPs) is necessary. Due to the heterogeneous variances of the indices, see 1.3.1 on page 10, between treatments, standard multiple test procedures like ANOVA, all-pairwise comparisons (MCA) [Tukey, 1953] and comparisons to control (MCC) [Dunnett, 1955] are not applicable [Rogers and Hsu, 2001]. These methods assume equal variances and normal distributed errors in the fixed-effects ANOVA model.

Fritsch and Hsu [1999] and Rogers and Hsu [2001] introduced variance estimators for the Simpson and the Shannon index respectively, which allow to perform MCAs or MCCs without assuming equal variances. Since these estimators depend on the assumption of multinomial distributed

species counts and are constructed for only one vector of observations in every community, they can not hold true for such data arising in the analyzed experiments, where one has several replicates per community. Due to this fact, I have to summarize the replications in one species count vector, allowing calculation of the diversity indices and the corresponding variance estimators by Rogers and Hsu [2001] and Fritsch and Hsu [1999].

This yields to an underestimation of the variance across the replicates, if these data show strong over-dispersion, see section 1.3.2 on page 13, indicating a higher variance than the multinomial distribution can describe. To handle such over-dispersed count data with heterogeneous variances between groups I use bootstrap methods to construct simultaneous confidence intervals without making any assumptions about the underlying distribution.

In the following I go more into details of the distributional assumptions for our data, the problem of variance heterogeneity and the advantages of bootstrap methods. Then I present several bootstrap confidence interval methods for the one-sample problem to analyze their performance in covering the true parameter of interest under a simulated over-dispersed multinomial distribution. Next, I show different bootstrap methods to construct simultaneous confidence intervals for the multiple sample design and study them under the same distributional assumptions. Finally I apply them to real data sets introduced in chapter 1.2.

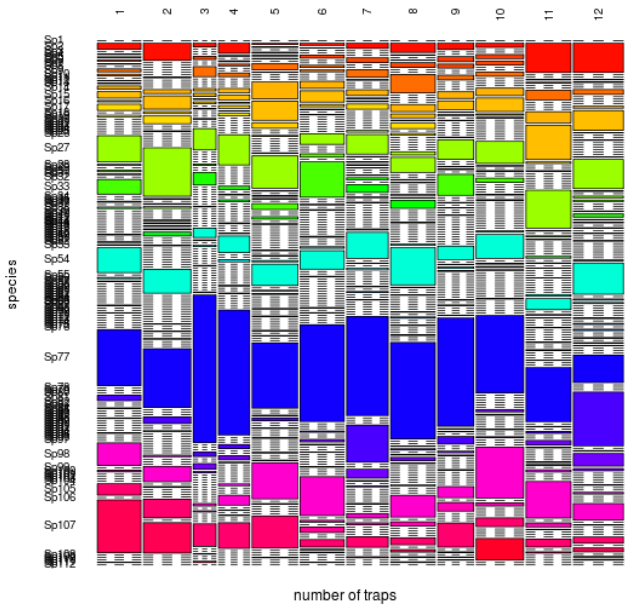
1.2 Motivating Example

In the following I will describe two example data sets based on real field trials. For both examples the statistical hypotheses are explained and further the data sets are graphically analysed using mosaicplots.

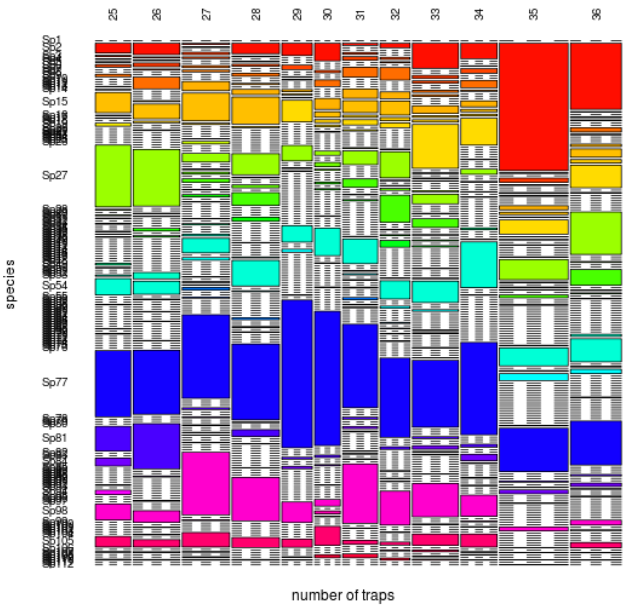
1.2.1 Example 1

The first example exhibits three different treatments to compare. These are one genetically modified line (GM), one near isogenic line (Ins) treated

with Baythroid and further one untreated, isogenic line. Per treatment 12 different traps were observed, where 112 species were counted within every trap. Number of individuals per trap range from 417 to 1148. The point of interest is proving safety of the genetically modified plants to the habitat species. This may be done by proving superiority of the GM treatment to the pesticide treatment or proving non-inferiority of the GM treatment to the untreated near isogenic plants. Since the corresponding hypotheses can not be ordered a priori one has to follow the *union-intersection principle*, see 1.4.2 on page 15, which leads to simultaneous confidence intervals. Superiority for the contrast (Pesticide - GM) means, the upper bound of the confidence interval for the difference is lower than a given superiority border $-\delta$. In contrast, non-inferiority for the contrast (Iso - GM) indicates, that the upper bound of the confidence interval for the difference is lower than a given non-inferiority border $+\delta$. Choosing the correct values for the superiority border $-\delta$ and the non-inferiority border $+\delta$ lies by the ecologist, who is familiar with the values of the biodiversity indices. The counts for species are plotted in figures from 1.1(a) on the following page to 1.2(a) on page 6.

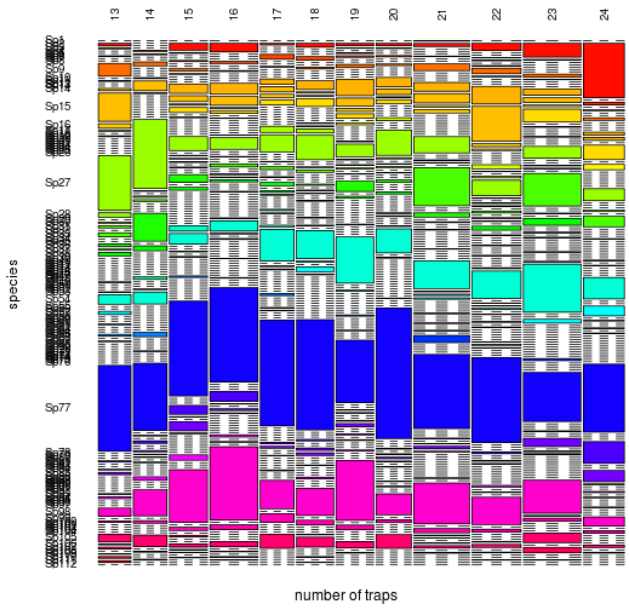


(a) GM line



(b) Ins line

Figure 1.1: Mosaicplot for species in GMO field and pesticide treated, near isogenic field – width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap



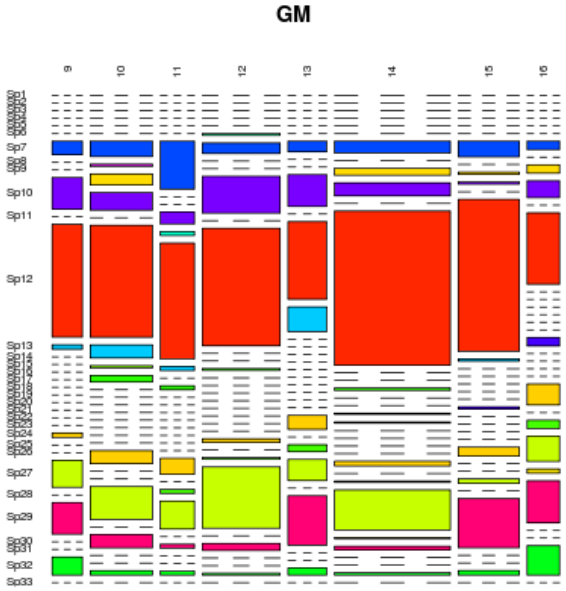
(a) Iso line

Figure 1.2: Mosaicplot for species in near isogenic, untreated field – width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap

1.2.2 Example 2

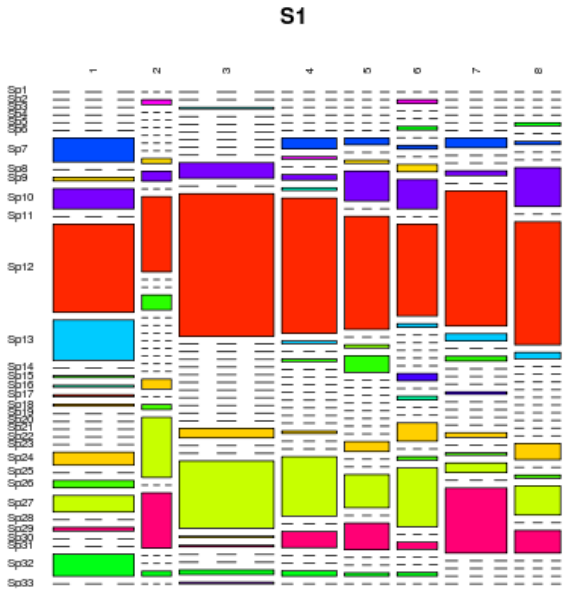
The second example is as well as the first example an ecological trial with genetically modified plants. The difference to the first example is more isogene varieties as well as less species, where only a few species are highly dominant and the rest are lowly abundant. In this trial one observes four different treatments to compare, which are presented in figures from 1.3(a) on the following page to 1.5(a) on page 9. One variety was a genetically modified line (GM), while the other three were conventional varieties (S1, S2, S3). Per variety eight plots were observed, each of them with one trap situated on. At all, 33 species were observed, while the number of individuals per trap ranges from 27 to 384.

Again, one has several possibilities to reject the null hypothesis that the GM plants are safe. The three possible alternative hypotheses are non-inferiority of GM treatment to S1 or S2 or S3.

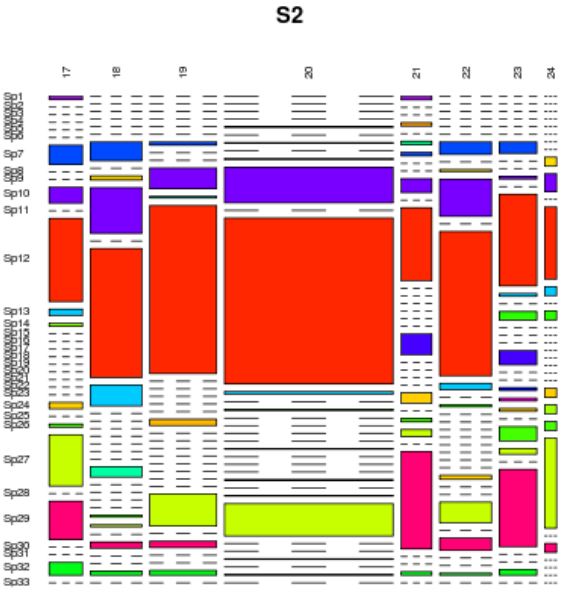


(a) GM field

Figure 1.3: Mosaicplot for species in GM field – width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap

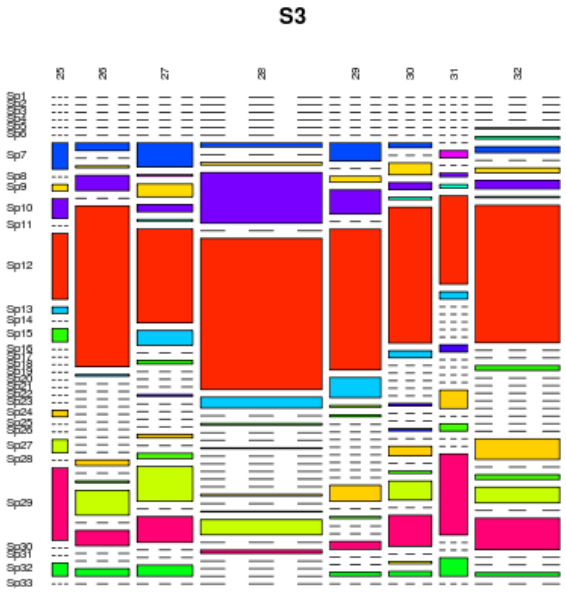


(a) S1 field



(b) S2 field

Figure 1.4: Mosaicplot for species in S1 and S2 field – width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap



(a) S3 field

Figure 1.5: Mosaicplot for species in S3 field – width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap

1.3 Difficulties in Comparing Diversity

During the statistical analysis of ecological field trials using biodiversity indices two main problems will arise. These are the heterogeneous variances of the indices between treatments as well as the distrubional assumptions including overdispersed species counts across replicates. In the next section I go more into the details of these issues.

1.3.1 Heterogeneous Variances

Let $\widehat{\theta}_{ij}$ be the estimated Shannon $\widehat{H'} = -\sum_{s=1}^{S} \hat{\pi}_s \ln \hat{\pi}_s$ or Simpson $\widehat{\varphi} = \sum_{s=1}^{S} \hat{\pi}_s^2$ diversity index in population $i,\ i = 1, ..., k$ and in trap or replicate $j,\ j = 1, ..., r$. Here $\hat{\pi}_s$ is the estimated proportion of species s. Further $\bar{\theta}_i$ is the point estimator $\bar{\theta}_i = \frac{1}{r} \sum_{j=1}^{r} \widehat{\theta}_{ij}$ for each population i.

Rogers and Hsu [2001] figured out, that the variances $\widehat{\sigma}_i^2$ of point estimators $\bar{\theta}_i$ of diversity indices may be unequal, even if all sample sizes are the same and even if all k populations have exactly the same diversity index. This depends on the calculation of the indices out of the probability vectors $\widehat{\boldsymbol{\pi}}_{\boldsymbol{i}} = \widehat{\pi}_{i1}, ..., \widehat{\pi}_{iS}$, which can be different for each community i, even if all point estimators $\bar{\theta}_i$ for the diversity indices $\widehat{\theta}_{ij}$ are the same, the estimated variances can be completely uneven.

In figure 1.6 on the following page and figure 1.7 on page 12 I present the distribution of Shannon's and Simpson's index based on a real data set. Here, the indices are calculated for every trap or replicate separately. These plots indicate variance heterogeneity as well as skew distributions.

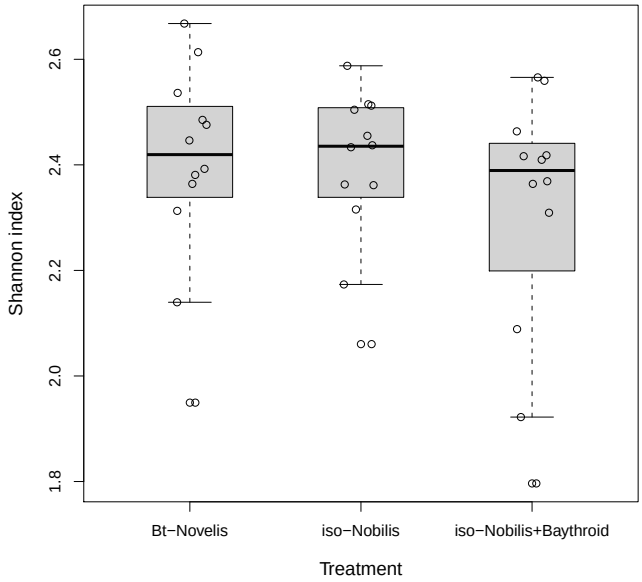
1.3.2 Distributional Assumptions

Rogers and Hsu [2001] describe the analyzed data in the following way. Let X_{ijs} be the sample count for species $s, s = 1, ..., S$ in community i and replicate j. Further, for every i one takes a sample of size $n_{ij} r_i$, where n_{ij} is the total sample size across all species for every replicates j, in other words $n_{ij} = \sum_{s=1}^{S} n_{ijs}$. If all replicates j are assumed to be from the same, evenly distributed population, one may summarize the counts for every species in one resulting count vector $\boldsymbol{X_i} = X_{i1}, ..., X_{iS}$ with corresponding vector of proportions $\boldsymbol{\pi_i} = \pi_{i1}, ..., \pi_{is}$ yielding to equation (1.3.2.1).

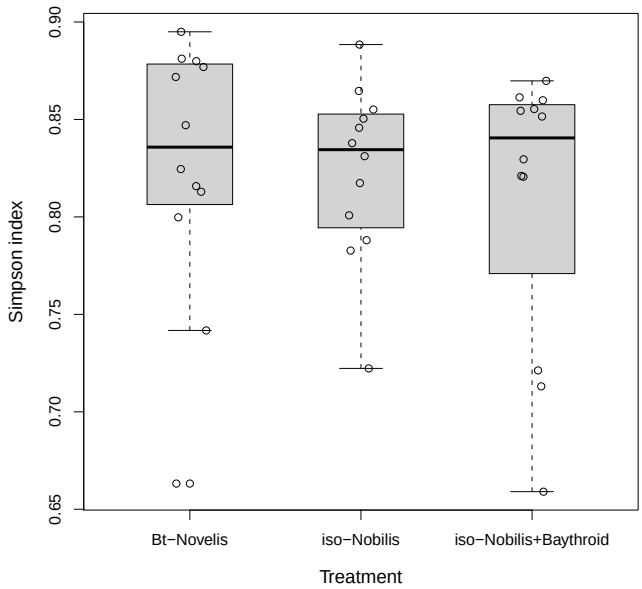
$$\boldsymbol{X_i} \sim \text{ Multinomial } (\boldsymbol{\pi_i}, n_{ij} r_i) \text{ independently for } i = 1, ..., k \qquad (1.3.2.1)$$

To take the replicates into account this leads to equation (1.3.2.2)

$$\boldsymbol{X_{ij}} \sim \text{ Multinomial } (\boldsymbol{\pi_i}, n_{ij}) \ \ iid \text{ for } j = 1, ..., r \qquad (1.3.2.2)$$

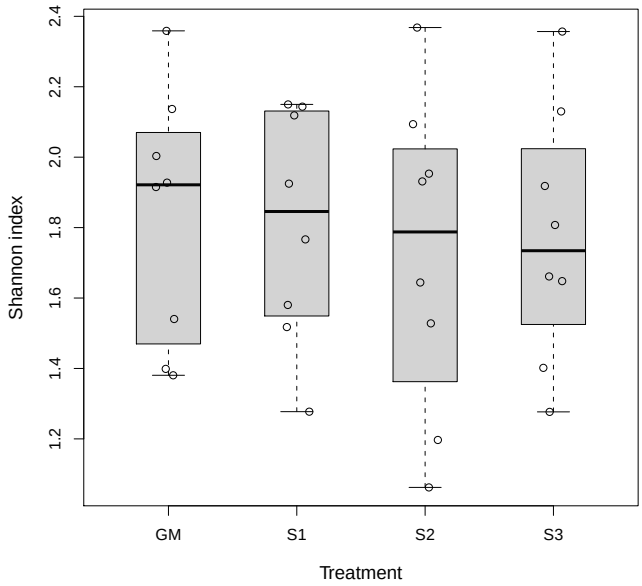


(a) Distribution of Shannon's index

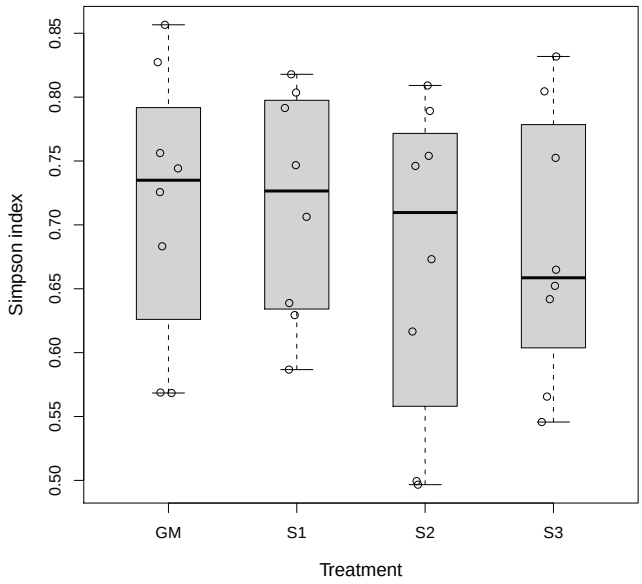


(b) Distribution of Simpson's index

Figure 1.6: Distribution of Biodiversity indices illustrated as Boxplots for the three different treatments: Bt-Novelis, Iso-Nobilis, Iso-Nobilis+Baythroid (from left to right). Indices are calculated for every trap separately



(a) Distribution of Shannon's index



(b) Distribution of Simpson's index

Figure 1.7: Distribution of Biodiversity indices illustrated as Boxplots for the four different treatments: GM, S1, S1, S3. Indices are calculated for every trap separately

From this follows, that Rogers and Hsu [2001] assume a multinomial distribution as described in section A.2 on page 76 with probability vector $\boldsymbol{\pi_i}$. The variance estimator $\widehat{\sigma}_i^2$ they use to construct simultaneous confidence intervals, see section 3.4 on page 36, depends on the estimated probability vector $\widehat{\boldsymbol{\pi}}_i$, being the same for every replicate j in community i.

Due to the assumption of multinomial distributed data, the variance estimator $\widehat{\sigma}_i^2$ according to Rogers and Hsu [2001] underestimates the variance for such kind of data arising in ecological field trials, see section 1.2. Here, one can observe a high variation of species counts across traps or replicates in one community. Figure 1.1(a) on page 5 presents species counts occurring in a real ecological field trial. Typically, species counts have a high variation across traps, as well as a high rate of zero counts.

This kind of high variation across replicates yields to an over-dispersed distribution (for over-dispersion: see section 1.3.2), meaning the observed variance exceeds the variance describable by the multinomial distribution. Therefore, I use a Dirichlet-multinomial [Mosimann, 1962] distribution to sample over-dispersed count data with different probability vectors $\boldsymbol{\pi_{ij}}$ for every replicate j. This yields to a higher variance than describable by the multinomial distribution, see appendix A.4 on page 77.

Over-Dispersion

According to McCullagh and Nelder [1989] over-dispersion occurs if the variance of the response Y exceeds the nominal variance of the postulated distribution.

In the following, let $\boldsymbol{\pi_{ij}} = \pi_{ij1}, ..., \pi_{ijS},\ s = 1, ..., S$ and $j = 1, ..., r$ be the probability vector of the multinomial distribution in community i. In case of the Dirichlet-multinomial distribution, I am using for simulation, $\boldsymbol{\pi_{ij}}$ is an outcome of a random vector $\boldsymbol{\Pi_{ij}}$, which follows a Dirichlet distribution, nearer described in section A.3 on page 77. Then, the resulting distribution is the Dirichlet-multinomial including an over-dispersion parameter. This leads to a richer class of distribution, which takes over-dispersion into account [Poortema, 1999].

1.4 Applied Statistical Methods

As explained before, there are several issues in the statistical analysis of ecological diversity trials. This leads to alternatives like the nonparametric bootstrap methods, which I introduce, among the multiple comparison requirements, in the following sections.

1.4.1 Bootstrap Methods

Due to the before described problems as heterogeneous variances and over-dispersed count data, I introduce here nonparametric bootstrap methods to construct confidence intervals without making any assumptions about the underlying distributions. The distribution for the confidence intervals is based on the observed counts, yielding to confidence intervals which take the over-dispersion into account.

The bootstrap confidence intervals described in section 3.1 on page 23 are applicable for one- and two-sample problems, while the methods in section 3.3 on page 30 are applicable for multiple comparisons using simultaneous confidence intervals. In ecological trials, as described in section 1.2, one may be interested in comparing more than two treatments with each other. The common goal could be proving superiority to one treatment or non-inferiority to another treatment. For example, superiority of the GMOs to the pesticide treatment or non-inferiority of the GMOs to the untreated plants. This yields to the comparison of one treatment group, i.e. GMO plants, against multiple control groups, i.e. pesticide and untreated plants. As the global null hypothesis is rejected if one of the multiple null hypotheses is rejected, this leads to the principle of *union-intersection tests*.

1.4.2 Multiple Comparisons

Union-Intersection Testing

The *union-intersection method* (UIT) is appropriate if the null hypothesis is expressed as an intersection.

Let θ be a population parameter yielding to the null and alternative hypotheses in 1.4.2.1

$$H_0 : \theta \in \Theta_0 \quad \text{and} \quad H_1 : \theta \in \Theta_0^c \tag{1.4.2.1}$$

with Θ_0 being some subset of the parameter space and Θ_0^c being its complement. For example, θ may denote the average change in a patient's blood pressure after taking a drug leading to the hypotheses $H_0 : \theta = 0$ and $H_1 : \theta \neq 0$.

The global null hypothesis can be written in the following form.

$$H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma, \tag{1.4.2.2}$$

where Γ is, depending on the problem, an arbitrary finite or infinite index.

Suppose, one is interested in testing the hypotheses $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$ and the rejection region for the test of $H_{0\gamma}$ is $\{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}$.

Thus the rejection region for the union-intersection test is

$$\bigcup_{\gamma \in \Gamma} \{\boldsymbol{x} : T_\gamma(\boldsymbol{x}) \in R_\gamma\}. \tag{1.4.2.3}$$

This indicates that the global H_0 is true, only if any $H_{o\gamma}$ is true for every γ. In case that one $H_{o\gamma}$ is rejected, the global H_0 is rejected as well [Casella, 2001].

In contrast to the *union-intersection testing* exists *intersection-union testing* (IUT) , which is applicable, if the global null hypothesis is only rejected if all individual null hypothesises are rejected (intersection of the alternative hypothesises), meaning IUT demands all significant level-α tests.

Controlling the FWE

On the basis of *union-intersection testing*, with more than one possibilities to reject the null hypothesis, one have to adjust for multiplicity. This yields to *Simultaneous Test Procedures* (STP's) , e.g. by Hochberg and Tamhane [1987], which control the *Familywise-Error-Rate* (FWE) by rejecting or accepting each of the γ hypotheses at a particular level α [Westfall and Young, 1993].

In section 3.4 on page 36 and 3.3 on page 30 I will introduce simultaneous confidence intervals, which control the FWE with adjusted quantiles from the multivariate normal distribution, or, similar to the method by Tukey [1953], from the maximum distribution of the bootstrapped test statistics.

Chapter 2

Measurement of Biodiversity

In the following chapter, I will describe three different indices for measuring biodiversity. These are two well known indices, the Shannon [Shannon and Weaver, 1949] and the Simpson index [Simpson, 1949], as well as a recently by Chao and Shen [2003] modified Shannon index considering unseen species in sample. This modified Shannon index is not included in the later following simulation results, since the corresponding variance estimator is not described adequately in the publication. However, it was possible to simulate this index with the method by Westfall and Young [1993], as the variance is estimated here from the ANOVA model. The resulting coverage probability showed only a slightly difference from the original Shannon index.

2.1 Diversity measures

There are several biodiversity indices which encapsulate different aspects of a samples' or a communities' diversity. However, it is not possible to combine all aspects of diversity in a single measure.

This yields to the distinction between species richness measures [McIntosh, 1967] and eveness/ dominance measures [Lloyd and Ghelardi, 1964]. Furthermore, there exist heterogeneity measures which combine species richness and eveness components.

Species richness measures are the simplest indices giving information on the number of species of a given taxon in a chosen assemblage. In contrast, eveness measures assess the departure of the observed pattern from the expected pattern and range from completely even species to several levels of uneven species [Magurran, 2004].

In the following I will present the indices applied in our analyzes. These are the Shannon [Shannon and Weaver, 1949] , the Simpson [Simpson, 1949] and the modified Shannon index by Chao and Shen [2003] considering unseen species in sample. Distinguishing them by classification into the diversity measure classes, the Shannon index belongs to the species richness measures, while the Simpson index is a dominance measure.

2.1.1 Shannon's index

The Shannon index is one of the most used indices. This is all the more remarkable, because different authors adverted to the disadvantages of this measure. The Shannon index assumes randomly sampled individuals from an infinitely large community and also that the sample includes all species [Magurran, 2004]. Calculation is done by equation (2.1.1.1). I use this index for the following methods, because of the common use in ecological area.

$$H' = -\sum_{s=1}^{S} \pi_s \ln \pi_s \tag{2.1.1.1}$$

with

π_s proportion of individuals in the sth species

In a sample the true value of π_s is unknown, therefore it is calculated by its maximum likelihood estimator $\widehat{\pi}_s = X_s/N$, where X_s is the sample count for species s and N is the total sample count across all species. A simple estimator for the Shannon H' index can be $\widehat{H}'$, where one replaces π with $\widehat{\pi}$ in equation 2.1.1.1. Since this estimator produces biased results one may calculate the index by procedure (2.1.1.2).

$$\widehat{H}' = -\sum_{s=1}^{S} \widehat{\pi}_s \ln \widehat{\pi}_s + \frac{S-1}{2N} - \frac{1-\sum_{s=1}^{S} \widehat{\pi}_s^{-1}}{12N^2} - \frac{\sum_{s=1}^{S} (\widehat{\pi}_s^{-1} - \widehat{\pi}_s^{-2})}{12N^3} - \ldots, \tag{2.1.1.2}$$

where S is the number of observed species. In practice using of the first two terms is enough due to the fact that the following are very small, see Fritsch and Hsu [1999].

The major disadvantage of the Shannon index is, that the sample could not include all species in the community. This leads to an increasing error with declining proportion of species in the sample. Another disadvantage of this index is confounding species richness and evenness, which are two different aspects of diversity. An increasing index value could show

greater species richness or greater evenness as well as both of them. This makes interpretation harder [Magurran, 2004].

In addition to this the Shannon index will always show the same result as long as the number of species and their proportional abundances are constant. That means the Shannon index will give the same result, if one sample features 10 species, each with 5 individuals, and another sample features 10 species, each with 10 individuals [Magurran, 2004].

2.1.2 Estimation of Shannon's index considering unseen species in sample

To consider unseen species in sample, Chao and Shen [2003] provided a different approach to the estimation of diversity. It is based on unequal probability sampling theory since species have different probabilities of being discovered in sample.

Chao and Shen [2003] combine the Horvitz-Thompson estimator and the concept of sample coverage. The Horvitz-Thompson estimator adjusts for missing species in an unequal probability sampling scheme by estimating the total population. It was first introduced by Horvitz and Thompson [1952]. The concept of sample coverage adjusts for the sample fraction of unseen species.

Applying the Horvitz-Thompson estimator to the estimation of Shannon's entropy results in equation (2.1.2.1)

$$\widehat{H}_{HT} = -\sum_{s=1}^{S} \frac{\pi_s log(\pi_s)}{1-(1-\pi_s)^n} I(A_s), \tag{2.1.2.1}$$

where $1-(1-\pi_s)^n$ is the probability of the sth species being included in the sample.

The estimator for the modified Shannon index, which combines the Horvitz-Thompson adjustment and the concept of sample coverage, is given in equation (2.1.2.2)

$$\widehat{H} = -\sum_{s=1}^{S} \frac{\widehat{C}\hat{\pi}_s \, log\,(\widehat{C}\hat{\pi}_s)}{1-(1-\widehat{C}\hat{\pi}_s)^N} I(A_s), \tag{2.1.2.2}$$

where $C = \sum_{s=1}^{S} \pi_s I[X_s > 0] \approx \widehat{C} = 1 - f_1/N$ and $f_1 = \sum_{s=1}^{S} I[X_s = 1]$. Here, f_1 represents the frequency of species with counts equal 1.

Further $I(A_s)$ is the indicator function for the sth unit being included in the sample, i.e. $I(A_s) = 1$ if the sth species has count $X_s > 0$ and $I(A_s) = 0$ if $X_s = 0$. In other words, the modified index by Chao and Shen [2003] adjusts for unseen species assuming them being equal to the number of singletons f_1.

2.1.3 Simpson's index

The Simpson index [Simpson, 1949] measures the probability that any two individuals in the sample are from the same species as shown in equation (2.1.3.1)

$$\varphi = \sum_{s=1}^{S} \pi_s^2, \tag{2.1.3.1}$$

where π_s is the proportion of individuals in the sth species. With increasing φ the diversity decreases. This leads to the often used expressions $1-\varphi$ or $1/\varphi$ [Magurran, 2004].

In the following I will use the corrected estimator in equation (2.1.3.2) according to Rogers and Hsu [2001]

$$\widehat{\varphi} = \frac{n}{n-1} \sum_{s=1}^{S} \hat{\pi}_s^2, \tag{2.1.3.2}$$

where n ist the total sample size across all species.

Chapter 3

Statistical Inference Methods

The following chapter deals with the applicated statistical methods. Main topic are the nonparametric bootstrap methods, which are later compared with the asymptotic methods. I describe and analyse, among the multiple comparison methods, several one-sample bootstrap methods. The simulations of the one-sample methods should give a first overview of how well bootstrap methods perform in this context. Due to this I skip the two-sample case and come then directly to the multiple comparison methods, which are the main topic of this thesis.

3.1 The Bootstrap Principle

The bootstrap is a data-based simulation method, which allows to produce statistical inferences for observed data. The term bootstrap derives from the phrase *to pull oneself up by one's bootstrap* based on one of the eighteenth century adventures of Baron Munchhausen written by Rudolph Erich Raspe. The Baron had fallen to the bottom of a deep lake. Just when it looked that all was lost, he thought to pick him up by his own bootstrap [Efron and Tibshirani, 1993]. The common ground of this adventure and the statistical method mentioned here, is, that both use their own base to come up.

One can distinguish between the parametric and the nonparametric bootstrap. In case of the parametric bootstrap, one estimates the distribution function $\hat{F}$ for the given data. Next step is to draw a sample, with the same sample size as the original sample, out of $\hat{F}$ to substitute the original sample. One repeats this step, for example 1000 times, and in every step one estimates the statistic of interest $\hat{\theta}^*$, i.e. the sample mean $\bar{x}$. Here and in the following, the star $(*)$ indicates the bootstrapped values. This yields a vector of sample means, which allows to construct confidence intervals in different ways for the sample mean or any other statistic of interest. The easiest way to construct a confidence interval is the percentile method which takes the 25th and 975th value out of these 1000 sample means for the lower and upper border of a 95% interval [Efron and Tibshirani, 1993].

In contrast, the nonparametric bootstrap is applicable if one can not make any assumptions about the underlying distribution function $\hat{F}$. Here, one draws with replacement out of the observed sample values. As before, one repeats this step several times with estimating the statistic of interest and constructs confidence intervals for the parameter of interest $\hat{\theta}$ in different ways, nearer described in chapter 3 on page 22 [Efron and Tibshirani, 1993].

In this thesis I will use the nonparametric bootstrap methods to estimate the desired biodiversity index, i.e. the Shannon H or the Simpson φ index. In the following, I will substitute the biodiversity index several times with the parameter of interest θ. In such kind of ecological trials with replicated samples one may estimate $\hat{\theta}_{ij}$ for every replication j resulting in a vector $\hat{\boldsymbol{\theta}}_i$ or, alternatively, by taking the sum for every species over all vectors of counts $\boldsymbol{x}_i = \sum_{j=1}^{r} \boldsymbol{x}_{ij}$, where i indicates the ith treatment group. Next step is to estimate $\hat{\boldsymbol{\pi}}_{\boldsymbol{i}} = \hat{\pi}_{i1}, ..., \hat{\pi_{iS}}$. On this probability vector $\hat{\boldsymbol{\pi}}_{\boldsymbol{i}}$ one may estimate the parameter of interest $\hat{\theta}_i$ with its corresponding estimated variance $\hat{\sigma}_{\theta_i}$.

I will use the second method in the following sections, where one draws with replacement vectors of counts $\boldsymbol{x}_{ij}$ j times for every bth bootstrap step. Then in every bootstrap step one takes the sum over all vectors of counts $\boldsymbol{x}^*_{ij}$, resulting in $\boldsymbol{x}^*_i$, and estimates $\hat{\theta}^*_i$ on this vector. This yields the distribution of $\hat{\theta}^*_i$, which one uses to construct confidence intervals for θ_i. I will use the before described method of taking the sum for all confidence intervals, except for the simultaneous confidence intervals according to Westfall and Young [1993] in section 3.3.1 on page 30.

The different methods for constructing bootstrap confidence intervals are presented in the following sections, where I will focus on the nonparametric bootstrap, since I know no applicable distributional assumptions for the data one observes in these kind of ecological trials. For further reading about bootstrap confidence intervals see Efron and Tibshirani [1993], Davison and Hinkley [1997] and Westfall and Young [1993].

For the simulation part of this thesis, the one-sample bootstrap confidence intervals are performed with the R [R Development Core Team,

2009] package *boot* [Canty and Ripley, 2009]. For the simultaneous confidence intervals the basic boot function in package *boot* is used for drawing with replacement.

3.2 One-Sample Bootstrap Confidence Intervals

For a better understanding of the construction of bootstrap confidence intervals, I first present the Normal- and Student's-t interval for the one-sample problem. Upon this, I come to the theory of bootstrap confidence intervals. The methods presented here are the bootstrap-t, the percentile and the BC_a interval.

3.2.1 Normal and Student's-t Interval

The standard confidence interval $[\hat{\theta} - z^{(1-\alpha)} \cdot \widehat{se}, \hat{\theta} - z^{(\alpha)} \cdot \widehat{se}]$ is constructed under assumption

$$Z = \frac{\hat{\theta} - \theta}{\widehat{se}} \dot{\sim} N(0,1), \tag{3.2.1.1}$$

whereas Student's-t interval $[\hat{\theta} - t_{n-1}^{(1-\alpha)} \cdot \widehat{se}, \hat{\theta} - t_{n-1}^{(\alpha)} \cdot \widehat{se}]$ assumes

$$Z = \frac{\hat{\theta} - \theta}{\widehat{se}} \dot{\sim} t_{n-1}. \tag{3.2.1.2}$$

The use of the t distribution implies normal distributed data and thereby does not adjust the confidence interval for skewness in the underlying population or other occurring errors, when $\hat{\theta}$ is not the sample mean $\bar{x}$ [Efron and Tibshirani, 1993]. In such cases using one of the following three methods is more applicable.

3.2.2 The Bootstrap-t Interval

By constructing a bootstrap-t interval it is not necessary to make normal theory assumptions. The distribution of Z is estimated directly from the data with

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{se}^*(b)}, \tag{3.2.2.1}$$

where $Z^*(b), b = 1, ..., B$ is estimated for every bootstrap sample one draws $\mathbf{x}^{*1}, \mathbf{x}^{*2}, ..., \mathbf{x}^{*B}$ and $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$ is the statistic of interest $\hat{\theta} = s(\mathbf{x})$ for the bootstrap sample $\mathbf{x}^{*b}$. Further $\widehat{se}^*(b)$ is the standard error for the bootstrap sample $\mathbf{x}^{*b}$. The αth percentile of $Z^*(b)$ is estimated by the value $\hat{t}^{(\alpha)}$ such that

$$\#\{Z^*(b) \leq \hat{t}^{(\alpha)}\}/B = \alpha, \tag{3.2.2.2}$$

where B is the number of bootstraps and # indicates for the number. For example, if $B = 1000$, the estimator for the 5% point is the $50th$ largest value of the $Z^*(b)$s while the 95% point is the $950th$ largest value.

This yields to the bootstrap-t interval

$$(\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \widehat{se}, \hat{\theta} - \hat{t}^{(\alpha)} \cdot \widehat{se}). \tag{3.2.2.3}$$

An important note is that one has to use enough bootstrap samples to estimate the interval like 1000, while 100 or 200 is not enough. Also the bootstrap-t interval is easier to calculate than the both in sections 3.2.3 and 3.2.4, but can give somewhat erratic results and is heavily influenced by some outlying data points.

As a generalization of the Student's-t method the bootstrap-t procedure is applicable for location statistics as the mean, median, or a sample quantile, which increases by a constant c, if each data value x_i is increased by constant c [Efron and Tibshirani, 1993].

3.2.3 The Percentile Interval

In case of the nonparametric percentile intervals one draws with replacement a sample from the original data b times. In every bth sample one

estimates $\hat{\theta}^*$. This leads to the distribution of the $\hat{\theta}^*$s, where one estimates the percentiles $\hat{\theta}_{\text{lo}}$ and $\hat{\theta}_{\text{up}}$ as described in equation (3.2.3.1 and 3.2.3.2).

$$\hat{\theta}_{\text{lo}} = \hat{\theta}^{*(\alpha)} = 100 \cdot \alpha^{th}\text{percentile of}\,\hat{\theta}^*\text{ 's distribution} \tag{3.2.3.1}$$

$$\hat{\theta}_{\text{up}} = \hat{\theta}^{*(1-\alpha)} = 100 \cdot 1 - \alpha^{th}\text{percentile of}\,\hat{\theta}^*\text{ 's distribution.} \tag{3.2.3.2}$$

Then the $1 - 2\alpha$ interval is defined by equation (3.2.3.3)

$$[\hat{\theta}_{\%,\text{lo}}, \hat{\theta}_{\%,\text{up}}] = [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]. \tag{3.2.3.3}$$

According to the definition $G^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$, the $100 \cdot \alpha$th percentile of the bootstrap distribution, the percentile interval is also writeable as equation (3.2.3.4)

$$[\hat{\theta}_{\%,\text{lo}}, \hat{\theta}_{\%,\text{up}}] = [\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]. \tag{3.2.3.4}$$

[Efron and Tibshirani, 1993].

3.2.4 The BC_a Method

The BC_a interval is an improved version of the percentile interval, which automatically corrects for bias in the plug-in estimate $\hat{\theta}$. It combines good properties in matching closely exact intervals and good coverage probabilities in different situations. However, the coverage accuracy can be erratic for small sample sizes. The abbreviation BC_α stands for bias-corrected and accelerated.

To construct the BC_a interval one has to modify the percentile interval

$$(\hat{\theta}_{\text{lo}}, \hat{\theta}_{\text{up}}) = (\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}) \tag{3.2.4.1}$$

with the acceleration $\hat{a}$ and the bias-correction $\hat{z}_0$, on which the percentiles depend. In equation 3.2.4.2 is the BC_a interval with intended coverage $1 - 2\alpha$ presented

$$BC_a : (\hat{\theta}_{\text{lo}}, \hat{\theta}_{\text{up}} = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}), \tag{3.2.4.2}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_o + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right) \tag{3.2.4.3}$$

$$\alpha_2 = \Phi\left(\hat{z}_o + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right), \tag{3.2.4.4}$$

with $\Phi(\cdot)$ = the standard normal cumulative distribution function and $z^{(\alpha)}$ = the 100αth percentile point of a standard normal distribution.

If $\hat{a}$ and $\hat{z}_0$ equal zero, the BC_a interval is the same as the percentile interval in 3.2.4.1 and for non-zero values the BC_a interval corrects for certain deficiencies of the standard and percentile methods. The computation of $\hat{a}$ and $\hat{z}_0$ is done by 3.2.4.5 and 3.2.4.6.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right) \tag{3.2.4.5}$$

Here, $\Phi^{-1}(\cdot)$ indicates the inverse function of a standard normal cumulative distribution function, e.g. $\Phi^{-1}(.95) = 1.645$, and $\#\{\}$ represents the number of $\hat{\theta}^*$ values lower than $\hat{\theta}$. In other words, $\hat{z}_0$ measures the median bias of $\hat{\theta}^*$, which is the discrepancy between the median of $\hat{\theta}^*$ and $\hat{\theta}$ in normal units. this means, that $\hat{z}_0$ equals zero, if exactly half of the $\hat{\theta}^*(b)$ values are less than or equal to $\hat{\theta}$.

The easiest way to compute the acceleration $\hat{a}$ is done by the jackknife values of a statistic $\hat{\theta} = s(\mathbf{x})$. Let $\mathbf{x}_i$ be the original sample with the ith value x_i deleted and $\hat{\theta}_{(i)} = s(\mathbf{x}_{(i)})$. Further, $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$, according to the jackknife method.

Then, the acceleration is computed by

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}}, \tag{3.2.4.6}$$

and refers to the rate of change of the standard error of $\hat{\theta}$, with respect to the true parameter θ.

There are two important theoretical advantages of the BC_a method, which are described in the following.

The first is the transformation respecting property, meaning that the BC_a method automatically chooses the best scale. If one changes the parameter θ to a function of θ, e.g. confidence intervals for $\sqrt{\theta}$, one obtains the intervals by taking square roots of the endpoints. Such a parameter change is a problem in the bootstrap t method, where one has to choose the proper scale for the interval.

The second advantage of the BC_a method is its accuracy in matching the true parameter θ. A central $1-2\alpha$ confidence interval $(\hat{\theta}_{\text{lo}}, \hat{\theta}_{\text{up}})$ is assumed to have probability α of not covering the true value θ from above or below.

$$\text{Prob}\{\theta < \hat{\theta}_{\text{lo}}\} = \alpha \quad \text{and} \quad \text{Prob}\{\theta > \hat{\theta}_{\text{up}}\} = \alpha \tag{3.2.4.7}$$

The BC_a interval is second-order accurate, meaning, that its error in matching goes to zero, at rate $1/n$, where n is the sample size and c_{lo} and c_{up} are two constants.

$$\text{Prob}\{\theta < \hat{\theta}_{\text{lo}}\} \doteq \alpha + \frac{c_{\text{lo}}}{n} \quad \text{and} \quad \text{Prob}\{\theta > \hat{\theta}_{\text{up}}\} \doteq \alpha + \frac{c_{\text{up}}}{n} \tag{3.2.4.8}$$

In comparison to this, the standard and percentile methods are only first-order accurate, which leads to larger errors in matching.

$$\text{Prob}\{\theta < \hat{\theta}_{\text{lo}}\} \doteq \alpha + \frac{c_{\text{lo}}}{\sqrt{n}} \quad \text{and} \quad \text{Prob}\{\theta > \hat{\theta}_{\text{up}}\} \doteq \alpha + \frac{c_{\text{up}}}{\sqrt{n}} \tag{3.2.4.9}$$

Here, the constants c_{lo}, c_{up} could be different to those in 3.2.4.8. In practice, second-order accuration would lead to much better approximations of exact endpoints, on condition, that some exist [Efron and Tibshirani, 1993].

3.3 Simultaneous Bootstrap Confidence Intervals

In the following sections I will show three different algorithms for the construction of simultaneous bootstrap confidence intervals. The first one is constructed according to Westfall and Young [1993], while the second one is similar to the first, except for the calculation of the groupwise diversity indices. Westfall and Young [1993] estimated the groupwise parameters of interest by assuming an ANOVA model. Here, the model is estimated for the diversity indices per replication, i.e. the traps. Also, I calculate the estimators in the second method by taking the sum over the replicated counts per group. From the resulting count vectors I estimate the diversity indices per group according to Fritsch and Hsu [1999] and Rogers and Hsu [2001] with the corresponding variance estimators. The third method is performed according to Besag et al. [1995] and developed in a Bayesian context. Mandel and Betensky [2008] published simultaneous bootstrap percentile intervals, which are similar to the one by Besag et al. [1995], except that the intervals by Mandel and Betensky [2008] adjust for ties between ranks in every bootstrap step. This method is not analysed in this thesis but may be of interest for the area of ecological field trials, too.

All of these three methods estimate the maximum distribution across the estimated contrasts for every bootstrap step. This maximum distribution yields multiplicity adjusted quantiles similar to Tukey [1953] for the first two methods. In case of the Bayesian method the maximum distribution leads to an multiplicity adjusted rank, which will be transformed back for the individual contrasts.

3.3.1 t-Statistic Based SCIs in the ANOVA Model [t_{max} lm]

The method of *single-step* simultaneous confidence intervals, as described by Westfall and Young [1993] controls, at least asymptotically, the familywise error rate. Accordingly to Westfall and Young [1993] one fits an ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{3.3.1.1}$$

and compute simultaneous confidence intervals from the bootstrap data sets $\hat{\epsilon}^*_{ij'}$, where the $\hat{\epsilon}^*_{ij}$ are a with replacement sample from the residuals $\hat{\epsilon}_{ij} = Y_{ij} - \bar{Y}_i$. Algorithm (1) demonstrates the construction of simultaneous confidence intervals in the heteroscedastic ANOVA model.

Algorithm 1 Bootstrap Simultaneous Pairwise intervals in the heteroscedastic ANOVA model (t_{max} lm)

1. Compute the index, i.e. Shannon's H' or Simpson's φ measure, of interest $\hat{\theta}$ for every replication $j,\ j = 1, ..., r$, separately. The number of replications here represents the number of traps in a field trial.
2. Fit an ANOVA model to the parameter of interest.
3. Draw a with replacement sample $\hat{\epsilon}^*_{ij}$ out of the residuals $\hat{\epsilon}_{ij}$ with unstratified resampling.
4. Compute the sample means $\bar{\epsilon}^*_i = \sum_{j=1}^{r} \hat{\epsilon}^*_{ij}/r_i,\ i = 1, ..., k$, as well as the residual mean square $(\hat{\sigma}^*)^2$ from the bootstrap data.
5. Compute the statistics
$$t^*_{ii'} = \frac{\bar{\epsilon}^*_i - \bar{\epsilon}^*_{i'}}{\left((\hat{\sigma}^2_i)^*/n_i + (\hat{\sigma}^2_{i'})^*/n_{i'}\right)^{1/2}}. \tag{3.3.1.2}$$
6. Compute and store $\max_{1 \leqslant i < i' \leqslant k} \mid t^*_{ii'} \mid$ for every boostrap step.
7. Repeat steps $3 - 6$ B bootstrap times.
8. $\widehat{Q}^{(\alpha)}$ is the $1 - \alpha$ empirical quantile of the B values $\max_{1 \leqslant i < i' \leqslant k} \mid t^*_{ii'} \mid$.
9. In case of one-sided intervals the lower $\widehat{Q}^{(\alpha)}_{min}$ is the α empirical quantile of the B values $\min_{1 \leqslant i < i' \leqslant k} t^*_{ii'}$ and the upper $\widehat{Q}^{(\alpha)}_{max}$ is the $1 - \alpha$ empirical quantile of the B values $\max_{1 \leqslant i < i' \leqslant k} t^*_{ii'}$ accordingly.

The resulting two-sided simultaneous confidence intervals are constructed in the following way

$$[L_L;\ L_U] = \hat{\theta}_i - \hat{\theta}_{i'} \pm \widehat{Q}^{(\alpha)}\ (\hat{\sigma}^2_i/n_i + \hat{\sigma}^2_{i'}/n_{i'})^{1/2}, \tag{3.3.1.3}$$

and accordingly the one-sided intervals are

$$[L_L;\ ] = \hat{\theta}_i - \hat{\theta}_{i'} + \widehat{Q}_{min}^{(\alpha)}\ (\hat{\sigma}_i^2/n_i + \hat{\sigma}_{i'}^2/n_{i'})^{1/2}, \tag{3.3.1.4}$$

and

$$[\ ; L_U] = \hat{\theta}_i - \hat{\theta}_{i'} + \widehat{Q}_{max}^{(\alpha)}\ (\hat{\sigma}_i^2/n_i + \hat{\sigma}_{i'}^2/n_{i'})^{1/2}, \tag{3.3.1.5}$$

where $\hat{\theta}_i$ is the estimated parameter of interest, i.e. the Shannon or Simpson index, out of the ANOVA model across all r replicates in the ith community.

3.3.2 t-Statistic Based SCIs with Summed up Counts [t_{max} sum]

For the t_{max} sum method we used the concept of centered statistics similar to Westfall and Young [1993], but without estimating an ANOVA model. This allowed us to apply this approach to summed up species counts in every treatment. Perform a non-parametric bootstrap, stratified by the levels of communities, with B denoting the number of bootstrap data sets. Further $(*)$ denotes the bootstrapped values. For each bootstrap data set, build the column wise sums of species counts within the observations of each treatment i, $\boldsymbol{Y_i} = (Y_{1i}, ..., Y_{Si})$, and from those calculate the group-wise indices $\hat{\theta}_i$ and the t-statistics

$$t_{ii'}^* = \frac{(\hat{\theta}_i^* - \hat{\theta}_{i'}^*) - (\hat{\theta}_i - \hat{\theta}_{i'})}{((\hat{\sigma}^2_{\hat{\theta}_i})^* + (\hat{\sigma}^2_{\hat{\theta}_{i'}})^*)^{1/2}} \tag{3.3.2.1}$$

between the group-wise diversity indices. Here, the variances $\hat{\sigma}^2$ are the estimated variances corresponding to the Shannon or Simpson index, see equation (3.4.2.3 on page 37) and (3.4.1.2 on page 37). The estimation of the t-statistic based SCIs on summed up counts is presented in algorithm (2 on the next page).

Algorithm 2 Bootstrap Simultaneous Pairwise intervals on the summed up counts (t_{max} sum)

1. Perform a non-parametric bootstrap stratified by the levels of communities.
2. Build the column wise sums of species within each treatment for every bootstrap sample b, yielding to the vector of species counts $\boldsymbol{Y_i^*} = \left(Y_{1i}^*, ..., Y_{Si}^*\right)$.
3. Estimate the group wise indices of interest, i.e. Shannon's $\widehat{H}_i$ or Simpson's $\widehat{\phi}_i$ index, denoted here with $\hat{\theta}_i^*$.
4. In every bootstrap sample, calculate the test statistic $t_{ii'}^*$ presented in equation (3.3.2.1)
5. Compute and store $\max_{1 \leqslant i < i' \leqslant k} \mid t_{ii'}^* \mid$ for every bootstrap step.
6. $\widehat{Q}^{(\alpha)}$ is the $1 - \alpha$ empirical quantile of the B values $\max_{1 \leqslant i < i' \leqslant k} \mid t_{ii'}^* \mid$.
7. In case of one-sided intervals the lower $\widehat{Q}_{min}^{(\alpha)}$ is the α empirical quantile of the B values $\min_{1 \leqslant i < i' \leqslant k} t_{ii'}^*$ and the upper $\widehat{Q}_{max}^{(\alpha)}$ is the $1 - \alpha$ empirical quantile of the B values $\max_{1 \leqslant i < i' \leqslant k} t_{ii'}^*$ accordingly.

The resulting two-sided simultaneous confidence intervals are constructed in the following way

$$[L_L;\ L_U] = \hat{\theta}_i - \hat{\theta}_{i'} \pm \widehat{Q}^{(\alpha)}\ (\hat{\sigma}^2_{\hat{\theta}_i} + \hat{\sigma}^2_{\hat{\theta}_{i'}})^{1/2}, \tag{3.3.2.2}$$

and the corresponding one-sided intervals by

$$[L_L;\ ] = \hat{\theta}_i - \hat{\theta}_{i'} + \widehat{Q}^{(\alpha)}_{min}\ (\hat{\sigma}^2_{\hat{\theta}_i} + \hat{\sigma}^2_{\hat{\theta}_{i'}})^{1/2}, \tag{3.3.2.3}$$

and

$$[\ ;L_U] = \hat{\theta}_i - \hat{\theta}_{i'} + \widehat{Q}^{(\alpha)}_{max}\ (\hat{\sigma}^2_{\hat{\theta}_i} + \hat{\sigma}^2_{\hat{\theta}_{i'}})^{1/2}, \tag{3.3.2.4}$$

where $\hat{\sigma}_i$ is the variance estimator according to the parameter of interest, see (3.4.2.3 on page 37) and (3.4.1.2 on page 37).

3.3.3 Percentile Based SCIs with Summed up Counts [rank-*perc*]

In a Bayesian context, Besag et al. [1995] describe an algorithm to derive simultaneous intervals based on an empirical joint distribution of the parameters of interest. Although proposed for a different purpose, I use it here to construct simultaneous percentile intervals, based on the joint empirical distribution of multiple differences of diversity indices.

Perform a non-parametric bootstrap, stratified by the levels of the i treatments, with B denoting the number of bootstrap data sets. For each bootstrap data set, build the column wise sums of species counts within the observations of each treatment i, $\boldsymbol{Y_i} = (Y_{1i}, ..., Y_{Si})$, and from those calculate the group-wise indices θ_i and the differences of interest between the group-wise diversity indices. Given that there are M differences of interest, δ_m, $m = 1, ..., M$. The results of this process can be written in a $(M \times B)$ matrix $\mathbf{\Psi}$, with elements ψ_{bm}, $b = 1, ..., B$, $m = 1, ..., M$. On the matrix Ψ, one applies Besag et al.s algorithm, which is presented in algorithm (3 on the following page). The derived region is two-sided for each parameter δ_m. Analogously, one-sided regions can be constructed.

Algorithm 3 Bootstrap Simultaneous Pairwise intervals based on an empirical joint distribution (rank-*perc*)

1. Order each of the M columns of $\mathbf{\Psi}$ separately. Results are the order statistics $\psi_m^{[b]}$, and the ranks u_{bm}, written in an $(B \times M)$ matrix $\mathbf{U}$.

2. Calculate the minimum and maximum over each of the B rows of $\mathbf{U}$, $u_b^{(min)} = min\left(u_{b1}, ..., u_{bm}, ..., u_{bM}\right)$, $u_b^{(max)} = max\left(u_{b1}, ..., u_{bm}, ..., u_{bM}\right)$, and then calculate $u_b^{(maxmin)} = max\left(B + 1 - u_b^{(min)}, u_b^{(max)}\right)$, for each $b = 1, ..., B$.

3. The vector $\mathbf{u}^{(maxmin)} = \left(u_1^{(maxmin)}, ..., u_b^{(maxmin)}, ..., u_B^{(maxmin)}\right)$ is again ordered, leading to order statistics $u^{[b]}$ and the corresponding ranks $\mathbf{r}^{(b)}$.

4. Let b^* denote the closest integer to $B(1 - \alpha)$. The quantile is then $t^* = u^{[b^*]}$, i.e. taking the b^*th value from the ordered sample of the folded empirical distribution of the maximum.

5. Finally, the confidence limits are constructed for each elementary parameter θ_m by taking $\left[\theta_m^{[B+1-t^*]}; \theta_m^{[t^*]}\right]$, i.e. the $B + 1 - t^*$th and t^*th value from the ordered sample of the joint empirical distribution obtained for δ_m.

3.4 Asymptotic SCIs Considering Heterogeneous Variances [asymSCI]

The following two asymptotic methods are created for the Shannon and the Simpson index as well by Fritsch and Hsu [1999] and Rogers and Hsu [2001]. They were primarily developed for multinomial distributed count data without replications. Both methods use special variance estimators for the indices, which should consider heterogeneous variances. The estimated quantile for the simultaneous confidence intervals are based on the multivariate normal distribution. This leads to multiplicity adjusted confidence intervals.

3.4.1 Calculation of the Simpson Index and the Corresponding Variance Under Heteroscedasticity

Suppose, one observes k communities, $i = 1, ..., k$, where n_i is the total sample size of community i, meaning the sum across all traps or replicates j, $j = 1, .., n$ and all species S in this community. Further, let $\boldsymbol{Y}_{is} = Y_{i1}, ..., Y_{iS}$ be the vector of the column wise sums of species counts within the observation of each treatment and $\widehat{\pi}_{is} \geq 0$. For every i let $\widehat{\boldsymbol{\pi}}_i = \widehat{\pi}_{i1}, ..., \widehat{\pi}_{iS}$ be the proportion of the ith community constituted by the sth specie.

Rogers and Hsu [2001] used the estimator in equation (3.4.1.1) for the Simpson index to construct simultaneous confidence intervals

$$\widehat{\varphi}_i = \frac{n_i}{n_i - 1} \sum_{s=1}^{S} \hat{\pi}_{is}^2, \tag{3.4.1.1}$$

where $n_i/n_i - 1$ is a correction term which leads to an unbiased estimator for this index.

The variance estimator for Simpson's index considering heterogeneous variances is provided by Rogers and Hsu [2001] and presented in equation (3.4.1.2)

$$\widehat{\sigma}^2_{\varphi_i} = 2 \cdot \frac{\left\{ \sum_{s=1}^{S} (\widehat{\pi}_{is})^2 + 2(n_i - 2) \sum_{s=1}^{S} (\widehat{\pi}_{is})^3 + (3 - 2n_1) \left(\sum_{s=1}^{S} (\widehat{\pi}_{is})^2 \right)^2 \right\}}{n_i(n_i - 1)} \tag{3.4.1.2}$$

3.4.2 Calculation of the Shannon Index and the Corresponding Variance Under Heteroscedasticity

Fritsch and Hsu [1999] derived simultaneous confidence intervals for the Shannon index with respect to heterogeneous variances. The situation for the Shannon index is more difficult than for the Simpson index since the point estimator for the Shannon index

$$\widehat{H}_i = -\sum_{s=1}^{S} \widehat{\pi}_{is} \log\left(\widehat{\pi}_{is}\right) \tag{3.4.2.1}$$

is negatively biased. Fritsch and Hsu [1999] showed in a simulation study, that introducing the bias-correction term $+\left(S_i - 1\right)/2n_i$ is a sufficient remedy:

$$\tilde{H}_i = \widehat{H}_i + \left(S_i - 1\right)/2n_i \tag{3.4.2.2}$$

with S_i being the number of observed species in the ith treatment.

The variance estimator for Shannon's index as used by Fritsch and Hsu [1999] is given in equation (3.4.2.3). Let $\widehat{\boldsymbol{\pi}}_i = (\widehat{\pi}_{i1}, ..., \widehat{\pi}_{iS})'$, i.e. a matrix with 1 column and S rows.

$$\widehat{\sigma}^2_{H_i} = \frac{\log(\widehat{\boldsymbol{\pi}}_i)' \left[\operatorname{diag}(\widehat{\boldsymbol{\pi}}_i) - \widehat{\boldsymbol{\pi}}_i \widehat{\boldsymbol{\pi}}_i'\right] \log(\widehat{\boldsymbol{\pi}}_i)}{n_i} \tag{3.4.2.3}$$

Construction of the Asymptotic SCIs

Applying equation (3.4.1.1) and (3.4.1.2) on the observed values $\mathbf{y}_i$ of all k communities, one yields a vector of estimators $\widehat{\boldsymbol{\varphi}}_{\boldsymbol{i}} = (\widehat{\varphi}_1, ..., \widehat{\varphi}_k)$, and variance estimators $\hat{\boldsymbol{\sigma}}_{\widehat{\boldsymbol{\varphi}}_{\boldsymbol{i}}} = \left(\hat{\sigma}_{\widehat{\varphi}_1}, ..., \hat{\sigma}_{\widehat{\varphi}_k}\right)$.

Rogers and Hsu [2001] derived simultaneous confidence intervals for Dunnett's Multiple Comparisons with Control (MCC) [Dunnett, 1955] and Tukey's All-Pairwise Comparisons (MCA) [Tukey, 1953] based on the quantiles of the multivariate normal distribution.

Simultaneous confidence intervals for the MCA for $L_{ii'}$ can be calculated with equation (3.4.2.4) for the Simpson index

$$\widehat{\varphi}_i - \widehat{\varphi}_{i'} \pm q_{2,\,1-\alpha;M,R}\sqrt{\widehat{\sigma}^2_{\widehat{\varphi}_i} + \widehat{\sigma}^2_{\widehat{\varphi}_{i'}}} \tag{3.4.2.4}$$

with $i = 1, ..., k$ and $i \neq i'$.

In equation (3.4.2.5) I present the SCIs for MCC

$$\widehat{\varphi}_i - \widehat{\varphi}_1 \pm q_{2,\,1-\alpha;M,R}\sqrt{\widehat{\sigma}^2_{\widehat{\varphi}_i} + \widehat{\sigma}^2_{\widehat{\varphi}_1}} \tag{3.4.2.5}$$

with the variance estimator presented in equation (3.4.1.2 on the previous page) and $i = 1$ indicating the control group as well as $i = 2, ..., k$ indicating the standards. The term $q^2_{1-\alpha;M,R}$ denotes the two-sided equicoordinate $1-\alpha$ quantile of an M-variate normal distribution with correlation matrix R.

For the Shannon index one can construct simultaneous confidence intervals for MCA in a similar way, see equation (3.4.2.6)

$$\widehat{H}_i - \widehat{H}_{i'} \pm q_{2,\,1-\alpha;M,R}\sqrt{\hat{\sigma}^2_{\widehat{H}_i} + \hat{\sigma}^2_{\widehat{H}_{i'}}} \tag{3.4.2.6}$$

and analogeously for MCC with equation (3.4.2.7)

$$\widehat{H}_i - \widehat{H}_1 \pm q_{2,\,1-\alpha;M,R}\sqrt{\hat{\sigma}^2_{\widehat{H}_i} + \hat{\sigma}^2_{\widehat{H}_1}} \tag{3.4.2.7}$$

with the variance estimator from equation (3.4.2.3 on the preceding page).

Calculation of the quantiles in equation (3.4.2.4) and (3.4.2.6) is performed in the statistical programming language `R` [R Development Core Team, 2009] via the package `mvtnorm` [Genz et al., 2008]. Required inputs are the probability $1-\alpha$, the correlation matrix $\boldsymbol{R}$ for the performed contrast, i.e. Tukeys MCA or Dunnetts MCC, and further the requested side to test.

The elements of the correlation matrix $\boldsymbol{R}$ for *comparisons to the control* $i = 1$: $\theta_i - \theta_1$ with $k-1$ rows and columns are shown in equation (3.4.2.8). $\rho_{ii'}$ is the correlation of $\theta_i - \theta_1$ with $\theta_{i'} - \theta_1$

$$\rho_{ii'} = \frac{1}{\sqrt{\left(1 + \frac{\sigma_{\widehat{\theta}_i}}{\sigma_{\widehat{\theta}_1}}\right)\left(1 + \frac{\sigma_{\widehat{\theta}_{i'}}}{\sigma_{\widehat{\theta}_1}}\right)}}, \tag{3.4.2.8}$$

where $\sigma_{(\widehat{\theta})}$ are the theoretical variances of the estimators $\widehat{\theta}$. For the estimated correlation matrix R one has to plug-in the estimated variances of the estimators.

In case of *all-pairwise comparisons*, where the correlation of two arbitrary differences $\theta_i - \theta_{i'}$ with $\theta_{i''} - \theta_{i'''}$ is denoted by $\rho_{(ii'),(i''i''')}$, one has to distinguish between two cases.

1. If there is no group common in the two differences (i.e., $\{i, i'\} \cap \{i'', i'''\} = \emptyset$), then $\rho_{(ii'),(i''i''')} = 0$.

2. If one common group is included in both differences with equal sign, e.g. $i = i''$ and the comparison $\theta_i - \theta_{i'}$ with $\theta_i - \theta_{i'''}$ is considered. Then correlation is given in equation (3.4.2.9)

$$\rho_{(ii'),(i''i''')} = \rho_{(ii'),(ii''')} = \frac{1}{\sqrt{\left(1 + \frac{\sigma_{\widehat{\theta}_{i'}}}{\sigma_{\widehat{\theta}_i}}\right)\left(1 + \frac{\sigma_{\widehat{\theta}_{i'''}}}{\sigma_{\widehat{\theta}_i}}\right)}} \tag{3.4.2.9}$$

   - If the common group enters the two differences with different sign, e.g. $\theta_i - \theta_{i'}$ with $\theta_{i'''} - \theta_i$, the correlation is negative with absolute value given in equation (3.4.2.9).

The correlation matrix $\boldsymbol{R}$ exhibits $k(k-1)/2$ rows and columns and as before the estimated variances of the estimators has to be plugged-in.

Chapter 4

Simulation Study

In this chapter, I give background knowledge to the data simulation process and further I present the simulation results.

4.1 Methods

In the following the data simulating process is described. First, I explain the different steps in the data generating process. Thereafter, I describe the settings for data generation, followed by several plots clarifying the simulated data sets.

4.1.1 Simulated distribution

The simulated data is generated out of the Dirichlet-multinomial distribution, see section A.4 on page 77, with parameters for the Dirichlet distribution taken from the geometrc series, see section A.3 on page 77. The data generating process is described in section 4.1.2.

4.1.2 Simulation steps

Data simulating an agricultural field trial with different species counted on several traps is constructed in the following three steps.

1. Use a vector $\boldsymbol{\alpha_i} = \alpha_{i1}, ..., \alpha_{iS}$ with parameters $\kappa_i \mid 0 < \kappa_i \leqslant 1$ and S for the number of species out of the geometric series, see A.1 on page 75. This vector $\boldsymbol{\alpha_{is}}$ represents the proportions of individuals in the ith community.

   For $\kappa_i = 1$ we yield no diversity, i.e. $\alpha_{i1} = 1$ and $\alpha_{i2}, ..., \alpha_{iS}$ equal zero. For $\kappa_i \to 0$ the proportions of the species are evenly distributed

2. The vector $\boldsymbol{\alpha_i}$ is for some settings multiplied with a constant $c > 1$, to reduce the overdispersion. This leads to more similar data sets in comparison with the observed example data sets.

3. Using these vectors $\boldsymbol{\alpha}_i$ as parameters for the Dirichlet distribution yields to the probability vectors $\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}_{ij11}, ..., \boldsymbol{\pi}_{ijs}$ with $j = 1, ..., r$ and $i = 1, ..., k$

4. These probability vectors $\boldsymbol{\pi}_{ij}$ represent the probabilities of the multinomial distribution for the jth replicate in the ith community. This yields a mixture multinomial distribution - the Dirichlet multinomial distribution - as described in Mosimann [1962] and in appendix A.4 on page 77.

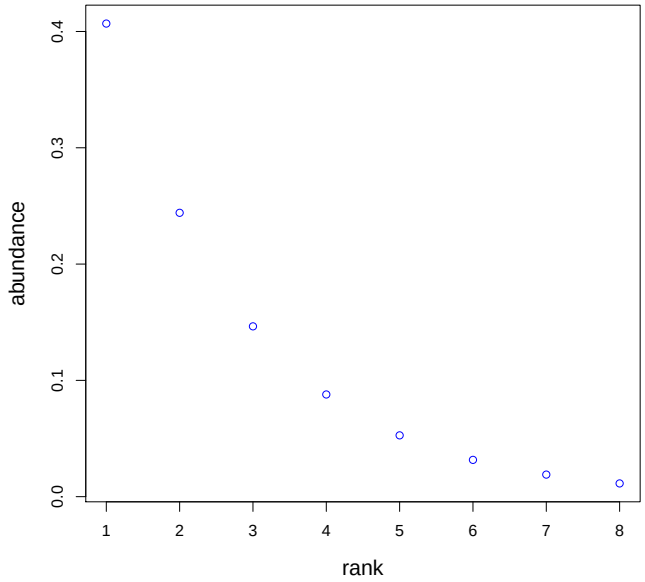
4.1.3 Simulation settings

The first step in the data generating process is producing a vector of species probabilities out of the geometric series. For this purpose I use four different settings for the parameters of the geometric series. These settings are slightly different between one-sample inference and multiple comparisons and are shown in tables from 4.1(a) on page 49 to 4.3(b) on page 52. For every simulation I use 3000 bootstrap steps as well as 1000 simulation steps. The confidence intervals are calculated with a confidence level of 95%.

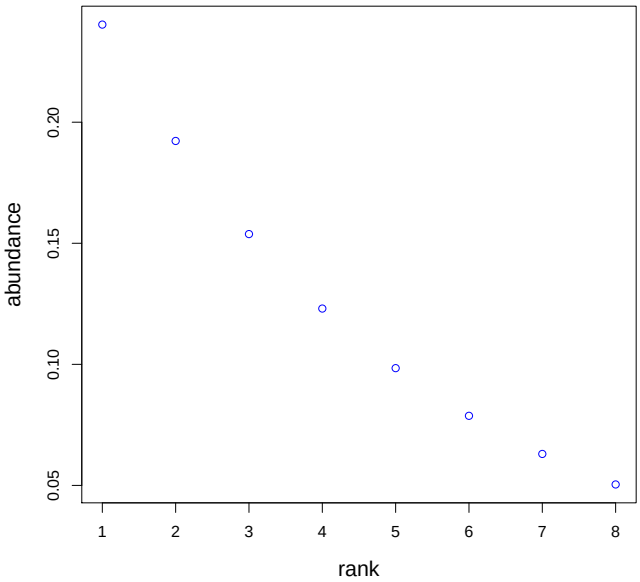
4.1.4 Simulated data

In the following I will present the different settings as plots to visualize the influence of the varying parameter settings. In figures from 4.1(a) on the next page to 4.2(b) on page 45 I show the proportions of individuals in a sample, determined by varying parameters of the geometric series. The simulated samples for the multiple comparisons are shown from figure 4.3(a) on page 46 to figure 4.4(b) on page 47.

In figures 4.4(a) and 4.4(b) the α_i-vector from the geometric series is multiplied with $c = 5$. With this adjustment, simulated data sets are more similar to the observed data sets, while the simulated data sets in figures 4.3(a) and 4.3(b) are extremely over-dispersed.

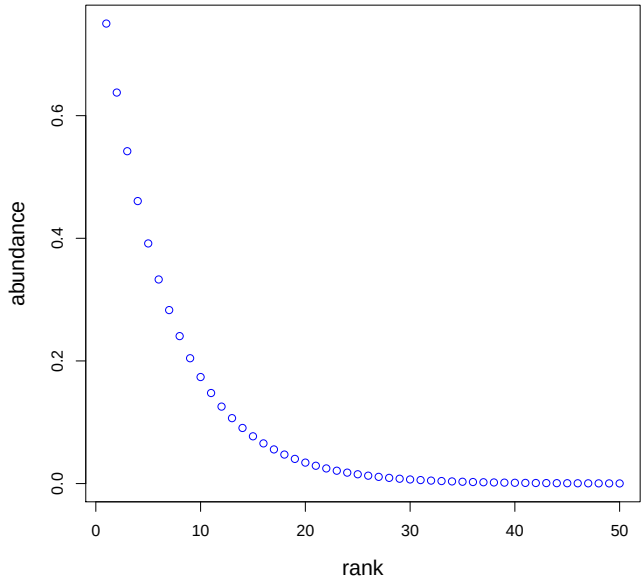


(a) Parameters k = 0.4, Species = 8

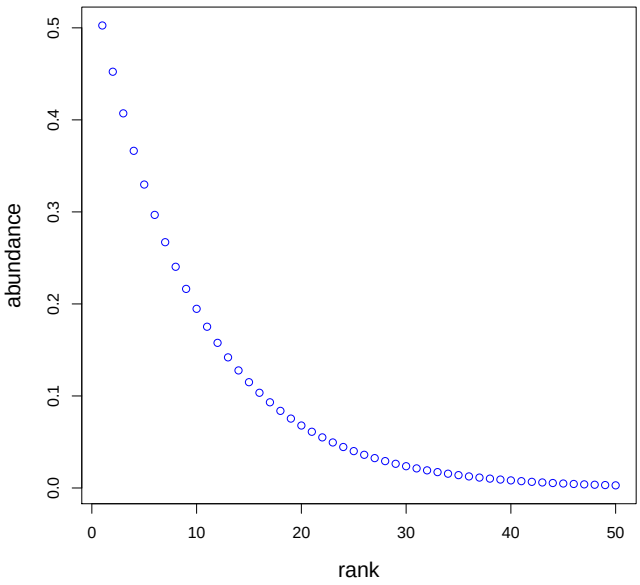


(b) Parameters k = 0.2, Species = 8

Figure 4.1: Pattern of true relative abundances of species taken from the geometric series.

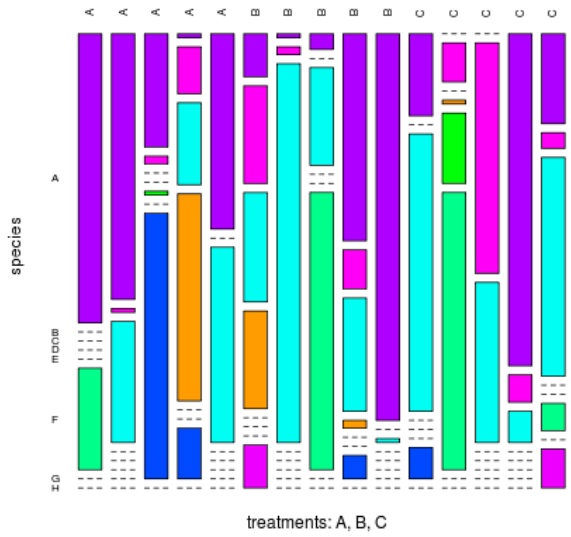


(a) Parameters k = 0.15, Species = 50, Species probabilities multiplied with factor 5

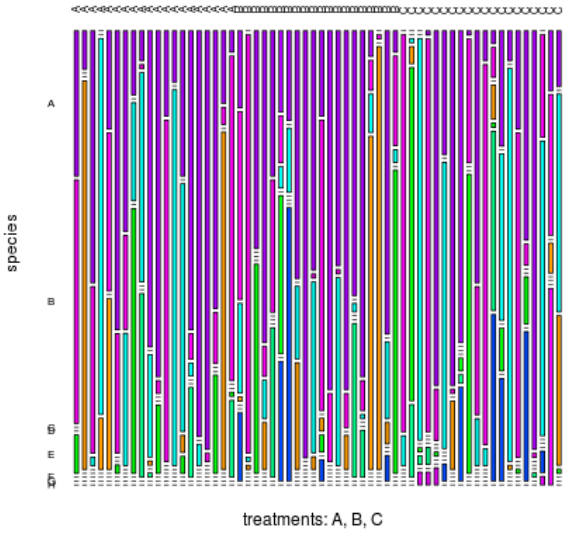


(b) Parameters k = 0.10, Species = 50, Species probabilities multiplied with factor 5

Figure 4.2: Pattern of true relative abundances of species taken from the geometric series.

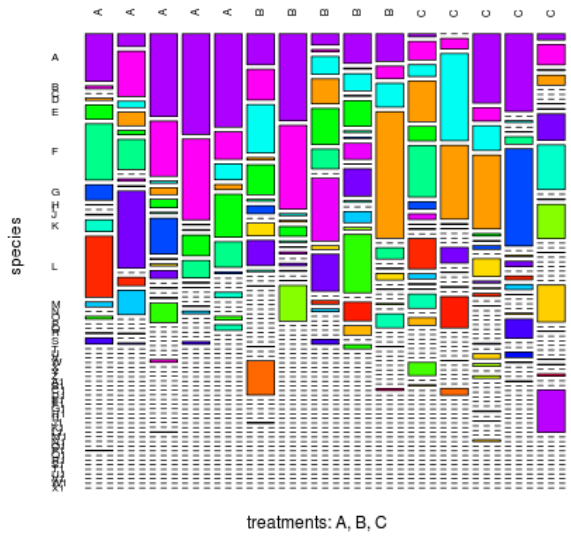


(a) Simulation Setting 1

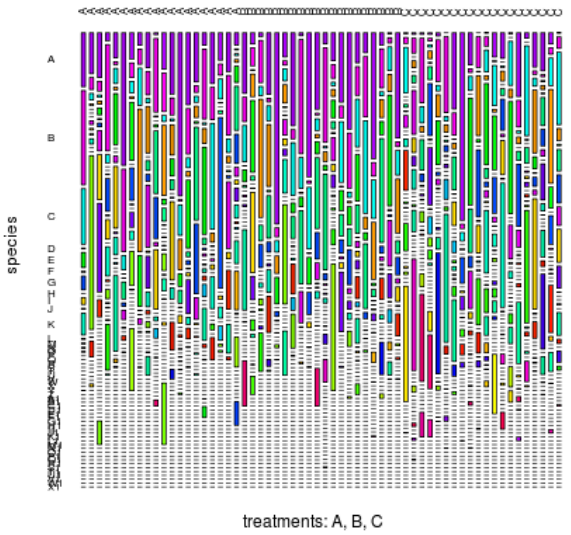


(b) Simulation Setting 2

Figure 4.3: Mosaicplot for species in simulated data set for simulation settings 1 + 2 – Width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species is not observed in one trap. The x-axis indicates the treatments and the y-axis the different species



(a) Simulation Setting 3



(b) Simulation Setting 4

Figure 4.4: Mosaicplot for species in simulated data set for simulation settings 3 + 4 – Width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species is not observed in one trap. The x-axis indicates the treatments and the y-axis the different species

4.2 Results

In the following, I will present the simulation results for the one-sample methods in section 3.1 on page 23 as well as for the multiple comparison methods in section 3.4 on page 36 and 3.3 on page 30 based on simulated data sets nearer described in subsection 4.1.2 on page 42.

4.2.1 One sample methods

The one sample results for Shannon's and Simpson's index are shown in tables 4.1(a) on the next page and 4.1(b) on the following page. Obviously, the methods perform better for the Simpson index than for the Shannon index, which can be explained by the biased estimator for the Shannon index. For both indices the coverage probability is nearer the desired level of 95% with rising sample size.

Furthermore, the coverage probabilty for the Boot-*perc* and the Boot-BC_a method is worth with increasing number of species and individuals in every replication. Under these cases the number of zero counts, as well as the number of species with small probabilities, rise up. But additionally, with increasing number of species and individuals, the parameters κ and c for the geometric series change. Therefore, an easy interpretation of the results is difficult to carry out.

Both percentile based methods, the Boot-*perc* method and the BC_a, exhibits lower coverage probability than the t-statistics based bootstrap-t method. The bootstrap-t methods comes with higher sample size nearer the desired coverage probability of 95%.

Table 4.1: One sample results for Shannon's and Simpson's index. covpl: lower coverage probability, covpu: upper coverage probability, covpt: total coverage probability. Global settings: 95% two sided Intervals, 3000 bootstraps, 1000 simulation steps. n = nr. of traps per treatment, κ & S = parameters for the geometric series, c = a constant, indiv = Individuals per trap.

(a) One sample results for Shannon's index

Shannon index								
Method	n	κ	c	S	indiv	covpl	covpu	covpt
Bootstrap-t	5	0.4	1	8	100	0.95	0.85	0.80
Boot-perc	5	0.4	1	8	100	1.00	0.21	0.21
Boot-BC_a	5	0.4	1	8	100	0.97	0.24	0.21
Bootstrap-t	20	0.4	1	8	100	0.96	0.94	0.90
Boot-perc	20	0.4	1	8	100	1.00	0.67	0.67
Boot-BC_a	20	0.4	1	8	100	0.96	0.87	0.83
Bootstrap-t	5	0.15	5	50	1000	1.00	0.69	0.69
Boot-perc	5	0.15	5	50	1000	1.00	0.00	0.00
Boot-BC_a	5	0.15	5	50	1000	1.00	0.01	0.01
Bootstrap-t	20	0.15	5	50	1000	0.98	0.85	0.83
Boot-perc	20	0.15	5	50	1000	1.00	0.17	0.17
Boot-BC_a	20	0.15	5	50	1000	0.98	0.51	0.49

(b) One sample results for Simpson's index

Simpson index								
Method	n	κ	c	S	indiv	covpl	covpu	covpt
Bootstrap-t	5	0.4	1	8	100	0.88	0.93	0.81
Boot-perc	5	0.4	1	8	100	1.00	0.54	0.54
Boot-BC_a	5	0.4	1	8	100	0.92	0.57	0.49
Bootstrap-t	20	0.4	1	8	100	0.94	0.97	0.91
Boot-perc	20	0.4	1	8	100	1.00	0.86	0.86
Boot-BC_a	20	0.4	1	8	100	0.94	0.93	0.87
Bootstrap-t	5	0.15	5	50	1000	0.98	0.86	0.84
Boot-perc	5	0.15	5	50	1000	1.00	0.06	0.06
Boot-BC_a	5	0.15	5	50	1000	0.99	0.08	0.08
Bootstrap-t	20	0.15	5	50	1000	0.95	0.96	0.91
Boot-perc	20	0.15	5	50	1000	1.00	0.59	0.59
Boot-BC_a	20	0.15	5	50	1000	0.95	0.86	0.82

4.2.2 Multiple comparison methods

The following tables from 4.2(b) on the next page to 4.3(b) on page 52 show the simulation results for the multiple comparison methods described in section 3.3 on page 30 and 3.4 on page 36 for the Tukey and Dunnett type in the two-sided case.

The most noticeable result is the poor performance of the asymptotic methods compared with the good perfomance of the bootstrap methods. Further the results indicate a better performance of the methods, which take the sum over the counts to estimate the Shannon index, than the method by Westfall and Young [1993], which estimates the Shannon index for every replicate seperately. For the Simpson index all bootstrap methods perform well. In detail, the method by Westfall and Young [1993] shows the best coverage probability, while the other bootstrap intervals are slightly more conservative.

In addition the tables 4.4(a) on page 53 and 4.4(b) on page 53 show the simulation results for the Dunnett type statistic in the one-sided case. The simulations are performed for the Shannon and the Simspon index respectively.

As well as in the two-sided case the performance of the asymptotic methods is very poor. Again, the coverage probability for the Shannon index is better if the indices are estimated out of the summed up vector, i.e. the t_{max} sum and the rank-*perc* method. These both methods give similar results, except that the rank-*perc* method performs better than the $t_{max}sum$ method with low numbers of species and individuals. The t_{max} lm method only performs well for the unbiased Simpson index.

Table 4.2: MCP results for Shannon's and Simpson's index – Dunnett contrast. covpl: lower coverage probability, covpu: upper coverage probability, covpt: total coverage probability. Global settings: 95% two sided Intervals, 3000 bootstrap steps, 1000 simulation steps. n = no. of traps per treatment, κ & S = parameters for the geometric series, c = a constant reducing overdispersion, indiv = Individuals per trap.

(a) MCP results for Shannon's index – Dunnett contrast

Shannon index									
Method	n	i	κ	c	S	indiv	covpl	covpu	covpt
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.97	0.96	0.93
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.99	0.96	0.95
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.98	0.97	0.95
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.53	0.41	0.07
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.97	0.97	0.94
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.99	0.56	0.55
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.97	0.97	0.94
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.49	0.47	0.07
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.96	0.95
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.88	0.88
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.96	0.95
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.51	0.30	0.02
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.94	0.92
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.99	0.22	0.21
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.99	0.93	0.92
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.47	0.33	0.03

(b) MCP results for Simpson's index – Dunnett contrast

Simpson index									
Method	n	i	κ	c	S	indiv	covpl	covpu	covpt
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.98	0.96	0.94
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.98	0.96	0.93
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.99	0.98	0.97
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.53	0.45	0.08
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.98	0.96	0.94
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.99	0.92	0.91
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.98	0.97	0.96
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.51	0.48	0.08
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.98	0.97
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.98	0.97	0.95
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.99	0.99
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.46	0.41	0.03
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.97	0.97	0.94
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.96	0.95
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.98	0.97
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.44	0.41	0.02

Table 4.3: MCP results for Shannon's and Simpson's index – Tukey contrast. covpl: lower coverage probability, covpu: upper coverage probability, covpt: total coverage probability. Global settings: 95% two-sided Intervals, 3000 bootstraps, 1000 simulation steps. n = no. of traps per treatment, κ & S = parameters for the geometric series, c = a constant reducing overdispersion, indiv = Individuals per trap

(a) MCP results for Shannon's index – Tukey contrast

Shannon index									
Method	n	i	κ	c	S	indiv	covpl	covpu	covpt
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.96	0.94	0.92
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.99	0.93	0.93
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.98	0.96	0.94
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.36	0.27	0.06
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.94	0.97	0.91
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.99	0.45	0.45
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.97	0.96	0.93
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.32	0.29	0.06
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.94	0.94
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.86	0.86
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.95	0.94
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.39	0.14	0.02
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.94	0.92
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.99	0.13	0.13
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.99	0.92	0.91
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.34	0.16	0.00

(b) MCP results for Simpson's index – Tukey contrast

Simpson index									
Method	n	i	κ	c	S	indiv	covpl	covpu	covpt
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.98	0.95	0.93
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.97	0.95	0.92
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.98	0.98	0.96
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.38	0.25	0.07
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.98	0.94	0.93
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.99	0.92	0.91
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.98	0.96	0.95
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.32	0.27	0.07
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.98	0.97	0.96
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.98	0.96	0.95
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.99	0.99	0.99
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.27	0.22	0.03
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.95	0.93
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.99	0.98	0.95
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.98	0.98	0.97
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.25	0.24	0.02

Table 4.4: MCP results for Shannon's and Simpson's index – Dunnett contrast – one-sided. covpu: upper coverage probability. Global settings: 95% one-sided Intervals, 3000 bootstraps, 1000 simulation steps. n = no. of traps per treatment, k & S = parameters for the geometric series, c = a constant reducing overdispersion, indiv = Individuals per trap.

(a) MCP results for Shannon's index – Dunnett contrast – one-sided

Shannon index							
Method	n	i	k	c	S	indiv	covpu
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.91
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.84
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.93
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.42
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.94
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.42
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.94
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.45
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.91
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.70
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.91
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.26
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.90
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.14
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.87
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.34

(b) MCP results for Simpson's index – Dunnett contrast – one-sided

Simpson index							
Method	n	i	k	c	S	indiv	covpu
t_{max} sum	5	3	0.4/0.4/0.2	1	8	100	0.91
t_{max} lm	5	3	0.4/0.4/0.2	1	8	100	0.92
rank-*perc*	5	3	0.4/0.4/0.2	1	8	100	0.95
asymSCI	5	3	0.4/0.4/0.2	1	8	100	0.45
t_{max} sum	20	3	0.4/0.4/0.2	1	8	100	0.92
t_{max} lm	20	3	0.4/0.4/0.2	1	8	100	0.90
rank-*perc*	20	3	0.4/0.4/0.2	1	8	100	0.95
asymSCI	20	3	0.4/0.4/0.2	1	8	100	0.45
t_{max} sum	5	3	0.15/0.15/0.1	5	50	1000	0.95
t_{max} lm	5	3	0.15/0.15/0.1	5	50	1000	0.94
rank-*perc*	5	3	0.15/0.15/0.1	5	50	1000	0.97
asymSCI	5	3	0.15/0.15/0.1	5	50	1000	0.41
t_{max} sum	20	3	0.15/0.15/0.1	5	50	1000	0.94
t_{max} lm	20	3	0.15/0.15/0.1	5	50	1000	0.94
rank-*perc*	20	3	0.15/0.15/0.1	5	50	1000	0.95
asymSCI	20	3	0.15/0.15/0.1	5	50	1000	0.45

4.2.3 Power of Bootstrap SCIs

In addition to the coverage probability, the power of the methods is an important point. Assume, a field trial is planned as a follow-up experiment of the trial shown in Example 2. The aim of the follow-up trial could be to prove that at least one of 3 treatments decreases biodiversity significantly compared to the control. The statistical hypotheses are shown in equation (4.2.3).

$$H_0 : \bigcap_{i=2}^{k} \theta_i - \theta_1 \geq 0 \tag{4.2.3.1}$$

$$H_A : \bigcup_{i=2}^{k} \theta_i - \theta_1 < 0 \tag{4.2.3.2}$$

One question to be solved before starting the experiment is: In what range should the number of replications be chosen, such that a present decrease in biodiversity can be shown with high probability, e.g. in 80% of the cases? Based on the data of the preliminary trial in Example 2, the following parameters are chosen to simulate power:

- Species $S = 30$
- Individuals per replicate $= 500$
- For control group $\kappa = 0.20$ and for treatments $\kappa = 0.25$
- The constant for reducing over-dispersion $c = 10$

Figure 4.5 on the following page shows the true relative abundances, calculated out of the geometric series, for the two different settings. The simulated data, which resembles the data of Example 2, shown in section 1.2 on page 3, is illustrated as a mosaicplot with 10 replications in figure 4.6 on page 56. In the following power analysis, the number of replications ranges from $j = 3$ to $j = 100$. Given that this single parameter setting and the assumption of a Dirichlet-multinomial distribution closely resembles reality, one can conclude from Figure 4.7 on page 57 that 20

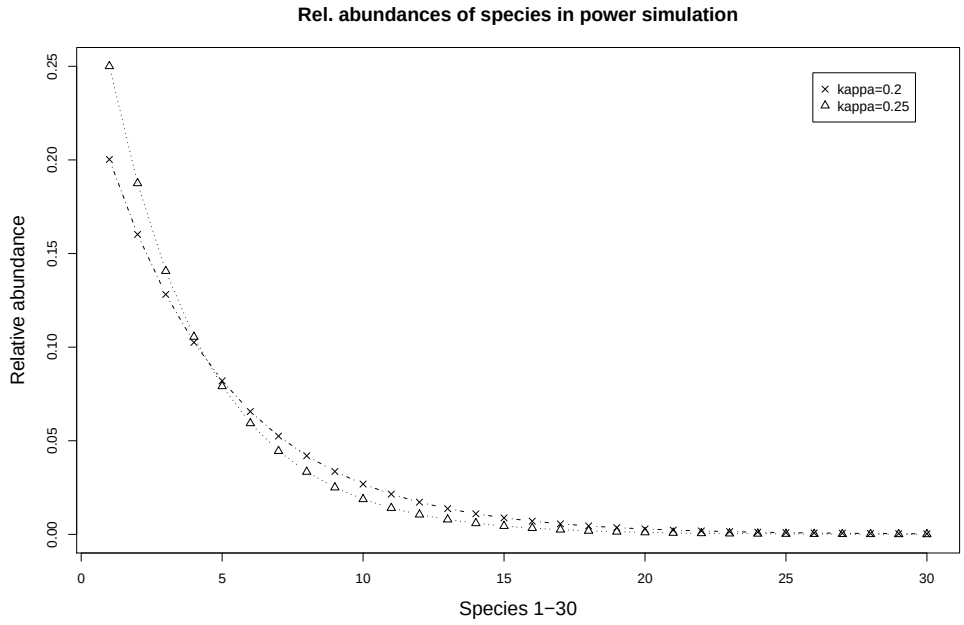


Figure 4.5: True relative abundances of species from the geometric series.

replications are sufficient, if one will analyse with the rank-*perc* method or the t_{max} sum method. With the t_{max} lm method one will only reach around 50% power with 20 traps per treatment.

None the less, the power simulation for ecological trials is questinable due to several non-influencable parameters. These are the number of species in every treatment, as well as the number of collected individuals. Further the relative abundance of the single species is an important factor. As these parameters strongly influence the power of the statistical inference, it will be difficult to determine the correct sample size.

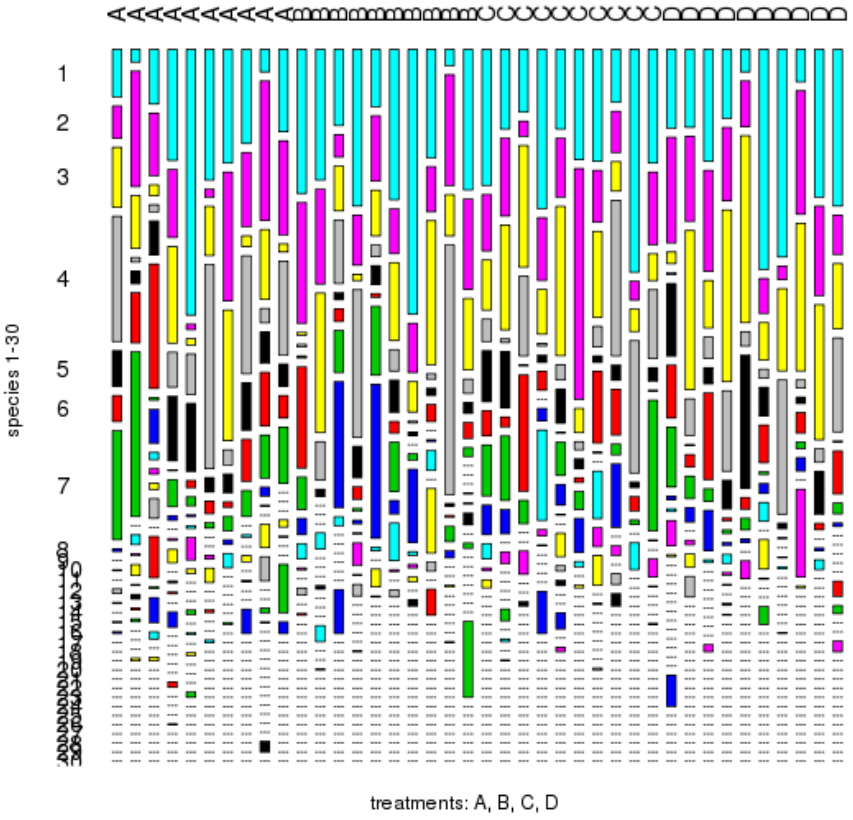


Figure 4.6: Mosaicplot for species in power analysis. Four groups with 10 replications in each – Width of bars is proportional to the total count in one trap, height of boxes within bar is proportional to the number of individuals of one species in one trap, dashed lines indicate that the species was not observed in one trap

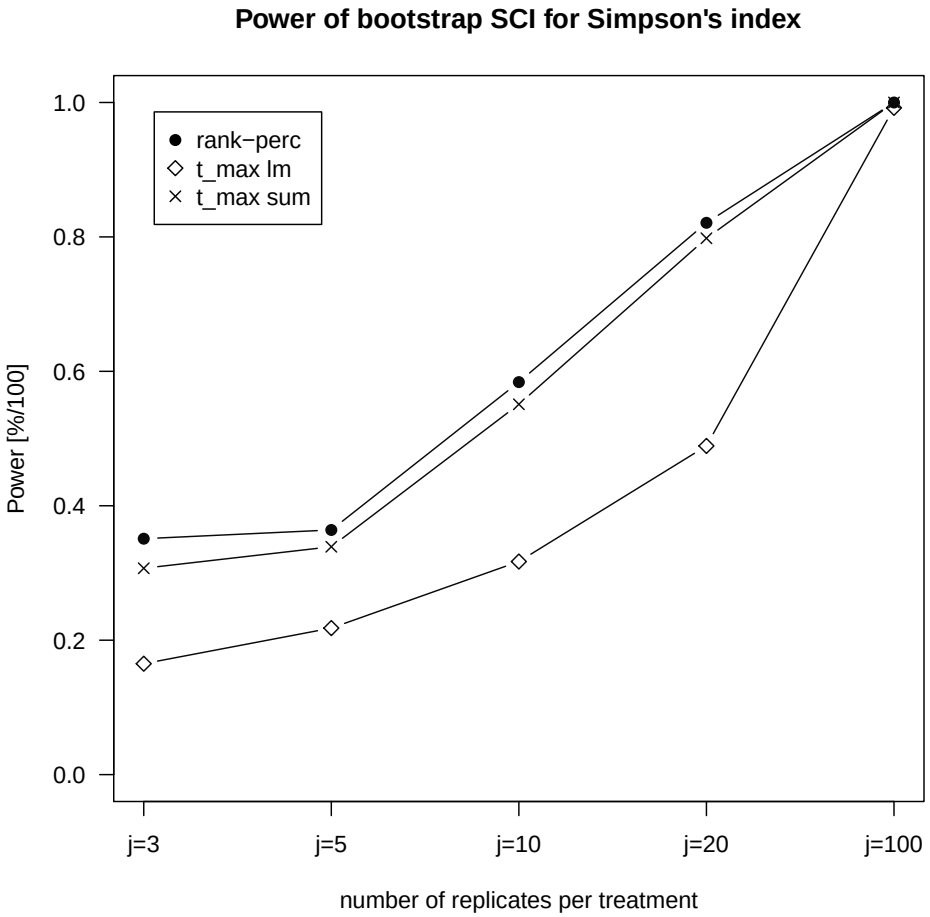


Figure 4.7: Power of the bootstrap SCIs for the Simpson index with 1000 simulation steps and 3000 bootstrap steps.

Chapter 5

Application to Real Data Sets

In this chapter I will apply the before introduced multiple comparison methods to real data sets. Both data sets were introduced in section 1.2.1 on page 3 and 1.2.2 on page 6. In the following, I will analyse them with the Shannon and the Simpson index respectively. The hypotheses were formulated before, but are repeated here in detail.

5.1 Application to data set No. I

In the following, I will analyse the data set no. 1 for the hypotheses in equation (5.1). The main aim is, as mentioned before, the proof of safety for the genetically modified organism (Bt line). This may be done with two separate hypotheses, where only one has to be accepted to prove safety. As these hypotheses can not be ordered a priori, one has to adjust for multiplicity. In detail, one hypothesis is superiority of the GM line to the pesticide treated, isogenic line (Iso-Bay), whereas the other hypothesis is non-inferiority of the GM line to the isogenic line (Iso). To prove these hypotheses, one has to determine a superiority border as well as a non-inferiority border.

Since it is hard to determine these borders without enough knowledge about biodiversity indices and the ecological background, I will leave this part open. In my opinion one has to discuss with ecologists to close this gap.

The statistically formulated hypotheses are presented in equation (5.1). To prove superiority of the GM line to the pesticide treated line the upper border of the SCI for the difference (BayIso - Bt) has to lie under a given superiority border $-\delta$. In contrast, the upper bound of the SCI for the non-inferiority (Iso - Bt) has to lie under a given non-inferiority border $+\delta$. The null hypothesis is not rejected, if both intervals cross the corresponding borders. If one interval lies with the upper bound under the corresponding superiority or non-inferiority border, the null hypothesis is rejected according to the *union-intersection principle*.

The resulting SCIs for the Shannon index are presented in figure 5.1 and the SCIs for the Simpson index are visible in figures from 5.2. In the following, I will define the test hypotheses for example no. 1, see equations (5.1).

$$H_0 : \bigcap_{i=2}^{k} \theta_i - \theta_1 \geq \delta_i \tag{5.1.0.1}$$

$$H_A : \bigcup_{i=2}^{k} \theta_i - \theta_1 < \delta_i \tag{5.1.0.2}$$

$$\delta_i = \begin{cases} positive, & \text{if } \delta_i \text{ is the non-inferiority border,} \\ negative, & \text{if } \delta_i \text{ is the superiority border.} \end{cases}$$

Here, δ_i is the corresponding superiority or non-inferiority border for the analysed contrast. Further θ_1 indicates the GMO treatment and θ_i, $i = 2, ..., k$ the standards (Iso, Bay-Iso). δ_i is positive for non-inferiority testing and negative for superiority testing, since I substract the treatment, i.e. GMO, from the standards, i.e. Iso or Bay-Iso. This yields the fact that negative differences indicate the cases one seeks for.

Regarding the resulting one-sided simultaneous confidence intervals leads to the conclusion, that superiority of GMO treatment to the pesticide treated, isogenic line can not be proven, except maybe for the asymptotic intervals for the Shannon index depending on the chosen superiority border. However, the simulation results showed, that the asymptotic intervals exhibit poor coverage probability leading to wrong decisions in statistical inference. In case of the other mcp methods no upper bound of an interval is lower than zero indicating the impossibility of lying under a given superiority border $-\delta$. In detail, the upper bounds of the SCIs for the Shannon index lay higher for the contrast (Iso - Bt), with estimators near zero, than the upper bounds for the contrast (Bay-Iso - Bt) with negative estimators. For the Simpson index the upper bounds for the contrasts are more similar, except for the t_{max} lm SCIs.

Further noticeable problems are the small intervals of the asymptotic methods for the Shannon and the Simpson index respectively, as well as the inverted estimators for the t_{max} lm SCIs in comparison to the other SCIs. Since, this method performs well for the Shannon index and for example no. 2, I do not assume a programming error. This abnormality may depend on the given example in conjunction with the Simpson index.

The proof of safety based on non-inferiority may be verified depending on the chosen non-inferiority border. As mentioned earlier I leave this part open due to the missing knowledge about the correct interpretation of differences for estimated diversity indices. This gap should be closed in discussion with ecologists.

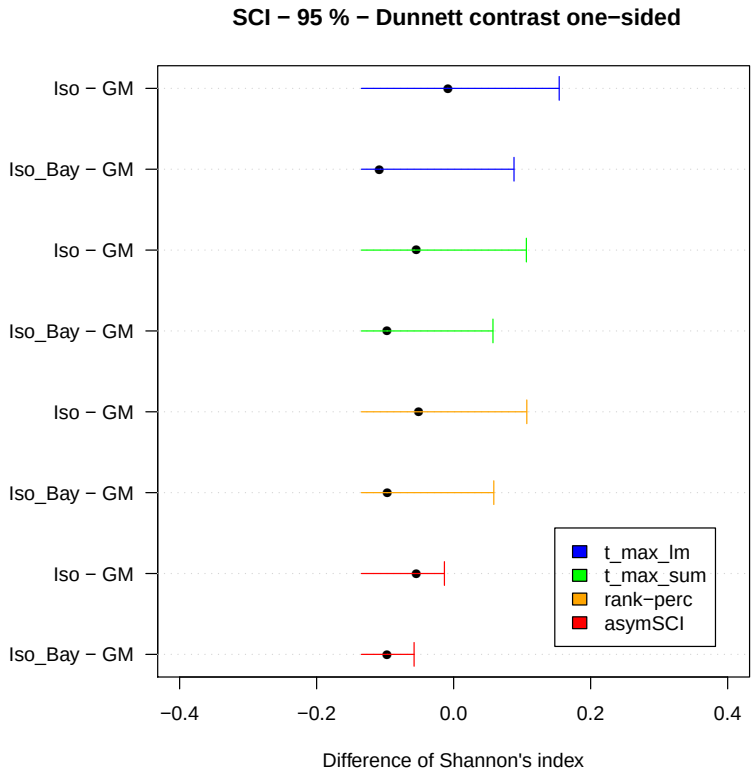


Figure 5.1: Simultaneous confidence intervals for Shannon's index in example no. 1. Methods described in section 3.3.1, 3.3.2, 3.3.3 and 3.4.2

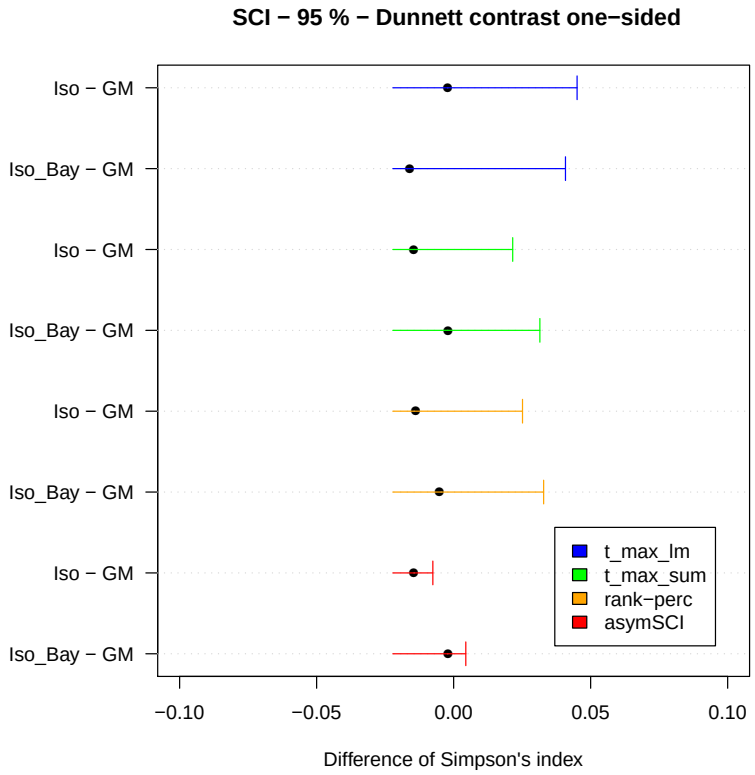


Figure 5.2: Simultaneous confidence intervals for Simpson's index in example no. 1. Methods described in section 3.3.1, 3.3.2, 3.3.3 and 3.4.1

5.2 Application to data set No. II

In the following, I will analyse the data set no. 2 for the hypotheses in equation (5.2). The analysis of the data set no. 2 is performed as in example no. 1 with the multiple comparison methods described in section 3.3 on page 30 and 3.4 on page 36. The resulting simultaneous confidence intervals are visible in picture 5.3 on page 65 for the Shannon index and in picture 5.4 on page 66 for the Simpson index.

As well as in example no. 1, the primary object is proving safety for the GMO, indicated here with the GM line. Again, there are multiple hypotheses to formulate, which can not be ordered a priori. This leads to the need of multiplicity adjusting methods. I will again use simultaneous confidence intervals, which allow to perform superiority as well as non-inferiority testing. For the following example one may use non-inferiority testing, since there is only one GM line, which will be compared with several isogenic lines (S1, S2, S3). The null hypothesis, i.e. the GM line is unsafe, will be rejected if one confidence interval for the difference (Iso - GM) lies with the upper bound under a given non-inferiority border $+\delta$, since a negative value for the difference indicates a higher diversity in the GMO treatment.

Again, we follow the *union-intersection principle* and reject the null hypothesis, if one interval lies with the upper bound under the given non-inferiority border. As ins example no. 1 I will leave the non-inferiority border determining part open, due to the problems discussed before. In the following equation (5.2) I will present the statistical hypotheses for example no.2. In this example I will only consider non-inferiority testing, since all standard treatments are untreated, isogenic lines. The main

goal is proving safety for the GMO treatment by proving non-inferiority of GMO to isogenic treatments.

$$H_0 : \bigcap_{i=2}^{k} \theta_i - \theta_1 \geq +\delta_i \tag{5.2.0.3}$$

$$H_A : \bigcup_{i=2}^{k} \theta_i - \theta_1 < +\delta_i \tag{5.2.0.4}$$

Regarding the resulting one-sided simultaneous confidence intervals in figure 5.3 on the following page and 5.4 on page 66 leads to the conclusion that the t-statistic based SCIs with summed up counts are more conservative than the other methods. The t-statistic based SCIs in the ANOVA model are similar to the percentile based SCIs based on summed up counts. The asymptotic methods are, as in example no. 1, highly liberal yielding wrong statistical conclusions. This fact is shown as well in the simulation part.

The proof of non-inferiority may be possible for the analyses depending on the chosen non-inferiority borders $-\delta$. Again, I will leave the border determining part open. In detail, the upper bounds of the SCIs for the contrasts (S1 - GM) and (S3 - GM) reach higher values than the upper bounds for the contrast (S2 - GM). All SCIs for the Shannon index lay under the difference of 0.5, except for the t-statistic based SCIs with summed up counts. Here, the upper bound for the contrasts (S1 - GM) and (S3 - GM) lays higher than 0.5. For the Simpson index all SCIs lay with the upper bound under the difference of 0.2. Again, the upper bound of the SCI for the t-statistic based SCIs, with summed up counts, lays higher than 0.2 for the contrast (S1 - GM).

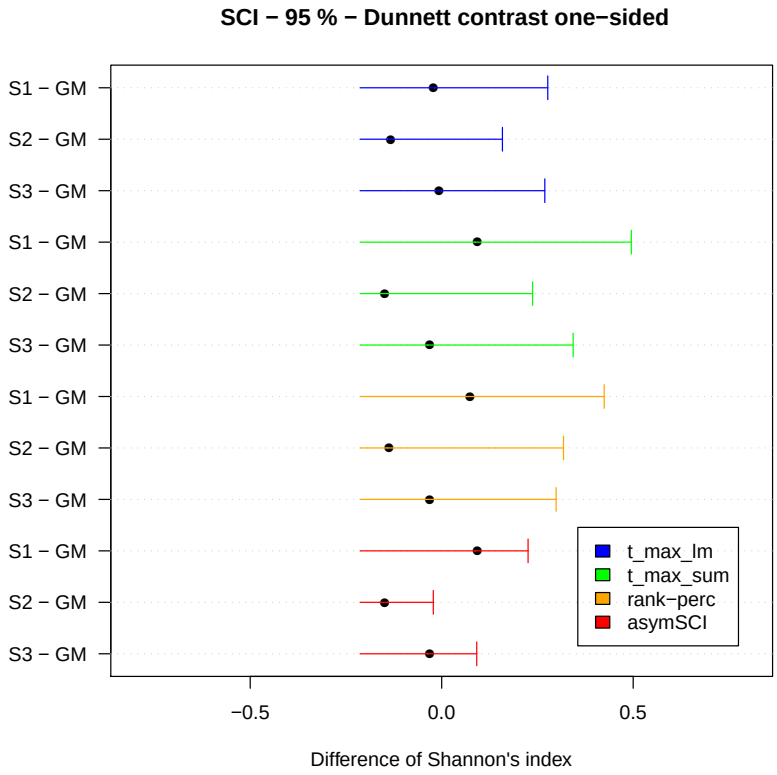


Figure 5.3: Simultaneous confidence intervals for Shannon's index in example no. 2. Methods described in section 3.3.1, 3.3.2, 3.3.3 and 3.4.2

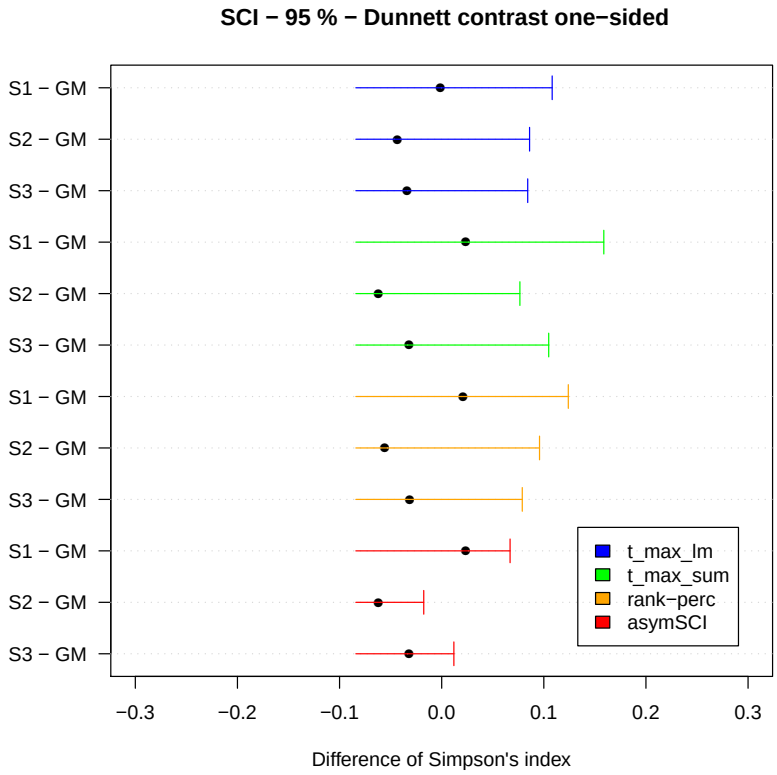


Figure 5.4: Simultaneous confidence intervals for Simpson's index in example no. 2. Methods described in section 3.3.1, 3.3.2, 3.3.3 and 3.4.1

Chapter 6

Conclusions

6.1 General Discussion

It is shown, that for multiple comparisons in ecological trials with replicated samples the asymptotic methods according to Fritsch and Hsu [1999] and Rogers and Hsu [2001] are not applicable, since they do not take the overdispersion into account. The estimated simultaneous confidence according to Fritsch and Hsu [1999] and Rogers and Hsu [2001] exhibit the same estimators and estimated variances as the simultaneous bootstrap intervals acoording to the t_{max} sum method. The only differences, which lead to smaller or larger intervals are the estimated quantiles. In case of the asymptotic methods the quantile is computed using the multivariate normal distribution and in case of the bootstrap methods the quantile is estimated out of the bootstrapped maximum distribution. This bootstrap maximum distribution takes the overdispersion into account and therefore leads to larger intervals with nearly correct coverage probability. The asymptotic methods show much too small intervals due to the normal approximated quantiles.

Further the estimated quantiles for the Shannon index show in case of the t_{max} lm method poor coverage probability. This may be explainable due to the biasness of the Shannon estimator [Magurran, 2004]. In contrast to the other two bootstrap methods, the estimators for the t_{max} lm method are calculated for each replication seperately. If the indices are estimated out of the summed up count vector, the estimation is more robust, since the number of zero counts is reduced. Since the bias correction terms for the Shannon index, see equation (2.1.1.2 on page 19), includes the number of species as well as the number of individuals ín sample, it corrects better with more observed species. The number of observed species increases with summed up counts, as one observes here more individuals in every species. In case of the row wise estimated indices the number of observed species is lower. This may first increase the bias of the Shannon index and second leads to a worse bias correction, due to the fact explained before.

One possibility to close this gap would be the estimation of the Shannon index according to Chao and Shen [2003]. But additional simulations

(not shown here) revealed no advantages of this estimation. The coverage probability of the simultaneous confidence intervals, studied here, was nearly equal with the Shannon index or the modified Shannon index by Chao and Shen [2003].

Additionally, the t_{max} lm method showed poorer coverage probability for the Shannon index with increasing sample size. This may be explainable by the estimation of smaller variances with higher sample size leading to the situation that the bias of the estimator is revealed more clearly in terms of insufficient coverage probability.

The assumption of Dirichlet-multinomial distributed counts in ecological trials was one possibility to simulate data sets. Nevertheless, this assumptions is also not the perfect solution, since the Dirichlet-multinomial distribution uses a common dispersion parameter for all species as well as an a-priori fixed number of individuals for every replication. However, the tendency to aggregate spatially or temporally (causing overdispersion) can be assumed to be species dependent and the total number of individuals is a random variable in reality. One possibility to solve this issue may be using the negative-binomial distribution to generate varying number of individuals per replicate.

To sum up, I would generally recommend to use the estimation of the indices out of the summed up counts, as this method is more robust against biased estimators. In addition the bootstrap methods, which use the estimation out of the summed up counts exhibit higher power, which will lead to smaller required sample sizes.

However, none of the methods studied here allows to include a covariable or a random block factor. As randomized block designs are often used in ecological trials, and covariables are often surveyed, this is a non negligible missing function. Further bootstrap methods demand modern computer hardware to be applied in acceptable time. But this is less problematic nowadays.

6.2 Extension and Outlook

The studied bootstrap methods for simultaneous confidence intervals may be applicable in a wider range of situation. For example in the field of genetics, where Hsu et al. [2006] published simultaneous bootstrap methods according to Beran [1988].

Despite the methods by Beran [1988], the simultaneous bootstrap confidence intervals according to Mandel and Betensky [2008] would be interesting to study in the field of diversity estimation. These intervals are similar to the one by Besag et al. [1995], except that they are adjusted for ties within one bootstrap step.

Further it may be of interest to extend the t_{max} lm method according to Westfall and Young [1993] to multifactorial designs. This would allow to take a block factor into account, which is often used in ecological field trials.

Bibliography

R. Beran. Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, 83(403):679–686, SEP 1988. ISSN 0162-1459.

J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic-systems. *Statistical Science*, 10(1):3–41, FEB 1995. ISSN 0883-4237.

Angelo Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2009. R package version 1.2-37.

G. and Berger R. L. Casella. *Statistical Inference*. Duxbury Press, 2 edition, 2001.

A. Chao and T. J. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10:429–443, 2003.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. URL `http://statwww.epfl.ch/davison/BMA/`. ISBN 0-521-57391-2.

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. ISSN 0162-1459.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

K. S. Fritsch and J. C. Hsu. Multiple comparison of entropies with application to dinosaur biodiversity., December 1999. ISSN 0006-341X.

A. Genz, F. Bretz, Te. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate Normal and t Distributions*, 2008. URL http://CRAN.R-project.org/package=mvtnorm. R package version 0.9-3.

Y. Hochberg and A.C. Tamhane. *Multiple Comparison Procedures*. Wiley, New York, 1987.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, pages 663–685, 1952.

J.C. Hsu, J.Y. Chang, and T. Wang. Simultaneous confidence intervals for differential gene expressions. *Journal of Statistical Planning and Inference*, 136(7):2182–2196, 2006.

N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate discrete distributions. 3rd ed.* Wiley Series in Probability and Statistics, 2005.

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Discrete Multivariate Distributions*. John Wiley & Sons, 1997.

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continous Multivariate Distributions*, volume 1. John Wiley & Sons, 2 edition, 2000.

M. Lloyd and R. J. Ghelardi. A table for calculating the "equitability" component of species diversity. *J. Anim. Ecol.*, 33:217–255, 1964.

A. E. Magurran. *Measuring Biological Diversity*. Blackwell Science Ltd, Malden, Oxford, Carlton, 2004.

M. Mandel and R. A. Betensky. Simultaneous confidence intervals based on the percentile bootstrap approach. *Computational Statistics & Data Analysis*, 52(4):2158–2165, Jan 2008.

R. M. May. *Patterns of species abundance and diversity*. Harvard University Press, Cambridge, MA, 1975.

P. McCullagh and P. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability 37. Chapman and Hall, 1989.

R. P. McIntosh. An index of diversity and the relation of certain concepts to diversity. *Ecology*, 48:392–404, 1967.

E. Merran, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley Series in Probability and Statistics. John Wiley & Sons, 3 edition, 2000.

J. E. Mosimann. On compound multinomial distribution, multivariate beta-distribution, and correlation among proportions. *BIOMETRIKA*, 49(1-2):65–82, 1962. ISSN 0006-3444.

K. Poortema. On modelling overdispersion of counts. *Stat. Neerl.*, 53(1): 5–20, 1999. doi: 10.1111/1467-9574.00094.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL http://www.R-project.org. ISBN 3-900051-07-0.

J. A. Rogers and J. C. Hsu. Multiple comparisons of biodiversity. *BIOMETRICAL JOURNAL*, 43(5):617–625, 2001. ISSN 0323-3847.

Frank Schaarschmidt. *BSagri: Statistical methods for safety assessment in agricultural field trials*, 2008. R package version 0.1-5.

Frank Schaarschmidt, Daniel Gerhard, and Martin Sill. *MCPAN: Multiple comparisons using normal approximation*, 2009. URL http://CRAN.R-project.org/package=MCPAN. R package version 1.1-9.

R. Schlittgen. *Einführung in die Statistik*. Lehr- und Handbücher der Statistik. Oldenburg Wissenschaftsverlag GmbH, Muenchen, 10 edition, 2003.

C. E. Shannon and W. Weaver. *The mathematical theory of communication*. University of Illinois Press, Urbana, IL, 1949.

E. H. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

J. W. Tukey. The problem of multiple comparisons. Dittoed manuscript of 396 pages. Department of Statistics, Princeton University, 1953.

P. H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing*. John Wiley & Sons, 1993.

Appendix A

Applicated Distributions

A.1 Geometric series

The geometric series splits the total number of Individuals into observed species by giving proportion k of the total number of individuals N to the first species. The second species gets proportion κ of the left number of individuals and the rest of the observed species are determined in the same way.

This yields to equation A.1.0.1, where the abundances of species are ranked from the most to the least abundant.

$$\pi_s = C_\kappa \kappa (1-\kappa)^{s-1}, \tag{A.1.0.1}$$

where κ is a constant determining the proportion of the remaining niche space occupied by a successively colonized species. Further $C_\kappa = [1-(1-\kappa)^S]^{-1}$ is a constant that insures that $\sum_{s=1}^{S} \pi_s = 1$. The geometric series is fully mathematically described in May [1975].

A.2 Multinomial distribution

Introductory example

Long before our Christian era people were using special animal bones to dice. These dices had four sides to lay on with corresponding probabilities $p_1 = p_2 = 0.4, p_3 = p_4 = 0.1$. The arising question is which distribution is applicable to model the number of throws with the different sides of the dice under condition of n - throws [Schlittgen, 2003].

Generally S random variables are multinominal distributed if the corresponding probability function reads as follows

$$p(n_1, ..., n_S) = n! \prod_{s=1}^{S} (\pi_s^{n_s} / n_s!) \tag{A.2.0.2}$$

with assumptions $n_1, n_2, ..., n_S \geq 0$ and $n_1 + n_2 + ... + n_S = n$ and is written as Multinomial$(n; \pi_1, ..., \pi_S)$.

Depending on the binomial marginal distribution with $k = 2$, follows for every random variable n_s the expectation value and variance

$$E[N_s] = n\pi_s \qquad \text{var}(N_s) = n\pi_s(1 - \pi_s).$$

The random variables X_i are correlated with covariance [Kotz et al., 1997]

$$\text{cov}(N_s, N_{s'}) = -n\pi_s\pi_{s'} \text{ for } s \neq s'.$$

Variate relationships

- The multinomial variate corresponds to the binominal variate $B(n, \pi)$ if $k = 2$ variables are described. Each N_s has the binominal distribution as the marginal distribution with parameters n, π_s [Merran et al., 2000].

A.3 Dirichlet distribution

The dirichlet distribution is the multivariate generalization of the beta distribution. Equation A.3.0.3 shows the probability density function of a Dirichlet random variable.

$$p(x_1, x_2, ..., x_S) = K \left(\prod_{s=1}^{S} x_s^{\alpha_s - 1} \right) \left(1 - \sum_{s=1}^{S} x_s \right)^{\alpha_{S+1} - 1}, \tag{A.3.0.3}$$

where

$$K = \frac{\Gamma(\sum_{s=1}^{S+1} \alpha_s)}{\prod_{s=1}^{S+1} \Gamma(\alpha_s)}$$

and $\alpha_s (s = 1, 2, ..., S + 1)$ are arbitrary positive real numbers [Kotz et al., 2000].

The Dirichlet distribution is applicable for the probability parameter calculation of a multinomial distribution to model the overdispersion. This yields a Dirichlet-multinomial with higher variance than describable by the multinomial distribution. The Dirichlet-multinomial distribution is described in appendix A.4 and in Mosimann [1962].

A.4 Dirichlet-Multinomial distribution

The Dirichlet-multinomial distribution is a compound multivariate distribution and is formed in the same way as compound univariate distributions, meaning assigning distributions to some (or all) parameters of a multivariate distribution. Let us consider

$$\text{Multinomial}(n; \pi_1, ..., \pi_S) \bigwedge_{\pi_1, ..., \pi_S} \text{Dirichlet}(\alpha_1, ..., \alpha_S), \tag{A.4.0.4}$$

leading to the probability mass function of this compound distribution

$$P(n_1, ..., n_S) = \frac{n!}{(\sum_{s=1}^{S} \alpha_s)^{[n]}} \prod_{s=1}^{S} \left\{ \frac{\alpha_s^{[n_s]}}{n_s!} \right\}, \tag{A.4.0.5}$$

where $n_s \geq 0$ and $\sum_{s=1}^{S} n_s = n$. This distribution is the multivariate generalization of the Beta-Binomial distribution discussed by Johnson et al. [2005], where

$$\text{Binomial}(n, \pi) \bigwedge_{\pi} \text{Beta}(\alpha, \beta)$$

The compound Dirichlet multinomial distribution has expectation and variance

$$E[N_s] = n\pi'_s \qquad \text{var}(N_s) = \left(\frac{n + \alpha_\bullet}{1 + \alpha_\bullet}\right) n\pi'_s(1 - \pi'_s), \tag{A.4.0.6}$$

and

$$\text{cov}(N_s, N_{s'}) = -\left(\frac{n + \alpha_\bullet}{1 + \alpha_\bullet}\right) n\pi'_s\pi'_{s'}, \tag{A.4.0.7}$$

where $\alpha_\bullet = \sum_{s=1}^{S} \alpha_s$ and $\pi'_s = \alpha_s/\alpha_\bullet$. Further $\sum_{s=1}^{S} \pi'_i = 1$.

Due to A.4.0.6 and A.4.0.7 we can see that the variance-covariance matrix of the Dirichlet-multinomial distribution is

$$\left(\frac{n + \alpha_\bullet}{1 + \alpha_\bullet}\right) \times \text{variance-covariance matrix of Multinomial}(n; \pi'_1, ..., \pi'_S). \tag{A.4.0.8}$$

Acknowledgements

I thank Prof. Dr. Ludwig A. Hothorn for giving me the opportunity to write my thesis in the field of biostatistics and for his helpful suggestions. Further I thank my supervisor Dr. Frank Schaarschmidt for his assistance, his help with programming problems and for technical as well as daily discussions. Also I thank Hannelore Visser for daily conversation and therewith involved motivation. Furthermore, I thank Prof. Dr. Armin Koch for taking me on in the Institute for Biometrics, Medical University Hannover, before finishing my thesis and the corresponding members of the institute for ideas and corrections of my thesis. Last, but not least, I thank my parents for supporting me, my sister Miriam for corrections and my wife Daniela and my dog Luna for making my closing time.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Hannover, den May 27, 2010 _____