
# Diplomarbeit

## Schätzer des Artenreichtums bei speziellen Erscheinungshäufigkeiten (species richness estimation)

vorgelegt von

Stefan Englert

betreut von

Prof. Dr. Michael Falk

13. Oktober 2009

Julius-Maximilians-Universität Würzburg

# Inhaltsverzeichnis

# 1 Einleitung und Motivation

Bei vielen Fragestellungen, in denen sich eine Grundgesamtheit in verschiedene Klassen unterteilt, ist weniger die relative Klassengröße als vielmehr die Anzahl der Klassen von Bedeutung. So interessiert sich beispielsweise der Biologe dafür, wie viele Spezien einer Gattung es gibt, der Numismatiker dafür, wie viele Münzen oder Münzprägestätten es in einer Epoche gab, der Informatiker dafür, wie viele unterschiedlichen Einträge es in einer sehr großen Datenbank gibt, der Programmierer dafür, wie viele Fehler eine Software enthält oder der Germanist dafür, wie groß der Wortschatz eines Autors war oder ist. Ganz allgemein haben diese Beispiele eine naheliegende Fragestellung gemeinsam: Wie viele Arten gibt es?

Um Einheitlichkeit zu wahren werden wir von nun an unabhängig vom Zusammenhang nicht von Spezien, Münzen, Einträgen oder Wörtern, sondern stets von der Anzahl an Arten oder Spezien sprechen.

Dieser Artenreichtum ist die einfachste und intuitivste Art und Weise eine Population oder Grundgesamtheit zu charakterisieren. Jedoch kann nur in Kollektiven, in denen die Gesamtanzahl der Bestandteile bekannt und relativ klein ist, die Anzahl der verschiedenen Spezien durch Erfassung aller bestimmt werden. In allen anderen Fällen ist es notwendig die Spezienanzahl durch Schätzungen zu bestimmen.

Mit dieser Thematik beschäftigte sich sogar die Zeitung *Die Zeit* vom 12. Februar 2009. Dort wird unter der Überschrift *Weißt du, wie viel Falter fliegen?* dargestellt, wie schwer es ist, aus den „1,75 Millionen Arten, die weltweit bisher bekannt sind," eine Schätzung für den Artenreichtum zu erhalten. Denn bis zum heutigen Tag wurde eine Vielzahl an verschiedenen Schätzmethoden entwickelt, sodass sich „je nach Schätzung (...) zwischen 10 und 90 Millionen Arten" ergeben.

Die gleichen Probleme treten in der Publikation *Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thougth* von Uwe K. Simon und Michael Weiß (2008) auf. In diesem Artikel wurden – vereinfacht dargestellt – drei verschiedene ribosomale (rDNA) Genregionen von vier verschiedenen Pilzarten nach Polymeraseketten-reaktion mit spezifischen Primern in das Bakterium *Escherichia coli* kloniert. Bei diesem Vorgang nimmt ein Bakterium nur jeweils ein Genfragment auf und vermehrt dieses durch Koloniebildung, weshalb auch selten vorkommende Kopien analysiert werden können, die bei direkter Sequenzierung ohne vorherige Klonierung unentdeckt blieben. Erfolgreich transformierte Kolonien wurden mit den PCR-Primern sequenziert. Anhand von Sequenzvergleichen konnte nun bestimmt werden, ob die jeweilige Kolonie die „Konsensus"-Variante (die am häufigsten auftretende Kopie innerhalb der aus vielen hintereinander angeordneten Kopien bestehenden rDNA) oder eine polymorphe Variante (mit einer oder mehreren Punktmutationen) trug. Für die genauen biologischen Hintergründe verweisen wir auf Simon und Weiß [SW08].

| Pilzart | Genregion | Anzahl Klonierungen | Spezien | Anteil Konsensusvariante |
|---|---|---|---|---|
| Davidiella tassiana | SSU | 50 | 37 | 28% |
| | ITS | 53 | 12 | 79% |
| | LSU | 51 | 30 | 41% |
| Mycosphaerella punctiformis | SSU | 60 | 31 | 50% |
| | ITS | 70 | 17 | 77% |
| | LSU | 55 | 20 | 65% |
| Phoma exigua var. exigua | SSU | 62 | 30 | 52% |
| | ITS | 56 | 10 | 80% |
| | LSU | 54 | 22 | 59% |
| Teratosphaeria microspora | SSU | 59 | 26 | 56% |
| | ITS | 55 | 11 | 82% |
| | LSU | 55 | 14 | 76% |

Tabelle 1.1: Daten von Uwe K. Simon und Michael Weiß

In Übereinstimmung mit unserer obigen Konvention werden die erhaltenen verschiedenen polymorphen Varianten inklusive des aufgenommenen Gens im Folgenden als Spezies bezeichnet.

Die von Uwe K. Simon und Michael Weiß gemessenen Werte finden sich die Tabelle 1.1 (vgl. auch Supplementary Table 1 und 2 zu Simon und Weiß [SW08]).

Aus den Häufigkeiten der einzelnen Spezies wurde nun die bei einer weiteren bzw. unendlich oft wiederholten Durchführung der Klonierungsvorgänge zu erwartende Anzahl an Spezies geschätzt, um so eine Vorstellung von der Gesamtvariabilität der betrachteten Genregion zu bekommen. Auch hier bestand die Problematik darin, den Schätzer zu finden, der unter diesen Rahmenbedingungen die besten Ergebnisse liefert.

Wer in einem konkreten Fall für eine spezielle Fragestellung den Artenreichtum einer Grundgesamtheit schätzen will, kann sich leicht überfordert vorkommen, denn er wird dabei mit einer Vielzahl von verschiedenen Modellen konfrontiert. Die daraus resultierenden Schätzverfahren liefern zwangsläfuig auch differente Ergebnisse. Jede dieser Methoden geht von gewissen Grundvoraussetzungen aus, die zweckmäßig erscheinen, eine spezifische Art der Rechtfertigung besitzen und in entsprechenden Situationen brauchbare und logisch nachvollziehbare Schätzwerte liefern.

Diese Diplomarbeit hat es sich deshalb zum Ziel gesetzt dem Leser ein besseres Verständnis dafür zu vermitteln, wie ausgewählte Schätzer des Artenreichtums bei den speziellen Verteilungen der Spezienhäufigkeiten, wie sie Tabelle 1.1 erwarten lässt, arbeiten, um so abschließend eine Empfehlung abgeben zu können, welches Schätzverfahren in dieser Anwendung von Herrn Simon die – in einem später noch festzulegenden Sinn – besten Ergebnisse liefert.

Um dies zu erreichen wird zuerst in Kapitel 2 eine allgemeine Einführung in die Vorgehensweise der Spezienschätzung gegeben und insbesondere für den hier anwendbaren Fall genauer dargestellt. Außerdem wird versucht die existierenden Schätzer der Spezienanzahl thematisch zu ordnen und zu kategorisieren. In Kapitel 3 werden dann die für die

Daten von Herrn Uwe Simon sinnvollen und durchführbaren Schätzer ausgewählt und die getroffenen Grundannahmen und Einschränkungen dargestellt. Des Weiteren wird für die hauptsächlich betrachteten Schätzer eine suffiziente Statistik aufgezeigt. In Kapitel 4 werden wir ausgewählte Schätzer des Artenreichtums theoretisch vorstellen und in Kapitel 5 im Rahmen einer Simulationsstudie dahingehend untersuchen, wie gut sie unter speziell vorgegebenen Rahmenbedingungen arbeiten. Mit den aus diesen Simulationen gewonnenen Erfahrungswerten zusammen mit den theoretischen Hintergründen wird dann ein für diese Situation optimales Schätzverfahren angegeben und anschließend in Kapitel 6 auf die Daten von Tabelle 1.1 angewandt. Im letzten Kapitel möchten wir dann noch einen Ausblick über die Thematik dieser Diplomarbeit hinaus geben.

# 2 Einführung in die Spezienanzahlschätzung

In diesem Kapitel wollen wir uns allgemein mit dem Konzept der Spezienanzahlschätzung beschäftigen.

Betrachtet wird eine Grundgesamtheit, die sich in verschiedene Klassen unterteilen lässt. Diese Klassen seien dadurch festgelegt, dass sich alle Elemente daraus eindeutig durch ein spezifisches Kennzeichen charakterisieren lassen und sich dadurch derselben Spezies zuordnen lassen. Wir nehmen an, dass diese Anzahl der verschiedenen Klassen oder Spezien endlich ist und die feste Größe $S$ hat. Ziel ist es also diesen Artenreichtum $S$ zu schätzen.

Wir werden dazu sogenannte multiple Capture-Recapture Experimente betrachten. In diesem Experimenttyp wird zuerst eine Stichprobe aus der Grundgesamtheit entnommen und dann die in dieser Stichprobe vorkommenden Spezien bestimmt. Anschließend wird die entnommene Stichprobe wieder der Grundgesamtheit zugeführt und aus dieser eine weitere Stichprobe gezogen und wieder die darin enthaltenen Spezien bestimmt und sowohl die zuvor schon einmal gemessenen als auch die neu festgestellten Spezien notiert. Dieses Verfahren wird bis zu einer festen Anzahl an Messdurchgängen $n$ wiederholt.

Viele Probleme bei der Spezienanzahlschätzung ergeben sich bereits bei der Konzeption des Capture-Recapture Experiments und somit vor der eigentlichen Datenerhebung. Normalerweise ist es schwer natürliche Messeinheiten zu finden, in denen die Spezien bestimmt werden sollen. Die Einteilung muss beliebig gewählt sein, gleichzeitig jedoch auch sinnvoll und durchführbar sein. Bei Spezienbestimmungen in Flächenbereichen besteht die Möglichkeit die Gesamtfläche in einzelne kleinere Bereiche aufzuteilen und diese jeweils als Messeinheit zu betrachten. Hier besteht jedoch wieder das Problem, wie groß diese kleineren Teilbereiche zu wählen sind. Bei anderen Experimenten können z. B. direkt einzelne oder eine feste Anzahl an Individuen eine Messeinheit bilden. Des Weiteren wird das Ziehen einer zufälligen Stichprobe dadurch erschwert, dass Individuen einer Spezies oft in Clustern vorkommen oder es Assoziationen zwischen verschiedenen Spezien gibt. Auf diese Problematik möchten wir jedoch nicht weiter eingehen und verweisen hierfür auf Heltshe und Forrester [HF83] sowie auf Smith und van Belle [SvB84]. Bei den Daten von Herrn Uwe Simon (siehe Kapitel 1) sind die Messeinheiten ohnehin schon natürlich vorgegeben, da die Pilzgene nacheinander einzeln kloniert wurden und demnach jeweils ein Gen als Messeinheit betrachtet werden muss. Außerdem kann davon ausgegangen werden, dass die unterschiedlichen Klonierungen sich gegenseitig nicht beeinflussen und zufällig entstehen.

Letztendlich erhält man nach Durchführung des Experiments für jeden Messdurchgang und jede Spezies die Häufigkeit ihres Erscheinens. Für unsere weiteren Betrachtungen benötigen wir jedoch nur Incidence-Daten, die dadurch charakterisiert sind, dass für

jede Spezies die Information der Anwesenheit oder Abwesenheit in dem entsprechenden Messdurchgang festgehalten wird.

Sei $i = 1, \ldots, n$ und $j = 1, \ldots, S$ und die zufällige Indikatorvariable wie folgt gegeben:

$$x_{ij} = \begin{cases} 1, & \text{falls die } j\text{-te Spezie im } i\text{-ten Messdurchgang gemessen wurde} \\ 0, & \text{sonst} \end{cases}$$

Die gemessenen Daten lassen sich z. B. als $n \times S$ Matrix $(x_{ij})$ schreiben. Der Stichprobenraum ist somit der Raum aller möglichen $2^{n \cdot S}$ solcher Matrizen.

**Beispiel 2.1.** *Sei $S = 5$ und $n = 3$. In den drei Messdurchgängen seien drei verschiedene Spezien wie folgt beobachtet worden:*

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Es ist nicht möglich diese Matrix direkt zu messen, da die Gesamtspezienanzahl $S$, die geschätzt werden soll, logischerweise unbekannt ist. Vielmehr kann nur der Teil der Matrix beobachtet werden, der aus den Spalten (Spezien) besteht, in denen mindestens ein $x_{ij} = 1$ ist und somit diese Spezies mindestens einmal bestimmt wurde. Diese Anzahl der beobachteten Spezien sei als $S_{obs}$ bezeichnet. Die Datenmatrix hat also die Größe $n \times S_{obs}$.

**Beispiel 2.2.** *Zu obigem Beispiel 2.1 lautet die Datenmatrix:*

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Wurde eine zufällige Stichprobe in dieser Weise gezogen und die zugehörige Datenmatrix bestimmt, so haben sich bisher in der Literatur sieben verschiedene Grundmodelle zur Beschreibung der Daten herausgebildet, die im Folgenden vorgestellt werden.

Es bezeichne nun $p_{ij}$ die Wahrscheinlichkeit, dass bei Messdurchgang $i$ Spezies $j$ nachgewiesen wird. In der restriktivsten und einfachsten Modellannahme, die üblicherweise als $M_0$ oder als *equal likely case* bezeichnet wird, wird angenommen, dass alle Spezien gleichwahrscheinlich sind und dass diese Wahrscheinlichkeit unabhängig vom Messdurchgang ist.

**Modell $M_0$:** Es gilt $p_{ij} = p$ für alle $i = 1, \ldots n$ und $j = 1, \ldots S$.

Ausgehend von diesem Modell können nun drei weitere Variationsmöglichkeiten betrachtet werden:

**Modell $M_t$:** Es wird angenommen, dass $p_{ij}$, also die Wahrscheinlichkeit, dass eine Spezies $j$ bei der Messdurchgang $i$ gemessen wird, unabhängig von $j$ ist, d.h. $p_{ij} = p_i$ für alle $j = 1, \ldots, S$. Für jeden Messdurchgang besitzen demnach alle Spezien die gleiche Erscheinungshäufigkeit. Diese kann jedoch mit der Stichprobenziehung variieren, z. B. sich in Abhängigkeit von der Zeit ändern.

**Modell** $M_b$**:** Die Wahrscheinlichkeit $p_{ij}$ ändert sich, wenn das Individuum in einem vorherigen Messdurchgang schon einmal gemessen wurde.

**Modell** $M_h$**:** Es gilt $p_{ij} = p_j$ für alle $i = 1, \ldots, n$, d.h. die Spezienwahrscheinlichkeiten unterscheiden sich, sind jedoch unabhängig vom jeweiligen Messdurchgang.

Durch Kombination dieser drei Modelle ergeben sich drei weitere Modellannahmen, nämlich $M_{tb}$, $M_{th}$ und $M_{tbh}$.

Wir werden im Folgenden ausschließlich das Modell $M_h$ betrachten und schließen folglich einen expliziten Einfluss der Messdurchgänge aus. Wir erlauben aber, dass die Erscheinungshäufigkeiten für die einzelnen Spezien variieren.

Für eine gegebene Stichprobe kann es im Grunde drei Hauptarten von Heterogenität geben:

1. Variation in den Spezien bezüglich ihrer Häufigkeit.

2. Variation der Häufigkeit in der Stichprobe durch Clusterung oder Abwesenheit.

3. Assoziation zwischen Spezien (Co-occurences).

Meistens wird davon ausgegangen, dass nur die erste Art der Heterogenität in der Stichprobe vorliegt und dass die anderen Arten der Heterogenität durch eine geeignete Ziehung und Randomisation der Stichprobe minimiert wurden und somit nicht weiter betrachtet werden müssen. Diese Annahme wollen wir auch treffen.

Wir betrachten ab jetzt Capture-Recapture Experimente mit genau einer Beobachtung pro Messdurchgang. Dieser Experimenttyp entspricht auch dem Versuchsaufbau von Uwe Simon.

Von ihm wurde wiederholt eine rDNA Genregion kloniert und die entstandene Spezies, diese entsprach in seinem Experiment der Zusammensetzung der Basenpaare, festgehalten. Dieses Verfahren wurde $n = 50$ bis 70-Mal wiederholt und jedes Mal die entstandene Spezies auf ihre Sequenz hin analysiert.

In obiger Schreibweise sähe eine mögliche Datenmatrix bei einer viermaligen Versuchsdurchfühung z. B. wie folgt aus:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Die Zeilensummen müssen sich dabei stets zu Eins aufsummieren, da dies genau der Individuenanzahl pro Messdurchgang entspricht. Es ist also nicht nötig die gesamte Datenmatrix zu speichern, es genügt vielmehr zeilenweise jeweils die Speziennummer anzugeben, die gemessen wurde

$$(1, 2, 3, 1).$$

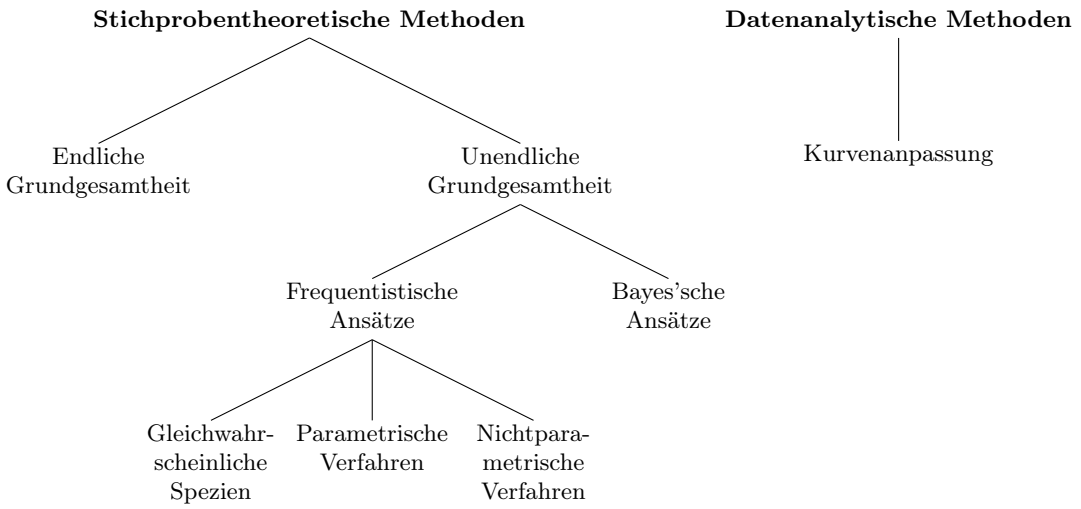


Abbildung 2.1: Kategorisierung der Herangehensweisen zur Bestimmung des Artenreichums

Wir können also die gemessene Stichprobe, die wir im Weiteren als Spezienliste bezeichnen werden, als zufälligen Zeilenvektor

$$(x_1, \ldots, x_n) \quad \text{mit } x_i \in \{1, \ldots, S\}$$

behandeln.

Unser Ziel ist es also aus einer derartig vorgegebenen Spezienliste die Gesamtanzahl der Spezien zu schätzen.

Als nächstes möchten wir eine allgemeine Einteilung und Kategorisierung der zu dieser Situation bisher in der Literatur entwickelten Schätzmethoden bezüglich ihrer statistischen Modelle vornehmen. Erst in Kapitel 3 werden wir genauer auf die Herangehensweisen und Methoden eingehen, die tatsächlich von uns betrachtet werden bzw. in unserem Fall überhaupt sinnvoll sind, um dann in Kapitel 4 die theoretischen Grundlagen zu diesen Schätzern herzuleiten.

Für das Problem, die Spezienanzahl einer Grundgesamtheit zu schätzen, existieren generell zwei grundlegend verschiedene Herangehensweisen:
Einerseits werden direkt die erhaltenen Daten modelliert, indem die Anzahl der beobachteten Spezien graphisch als eine monotone Funktion in irgendeiner Messeinheit dargestellt wird. Diese wird dann durch fest vorgegebene Funktionen approximiert um so, beispielsweise durch deren asymptotisches Verhalten, eine Spezienanzahlschätzung zu erhalten.

Andererseits wird mit stichprobentheoretischen Methoden versucht die Spezienerscheinungen zu modellieren. Hierbei wird weiter unterschieden, ob die Gesamtpopulation endlich und bekannt ist oder unendlich ist. Endliche Populationen treten dabei zum Beispiel bei sehr großen Datensätzen auf, in denen zwar die Gesamtanzahl der Elemente leicht bestimmt werden kann, aber eine Einzeluntersuchung jedes Eintrags nicht wirtschaftlich oder nicht durchführbar ist. Wesentlich häufiger treten so große Grundpopulationen auf, dass die Anzahl der Individuen nicht mehr angegeben werden kann. Diese werden als unendlich große Populationen betrachtet. Um diese zu beschreiben werden sowohl frequentistische als auch Bayes'sche Ansätze verwendet. Während bei Letzteren eine prior Verteilung der Spezienanzahl und der Erscheinungswahrscheinlichkeiten angenommen

wird und dann aus der posterior Verteilung der Spezienanzahlen unter den beobachteten Spezienhäufigkeiten heraus die Spezienanzahlen geschätzt wird, werden bei frequentistischen Ansätzen entweder parametrische oder nichtparametrische Modelle direkt auf die gemessenen Daten angewandt. Oftmals wird auch immer noch angenommen, dass alle Spezien gleichwahrscheinlich sind („equal class size“ oder Modell $M_0$). Dies mag auch historisch bedingt sein, da von diesem Modell, das etwa im Jahre 1958 von Darroch genauer studiert wurde, die gesamte Entwicklung der Spezienschätztheorie ausging.

Allen Schätzern ist dabei gemeinsam, dass die Anzahl der in einem Experiment gefundenen Spezien immer als untere Schranke des Artenreichtums betrachtet werden muss.

Die beiden verschiedenen Methodenansätze zur Schätzung des Artenreichtums sind in Abbildung 2.1 schematisch dargestellt.

# 3 Anpassung der Schätzung des Artenreichtums auf den konkreten Fall der Daten von Herrn Uwe Simon

In diesem Kapitel wollen wir uns damit beschäftigen, welche der Herangehensweisen zur Schätzung des Artenreichtums (vgl. Abbildung 2.1) im Falle von Capture-Recapture Experimenten mit genau einer Beobachtung pro Messdurchgang in unserem konkreten Fall der Klonierung ribosomaler Genregionen anwendbar sind und hierfür suffiziente Statistiken herleiten.

Zuerst wollen wir untersuchen, inwieweit stichprobentheoretische Methoden bei diesen Daten angewendet werden können.

Da prinzipiell die Klonierungsvorgänge beliebig oft wiederholt werden können, muss die Grundgesamtheit als unendlich groß betrachtet werden. Wir werden also nur Verfahren für unendliche Grundgesamtheiten betrachten.

Verdeutlicht man sich die Häufigkeitsverteilungen der Spezienerscheinungen für die Kombinationen aus Pilzart und Genregion, wie dies in Tabelle 1.1 und beispielhaft für die Pilzart *Phoma exigua var. exigua* und Genregion *LSU* in Abbildung 3.1 dargestellt ist, so erkennt man sofort, dass diese einen sehr speziellen und stark schiefen Verlauf aufweisen. Die Konsensusvariante aller Pilzarten, in Abbildung 3.1 als Speziennummer 1 bezeichnet, hat einen Anteil von 28% bis 82%, während alle anderen Spezien mit weitaus geringeren Häufigkeiten entstehen.

Um nun einen Bayes'schen Ansatz zu verfolgen würde man eine prior Struktur des unbekannten Parameters $S$, der Spezienanzahl, und der Erscheinungswahrscheinlichkeiten $p_1, \ldots, p_S$ annehmen. So wird beispielsweise in Boender und Kan [BRK87] angenommen, dass für gegebenes $S$ die Wahrscheinlichkeiten einer symmetrischen Dirichletverteilung mit Glättungsparameter $\alpha$ unterliegen. Die prior Verteilung der Spezienanzahlen wird beliebig gewählt. Ausgehend davon wird im selben Artikel die posteriori Verteilung der Anzahl der Spezien entwickelt und so eine Schätzung der Spezienanzahl ermöglicht. Dabei zeigt sich, dass diese Schätzung sehr stark von dem als bekannt vorausgesetzten $\alpha$ abhängt. Dieses $\alpha$ oder die gesamte prior Struktur müsste nun durch Expertenwissen bekannt sein, damit dieser Bayes'sche Ansatz in dieser Weise durchgeführt werden kann. Da dieses biologische Hintergrundwissen nicht verfügbar ist und Bayes'sche Schätzer generell einen beträchtlichen Rechenaufwand benötigen, werden im Folgenden keine Bayes'schen Ansätze betrachtet.

Bei den frequentistischen Ansätzen werden parametrische Verfahren, die den Zellwahrscheinlichkeiten der beobachteten Spezien eine funktionale Form unterstellen und so eine Schätzung der Gesamtanzahl der Spezien basierend auf dieser Form vornehmen, nur sehr kurz dargestellt, da generell ihre Anpassungsgüte in unserem Fall ehr gering ist. Die-

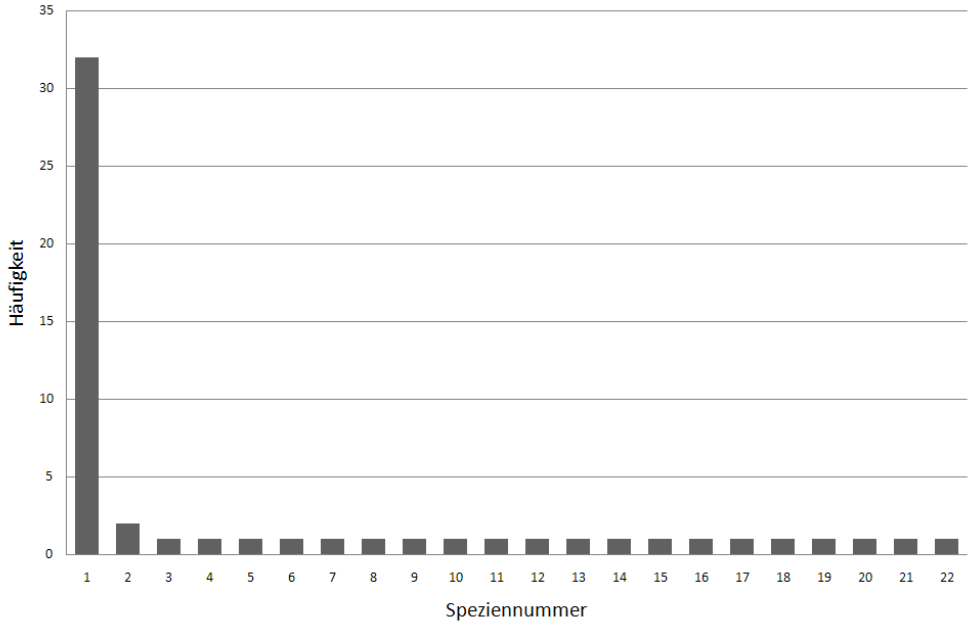


Abbildung 3.1: Verteilung der Spezienhäufigkeiten am Beispiel der Pilzart *Phoma exigua var. exigua* und Genregion *LSU*

se Schätzer können sich aufgrund der fest vorgegebenen Form nicht an die speziellen Gegebenheiten der Daten von Herrn Uwe Simon anpassen.

Der Fall der gleichwahrscheinlichen Spezien (Modell $M_0$) wird aus historischen Gründen am Schätzer von Darroch aus dem Jahre 1958 genauer dargestellt. Dieses Modell $M_0$ war Ausgangspunkt für die weitere Entwicklung der Spezienschätztheorie. Allerdings können in Anbetracht der Verteilung aus Abbildung 3.1, aus der ersichtlich wird, dass die Konsensusvariante wohl nicht gleichwahrscheinlich zu den übrigen Spezien ist, keine sonderlich guten Schätzungen erwartet werden.

Die Schwierigkeit der parametrischen bzw. der bayesianischen Schätzer liegt in der Wahl einer geeigneten Funktion bzw. prior Verteilung. Auch wenn zwei verschiedene Modelle jeweils die Daten sehr gut anpassen, können sie stark unterschiedliche Schätzwerte liefern. Ein Modell, das eine gute Anpassung an die Daten liefert, resultiert also nicht unbedingt in einem guten Wert für die Spezienanzahl. Diese Bedenken haben dazu geführt, dass nichtparametrische Verfahren entwickelt wurden, die es vermeiden Annahmen über die Spezienerscheinungswahrscheinlichkeit $p$ zu treffen. Stattdessen wird wie in Burnham und Overton [BO78] angenommen, dass $p$ eine zufällige Größe mit Verteilungsfunktion $F$ ist, sodass über diese gemittelt werden kann.

Wir geben im Folgenden eine suffiziente Statistik für nichtparametrische Verfahren an. Dazu müssen wir jedoch zunächst einige Bezeichnungen einführen.

Wie bereits in Kapitel 2 erwähnt, lassen sich die gemessenen Spezien zeilenweise in Form einer Liste $\boldsymbol{x} = (x_1, \ldots, x_n)$ mit $x_i \in \{1, \ldots, S\}$ angeben. Es bezeichne nun $\mathcal{X}$ den Raum aller Spezienlisten $\boldsymbol{x}$ und $\mathcal{B} = \mathcal{P}(\mathcal{X})$, wobei mit $\mathcal{P}(\mathcal{X})$ die Potenzmenge von $\mathcal{X}$ bezeichnet sei.

Des Weiteren seien $X_j$, die Anzahl der Spezies $j$, und $F_k$, die Anzahl der Spezien die

genau $k$-Mal vorkommen, wie folgt definiert:

$$X_j = |\{x_i = j; i = 1, \ldots, n\}|, \quad j = 1, \ldots, S$$

$$F_k = |\{X_j = k; j = 1, \ldots, S\}|, \quad k = 1, \ldots, n$$

Weiter sei $\mathcal{Y}$ der Raum aller Frequenzen der Frequenzen der Spezienhäufigkeiten, d.h. die Menge aller $(F_1, \ldots, F_n)$ mit $\sum_{k=1}^{n} k \cdot F_k = n$. Außerdem soll gelten $\mathcal{G} = \mathcal{P}(\mathcal{Y})$.

**Satz 3.1.** *Für nichtparametrische Spezienschätzungen ist*

$$\begin{aligned} T: \quad (\mathcal{X}, \mathcal{B}) \quad & \to (\mathcal{Y}, \mathcal{G}) \\ (x_1, \ldots, x_n) \quad & \mapsto (|\{|\{x_i = j; i = 1, \ldots, n\}| = k; j = 1, \ldots, S\}|)_{k=1,\ldots,n} \\ & \mapsto (|\{X_j = k; j = 1, \ldots, S\}|)_{k=1,\ldots,n} \\ & \mapsto (F_k)_{k=1,\ldots,n} = (F_1, \ldots, F_n) \end{aligned}$$

*suffizient für die Verteilungsannahme $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, wobei $\Theta$ die Menge aller möglichen Verteilungsfunktionen $F$ der Spezienerscheinungswahrscheinlichkeiten bezeichne.*

*Beweis.* Gemäß Falk [Fal07] genügt es zu zeigen, dass für alle $B \in \mathcal{B}$ die bedingte Wahrscheinlichkeit (bzgl. $P_\vartheta$) von $B \in \mathcal{B}$ bei gegebenem $T$ unabhängig von $\vartheta \in \Theta$ ist.

Es ist für $\boldsymbol{x} = (x_1, \ldots, x_n)$ und $\boldsymbol{p} = (p_1, \ldots, p_S)$, da im Modell $M_h$ kein Zeiteffekt einfließt und somit die Reihenfolge der gezogenen Spezien keine Rolle spielt:

$$P(\boldsymbol{x}|\boldsymbol{p}) = \prod_{j=1}^{S} \underbrace{p_j^{X_j} \cdot (1-p_j)^{n-X_j}}_{\text{Wkt. Spezies } j \text{ erscheint } X_j\text{-Mal und } n-X_j\text{-Mal nicht}}$$

Für nichtparametrische Spezienschätzungen nimmt man an, dass die $p_j$ einer gemeinsamen Verteilungsfunktion $F$ entstammen. Eine dieser sei $\vartheta \in \Theta$. Damit lässt sich dies nun durch Mittelung über $p$ schreiben als:

$$P_\vartheta(\boldsymbol{x}) = \prod_{j=1}^{S} \left\{ \int_0^1 p^{X_j} \cdot (1-p)^{n-X_j} \, d\vartheta(p) \right\}$$

Bildet man das Produkt nicht über die Frequenzen $X_j$ der Ereignisse, sondern über $F_i$, die Frequenzen der Frequenzen, und beachtet, dass $F_0 = S - \sum_{i=1}^{n} F_i$ gilt, so bekommt man

$$\begin{aligned} &= \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{F_0} \cdot \prod_{i=1}^{n} \left\{ \int_0^1 p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\}^{F_i} \\ &= \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{S-\sum_{i=1}^{n} F_i} \cdot \prod_{i=1}^{n} \left\{ \int_0^1 p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\}^{F_i} \\ &= \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{S} \cdot \\ &\qquad \prod_{i=1}^{n} \left[ \left\{ \int_0^1 p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\} / \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\} \right]^{F_i}. \end{aligned}$$

Für das Ereignis $E_i$, dass eine feste Spezies genau $i$-Mal erscheint, gilt wieder durch Mittelung über $p$:

$$P_\vartheta(E_i) \quad = \quad \int_0^1 \binom{n}{i} p^i \cdot (1-p)^{n-i} \, d\vartheta(p). \tag{3.1}$$

Somit erhält man, da die Fangwahrscheinlichkeiten multinomial verteilt sind:

$$\begin{aligned}
&P_\vartheta(\{T = (F_1, \ldots, F_n)\}) \\
&= \binom{S}{F_0 \cdots F_n} \cdot \prod_{i=0}^{n} (P_\vartheta(E_i))^{F_i} \\
&= \underbrace{\binom{S}{F_1} \cdot \binom{S-F_1}{F_2} \cdot \ldots \cdot \binom{S-F_1-\cdots-F_{n-1}}{F_n}}_{=:\alpha} \cdot \prod_{i=0}^{n} (P_\vartheta(E_i))^{F_i} \\
&= \alpha \cdot \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{F_0} \cdot \prod_{i=1}^{n} \left\{ \int_0^1 \binom{n}{i} p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\}^{F_i}
\end{aligned}$$

Nun verwendet man wieder $F_0 = S - \sum_{i=1}^n F_i$.

$$\begin{aligned}
&= \alpha \cdot \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{S} \cdot \\
&\qquad \prod_{i=1}^{n} \left[ \left\{ \int_0^1 \binom{n}{i} p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\} / \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\} \right]^{F_i} \\
&= \alpha \cdot \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\}^{S} \cdot \prod_{i=1}^{n} \binom{n}{i}^{F_i} \cdot \\
&\qquad \prod_{i=1}^{n} \left[ \left\{ \int_0^1 p^i \cdot (1-p)^{n-i} \, d\vartheta(p) \right\} / \left\{ \int_0^1 (1-p)^n \, d\vartheta(p) \right\} \right]^{F_i}
\end{aligned}$$

Damit können wir zeigen, dass für jedes $B \in \mathcal{B}$ die bezüglich $P_\vartheta$ gebildete bedingte Wahrscheinlichkeit von $B$ unter dem Ereignis $\{T = \boldsymbol{F}\}$ mit $\boldsymbol{F} = (F_1, \ldots, F_n)$ unabhängig von dem Parameter $\vartheta \in \Theta$ ist:

$$\begin{aligned}
P_\vartheta(B | \{T = \boldsymbol{F}\}) \quad &= \quad \frac{P_\vartheta(B \cap \{T = \boldsymbol{F}\})}{P_\vartheta(\{T = \boldsymbol{F}\})} \\
&= \quad \frac{\sum_{\boldsymbol{x} \in B \cap \{T = \boldsymbol{F}\}} P_\vartheta(\boldsymbol{x})}{P_\vartheta(\{T = \boldsymbol{F}\})} \\
&\overset{s.o.}{=} \quad \frac{\sum_{\boldsymbol{x} \in B \cap \{T = \boldsymbol{F}\}} 1}{\alpha \cdot \prod_{i=1}^{n} \binom{n}{i}^{F_i}} \\
&= \quad \frac{|B \cap \{T = \boldsymbol{F}\}|}{\alpha \cdot \prod_{i=1}^{n} \binom{n}{i}^{F_i}}
\end{aligned}$$

□

Der Satz 3.1 ermöglicht es uns die weiteren Ergebnisse kompakter zu schreiben und es kann so, insbesondere im Hinblick auf die Simulationen, Speicherplatz und Rechenzeit

gespart werden, da es nicht mehr notwendig ist die gesamte Spezienliste zu speichern. Stattdessen muss nur der Vektor $(F_1, \ldots, F_n)$ abgespeichert werden, in dem ein Großteil der Einträge gleich Null ist. Außerdem rechtfertigt dieser Satz die Nutzung von $\boldsymbol{F} = (F_1, \ldots, F_n)$ in nichtparametrischen Schätzern.

Die zuvor erwähnten Bedenken zu parametrischen Verfahren haben auch dazu geführt, dass wir hauptsächlich nichtparametrische Analysemethoden betrachten werden. Von den bisher in der Literatur verfügbaren nichtparametrischen Schätzern kann dabei in dieser Diplomarbeit jedoch nur eine Auswahl untersucht werden.

Einerseits wird der Schätzer von Chao analysiert, der so konzipiert ist, dass er eine untere Schranke der Spezienanzahl schätzt. Des Weiteren wird ein Schätzer vorgestellt, der den Sample Coverage, den Anteil der beobachteten Spezien an den Gesamtspezien, verwendet, um die Spezienanzahl zu schätzen. Außerdem betrachten wir zwei Schätzer, die die Jackknife Methode verwenden. Diese Schätzer wurden von Uwe Simon und Michael Weiß verwendet um die Spezienanzahlen in *Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thougth* zu schätzen.

Es existiert auch ein nichtparametrischer Schätzer des Artenreichtums über einen Bootstrap-Ansatz. Dieser Schätzer disqualifiziert sich jedoch in der von uns betrachteten Anwendung, da die Spezienschätzung durch die doppelte Stichprobengröße beschränkt ist und diese Einschränkung in vielen Situationen, die später in Simulationen genauer betrachteten werden, nicht zutreffend ist. Weitere Informationen zu diesem Schätzer finden sich in Smith und van Belle [SvB84].

Alternativ zu den stichprobentheoretischen Methoden existieren auch datenanalytische Methoden. Hier werden die beobachteten Spezien graphisch als eine monotone Funktion in irgendeiner Messeinheit dargestellt (Species Accumulation Curve) und an diese eine fest vorgegebene Funktion approximiert. In dem von uns betrachteten Fall ist es sinnvoll als diese Messeinheiten die Anzahl der klonierten Gene zu wählen, an die wir eine Species Accumulation Curve nach Mao anpassen werden. Aus der Vielzahl an Funktionen, die für eine Approximation dieser Kurve in Frage kommen, wird die Michaelis Menten Funktion verwendet, bei der es sich nach Raaijmakers [Raa87] um eine der gebräuchlichsten Funktionen für diese Anwendung handelt.

# 4 Theoretische Betrachtung ausgewählter Schätzer der Spezienanzahl

Um im Folgenden die in Kapitel 3 ausgewählten Schätzer des Artenreichtums theoretisch vorstellen zu können wurde eine einheitliche Notation gewählt, die sich Tabelle 4.1 entnehmen lässt.

Die theoretischen Betrachtungen der Schätzverfahren erfolgen schon im Hinblick auf die Anwendung der aufgezeigten Schätzer auf spezielle Verteilungen der Erscheinungshäufigkeiten, wie sie beispielsweise bei Klonierungen von ribosomalen Genregionen vorliegen. Einige Schätzer werden deshalb speziell auf diesen Fall hin angepasst und es werden Erweiterungen der Schätzer auf allgemeinere Modelle, wie zum Beispiel auf $M_t$ oder $M_{th}$, ausgespart.

| | |
|---|---|
| $S$ | Gesamtanzahl der Spezien im Kollektiv |
| $p_j$ | Wahrscheinlichkeit der Spezies $j$ im Kollektiv, $j = 1, \ldots, S$ |
| $S_{obs}$ | In der Stichprobe beobachtete Spezienanzahl |
| $X_j$ | Anzahl der Individuen (Frequenz) der Spezies $j$ in der Stichprobe, $j = 1, \ldots, S$. (nur die Spezien mit $X_j > 0$ wurden beobachtet) |
| $m$ | Maximale Anzahl der Individuen einer Spezies |
| $F_k$ | Anzahl der genau $k$-Mal in der Stichprobe vorkommenden Spezien, $k = 0, 1, \ldots, m$. $F_k = \lvert\{X_j = k; j = 1, \ldots, S\}\rvert$ ($F_0$ bezeichnet die Anzahl der nicht beobachteten Spezien) |
| $n$ | Anzahl der Messdurchgänge, $n = \sum_{j=1}^{S} X_j = \sum_{k=1}^{n} k \cdot F_k$ |

Tabelle 4.1: Verwendete Notation

Des Weiteren werden Schätzwerte zu bestehenden Parametern durch ein Dach dargestellt, beispielsweise $\hat{S}$, und Vektoren durch Fettschreibung gekennzeichnet. Bei Vektoren wie zum Beispiel $\boldsymbol{X}$ ergibt sich hierbei die passende Länge stets aus dem Kontext.

## 4.1 Schätzer im Falle gleichwahrscheinlicher Spezien

Das erste betrachtete Schätzverfahren des Artenreichtums ist das nach Darroch [Dar58] aus dem Jahr 1958.

Dieses Verfahren beschäftigt sich mit der Gleichwahrscheinlichkeit aller Spezien, also mit den Modellen $M_0$ und $M_t$. Dabei wird ein allgemeinerer Fall betrachtet, in dem angenommen wird, dass eine feste Anzahl an Elementen $a_i$, $i = 1, \ldots, n$ je Messdurchgang gemessen wird. Für den hier zugrundeliegenden Fall von einem Element pro Messdurchgang lässt sich die Herleitung des von Darroch postulierten Schätzers anders und kürzer durchführen.

Nimmt man nämlich an, dass alle Spezien gleichwahrscheinlich sind, so reduziert sich das Spezienschätzproblem auf den Parameter $S$, da stets

$$\sum_{j=1}^{S} p_j = \sum_{j=1}^{S} p = 1$$

gilt, und somit $p = 1/S$ gelten muss.

**Lemma 4.1.** *Die Wahrscheinlichkeit, dass eine Spezienliste genau $S_{obs}$ verschiedene Spezien enthält, ergibt sich für gleichwahrscheinliche Spezien zu*

$$\frac{1}{S^n} \frac{S!}{(S - S_{obs})!} \cdot S_{obs}^{n-S_{obs}} \frac{n!}{S_{obs}!}.$$

*Beweis.* Da alle Spezien gleichwahrscheinlich sind, haben alle $S^n$ verschiedenen Spezienlisten die gleiche Wahrscheinlichkeit und somit handelt es sich um ein Laplace-Experiment mit $|\Omega| = S^n < \infty$. Dabei gibt es $\binom{S}{S_{obs}}$ Möglichkeiten, genau $S_{obs}$ verschiedene Spezien auszuwählen. Für die verbleibenden $n - S_{obs}$ Messdurchgänge kann jeweils eine beliebige der beobachteten Spezien gemessen werden ($S_{obs}^{n-S_{obs}}$ Möglichkeiten). Da zusätzlich die Reihenfolge keine Rolle spielt, ergeben sich insgesamt $\binom{S}{S_{obs}} \cdot S_{obs}^{n-S_{obs}} \cdot n!$ Möglichkeiten eine Spezienliste mit genau $S_{obs}$ verschiedenen Spezien zu erhalten. Die gesuchte Wahrscheinlichkeit berechnet sich also zu

$$\begin{aligned} \frac{\binom{S}{S_{obs}} \cdot S_{obs}^{n-S_{obs}} \cdot n!}{S^n} &= \frac{\frac{S!}{(S-S_{obs})! \cdot S_{obs}!} \cdot S_{obs}^{n-S_{obs}} \cdot S_{obs}! \frac{n!}{S_{obs}!}}{S^n} \\ &= \frac{1}{S^n} \frac{S!}{(S - S_{obs})!} \cdot S_{obs}^{n-S_{obs}} \frac{n!}{S_{obs}!}. \end{aligned}$$

□

Betrachtet man nun die in Lemma 4.1 berechnete Wahrscheinlichkeit als Likelihoodfunktion $L(S)$ des gesuchten Parameters $S$, so lässt sich ein Schätzer der Spezienanzahl durch einen Maximum-Likelihood-Ansatz bezüglich dieser Variablen gewinnen.

**Satz 4.2** (Schätzer Darroch)**.** *Ein Schätzer des Artenreichtums im Falle gleichwahrscheinlicher Spezien ist die kleinste Lösung für $S$ größer oder gleich $S_{obs}$ von*

$$(S - 1)^n = S^{n-1}(S - S_{obs}).$$

*Beweis.* Da die betrachtete Likelihoodfunktion $L(S)$ nur für natürliche Zahlen $S$ definiert ist, ist im Hinblick auf einen Maximum-Likelihood-Ansatz der Wert gesucht, für den die

Änderung $\Delta L(S) = L(S) - L(S-1)$ eine Nullstelle hat. Der Fehler, den man hierdurch begeht, ist kleiner als eine Einheit.

$$\begin{aligned} & & L(S) &= L(S-1) \\ \Leftrightarrow & \frac{1}{S^n} \cdot \frac{S!}{(S-S_{obs})!} &= & \frac{1}{(S-1)^n} \cdot \frac{(S-1)!}{(S-1-S_{obs})!} \\ & \Leftrightarrow \quad (S-1)^n &= & S^{n-1}(S-S_{obs}) \\ & \Leftrightarrow \quad 0 &= & (S-1)^n - S^{n-1}(S-S_{obs}) \end{aligned}$$

Außerdem sind Schätzer des Artenreichtums stets durch $S_{obs}$ nach unten beschränkt. □

Die Gleichheit $S = S_{obs}$ gilt hierbei nur für $(S_{obs}-1)^n = 0 \Rightarrow S_{obs} = 1$, d.h. wenn nur genau eine Spezies gemessen wurde. Für den Fall, dass alle Spezien nur genau einmal vorkommen ($S_{obs} = n$), folgt $S = \infty$. Es ist klar, dass diese beiden Extremfälle in der Realität sehr selten auftreten. In allen anderen Fällen gilt mit $g(S) := (S-1)^n - S^{n-1}(S-S_{obs})$,

$$g(S_{obs}) = (S_{obs}-1)^n > 0$$

und zusammen mit dem Binomischen Lehrsatz

$$\lim_{S\to\infty} g(S) = \lim_{S\to\infty} \sum_{k=0}^{n-1} \binom{n}{k} S^k (-1)^{n-k} + S^{n-1} S_{obs} = -\infty,$$

sodass nach dem Zwischenwertsatz mindestens eine Nullstelle von $g(S)$ und somit Lösung der Gleichung aus Satz 4.2 existiert.

In der Anwendung wird, da die Anzahl der Spezien eine natürliche Zahl ist, für gewöhnlich der Ganzteil der Lösung aus Satz 4.2 verwendet.

## 4.2 Parametrische Schätzmethoden

In parametrischen Ansätzen wird den multinomial verteilten Zellwahrscheinlichkeiten eine funktionale Form unterstellt und die Schätzung der Gesamtanzahl der Zellen wird basierend auf dieser Form durchgeführt. Es wird also angenommen, dass $p_j = f(j)$ gilt für $j = 1, \ldots, S$ mit einer gewissen Funktion $f$.

Insofern kann man auch den Fall der gleichwahrscheinlichen Spezien als einen Spezialfall einer parametrischen Schätzmethode ansehen indem man $f \equiv p$ wählt.

Da die Zellhäufigkeiten des von uns hauptsächlich betrachteten Falls einen sehr speziellen Verlauf haben lassen sich parametrische Methoden nur bedingt anwenden und werden deshalb auch nur sehr kurz skizziert.

### 4.2.1 Schätzer von McNeil und Zipf's law

Eine parametrische Schätzmethode des Artenreichtums stammt von McNeil. Dabei wird für die Ereigniswahrscheinlichkeit $p_i$ ein gesetzmäßiger Verlauf, der auch als *Zipf's law*

bezeichnet wird, angenommen.

$$p_i = \left( i \sum_{j=1}^{S} j^{-1} \right)^{-1}, \quad i = 1, \ldots, S$$

Eine praktische Anwendung findet Zipf's law in der Linguistik, nämlich dann, wenn die Häufigkeit von Wörtern in einem Text beschrieben werden soll. Folglich wird das parametische Verfahren von McNeil hauptsächlich als Grundlage zur Schätzung der Größe des Wortschatzes eines Autors verwendet. Aufgrund der Popularität in diesem Bereich wird es auch häufig bei anderen Gelegenheiten verwendet.

**Satz 4.3** (Schätzer McNeil)**.** *Erfüllen die Zellhäufigkeiten die Bedingung („Zipf's law“)*

$$p_i = \left( i \sum_{j=1}^{S} j^{-1} \right)^{-1}, \quad i = 1, \ldots, S$$

*so erhält man einen Schätzer $\hat{S}$ für $S$ als Lösung $S^*$ von*

$$S_{obs} = S^* \left\{ 1 - E_2 \left[ \frac{n}{S^* \left( \sum_{i=1}^{S^*} i^{-1} \right)} \right] \right\}.$$

*Hierbei ist*

$$E_2(t) = \int_1^\infty \exp(-tx) x^{-2} \, dx.$$

*Beweis.* Siehe McNeil [McN73]. □

### 4.2.2 Parametrische Schätzmethode durch eine inverse Gauß-Verteilung

Bei einem anderen parametrischen Ansatz, nämlich bei der Schätzmethode durch eine inverse Gauß-Verteilung, wird angenommen, dass man das Histogramm der Zellwahrscheinlichkeiten durch eine Wahrscheinlichkeitsdichtefunktion beschreiben oder wenigstens approximieren kann. Die Schätzung der Anzahl der Zellen basiert auf den Parametern dieser Dichtefunktion.

Diese Methode geht auf Sichel (1986) zurück. Er nutzte die inverse Normalverteilung (inverse Gauß-Verteilung) um das Histogramm zu approximieren. Diese hat die Wahrscheinlichkeitsdichte

$$f_{iG}(p; \theta_1, \theta_2) = \frac{\theta_1 \theta_2^{1/2}}{2\pi^{1/2} p^{3/2}} \exp\left[ \theta_1 - \frac{p}{\theta_2} - \frac{\theta_1^2 \theta_2}{4\, p} \right], \quad p > 0,\ \theta_1 \geq 0 \text{ und } \theta_2 > 0.$$

**Satz 4.4** (Schätzer Inverse Gaussian)**.** *Lässt sich das Histogramm der Zellwahrscheinlichkeiten durch eine inverse Gauß-Verteilung approximieren, so erhält man hieraus einen Schätzer der Spezienanzahl durch*

$$\hat{S} = \frac{2}{\hat{\theta_1}\hat{\theta_2}}.$$

*Dabei ist*

$$\hat{\theta_1}\hat{\theta_2} = (1 + r)\frac{\log(nr/F_1)}{nr}$$

*wobei $r$ die Lösung von*

$$(1+r)\log(r) - \left[2\left(\frac{n}{S_{obs}}\right) - \log\left(\frac{n}{F_1}\right)\right] r + 2\left(\frac{F_1}{S_{obs}}\right) + \log\left(\frac{n}{F_1}\right) = 0$$

*ist.*

*Beweis.* Siehe Bunge und Fritzpatrick [BFH95] oder Sichel [Sic86]. □

Unter den beiden hier vorgestellten parametrischen Modellen besitzt das Schätzverfahren basierend auf einer inversen Gauß-Verteilung klare Vorteile. Es ist etwas flexibler, da die Zellwahrscheinlichkeiten nicht exakt vorgegeben sind sondern deren parametrischer Verlauf aus den Daten geschätzt wird. Wir werden deshalb nur dieses später in den Simulationen in Kapitel 5 weiter betrachten.

## 4.3 Nichtparametrische Verfahren

Wenn eine Gleichwahrscheinlichkeit aller Klassen nicht angenommen werden kann, so wurde bisher davon ausgegangen, dass dann die Erscheinungs- oder Zellhäufigkeiten einem speziellen parametrischen Verlauf unterliegen. Diese Zusatzannahme wollen wir nun aufgeben und nichtparametrische Verfahren betrachten, in denen die Erscheinungswahrscheinlichkeiten $p$ zufällige Größen mit Verteilungsfunktion $F$ sind.

### 4.3.1 Schätzer einer unteren Schranke der Spezienanzahl nach Chao

Der erste betrachtete nichtparametrische Schätzer wurde als Schätzer für eine untere Schranke der Spezienanzahl entwickelt. Dieser eignet sich besonders für „sparse Data“, d.h. für Datensätze mit vielen sehr selten vorkommenden Spezien in denen die meiste Information in $S_{obs}$, $F_1$ und $F_2$ enthalten ist.

**Satz 4.5** (Schätzer Chao)**.** *Ist $p$ eine zufällige Größe mit Verteilungsfunktion $F$ und gilt $E(F_2) \neq 0$, so lässt sich zeigen*

$$E(F_0) \geq \left(\frac{n-1}{n}\right) \frac{E(F_1)^2}{2\,E(F_2)}.$$

*Beweis.* Im Folgenden sei $E(F_2) \neq 0$. Aus Gleichung (3.1) aus dem Beweis zu Satz 3.1 erhält man folgenden Erwartungswert

$$E(\text{„eine feste Spezies } j \text{ erscheint genau } k \text{ Mal“}) = \int_0^1 \binom{n}{k} p^k \cdot (1-p)^{n-k}\, dF(p),$$

der unabhängig von $j$ ist.

$\Rightarrow$

$$\begin{aligned} E(F_k) &= \sum_{j=1}^{S} E(\text{„eine feste Spezies } j \text{ erscheint genau } k \text{ Mal“}) \\ &= \sum_{j=1}^{S} \int_0^1 \binom{n}{k} p^k \cdot (1-p)^{n-k} \, dF(p) \\ &= S \cdot \int_0^1 \binom{n}{k} p^k \cdot (1-p)^{n-k} \, dF(p) \end{aligned} \tag{4.1}$$

Da $F_k$ nur endlich viele Werte annehmen kann, existiert der in Gleichung (4.1) berechnete Erwartungswert. Außerdem gilt nach der Cauchy-Bunjakowski-Schwarz-Ungleichung[1] für Wahrscheinlichkeitsintegrale

$$\begin{aligned} &\left[\int (1-p)^n \, dF(p)\right] \cdot \left[\int p^2 \cdot (1-p)^{n-2} \, dF(p)\right] \\ &= \left[\int \left((1-p)^{\frac{n}{2}}\right)^2 \, dF(p)\right] \cdot \left[\int \left(p \cdot (1-p)^{\frac{n}{2}-1}\right)^2 \, dF(p)\right] \\ &\geq \left[\int (1-p)^{\frac{n}{2}} \cdot p \cdot (1-p)^{\frac{n}{2}-1} \, dF(p)\right]^2 \\ &= \left[\int p \cdot (1-p)^{n-1} \, dF(p)\right]^2. \end{aligned}$$

$\Leftrightarrow$

$$\begin{aligned} &\binom{n}{0}^{-1} \cdot S \cdot \left[\int \binom{n}{0} (1-p)^n \, dF(p)\right] \cdot \binom{n}{2}^{-1} \cdot S \cdot \left[\int \binom{n}{2} p^2 \cdot (1-p)^{n-2} \, dF(p)\right] \\ &\geq \left[\binom{n}{1}^{-1} \cdot S \cdot \int \binom{n}{1} p \cdot (1-p)^{n-1} \, dF(p)\right]^2 \end{aligned}$$

$\overset{(4.1)}{\Leftrightarrow}$

$$E(F_0) \cdot E(F_2) \cdot \left(\frac{n \cdot (n-1)}{2}\right)^{-1} \geq n^{-2} \cdot E(F_1)^2$$

$\Leftrightarrow$

$$E(F_0) \geq \left(\frac{n-1}{n}\right) \frac{E(F_1)^2}{2 \, E(F_2)}$$

□

*Bemerkung* 4.6. Die in diesem Satz 4.5 verwendete Voraussetzung $E(F_2) \neq 0$ ist für endliche Gesamtkollektive ($S < \infty$) identisch zu $S \neq 1$. Nimmt man nämlich an, dass es mindestens zwei verschiedene Spezien gibt mit Häufigkeiten $p_k$ und $p_l$ mit $0 < p_k < 1$ und $0 < p_l < 1$ und dass $E(F_2) = 0$ gilt, so folgt

$$\begin{aligned} E(F_2) &= \sum_{i=1}^{S} \underbrace{\binom{n}{2} p_i^2 \cdot (1-p_i)^{n-2}}_{\geq 0} = 0 \\ &\Rightarrow \forall_{i=1}^{S} \binom{n}{2} p_i^2 \cdot (1-p_i)^{n-2} = 0, \end{aligned}$$

[1] Ein Beweis dieser Ungleichung findet sich beispielsweise in Werner [Wer06, Seite 315].

was einen Widerspruch zur Annahme liefert, da beispielsweise für $p_k$ gilt $\binom{n}{2} p_k^2 \cdot (1 - p_k)^{n-2} \neq 0$. Die umgekehrte Implikation $S = 1 \Rightarrow E(F_2) = 0$ ist klar.

Es wäre nun möglich zusammen mit $S = S_{obs} + F_0$ einen Schätzer einer unteren Schranke der Spezienanzahl dadurch anzugeben, dass man die Größe $F_0$ zusammen mit Satz 4.5 schätzt, indem man die entsprechenden Erwartungswerte durch die tatsächlich gemessenen Größen ersetzt.

$$\hat{S}_{chao} = S_{obs} + \frac{n-1}{n} \frac{F_1^2}{2\, F_2}$$

Manchmal wird zusätzlich noch $n$ als groß angesehen, sodass man $\frac{n-1}{n}$ approximativ durch 1 ersetzten kann.

Die Fähigkeiten dieses Schätzers eine untere Grenze des Artenreichtums zu schätzen ist, wie in Chao [Cha84] gezeigt wurde, bereits sehr ermutigend, insbesondere wenn $S_{obs}$, $F_1$ und $F_2$ einen Großteil der Informationen aus der Stichprobe enthalten.

Wir können nun den Erwartungswert von $\hat{S}_{chao}$ abschätzen, da $1/x$ eine konvexe Funktion für $x \in \mathbb{R}^+$ ist und somit die Jensensche Ungleichung[2] angewendet werden kann. Weiter verwenden wir, dass die Zufallsvariable $F_k$ nur endlich viele Werte annimmt und somit einen endlichen Erwartungswert und eine endliche Varianz besitzt und folglich quadratintegrierbar ist. Wir können also auch die Cauchy-Schwarz Ungleichung für Erwartungswerte[3] angewenden und erhalten

$$\begin{aligned}
E(\hat{S}_{chao}) &= E(S_{obs}) + \frac{n-1}{n} \cdot E\left(\frac{F_1^2}{2\, F_2}\right) \\
&= S - E(F_0) + \frac{n-1}{n} \cdot E\left(F_1^2 \cdot \frac{1}{2\, F_2}\right) \\
&= S - E(F_0) + \frac{n-1}{n} \cdot E(F_1^2) \cdot E\left(\frac{1}{2\, F_2}\right) + \\
&\quad \frac{n-1}{n} \cdot Cov\left(F_1^2, \frac{1}{2\, F_2}\right) \\
&\overset{\text{Cauchy-Schwarz}}{\geq} S - E(F_0) + \frac{n-1}{n} \cdot E(F_1)^2 \cdot E\left(\frac{1}{2\, F_2}\right) + \\
&\quad \frac{n-1}{n} \cdot Cov\left(F_1^2, \frac{1}{2\, F_2}\right) \\
&\overset{\text{Jensensche Ungleichung}}{\geq} S - E(F_0) + \frac{n-1}{n} \cdot E(F_1)^2 \cdot \frac{1}{E(2\, F_2)} + \\
&\quad \frac{n-1}{n} \cdot Cov\left(F_1^2, \frac{1}{2\, F_2}\right).
\end{aligned}$$

[2] Jensensche Ungleichung: Sei $I \subseteq \mathbb{R}$ ein Intervall. Ist $f : I \to \mathbb{R}$ eine konvexe Funktion und $X : \Omega \to I$ eine Zufallsvariable deren Erwartungswert existiert, dann gilt $f(E(X)) \leq E(f(X))$.
Ein Beweis hierfür findet sich in Irle [Irl05, Seite 140].

[3] Ein Beweis dieser Ungleichung findet sich beispielsweise in Irle [Irl05, Seite 133].

Zusätzlich gilt im equal likely Fall

$$\begin{aligned}\frac{E(F_1)^2}{2\,E(F_2)} &\overset{(4.1)}{=} \frac{\left(n\cdot p\cdot(1-p)^{n-1}\cdot S\right)^2}{\frac{(n-1)\cdot n}{2}\cdot 2\cdot p^2\cdot(1-p)^{(n-2)}\cdot S}\\ &= \frac{n}{n-1}(1-p)^n\cdot S\\ &= \frac{n}{n-1}E(F_0).\end{aligned}$$

D.h. obige Ungleichung aus Satz 4.5 aus der der Schätzer hergeleitet wurde wird im Falle gleichwahrscheinlicher Spezien zu einer Gleichung und die untere Schranke wird angenommen. In diesem Fall gilt dann

$$E(\hat{S}_{chao}) \geq S + \frac{n-1}{n}\cdot Cov\left(F_1^2, \frac{1}{2\,F_2}\right),$$

sodass, falls $Cov\left(F_1^2, \frac{1}{2\,F_2}\right) \geq 0$ gilt, der untere Schranken-Schätzer mit einem Bias behaftet ist und i. A. keine untere Schranke geschätzt wird.

Da außerdem der Schätzer *Chao* nur die Informationen aus $F_1$ und $F_2$ enthält, muss diese equal likely Annahme nicht für alle Spezien zutreffen, sondern es genügt, dass diese für alle „seltenen“, d.h. nicht häufiger als zwei Mal vorkommenden Spezien erfüllt ist.

Insofern ist diese Erkenntnis auch für die von uns betrachteten Erscheinungshäufigkeiten von Belang, da, obwohl unsere Daten sehr weit vom equal likely Fall entfernt zu sein scheinen, diese Gleichwahrscheinlichkeitsannahme für die seltenen Spezien durchaus eine gültige Modellannahme ist.

Um diese Problematik zu umgehen wird, beispielsweise in der Onlinehilfe zu EstimateS [Col05], eine Bias-Corrected Version des Schätzers von Chao vorgeschlagen

$$\hat{S} = S_{obs} + \frac{F_1(F_1-1)}{2(F_2+1)}.$$

Diese hat den Vorteil, dass sie auch für den Fall $F_2 = 0$ berechenbar bleibt.

Falls $F_2 \neq 0$ gilt jedoch stets

$$\frac{F_1(F_1-1)}{2(F_2+1)} \leq \frac{F_1^2}{2F_2}.$$

Entsprechendes gilt, falls die Erwartungswerte $E(F_1)$ und $E(F_2)$ verwendet werden, sodass diese vorgeschlagene Bias-Corrected Version des Schätzers generell kleiner ist als die ursprüngliche.

Persönlicher E-Mail Kontakt mit Frau Prof. Dr. Anne Chao vom 20. Mai 2009 zu diesem Thema:

> The non-bias-corrected estimator is theoretically a lower bound for all situations. However, in homogeneous case (that is, all species have the same abundances; or all species are equally-likely) the lower bound is equal to the true number of species (the bound is attained). Thus, when we estimate it from sample data, the estimate in the homogeneous case slightly overestimates (especially when $F_2 = 0$). Under a homogeneous case, we can assess the

bias, and get that bias-corrected estimator (and this is why the bias-corrected one is less than the un-corrected one). The bias-corrected estimator is valid only for the homogeneous case. In non-homogeneous case, bias-correction is not needed.

Wir wollen nun eine Bias Korrektur vornehmen, die unabhängig von der Annahme gleicher Spezienhäufigkeiten und nur unter den gewöhnlichen Annahmen für nichtparametrische Schätzverfahren des Artenreichtums durchgeführt werden kann, da, wie wir vorher gezeigt haben, eine Bias Korrektur mitunter auch im Falle ungleicher Spezienhäufigkeiten notwendig ist, z. B. wenn nur die seltenen Spezien gleichwahrscheinlich sind.

Sei nun $E(F_2) > 0$ angenommen.

$$\begin{aligned} E(F_0) &\overset{\text{Satz 4.5}}{\geq} \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \\ &= \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0) + \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 = 0) \\ &\geq \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0) \end{aligned}$$

Grundidee der Bias Korrektur ist es nun keinen Schätzer für $E(F_0)$ sondern für

$$\frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0)$$

zu entwickeln. Dabei handelt es sich dann, wie wir später nachweisen werden, wirklich um einen unverzerrten Schätzer einer unteren Schranke der Spezienanzahl, sodass wir von einer Bias Korrektur sprechen können. Dazu benötigen wir jedoch noch ein zusätzliches Lemma, das zum Beispiel im Web Appendix zu Rivest und Baillareon [RB07] angegeben ist.

**Lemma 4.7.** *Seien $(X_1, X_2, X_3, X_4)$ zufällige Variablen mit einer Multinomialverteilung mit Parametern $N$ und $(p_1, p_2, p_3, p_4)$, mit $\sum_{i=1}^{4} p_i = 1$. Falls $t_1$, $t_2$ und $t_3$ nichtnegative ganze Zahlen sind, dann ist*

$$\begin{aligned} &E\left(\frac{X_1(X_1-1)\dots(X_1-t_1+1)X_2(X_2-1)\dots(X_2-t_2+1)}{(X_3+1)\dots(X_3+t_3)}\right) \\ &= \frac{p_1^{t_1} p_2^{t_2}}{p_3^{t_3}} \frac{N!}{(N-t_1-t_2+t_3)!} \cdot \\ &\qquad \left(1 - \sum_{k=0}^{t_3-1} \frac{(N-t_1-t_2+t_3)!}{k!(N-t_1-t_2+t_3-k)!} p_3^k (1-p_3)^{N-t_1-t_2+t_3-k}\right). \end{aligned}$$

**Korollar 4.8.** *Es gilt*

$$\frac{E(F_1)^2}{2\,E(F_2)} \cdot \left(1 - \left(1 - \frac{E(F_2)}{S}\right)^{S-1}\right) = E\left(\frac{F_1(F_1-1)}{2(F_2+1)}\right).$$

*Beweis.* Wir setzen in Lemma 4.7 $N = S$, $X_1 = F_1$, $X_2 = F_0$ und $X_3 = F_2$. Gleichzeitig wählen wir $X_4$ als Komplement von $X_1$, $X_2$ und $X_3$. Aus den Beweisen von Satz 3.1 und

Satz 4.5 wissen wir, dass die Wahrscheinlichkeit, dass eine beliebige aber feste Spezies genau $k$ Mal erscheint, gegeben ist als $\int_0^1 \binom{n}{k} p^k \cdot (1-p)^{n-k}\, dF(p) = \frac{E(F_k)}{S}$. Außerdem entstehen die einzelnen Spezien unabhängig voneinander, sodass $(X_1, X_2, X_3, X_4)$ multinomialverteilt ist mit $p_1 = \frac{E(F_1)}{S}$, $p_2 = \frac{E(F_0)}{S}$ und $p_3 = \frac{E(F_2)}{S}$. Durch die Wahl von $X_4$ ist dabei die Zusatzbedingung $\sum_{i=1}^4 p_i = 1$ erfüllt. Nun erhält man mit $t_1 = 2$, $t_2 = 0$ und $t_3 = 1$ die Behauptung direkt aus Lemma 4.7. □

Zusätzlich wissen wir aus den Beweisen von Satz 3.1 und Satz 4.5, dass die Wahrscheinlichkeit, dass eine beliebige aber feste Spezies genau zwei Mal erscheint, gegeben ist als $\int_0^1 \binom{n}{2} p^2 \cdot (1-p)^{n-2}\, dF(p) = \frac{E(F_2)}{S}$. Somit ist $1 - \frac{E(F_2)}{S}$ die Wahrscheinlichkeit, dass diese Spezies nicht genau zwei Mal vorkommt. Nimmt man nun an, dass die einzelnen Spezien unabhängig voneinander entstehen, so gilt

$$P(F_2 = 0) = \left(1 - \frac{E(F_2)}{S}\right)^S$$

und somit

$$P(F_2 \neq 0) = 1 - \left(1 - \frac{E(F_2)}{S}\right)^S.$$

Damit lässt sich nun ein Schätzer für den oben angegebenen Term entwickeln

$$\begin{aligned}
\frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0) \quad &= \quad \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot \left(1 - \left(1 - \frac{E(F_2)}{S}\right)^S\right) \\
&\geq \quad \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot \left(1 - \left(1 - \frac{E(F_2)}{S}\right)^{S-1}\right) \\
&\overset{\text{Korollar 4.8}}{=} \quad \frac{n-1}{n} \cdot E\left(\frac{F_1(F_1-1)}{2(F_2+1)}\right).
\end{aligned}$$

Insgesamt können wir nun den folgenden Satz zeigen.

**Satz 4.9** (Schätzer Chao Bias Corrected)**.** *Der nichtparametrische Schätzer des Artenreichtums*

$$S_{obs} + \frac{n-1}{n} \cdot \frac{F_1(F_1-1)}{2(F_2+1)}$$

*ist unverzerrt als Schätzer einer unteren Schranke der Spezienanzahl.*

*Beweis.*

$$\begin{aligned}
E\left(S_{obs} + \frac{n-1}{n} \cdot \frac{F_1(F_1-1)}{2(F_2+1)}\right) \quad &\leq \quad E(S_{obs}) + \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0) \\
&\leq \quad E(S_{obs}) + \frac{n-1}{n} \cdot \frac{E(F_1)^2}{2\,E(F_2)} \\
&\overset{\text{Satz 4.5}}{\leq} \quad E(S - F_0) + E(F_0) \\
&= \quad S.
\end{aligned}$$

□

Interessant ist, dass dieser Schätzer bis auf den Faktor $\frac{n-1}{n}$ mit der in EstimateS [Col05] vorgeschlagenen Bias-Corrected Version des Schätzers von Chao übereinstimmt und dass dieser Schätzer im Gegensatz zu dem ursprünglichen Schätzer von Chao auch im Fall $F_2 = 0$ berechenbar bleibt.

Es ist nun also auch möglich eine untere Schranke der Spezienanzahl im Falle von $F_2 = 0$ anzugeben. Satz 4.9 trifft jedoch nur eine Aussage über den Erwartungswert des *Schätzers Chao Bias Corrected*, sodass Ergebnisse mit $F_2 = 0$ weiterhin kritisch betrachtet werden müssen.

Nach unserer Herleitung und den anfangs geäußerten Bedenken bezüglich des Schätzers nach Chao für die von uns betrachteten Daten empfehlen wir den Schätzer aus Satz 4.9 in allen Gelegenheiten, bei denen man eine untere Schranke der Artenvielfalt schätzen möchte.

### 4.3.2 Schätzer über Sample Coverage

Im Folgenden wollen wir uns mit einer weiteren nichtparametrischen Schätztechnik beschäftigen, die den Artenreichtum über den *Sample Coverage* $C$ schätzt. Hierunter versteht man die Summe der Wahrscheinlichkeiten der beobachteten Spezien.

$$C = \sum_{i=1}^{S} p_i \, I[X_i > 0],$$

wobei $I[A]$ die gewöhnliche Indikatorfunktion bezeichnet.

Im Falle gleichwahrscheinlicher Spezien gilt $p_i = p = \frac{1}{S}$ für alle $i = 1, \ldots, S$ und somit ergibt sich $C = S_{obs}/S$. In diesem Fall ließe sich der Artenreichtum wie folgt schätzen

$$\hat{S}_1 = \frac{S_{obs}}{\hat{C}}. \tag{4.2}$$

Diese Idee wollen wir nun mit den folgenden beiden Sätzen verallgemeinern.

**Satz 4.10.** *Angenommen man zieht eine Spezienliste der Größe $n$ aus einer Population mit $S$ Spezien mit Häufigkeiten $\boldsymbol{p} = (p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$ und es bezeichne $\bar{p}$ den Mittelwert, $\bar{\boldsymbol{p}} = (\bar{p}, \ldots, \bar{p})$ den Mittelwertsvektor und $\gamma$ den Stichprobenvariationskoeffizienten von $\boldsymbol{p}$, so gilt*

$$\frac{E(S_{obs})}{E(C)} = S - \frac{n \cdot (1-\bar{p})^{n-1}}{E(C)} \cdot \gamma^2 + o(|\boldsymbol{p} - \bar{\boldsymbol{p}}|^2).$$

*Beweis.* Für festes $\boldsymbol{p}$ erhält man

$$\begin{aligned} E(S_{obs}) &= S - \sum_{i=1}^{S} E(I[X_i = 0]) \\ &= S - \sum_{i=1}^{S} (1-p_i)^n \end{aligned}$$

und

$$\begin{aligned} E(C) &= E\left(\sum_{i=1}^{S} p_i\, I[X_i > 0]\right) \\ &= \sum_{i=1}^{S} p_i\, E(I[X_i > 0]) \\ &\overset{X_i \geq 0}{=} 1 - \sum_{i=1}^{S} p_i\, E(I[X_i = 0]) \\ &= 1 - \sum_{i=1}^{S} p_i \cdot (1-p_i)^n \end{aligned}$$

$\Rightarrow$

$$\begin{aligned} \frac{E(S_{obs})}{E(C)} &= \frac{S - \sum_{i=1}^{S}(1-p_i)^n}{1 - \sum_{i=1}^{S} p_i(1-p_i)^n} \\ &= \frac{S \cdot \left(1 - \sum_{i=1}^{S} p_i(1-p_i)^n\right) + \overbrace{S \cdot \sum_{i=1}^{S} p_i(1-p_i)^n - \sum_{i=1}^{S}(1-p_i)^n}^{=:g(\boldsymbol{p})}}{E(C)} \\ &= S + \frac{g(\boldsymbol{p})}{E(C)} \end{aligned} \tag{4.3}$$

Wir entwickeln nun $g(\boldsymbol{p})$ bei $\bar{\boldsymbol{p}}$ mehrdimensional nach Taylor.

$$\begin{aligned} g(\boldsymbol{p}) &= g(\bar{\boldsymbol{p}}) + \sum_{i=1}^{S} \left.\frac{\partial g(\boldsymbol{p})}{\partial p_i}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} (p_i - \bar{p}) + \frac{1}{2} \cdot \sum_{i=1}^{S} \left.\frac{\partial^2 g(\boldsymbol{p})}{\partial p_i^2}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} (p_i - \bar{p})^2 \\ &\quad + \sum_{i<j} \left.\frac{\partial^2 g(\boldsymbol{p})}{\partial p_i\, \partial p_j}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} (p_i - \bar{p})(p_j - \bar{p}) + o(|\boldsymbol{p} - \bar{\boldsymbol{p}}|^2) \end{aligned} \tag{4.4}$$

Zunächst ist $\bar{p} = \frac{1}{S}$. Damit ist der nullte Entwicklungsgrad

$$g(\bar{\boldsymbol{p}}) = 0$$

und die partiellen Ableitungen erster Ordnung von $g$ an der Stelle $\bar{\boldsymbol{p}}$ sind unabhängig von $\boldsymbol{p}$. Genauer gilt für alle $i = 1, \ldots, S$

$$\begin{aligned} \left.\frac{\partial g(\boldsymbol{p})}{\partial p_i}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} &= S \cdot (1-p_i)^n - S \cdot p_i \cdot n \cdot (1-p_i)^{n-1} + n \cdot (1-p_i)^{n-1}\big|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} \\ &= S \cdot \left(1 - \frac{1}{S}\right)^n. \end{aligned} \tag{4.5}$$

Somit verschwindet wegen $\sum_{i=1}^{S}(p_i - \bar{p}) = 0$ auch der erste Entwicklungsgrad. Da außerdem nach Gleichung (4.5) $\frac{\partial g(\boldsymbol{p})}{\partial p_i}$ unabhängig von $p_j$ ist für $j \neq i$ und alle $i = 1, \ldots, S$, verschwinden alle gemischten Ableitungen. Ebenso folgt nach einiger Rechnung

$$\begin{aligned} \left.\frac{\partial^2 g(\boldsymbol{p})}{\partial p_i^2}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}} &= -2nS \cdot (1-\bar{p})^{n-1} \\ &= \frac{-2n \cdot (1-\bar{p})^{n-1}}{\bar{p}}. \end{aligned}$$

Setzt man nun diese Ergebnisse in Gleichung (4.4) ein, so erhält man

$$\begin{aligned} g(\boldsymbol{p}) &= \frac{1}{2}\frac{-2n\cdot(1-\bar{p})^{n-1}}{\bar{p}}\sum_{i=1}^{S}(p_i-\bar{p})^2 + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2) \\ &= -n\cdot(1-\bar{p})^{n-1}\frac{\frac{1}{S}\sum_{i=1}^{S}(p_i-\bar{p})^2}{\bar{p}^2} + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2) \\ &= -n\cdot(1-\bar{p})^{n-1}\gamma^2 + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2). \end{aligned}$$

Insgesamt ergibt sich nun aus Gleichung (4.3) die Behauptung. □

**Satz 4.11** (Schätzer über Sample Coverage)**.** *Mit den Bezeichnungen aus Satz 4.10 gilt*

$$S = \frac{E(S_{obs})}{E(C)} + \frac{E(F_1)}{E(C)}\cdot\gamma^2 + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2).$$

*Beweis.* Nach Satz 4.10 genügt es zu zeigen, dass $\frac{E(F_1)}{E(C)}\cdot\gamma^2 = \frac{n\cdot(1-\bar{p})^{n-1}}{E(C)}\cdot\gamma^2 + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2)$ $\Leftrightarrow E(F_1)\cdot\gamma^2 = n\cdot(1-\bar{p})^{n-1}\cdot\gamma^2 + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2)$ gilt.

$$\begin{aligned} E(F_1) &= \sum_{i=1}^{S} E(\text{"Spezies } i \text{ erscheint genau ein Mal"}) \qquad (4.6) \\ &= \sum_{i=1}^{S}\binom{n}{1}p_i\cdot(1-p_i)^{n-1} \\ &= \sum_{i=1}^{S} n\cdot p_i\cdot(1-p_i)^{n-1} \end{aligned}$$

Sei nun $h(\boldsymbol{p}) := \sum_{i=1}^{S} n\cdot p_i\cdot(1-p_i)^{n-1}$. Wir entwickeln nun $h$ mehrdimensional nach Taylor um den Punkt $\bar{\boldsymbol{p}}$.

$$h(\boldsymbol{p}) = h(\bar{\boldsymbol{p}}) + \sum_{i=1}^{S}\left.\frac{\partial h}{\partial p_i}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}}\cdot(p_i-\bar{p}) + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|) \qquad (4.7)$$

Dabei ist wegen $\bar{p} = \frac{1}{S}$

$$h(\bar{\boldsymbol{p}}) = n\cdot(1-\bar{p})^{n-1}.$$

Zudem ist

$$\frac{\partial h}{\partial p_i} = n\cdot(1-p_i)^{n-1} - n\cdot p_i\cdot(n-1)\cdot(1-p_i)^{n-2}$$

identisch an der Stelle $\bar{\boldsymbol{p}}$ für alle $i = 1,\ldots,S$, sodass

$$\sum_{i=1}^{S}\left.\frac{\partial h}{\partial p_i}\right|_{\boldsymbol{p}=\bar{\boldsymbol{p}}}\cdot(p_i-\bar{p}) = 0$$

folgt. Damit vereinfacht sich Gleichung (4.7) zu

$$h(\boldsymbol{p}) = n\cdot(1-\bar{p})^{n-1} + o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|). \qquad (4.8)$$

Da außerdem

$$\gamma^2 = \frac{\frac{1}{S}\sum_{i=1}^{S}(p_i-\bar{p})^2}{\bar{p}^2} = o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|) \qquad (4.9)$$

gilt, folgt die Behauptung aus der Definition von $h$ und Gleichung (4.6) wie folgt:

$$\begin{aligned} E(F_1)\cdot\gamma^2 &= h(\boldsymbol{p})\cdot\gamma^2 \\ &\overset{(4.8)}{=} \left(n\cdot(1-\bar{p})^{n-1}+o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|)\right)\cdot\gamma^2 \\ &\overset{(4.9)}{=} n\cdot(1-\bar{p})^{n-1}\cdot\gamma^2+o(|\boldsymbol{p}-\bar{\boldsymbol{p}}|^2) \end{aligned}$$

□

Um nun aus Satz 4.11 einen Schätzer des Artenreichtums zu konstruieren geht man davon aus, dass der Fehlerterm zu vernachlässigen ist.

$$S \approx \frac{E(S_{obs})}{E(C)} + \frac{E(F_1)}{E(C)}\cdot\gamma^2$$

Mit der folgenden Darstellung

$$E\left[\sum_{i=1}^{n} i(i-1)\frac{F_i}{n(n-1)}\right] = \sum_{i=1}^{S} p_i^2,$$

die beispielsweise in Good und Toulmin [GT56] gefunden werden kann, lässt sich dabei der Variationskoeffizient alternativ schreiben als

$$\gamma^2 = \frac{\sum_{i=1}^{S}(p_i-\bar{p})^2/S}{\bar{p}^2} = S\sum_{i=1}^{S} p_i^2 - 1 = S\left[\sum_{i=1}^{n} i(i-1)\frac{E(F_i)}{n(n-1)}\right] - 1.$$

Um nun einen Schätzer für $\gamma^2$ zu erhalten benötigt man einen Schätzer für dieses $S$. Hier bietet sich der Schätzer aus dem equal likely case $\hat{S}_1$ an, sodass man insgesamt den folgenden Schätzer erhält

$$\hat{\gamma}^2 = \max\left\{\hat{S}_1\left[\sum_{i=1}^{n} i(i-1)\frac{F_i}{n(n-1)}\right] - 1, 0\right\}.$$

Die darin vorkommende Summe lässt sich verkürzen und man kann alternativ nur bis $m$, der maximalen Anzahl der Individuen einer Spezies, summieren, da $F_i = 0$ für $i > m$ gilt.

Zusätzlich benötigen wir einen Schätzer für den Sample Coverage. Hier bietet sich das Verhältnis der Spezien, die sich wiederholt haben, zur Gesamtanzahl an.

$$\hat{C} = \frac{2\,F_2 + 3\,F_3 + \cdots + n\,F_n}{n} = 1 - \frac{F_1}{n}$$

Dieser Schätzer für $C$ wurde ursprünglich von Turing vorgeschlagen und wird beispielsweise in Good und Toulmin [GT56], Robbins [Rob68] und Esty [Est78] diskutiert.

Aus obigen Formeln ergibt sich somit der folgende Schätzer

$$\hat{S}_2 = \frac{S_{obs}}{\hat{C}} + \frac{n(1-\hat{C})}{\hat{C}}\hat{\gamma}^2.$$

### 4.3.3 Modifizierter Schätzer über einen Sample Coverage Ansatz

Die Güte des im vorherigen Abschnitt entwickelten Schätzers ist hauptsächlich dadurch bestimmt, ob der Fehlerterm in Satz 4.11 vernachlässigt werden kann.

Die Taylorapproximation wird umso besser je näher $\boldsymbol{p}$ an $\bar{\boldsymbol{p}}$ liegt, sodass $|\boldsymbol{p} - \bar{\boldsymbol{p}}|$ klein ist.

Diesen Sachverhalt wollen wir nutzen, um in diesem Abschnitt den Schätzer aus dem Sample Coverage Ansatz so zu modifizieren, dass er für unsere konkrete Anwendung bessere Ergebnisse liefert.

Die Idee, die wir hierbei verfolgen, wird nun kurz beschrieben.

Führt man ein multiples Capture-Recapture-Experiment mit $n$ Messdurchgängen durch, so wird es eine gewisse Anzahl an Spezien geben, die sehr häufig gemessen werden. Es kann also davon ausgegangen werden, dass diese Spezien in jedem vergleichbaren Experiment mindestens einmal bestimmt werden. Deren Anzahl sei nun als $S_{oft}$ bezeichnet. In dieser Weise ist es möglich die Spezien in häufig und selten vorkommende Spezien zu unterteilen

$$S_{obs} = S_{oft} + S_{selten},$$

wobei $S_{selten}$ die Anzahl der weniger häufig vorkommenden Spezien bezeichne. Es ist klar, dass die Erscheinungswahrscheinlichkeiten $p$ zwischen den Spezien aus $S_{oft}$ und $S_{selten}$ sehr stark variieren. Schwieriger ist jedoch die Festlegung, welche Spezien zu den häufigen Spezien gezählt werden sollen.

In unserem speziell betrachteten Fall wählen wir als $S_{oft}$ nur die Konsensusvariante, die im Normalfall auch der häufigsten Spezies entspricht. Von dieser kann angenommen werden, dass sie in jedem vergleichbaren Experiment nachgewiesen wird, da die Entstehungshäufigkeit dieser Art in den zwölf von Herrn Uwe Simon durchgeführten Experimenten bei 28% bis 82% lag und sie in jedem mindestens 10-Mal entstanden ist (vgl. Tabelle 1.1 und Abbildung 3.1). Wir erhalten also $S_{oft} = 1$.

$$S_{obs} = 1 + S_{nichtKonsensus}$$

Statt nun, wie im Schätzer über Sample Coverage direkt einen Schätzer für $S$ über $S_{obs}$ zu entwickeln, schätzen wir $S$ über $S_{nichtKonsensus}$ analog und addieren den festen Wert $S_{oft} = 1$ hinzu.

Durch diese Aufteilung entnehmen wir den Daten einen Großteil der Heterogenität, da die Konsensusvariante gewöhnlich eine sehr viel größere Wahrscheinlichkeit besitzt als die übrigen Arten. Der Mittelwert der Erscheinungshäufigkeiten der seltenen Spezien liegt somit näher an den Erscheinungshäufigkeiten der weiteren noch nicht beobachteten Spezien, sodass der Ausdruck $|\boldsymbol{p} - \bar{\boldsymbol{p}}|$ klein ist. Dadurch wird, wie oben erwähnt, die Taylorapproximation besser und somit der erhaltene Schätzer zuverlässiger.

Wir betrachten nun den etwas allgemeineren Fall, dass wir als $S_{oft}$ diejenige Spezies oder diejenigen Spezien wählen, die am häufigsten vorkommen ($m$-Mal). Für unsere Anwendung ist dies für gewöhnlich nur die Konsensusvariante, sodass diese Verallgemeinerung identisch zum vorher betrachteten Fall ist.

$$S_{obs} = F_m + S_{selten}$$

Analog zu vorher erhalten wir den Schätzer (Ersetzung von $S_{obs}$ durch $S_{obs} - F_m$ und $n$ durch $n - (F_m \cdot m)$)

$$\hat{S}_{mod} = F_m + \frac{S_{obs} - F_m}{\hat{C}} + \frac{(n - (F_m \cdot m))(1 - \hat{C})}{\hat{C}} \hat{\gamma}^2$$

mit

$$\hat{\gamma}^2 = \max \left\{ \hat{S}_1 \left[ \sum_{i=1}^{m-1} i(i-1) \frac{F_i}{\big(n - (F_m \cdot m)\big)\big((n - (F_m \cdot m)) - 1\big)} \right] - 1, 0 \right\}$$

und

$$\hat{S}_1 = \frac{S_{selten}}{\hat{C}} = \frac{S_{obs} - F_m}{\hat{C}}$$

und

$$\hat{C} = 1 - \frac{F_1}{n_{selten}} = 1 - \frac{F_1}{n - (F_m \cdot m)}.$$

Hierbei erkennt man, dass dieser Schätzer dann, wenn alle seltenen Spezien nur einmal vorkommen, nicht berechenbar ist, da dann $\hat{C} = 0$ folgt. Dies ist insofern logisch, da immer dann, wenn sich keine seltene Spezies wiederholt, nicht ausgeschlossen werden kann, dass es eine beliebig große Anzahl an seltenen Spezien gibt. Es kann somit keinen endlichen Schätzer des Artenreichtums geben.

### 4.3.4 Schätzer über den Jackknife Ansatz

Die Jackknife-Methode wurde als allgemeine Methode zur Verringerung des Bias eines verzerrten Schätzers erstmals von Quenouille im Jahre 1949 vorgestellt. Der Name *Jackknife* stammt dabei von der anfänglichen Bezeichnung „jack of all trades“ (übersetzt in etwa *Alleskönner*). Man glaubte, dass diese Methode an allen möglichen Stellen eingesetzt werden kann.

Sei $F$ eine Verteilungsfunktion und $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe mit $X_i \overset{\text{iid}}{\sim} F$ für $i = 1, \ldots, n$. Weiter sei $f = f(\boldsymbol{X})$ eine Schätzfunktion eines unbekannten Parameters $\theta$ der Verteilungsfunktion $F$.

Von einer konkreten Realisation der Stichprobe $\boldsymbol{x} = (x_1, \ldots, x_n)$ ausgehend nimmt man an, dass $\hat{\theta} = f(\boldsymbol{x})$ eine relativ gute Approximation von $\theta$ ist. Um den Jackknife Schätzer zu erhalten, der den Bias von $\hat{\theta}$ reduzieren soll, führt man die folgenden drei Schritte durch:

1. Entferne eine der Beobachtungen. Dies sei $x_i$.

2. Berechne die Schätzung von $\theta$ basierend auf $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ und bezeichne sie mit $\hat{\theta}_{-i} = f(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

3. Berechne den Pseudowert $\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$.

Diese Schritte werden $n$-Mal wiederholt für $i = 1, \ldots, n$. Der Jackknife Schätzer ist dann gegeben als

$$\hat{\theta}_{jack} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i.$$

Diese Schätzung ist als „first-order jackknife“ bekannt. Auch wenn im Rahmen dieser Arbeit nicht ausführlich auf die statistischen Hintergründe der Jackknife-Methode eingegangen werden kann, wollen wir doch kurz motivieren, dass diese Methode hilft den Bias zu reduzieren.

Dazu sei vorausgesetzt, dass der Erwartungswert von $f$ bzw. der Zufallsvariablen $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ eine Entwicklung in $\frac{1}{n}$ besitzt. Es existiere also eine Darstellung

$$E(\hat{\theta}) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots, \tag{4.10}$$

wobei die $a_i \in \mathbb{R}$ unabhängig von $n$ sind.

Aus dieser Voraussetzung folgt sofort

$$E(\hat{\theta}_{-i}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \cdots$$

unabhängig von $i$. Damit gilt

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{a_1}{n} + \frac{a_2}{n^2} + \cdots = \mathcal{O}\left(\frac{1}{n}\right)$$

und

$$\begin{aligned}
\text{Bias}(\hat{\theta}_{jack}) &= E(\hat{\theta}_{jack}) - \theta \\
&= \frac{1}{n} \sum_{i=1}^{n} E(n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}_{-i}) - \theta \\
&= nE(\hat{\theta}) - (n-1)\frac{1}{n} \sum_{i=1}^{n} E(\hat{\theta}_{-i}) - \theta \\
&= n\theta + a_1 + \frac{a_2}{n} - (n-1)\theta - a_1 - \frac{a_2}{n-1} - \theta + \mathcal{O}\left(\frac{1}{n^2}\right) \\
&= -\frac{a_2}{n \cdot (n-1)} + \mathcal{O}\left(\frac{1}{n^2}\right) \\
&= \mathcal{O}\left(\frac{1}{n^2}\right).
\end{aligned}$$

Während also der Bias von $\hat{\theta}$ von der Größenordnung $\mathcal{O}\left(\frac{1}{n}\right)$ ist, ist der von $\hat{\theta}_{jack}$ nur noch von Ordnung $\mathcal{O}\left(\frac{1}{n^2}\right)$. Dies rechtfertigt die Verwendung der Jackknife-Methode zur Verringerung des Bias eines verzerrten Schätzers.

Nun wollen wir diese Methode nutzen um einen Schätzer des Artenreichtums zu bestimmen.

**Satz 4.12** (Schätzer Jackknife 1)**.** *Betrachtet man die beobachtete Spezienanzahl $S_{obs}$ als verzerrte Schätzung des Artenreichtums $S$, die die Voraussetzung aus Gleichung (4.10) erfüllt, so erhält man nach der Jackknife-Methode den folgenden Schätzer des Artenreichtums*

$$\hat{S} = S_{obs} + \frac{n-1}{n} F_1.$$

*Beweis.* Es wird die Anzahl der Spezien $S = \theta$ basierend auf $n$ unabhängigen Beobachtungen geschätzt. Es wird nach Voraussetzung angenommen, dass die beobachtete Spezienanzahl $S_{obs} = \hat{\theta}$ eine relativ gute Approximation ist. Als $f$ hat man folglich die Spezienzählfunktion gewählt. Die Entfernung einer Beobachtung hat nur dann einen Einfluss auf die Spezienzählfunktion, wenn es sich um eine Spezies handelt, die genau einmal vorkommt. Somit ergibt sich genau $F_1$-Mal der Wert $S_{obs} - 1$ und $n - F_1$-Mal der Wert $S_{obs}$ bei den $n$ Jackknife-Schätzungen.

$$\begin{aligned}
\hat{\theta}_{jack} &= \frac{1}{n}\sum_{i=1}^{n}\Big(n\hat{\theta} - (n-1)\hat{\theta}_{-i}\Big) \\
&= \frac{1}{n}\Big(\big(n \cdot S_{obs} - (n-1)\cdot(S_{obs}-1)\big)\cdot F_1 + \big(n \cdot S_{obs} - (n-1)\cdot S_{obs}\big)\cdot(n - F_1)\Big) \\
&= S_{obs} + \frac{n-1}{n}F_1
\end{aligned}$$

□

Satz 4.12 zeigt, dass es sich bei dem aus der Jackknife-Methode hergeleiteten Schätzer der Spezienanzahl um eine Funktion der nur einmal vorkommenden Spezien handelt. Der Schätzer steigt für jede seltene – nur in einem Messdurchgang vorkommende – Spezies um den Wert $(n-1)/n$.

Unter der Voraussetzung aus Gleichung (4.10) lassen sich Jackknife-Schätzungen $k$-ter Ordnung, d.h. mit einem Bias der Ordnung $n^{-k-1}$, bestimmen. Diese sind zum Beispiel in Burnham und Overton [BO78] oder in Schucany et al. [SGO71] angegeben:

$$\hat{\theta}_{jack_k} = \frac{1}{k!}\sum_{j=0}^{k}\hat{\bar{\theta}}_{(j)}(-1)^j\binom{k}{j}(n-j)^k, \tag{4.11}$$

wobei $\hat{\bar{\theta}}_{(j)}$ der Mittelwert der Schätzungen ist, wenn man Gruppen der Größe $j$ entfernt.

Dadurch lassen sich Schätzer des Artenreichtums höherer Ordnung bestimmen. Wir wollen uns hier jedoch auf die Ordnung 2 beschränken.

**Satz 4.13** (Schätzer Jackknife 2)**.** *Betrachtet man die beobachtete Spezienanzahl $S_{obs}$ als verzerrte Schätzung des Artenreichtums $S$, die die Voraussetzung aus Gleichung (4.10) erfüllt, so erhält man nach der Jackknife-Methode den folgenden Schätzer des Artenreichtums zweiter Ordnung*

$$\hat{S} = S_{obs} + \left(\frac{F_1(2n-3)}{n} - \frac{F_2(n-2)^2}{n(n-1)}\right)$$

*Beweis.* Der Beweis ergibt sich aus Darstellung (4.11) und analogen Überlegungen wie im Beweis zu Satz 4.12 durch elementare Rechnungen. □

## 4.4 Datenanalytische Methoden

Nachdem wir bisher nur stichprobentheoretische Methoden zur Bestimmung der Artenvielfalt betrachtet haben, wollen wir uns nun datenanalytischen Methoden zuwenden.

Dabei wird nicht versucht die Daten durch statistische Modelle zu beschreiben, sondern es werden monotone Funktionen bestimmt, die die Anzahl der erwarteten beobachteten Spezien für irgendeine Größe oder Anzahl der Messeinheiten darstellen. Diese Messeinheiten können dabei sowohl kontinuierlich als auch diskret gewählt werden und z. B. Zeiteinheiten, Quadranten (kleinere Flächenabschnitte) oder Individuen sein. Solche Funktionen werden in diesem Kontext als *Species Accumulation Curves* bezeichnet und sollen im nächsten Teilabschnitt näher untersucht werden. Anschließend werden diese Kurven durch fest vorgegebene Funktionen approximiert, um so, beispielsweise durch deren asymptotisches Verhalten, eine Schätzung des Artenreichtums zu erhalten.

### 4.4.1 Species Accumulation Curve

Eine Species Accumulation Curve ist, wie gerade erwähnt, eine monoton steigende Funktion bezüglich des Aufwandes und verläuft typischerweise asymptotisch. Sie wird einerseits dazu verwendet den minimalen Aufwand zu schätzen, um eine bestimmte Spezienanzahl zu erreichen, andererseits kann durch sie der Artenreichtum geschätzt werden. In der zweiten Weise wollen wir die Species Accumulation Curve verwenden.

Als Messeinheit ist in unserer Anwendung die Individuenanzahl, d.h. die Anzahl der klonierten Individuen, natürlich vorgegeben. Gesucht ist also eine Funktion, die die erwartete Anzahl der verschiedenen Genvarianten als Funktion der Anzahl der klonierten Gene beschreibt.

Das einfache Poolen der Stichprobe gemäß ihres Erscheinens und das graphische Auftragen derselben wird zwar eine Species Accumulation Curve liefern, diese wird aber, aufgrund von einfachen stochastischen Effekten, keine glatte Kurve sein. Glatte Species Accumulation Curves könnten durch Resampling der gemessenen Stichprobe erhalten werden, indem man für jede Anzahl der Messdurchgänge den Mittelwert der vorkommenden Spezienanzahl in allen Teilmengen der Gesamtstichprobe eben dieser Größe als Schätzgröße der erwarteten Spezienanzahl verwendet. Dieses Verfahren ist jedoch für große $n$ sehr rechenintensiv, weshalb eine alternative Vorgehensweise gewählt wird. Wir werden deshalb im Folgenden eine einfacher zu berechnende wahrscheinlichkeitstheoretische geschlossene Form der Species Accumulation Curve entwickeln.

Die erwartete Anzahl der Spezien in $h$ Messdurchgängen ist die Summe der Wahrscheinlichkeiten, dass eine Spezies nicht abwesend von allen $h$ Quadranten ist. Unterstellt man den Erscheinungshäufigkeiten der einzelnen Spezien eine Bionomialverteilung und wird mit $\tau$ die Species Accumulation Curve bezeichnet, so erhält man

$$\tau(h) = \sum_{j=1}^{S} \left[1 - (1 - p_j)^h\right]. \tag{4.12}$$

**Satz 4.14.** *Eine erwartungstreue Schätzung der Species Accumulation Curve ist gegeben durch*

$$\hat{\tau}(h) = S_{obs} - \sum_{i=1}^{n} \alpha_{ih} F_i,$$

*wobei*

$$\alpha_{ih} = \begin{cases} \frac{(n-i)!(n-h)!}{n!(n-i-h)!}, & \textit{falls } i + h \leq n \\ 0, & \textit{falls } i + h > n. \end{cases}$$

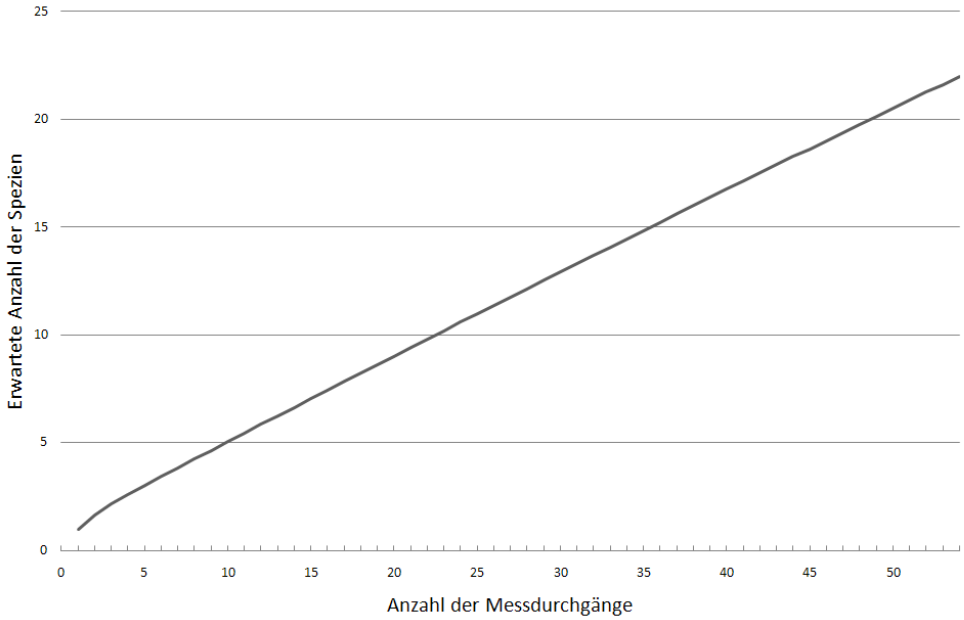


Abbildung 4.1: Species Accumulation Curve am Beispiel der Pilzart *Phoma exigua var. exigua* und Genregion *LSU*

*Beweis.* Es bezeichne nun $\alpha_{ih}$ die Wahrscheinlichkeit, dass bei $h$ ausgewählten Individuen genau $n-i$ ausgewählt wurden, in denen eine Spezies die $i$-Mal vorkommt nicht auftaucht. Demnach ist $(1-\alpha_{ih})$ die Wahrscheinlichkeit dafür, dass mindestens ein Individuum einer Spezies die genau $i$-Mal vorkommt in den $h$ ausgewählten Messeinheiten gemessen wird. Demnach erhält man durch Änderung der Summation

$$\begin{aligned} E(\tau(h)) &= E\left(\sum_{j=1}^{S}\left[1-(1-p_j)^h\right]\right) \\ &= E\left(\sum_{i=1}^{n}(1-\alpha_{ih})F_i\right) \end{aligned}$$

Nun verwendet man $\sum_{i=1}^{n} F_i = S_{obs}$.

$$= E\left(S_{obs} - \sum_{i=1}^{n}\alpha_{ih}F_i\right).$$

Beim Ziehen von $h$ Individuen aus $n$ Möglichen handelt es sich um ein Laplace-Experiment mit $|\Omega| = \binom{n}{h}$, sodass sich die Wahrscheinlichkeit $\alpha_{ih}$ für $n-j \geq h$ wie folgt berechnen lässt:

$$\alpha_{ih} = \frac{\binom{n-i}{h}}{\binom{n}{h}} = \frac{(n-i)!(n-h)!}{n!(n-i-h)!}$$

Für $n-j < h$ ergibt sich stets $\alpha_{ih} = 0$, da dann eine solche Auswahl nicht möglich ist. □

*Bemerkung* 4.15. Man beachte, dass in Satz 4.14 $\alpha_{ih} = \alpha_{hi}$ gilt und dass für $h = n$ stets $a_{jn} = 0$ folgt und somit $\hat{\tau}(n) = S_{obs}$ gilt.

### 4.4.2 Schätzer des Artenreichtums über Kurvenanpassung

Um nun aus einer Species Accumulation Curve einen Schätzer des Artenreichtums zu erhalten, wird an diese eine – typischerweise asymptotische – Funktion angepasst. Es gibt eine Vielzahl solcher Funktionen, die hierfür in Frage kommen. Wir werden jedoch ausschließlich die sog. Michaelis-Menten Funktion verwenden. Denn wie bei Raaijmakers [Raa87] genauer ausgeführt ist, handelt es sich dabei um eine der gebräuchlichsten und zweckmäßigsten Funktionen für die hier betrachtete Fragestellung. Die Michaelis-Menten Funktion hat die folgende Darstellung

$$v(s) = \frac{\alpha s}{(s+\beta)},$$

wobei $v$ die abhängige und $s$ die unabhängige Variable ist, und $\alpha$ und $\beta$ Parameter oder Konstanten sind.

In unserem Fall soll also $v(s)$ den Wert der Species Accumulation Curve bei einer bestimmten Anzahl an Messdurchgängen $s$ approximieren. Hierzu ist es notwendig die Konstanten $\alpha$ und $\beta$ zu schätzen. Wegen

$$\lim_{s\to\infty} v(s) = \alpha \tag{4.13}$$

entspricht $\alpha$ dem asymptotischen Wert der Michaelis-Menten Funktion und somit dem geschätzten maximalen Wert der Species Accumulation Kurve. Wir können also $\alpha$ als Schätzwert des Artenreichtums verwenden. Bei dem Wert $\beta$ handelt es sich um die Michaeliskonstante, die in unserer Anwendung von geringerem Interesse ist.

Eine Möglichkeit diese beiden Koeffizienten zu schätzen besteht darin, die Gleichung in eine Form $Y = a + bX$ zu bringen und dann die Koeffizienten $a$ und $b$ durch eine lineare Regression zu schätzen. Die Michaelis-Menten Funktion ist beispielsweise gleichbedeutend zu

$$\begin{aligned}
\frac{1}{v} &= \frac{1}{\alpha} + \frac{(\beta/\alpha)}{s} \\
\frac{s}{v} &= \frac{\beta}{\alpha} + \frac{s}{\alpha} \\
\frac{v}{s} &= \frac{\alpha}{\beta} - \frac{v}{\beta} \\
v &= \alpha - \frac{\beta v}{s}.
\end{aligned}$$

Betrachten wir nun den Fall, dass wir eine Michaelis-Menten Funktion an eine Menge an Datenpunkten $\{(s_i, v_i) \,:\, i = 1, \ldots, n\}$ anpassen möchten. Wir nehmen an, dass jede Beobachtung $v_i$ sich als Michaelis-Menten Funktion von $s_i$ zusammen mit einem additiven Fehlerterm $\tilde{\epsilon}_i$ beschreiben lässt. Genaugenommen ist unser Modell, dass $v_1, \ldots, v_n$ Realisationen von Zufallsvariablen $V_1, \ldots, V_n$ sind, die

$$V_i = \frac{\alpha s_i}{(s_i+\beta)} + \tilde{\epsilon}_i$$

erfüllen, wobei die $\tilde{\epsilon}_i$ unabhängige Fehlerterme mit $E(\tilde{\epsilon}) = 0$ bezeichnen.

Würden wir nun eine der oben vorgeschlagenen Transformationen hin zu einer linearen Form durchführen, so würde der zugehörige Fehlerterm mittransformiert werden. Somit

erhält man durch jede Transformation andere Fehlerterme und eine andere Fehlerstruktur. Zusätzlich sind dadurch die Schätzer des Artenreichtums, die man durch Gleichung (4.13) erhält, nicht identisch. Die erhaltene Lösung hängt also stark davon ab, welche Struktur man dem Fehler im Modell unterstellt.

Wir verfolgen die Annahme von Raaijmakers [Raa87] die den Daten einen konstanten Variationskoeffizienten unterstellt. Wir verwenden also das Modell

$$V_i = \frac{\alpha s_i}{(s_i + \beta)} + \frac{s_i}{s_i + \beta}\epsilon_i. \tag{4.14}$$

Dabei sind die $\epsilon_i$ unabhängig normalverteilt mit Mittelwert 0 und Varianz $\sigma^2$. Die Schätzer von $\alpha$ und $\beta$ entwickeln wir nun nicht durch eine Transformation auf eine lineare Form, sondern durch einen Maximum-Likelihood Ansatz, wodurch die vorgegebene Fehlerstruktur erhalten bleibt.

**Satz 4.16.** *Es seien $s_1, \ldots, s_n$ gegeben und $v_1, \ldots, v_n$ Realisationen von Zufallsvariablen $V_1, \ldots, V_n$, die*

$$V_i = \frac{\alpha s_i}{(s_i + \beta)} + \frac{s_i}{s_i + \beta}\epsilon_i, \quad i = 1, \ldots, S \tag{4.15}$$

*erfüllen, mit $\epsilon_i$ unabhängig normalverteilt mit Mittelwert 0 und Varianz $\sigma^2$. Daraus ergeben sich die Maximum-Likelihood Schätzer von $\alpha$ und $\beta$ mit $X_i = v_i/s_i$ und $Y_i = v_i$ zu*

$$\begin{aligned} \hat{\alpha} &= \bar{Y} + \hat{\beta}\bar{X} \\ \hat{\beta} &\approx \frac{\bar{X}S_{yy} - \bar{Y}S_{xy}}{\bar{Y}S_{xx} - \bar{X}S_{xy}}, \end{aligned}$$

*wobei $S_{xx}$, $S_{yy}$ und $S_{xy}$ die Summe der Quadrate oder Kreuzprodukte von $X_i - \bar{X}$ und $Y_i - \bar{Y}$ bezeichnen und $\bar{X}$ bzw. $\bar{Y}$ den Mittelwert der $X_i$ bzw. $Y_i$ bezeichne.*

*Beweis.* Die zu Modell (4.15) gehörende Likelihood Funktion ist

$$L = \prod_{i=1}^{n} \left( \frac{\sqrt{2\pi\sigma^2}s_i}{s_i + \beta} \right)^{-1} \exp\left\{ -\frac{1}{2} \left( \frac{v_i - \frac{\alpha s_i}{s_i+\beta}}{\sigma \frac{s_i}{s_i+\beta}} \right)^2 \right\}.$$

Es ist nun einfacher nicht direkt $L$ zu maximieren, sondern $\ln(L)$. Da der natürliche Logarithmus eine monotone Funktion ist nehmen beide an derselben Stelle ihr Maximum an.

$$-\ln(L) = \left(\frac{n}{2}\right)\ln(2\pi) + \left(\frac{n}{2}\right)\ln(\sigma^2) - \sum_{i=1}^{n} \ln\left(\frac{s_i + \beta}{s_i}\right) + \sum_{i=1}^{n} \frac{(v_i + \beta v_i/s_i - \alpha)^2}{2\sigma^2}$$

Beziehungsweise nach Substitution von $X_i = v_i/s_i$ und $Y_i = v_i$

$$= \left(\frac{n}{2}\right)\ln(2\pi) + \left(\frac{n}{2}\right)\ln(\sigma^2) - \sum_{i=1}^{n} \ln\left(1 + \frac{\beta X_i}{Y_i}\right) + \sum_{i=1}^{n} \frac{(Y_i + \beta X_i - \alpha)^2}{2\sigma^2}$$

Nun erhält man durch elementare Rechnungen, wobei $S_{xx}$, $S_{yy}$ und $S_{xy}$ die Summe der Quadrate oder Kreuzprodukte von $X_i - \bar{X}$ und $Y_i - \bar{Y}$ bezeichnen und $\bar{X}$ bzw. $\bar{Y}$ den

Mittelwert der $X_i$ bzw. $Y_i$ bezeichne.

$$\begin{aligned} \frac{\partial(-\ln(L))}{\partial\alpha} &= 0 \\ \Leftrightarrow \sum_{i=1}^{n}(Y_i + \beta X_i - \alpha) &= 0 \\ \Leftrightarrow \bar{Y} + \beta\bar{X} &= \alpha, \end{aligned}$$

$$\begin{aligned} \frac{\partial(-\ln(L))}{\partial\sigma} &= 0 \\ \Leftrightarrow -n\sigma^2 + \sum_{i=1}^{n}\left(Y_i + \beta X_i - (\bar{Y} + \beta\bar{X})\right)^2 &= 0 \\ \Leftrightarrow \sigma^2 &= \frac{S_{yy} + 2\beta S_{xy} + \beta^2 S_{xx}}{n}, \end{aligned}$$

$$\begin{aligned} \frac{\partial(-\ln(L))}{\partial\beta} &= 0 \\ \Leftrightarrow S_{xy} + \beta S_{xx} &= (S_{yy} + 2\beta S_{xy} + \beta^2 S_{xx})\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{Y_i + \beta X_i}. \end{aligned}$$

Man könnte nun durch einen geeigneten Algorithmus eine Lösung für $\beta$ aus der letzten Gleichung bestimmen. Verwendet man jedoch die Approximation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{Y_i + \beta X_i} \approx \frac{\bar{X}}{\bar{Y} + \beta\bar{X}},$$

so ergeben sich die im Satz angegebenen direkten Formen. □

Hat man nun eine Species Accumulation Kurve wie in Abschnitt 4.4.1 gegeben und passt man an diese Daten eine Michaelis-Menten Funktion an, so ist es mit Satz 4.16 und Gleichung (4.13) möglich einen Schätzwert der Spezienanzahl dadurch zu bestimmen, dass man

$$\hat{S} = \hat{\alpha}$$

setzt.

Wie ausgeführt lässt sich der Artenreichtum mit zwei verschiedenen Herangehensweisen schätzen. Dabei hat die datenanalytische Methode (mit gleichzeitiger Kurvenanpassung) gegenüber der stichprobentheoretischen Betrachtungsweise folgende Pluspunkte.

**Vorteile**

- Es sind keine Annahmen über die Struktur der Spezienhäufigkeiten notwendig.
- Datenanalytische Methoden können – auch wenn wir dies hier nicht dargestellt haben – sowohl auf kontinuierliche als auch auf diskrete Einteilungen des Aufwandes zur Bestimmung der Stichprobe angewendet werden.

Andererseits sind auch gewisse Schwachstellen zu berücksichtigen.

**Nachteile und Bedenken**

- Es muss ausreichend viel Datenmaterial zur Verfügung stehen, um die Species Accumulation Curve zu erstellen.

- Verschiedene Funktionen können sich gut an die Daten anpassen und trotzdem sind die erhaltenen Schätzwerte sehr unterschiedlich.

- Eine gute Anpassung lässt nicht unbedingt auf eine gute Extrapolationseigenschaft und somit auf gute Schätzwerte schließen.

# 5 Simulationsstudie

Um die Arbeitsweise der jeweiligen Schätzer, die soeben theoretisch hergeleitet wurden, überprüfen zu können wurden Simulationen bei verschiedenen vorgegebenen Rahmenbedingungen durchgeführt.

Bei den hier durchgeführten Simulationen war jeweils die verwendete Anzahl der Spezien bekannt, sodass der Wert der Schätzung direkt mit dieser verglichen werden konnte.

Diese Simulationen sollen uns insbesondere dabei helfen, die anschließend in Kapitel 6 durchgeführten Schätzungen des Artenreichtums für die von Uwe Simon erhobenen Daten besser zu verstehen und ihre Aussagekraft beurteilen zu können. Die in diesem Simulationsteil verwendeten Verteilungen der Spezienhäufigkeiten wurden deshalb speziell an diese Daten angepasst. Wir verzichten in diesem Kapitel noch auf konkrete Empfehlungen, welcher Schätzer bei welcher Situation verwendet werden soll, und werden dies in Abschnitt 6.1 nachholen.

## 5.1 Programm in Mathematica Version 6

Es existieren zwar bereits die beiden Computerprogramme *SPADE* [CS03] und *EstimateS* [Col05], die vollautomatisch aus einer Spezienkonfiguration eine Vielzahl an Schätzern berechnen. Diesen fehlen jedoch immer Funktionen, die wir im Rahmen dieser Simulationsstudie verwenden möchten. So können beispielsweise parametrische Schätzer, wie wir sie in Abschnitt 4.2 vorgestellt haben, in diesen Programmen nicht berechnet werden oder es fehlt die Möglichkeit die Ergebnisse zur weiteren Verarbeitung in andere Programme zu exportieren.

Diese Überlegungen haben dazu geführt, ein Programm in `Mathamatica 6` zu erstellen, das alle für diese Simulationsstudie notwendigen Funktionen enthält und somit die zu diesem Thema bestehende Software sinnvoll ergänzt. Zudem ist es so möglich alle durchgeführten Programmabläufe transparent darzustellen. Das gesamte Programm befindet sich auf CD als Beilage zu dieser Diplomarbeit. Des Weiteren hat sich gezeigt, dass das Programm, auch wenn der Funktionsumfang durch die in Kapitel 2 und 3 gemachten Einschränkungen deutlich geringer ist als bei beiden oben erwähnten Programmen, überraschend schnell arbeitet. Die Berechnung der Schätzer des Artenreichtums aus einer gegebenen Spezienliste erfolgt in den betrachteten Fällen im Millisekundenbereich.

Zudem hat das für diese Diplomarbeit entwickelte Programm den Vorteil, dass es nicht erst extra installiert werden muss, sondern als weiteres Package in `Mathematica` eingebunden und direkt verwendet werden kann. Um die Fehleranfälligkeit des Programms zu minimieren wurde es in einzelne getrennte Module aufgeteilt, die verschiedene Aufgaben übernehmen. Für jedes Modul wurde eine eigene Mathematicafunktion programmiert,

die die geforderten Befehle ausführt. Somit lässt sich der Programmablauf wie folgt gliedern:

**SpezienSpezifikationen** Modul zum Einlesen und Überprüfen einer gegebenen Spezienliste und Berechnung von verschiedenen Kenngrößen der Stichprobe, wie beispielsweise $S_{obs}$, $n$ und insbesondere $F_1, \ldots, F_n$, die nach Satz 3.1 suffizient für nichtparametrische Ansätze sind. Als Ausgabe erhält man eine Mathematicaliste mit Kenngrößen der Spezienliste, den Spezienspezifikationen.

**SchaetzerBerechung** Berechnung von verschiedenen stichprobentheoretischen Schätzern aus der Liste der Spezienspezifikationen. Unter den berechneten Schätzern befinden sich alle in Kapitel 4 vorgestellten bis auf den Schätzer nach McNeil, der aus den in diesem Kapitel angegebenen Gründen nicht mitberücksichtigt wurde. Das Ergebnis dieser Prozedur ist eine Mathematicaliste, die die Werte aller berechneten Schätzer enthält.

**SchaetzerSpeciesAccumulationCurve** Hier wird die Berechnung der Species Accumulation Curve sowie des daraus resultierenden Schätzers Michaelis-Menten nach der in Abschnitt 4.16 vorgestellten Methode über Maximum-Likelihood durchgeführt. Man erhält als Ausgabe eine Liste für die Species Accumulation Curve zusammen mit dem berechneten Schätzer.

**Auswertungsprogramm** Dieses Teilmodul ist für eine Schätzung des Artenreichtums aus einer vorgegebenen Spezienliste nicht notwendig. Es führt lediglich alle in dieser Diplomarbeit betrachteten Simulationen durch.

## 5.2 Modellannahmen und Ablauf der Simulationen

In den folgenden Simulationen betrachten wir – wie in den vorherigen Kapiteln auch – ausschließlich multiple Capture-Recapture Experimente mit genau einer Beobachtung pro Messeinheit, d.h. es wird wiederholt mit Zurücklegen aus einer Grundgesamtheit mit einer festen Anzahl an Spezien gezogen.

In den Simulationen unterstellen wir dem Kollektiv eine vorher fest vorgegebene Verteilung der Spezienhäufigkeiten. Die von Uwe Simon ermittelten Daten legen es dabei nahe, dass es eine sehr häufig vorkommende Spezies gibt und eine Vielzahl an weniger häufigen Spezien, vgl. Abbildung 3.1. Die häufigste Spezies, die der Konsensusvariante entspricht, entsteht dabei gemäß Tabelle 1.1 mit einer Wahrscheinlichkeit von 0.28 bis 0.82.

Wir betrachten deshalb im Simulationsteil Verteilungen der Spezienhäufigkeiten mit einer häufigen Spezies, deren Anteil in drei Kategorien zu 0.4, 0.6 und 0.8 gewählt wurde.

Die Stichprobengröße $n$ wurde zu 50, 60, 70, 100, 150, 200 und 250 gewählt. Einerseits soll die Größenordnung von 60, die der Anzahl der Klonierungen in den Versuchen von Uwe Simon entspricht, genauer untersucht werden. Andererseits soll eine Prognose auf größere Datensätze möglich sein. Dies soll es auch ermöglichen Stichprobengrößen angeben zu können, um einen gewissen Artenreichtum mit einer notwendigen Sicherheit vorhersagen zu können.

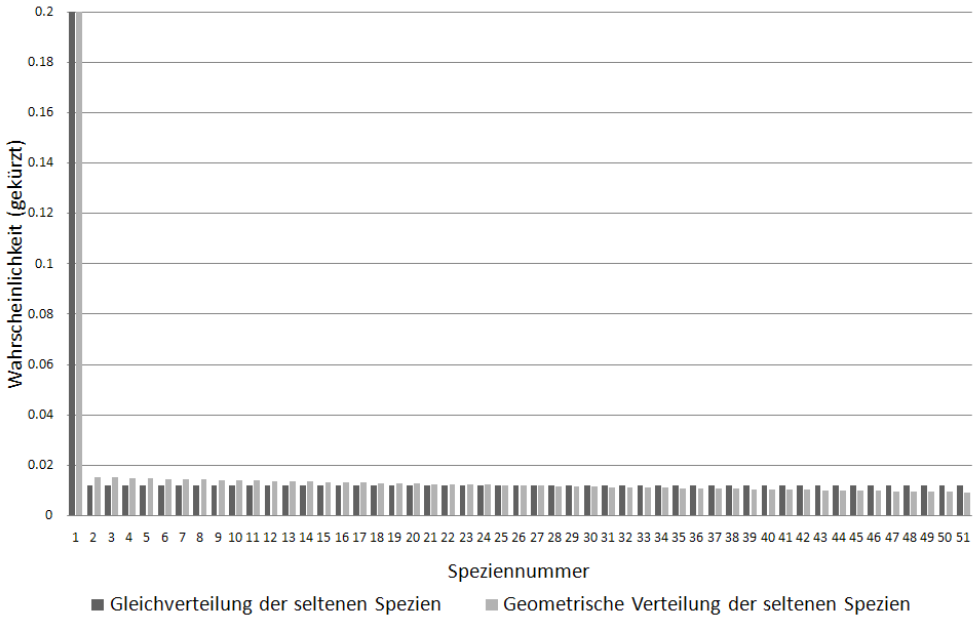


Abbildung 5.1: Verteilungen der Spezienhäufigkeiten wie sie in den Simulationen verwendet werden für $S = 51$ und $p = 0.4$.

In den Simulationen werden Kollektive mit einer festen Spezienanzahl $S$ betrachtet. Diese wurde zu 51, 101, 201, 401 und 601 im Falle einer häufigen Spezies mit einer Wahrscheinlichkeit von 0.4 bzw. 0.6 gewählt und zu 26, 51, 101, 201 und 401 im Falle einer häufigen Spezies mit einer Wahrscheinlichkeit von 0.8 gewählt. Somit ergeben sich 25, 50, 100, 200, 400 bzw. 600 Spezien, die nicht der häufigsten Spezies entsprechen. Diese Festlegung geschah in Rücksprache mit Herrn Uwe Simon, da in diesem Bereich die vermutete Anzahl der polymorphen Varianten bei seinen Klonierungsvorgängen lag.

Den $S - 1$ weniger häufigen Spezien wurde in unseren Modellen eine parametrische Verteilung unterstellt. Wir haben uns für die diskrete Gleichverteilung sowie eine truncated geometrische Verteilung mit sehr kleinem Koeffizienten ($p = \lambda = 0.01$) entschieden. Diese beiden sind beispielhaft für $S = 51$ und $p = 0.4$ in Abbildung 5.1 dargestellt.

Die Gleichverteilung wurde als einfachste Beschreibung der Wahrscheinlichkeiten der seltenen Spezien bzw. derjenigen Spezien, die nicht der Konsensusvariante entsprechen, gewählt. Dabei wird jeder Spezies die gleiche Wahrscheinlichkeit zugewiesen. Diese Annahme ist auch insofern sinnvoll, da man davon ausgehen kann, dass sich die einzelnen Basenpaare einer betrachteten Genregion bei einem Klonierungsvorgang zufällig und mit gleicher Wahrscheinlichkeit ändern.

Die geometrische Verteilung wurde deshalb gewählt, da sie einen monoton fallenden Verlauf besitzt und sich im Grenzwert $\lambda \to 0$ wieder die Gleichverteilung ergibt. Sie ist somit die logische Fortsetzung aus obiger Annahme einer Gleichverteilung. Außerdem kann so untersucht werden, wie die Schätzer darauf reagieren wenn die Spezienerscheinungshäufigkeiten variieren. Die Daten, die von Herrn Uwe Simon gemessen wurden, legen es dabei nahe, dass diese Variation eher gering ist, da sehr viele seltene Spezien gemessen wurden und deshalb die Erscheinungshäufigkeiten nicht zu schnell vernachlässigbar sein sollten. Der Koeffizient der geometrischen Verteilung wurde deshalb bewusst klein gewählt, da die Daten von Herrn Uwe Simon keine starken Schwankungen bei den Wahrscheinlichkei-

ten der seltenen Spezien vermuten lassen und da so auch bei größerem Artenreichtum die Wahrscheinlichkeit eine feste Spezies $k$ zu bestimmen für wachsendes $k$ nicht zu schnell vernachlässigbar wird.

In einer weiteren Untersuchung wurde auch kurz das Querschnittsverhalten, d.h. das Verhalten der Schätzer bei einer Variation des Glättungsparameters $\lambda$ der truncated geometrischen Verteilung untersucht. Hierzu wurde der Koeffizient $\lambda$ im Bereich von 0 bis 0.2 variiert.

In den Simulationen wurde nun für jede Kombination aus angenommener Verteilung der seltenen Spezien, festgelegter Spezienanzahl und Anzahl der Messdurchgänge mit Hilfe von Pseudozufallszahlen eine Spezienliste der Länge $n$ konstruiert, indem $n$ Realisationen aus der so erhaltenen Verteilung der Spezienhäufigkeiten simuliert und zu einer Spezienliste zusammengefasst wurden. Dieses Ziehen mit Zurücklegen entspricht einer Durchführung eines multiplen Capture-Recapture Experimentes mit $n$ Messdurchgängen. Diese Prozedur wurde 1'000-Mal wiederholt und in jedem dieser Durchläufe wurden für die erhaltene Spezienliste alle betrachteten Schätzer des Artenreichtums bestimmt. Unter gewissen Umständen, z. B. wenn eine Stichprobe keine Spezies genau zwei Mal enthält ($F_2 = 0$), lassen sich bestimmte Schätzer nicht berechnen. Dies tritt z. B. im ursprünglichen von Chao vorgeschlagenen Schätzer auf, da hier $F_2$ im Nenner steht. Solche Ereignisse werden als *fehlgeschlagen* markiert. Zu jedem Schätzer wurde notiert, wie oft dieser in den 1'000 Simulationsdurchläufen nicht berechnet werden konnte. Diese fehlenden Werte wurden nicht durch den Wert eines erneuten Durchlaufs der Simulation ersetzt.

Die Auswertung der Güte der Schätzer des Artenreichtums erfolgt für jeden getrennt über die Anzahl der nicht berechenbaren Schätzer („Fehlgeschlagen"), den (mittleren) Bias

$$\frac{1}{1000}\sum_{i=1}^{1000}\hat{S}_i - S$$

und den RootMSE (root mean squared error)

$$\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{S}_i - S)^2},$$

wobei die Summen entsprechend verkürzt werden, falls die entsprechenden Schätzer nicht berechnet werden können.

## 5.3 Ergebnisse und Auswertung der Simulationen

Die folgenden Tabellen 5.1 bis 5.6 fassen Ausschnitte aus den Ergebnissen der Simulationen systematisch zusammen. Die Ergebnisse sind dabei stets auf die zweite Nachkommastelle gerundet. Diese und alle weiteren für diese Diplomarbeit durchgeführten Berechnungen finden sich aus Platzgründen als Anhang auf CD.

Im Folgenden bezeichne nun $p$ stets den Anteil der häufigen Spezies.

| Schätzername $S$ | $n$ | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| **51** | 50 | *Fehlgeschlagen* | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -21.38 | -26.73 | 7.78 | -1.36 | 9.84 | -11.09 | 0.10 | -1.52 |
| | | *RootMSE* | 22.02 | 26.90 | 31.39 | 20.55 | 29.38 | 12.43 | 9.36 | 11.30 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -11.61 | -14.88 | 2.09 | -0.45 | 2.57 | 2.77 | 9.86 | 12.01 |
| | | *RootMSE* | 12.20 | 15.14 | 11.53 | 9.59 | 9.55 | 6.26 | 13.92 | 14.39 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -3.52 | -4.51 | 0.76 | -0.09 | 0.58 | 6.24 | 4.06 | 16.36 |
| | | *RootMSE* | 3.97 | 4.87 | 4.34 | 3.76 | 3.20 | 7.29 | 8.66 | 16.77 |
| **101** | 50 | *Fehlgeschlagen* | 0 | 0 | 28 | 0 | 14 | 0 | 0 | 0 |
| | | *Bias* | -63.82 | -72.73 | 24.74 | -3.48 | 33.67 | -52.11 | -33.83 | -31.90 |
| | | *RootMSE* | 64.31 | 72.84 | 82.96 | 57.42 | 89.27 | 52.51 | 35.42 | 36.89 |
| **101** | 100 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -45.01 | -53.85 | 7.68 | -0.78 | 11.00 | -22.00 | 0.79 | 6.42 |
| | | *RootMSE* | 45.55 | 54.01 | 34.46 | 27.6 | 33.41 | 23.37 | 13.27 | 18.72 |
| **101** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -24.02 | -29.77 | 2.06 | -0.32 | 2.91 | 5.80 | 20.18 | 33.54 |
| | | *RootMSE* | 24.56 | 30.03 | 14.54 | 13.39 | 12.59 | 9.75 | 24.49 | 35.68 |
| **401** | 50 | *Fehlgeschlagen* | 0 | 0 | 329 | 0 | 323 | 0 | 0 | 0 |
| | | *Bias* | -355.48 | -369.04 | -75.40 | -137.37 | -56.59 | -343.96 | -318.25 | -306.38 |
| | | *RootMSE* | 355.66 | 369.07 | 157.95 | 193.47 | 154.91 | 344.02 | 318.42 | 307.67 |
| **401** | 100 | *Fehlgeschlagen* | 0 | 0 | 19 | 0 | 17 | 0 | 0 | 0 |
| | | *Bias* | -319.97 | -341.04 | 127.17 | -2.51 | 142.87 | -292.70 | -245.62 | -197.88 |
| | | *RootMSE* | 320.24 | 341.09 | 391.14 | 242.31 | 390.59 | 292.85 | 246.08 | 202.14 |
| **401** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -263.56 | -292.42 | 24.06 | -4.49 | 36.68 | -207.39 | -132.37 | -22.49 |
| | | *RootMSE* | 263.95 | 292.51 | 121.34 | 102.23 | 118.45 | 207.77 | 133.93 | 56.68 |

Tabelle 5.1: Simulationsergebnisse für gleichwahrscheinliche seltene Spezien und $p = 0.4$

| Schätzername $S$ | $n$ | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| **51** | 50 | *Fehlgeschlagen* | 0 | 2 | 58 | 0 | 36 | 0 | 0 | 0 |
| | | *Bias* | -31.51 | -33.38 | 11.59 | -2.41 | 17.20 | -20.32 | -9.93 | -21.98 |
| | | *RootMSE* | 31.73 | 33.50 | 40.23 | 28.50 | 46.45 | 21.01 | 13.02 | 23.43 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -20.81 | -22.34 | 5.75 | -0.23 | 5.77 | -4.52 | 6.16 | -4.67 |
| | | *RootMSE* | 21.11 | 22.55 | 23.89 | 15.59 | 17.91 | 7.38 | 11.44 | 9.27 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -8.95 | -9.95 | 1.41 | -0.14 | 1.54 | 6.08 | 9.19 | 9.53 |
| | | *RootMSE* | 9.31 | 10.27 | 7.73 | 6.71 | 6.27 | 7.89 | 13.04 | 10.92 |
| **101** | 50 | *Fehlgeschlagen* | 0 | 0 | 184 | 0 | 161 | 0 | 0 | 0 |
| | | *Bias* | -78.98 | -81.60 | 9.45 | -14.26 | 18.92 | -65.60 | -51.23 | -65.88 |
| | | *RootMSE* | 79.12 | 81.67 | 63.50 | 54.45 | 69.42 | 65.89 | 52.10 | 66.81 |
| **101** | 100 | *Fehlgeschlagen* | 0 | 0 | 4 | 0 | 1 | 0 | 0 | 0 |
| | | *Bias* | -63.97 | -66.80 | 21.76 | 0.57 | 26.08 | -40.21 | -18.92 | -35.42 |
| | | *RootMSE* | 64.19 | 66.92 | 76.82 | 49.62 | 73.05 | 40.98 | 22.73 | 38.32 |
| **101** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -42.20 | -44.55 | 5.64 | 0.66 | 7.62 | -8.52 | 13.14 | 4.38 |
| | | *RootMSE* | 42.50 | 44.75 | 26.06 | 22.32 | 23.59 | 12.04 | 19.44 | 14.24 |
| **401** | 50 | *Fehlgeschlagen* | 0 | 0 | 628 | 0 | 626 | 0 | 0 | 0 |
| | | *Bias* | -377.09 | -380.39 | -229.29 | -241.55 | -215.66 | -362.14 | -344.44 | -360.88 |
| | | *RootMSE* | 377.13 | 380.41 | 239.42 | 253.17 | 228.14 | 362.20 | 344.59 | 361.16 |
| **401** | 100 | *Fehlgeschlagen* | 0 | 0 | 151 | 0 | 141 | 0 | 0 | 0 |
| | | *Bias* | -356.37 | -361.29 | 46.34 | -60.31 | 66.73 | -325.54 | -291.58 | -308.07 |
| | | *RootMSE* | 356.44 | 361.33 | 247.02 | 211.83 | 257.74 | 325.67 | 291.91 | 308.93 |
| **401** | 200 | *Fehlgeschlagen* | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -320.38 | -327.18 | 77.07 | -2.81 | 89.99 | -262.46 | -204.10 | -197.15 |
| | | *RootMSE* | 320.49 | 327.24 | 290.80 | 173.33 | 283.03 | 262.73 | 204.93 | 200.31 |

Tabelle 5.2: Simulationsergebnisse für gleichwahrscheinliche seltene Spezien und $p = 0.6$

| Schätzername | | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| $S$ | $n$ | | | | | | | | | |
| **26** | 50 | *Fehlgeschlagen* | 0 | 514 | 273 | 0 | 217 | 0 | 0 | 0 |
| | | *Bias* | -15.67 | -14.92 | 0.80 | -3.12 | 4.85 | -9.95 | -4.57 | -15.69 |
| | | *RootMSE* | 15.82 | 14.98 | 15.32 | 14.83 | 18.84 | 10.78 | 7.90 | 16.07 |
| **26** | 100 | *Fehlgeschlagen* | 0 | 680 | 12 | 0 | 1 | 0 | 0 | 0 |
| | | *Bias* | -10.14 | -8.69 | 5.38 | -0.49 | 5.53 | -2.15 | 3.26 | -9.39 |
| | | *RootMSE* | 10.37 | 8.78 | 16.77 | 10.56 | 14.56 | 4.65 | 7.45 | 10.06 |
| **26** | 200 | *Fehlgeschlagen* | 0 | 1000 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -4.06 | | 1.77 | -0.12 | 1.38 | 3.01 | 4.75 | -2.37 |
| | | *RootMSE* | 4.50 | | 6.98 | 4.76 | 4.91 | 4.60 | 7.83 | 3.87 |
| **51** | 50 | *Fehlgeschlagen* | 0 | 403 | 411 | 0 | 393 | 0 | 0 | 0 |
| | | *Bias* | -39.87 | -39.27 | -12.89 | -18.82 | -8.39 | -32.84 | -25.79 | -39.42 |
| | | *RootMSE* | 39.95 | 39.31 | 25.74 | 27.48 | 25.12 | 33.18 | 26.78 | 39.67 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 367 | 54 | 0 | 37 | 0 | 0 | 0 |
| | | *Bias* | -32.50 | -31.77 | 12.72 | -2.33 | 17.12 | -20.19 | -9.61 | -30.25 |
| | | *RootMSE* | 32.64 | 31.84 | 41.75 | 29.17 | 45.35 | 20.98 | 13.06 | 30.76 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 586 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -21.52 | -19.63 | 6.13 | -0.14 | 5.90 | -4.67 | 6.04 | -15.53 |
| | | *RootMSE* | 21.76 | 19.72 | 26.78 | 16.46 | 19.35 | 7.66 | 11.67 | 16.72 |
| **201** | 50 | *Fehlgeschlagen* | 0 | 347 | 812 | 0 | 805 | 0 | 0 | 0 |
| | | *Bias* | -189.24 | -188.69 | -148.73 | -150.80 | -143.53 | -180.96 | -172.03 | -188.40 |
| | | *RootMSE* | 189.26 | 188.70 | 151.43 | 153.48 | 146.61 | 181.04 | 172.23 | 188.48 |
| **201** | 100 | *Fehlgeschlagen* | 0 | 176 | 389 | 0 | 374 | 0 | 0 | 0 |
| | | *Bias* | -180.02 | -179.89 | -46.41 | -76.61 | -34.60 | -163.17 | -146.36 | -175.57 |
| | | *RootMSE* | 180.05 | 179.92 | 87.53 | 102.36 | 86.16 | 163.33 | 146.75 | 175.77 |
| **201** | 200 | *Fehlgeschlagen* | 0 | 68 | 27 | 0 | 18 | 0 | 0 | 0 |
| | | *Bias* | -162.60 | -162.94 | 52.00 | -11.04 | 61.96 | -130.92 | -101.65 | -145.74 |
| | | *RootMSE* | 162.67 | 163.00 | 161.16 | 106.07 | 167.58 | 131.26 | 102.62 | 146.33 |

Tabelle 5.3: Simulationsergebnisse für gleichwahrscheinliche seltene Spezien und $p = 0.8$

| Schätzername $S$ | $n$ | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| **51** | 50 | *Fehlgeschlagen* | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -21.59 | -26.85 | 10.25 | -0.59 | 10.57 | -11.23 | 0.02 | -1.95 |
| | | *RootMSE* | 22.24 | 27.03 | 37.18 | 20.88 | 29.42 | 12.59 | 9.51 | 11.54 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -11.89 | -15.08 | 2.15 | -0.49 | 2.24 | 2.48 | 9.66 | 11.30 |
| | | *RootMSE* | 12.45 | 15.33 | 12.06 | 9.85 | 9.29 | 6.02 | 13.68 | 13.59 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -3.74 | -4.74 | 0.70 | -0.19 | 0.45 | 6.05 | 4.15 | 15.79 |
| | | *RootMSE* | 4.21 | 5.11 | 4.49 | 3.92 | 3.34 | 7.17 | 8.74 | 16.25 |
| **101** | 50 | *Fehlgeschlagen* | 0 | 0 | 22 | 0 | 12 | 0 | 0 | 0 |
| | | *Bias* | -64.82 | -73.22 | 21.94 | -7.72 | 26.80 | -52.98 | -35.19 | -34.53 |
| | | *RootMSE* | 65.24 | 73.32 | 85.43 | 54.59 | 85.13 | 53.35 | 36.62 | 38.70 |
| **101** | 100 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -46.56 | -54.79 | 2.81 | -5.22 | 4.78 | -24.01 | -2.31 | 1.65 |
| | | *RootMSE* | 47.04 | 54.94 | 32.63 | 26.87 | 28.94 | 25.25 | 13.27 | 16.97 |
| **101** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -26.44 | -31.63 | -0.87 | -3.28 | -0.20 | 2.50 | 16.32 | 26.87 |
| | | *RootMSE* | 26.92 | 31.87 | 14.63 | 13.85 | 12.24 | 8.14 | 21.19 | 29.34 |
| **401** | 50 | *Fehlgeschlagen* | 0 | 0 | 124 | 0 | 101 | 0 | 0 | 0 |
| | | *Bias* | -359.18 | -370.61 | -172.29 | -228.06 | -155.63 | -347.21 | -324.38 | -317.87 |
| | | *RootMSE* | 359.31 | 370.63 | 214.71 | 251.33 | 206.38 | 347.28 | 324.55 | 318.80 |
| **401** | 100 | *Fehlgeschlagen* | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 |
| | | *Bias* | -330.54 | -346.00 | -150.97 | -188.61 | -149.83 | -303.78 | -266.06 | -244.03 |
| | | *RootMSE* | 330.71 | 346.04 | 209.79 | 222.10 | 204.95 | 303.92 | 266.50 | 246.03 |
| **401** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -290.91 | -307.59 | -164.75 | -175.59 | -164.94 | -241.50 | -191.15 | -156.86 |
| | | *RootMSE* | 291.09 | 307.65 | 175.28 | 183.56 | 173.55 | 241.80 | 192.20 | 160.32 |

Tabelle 5.4: Simulationsergebnisse für geometrisch verteilte seltene Spezien und $p = 0.4$

| Schätzername $S$ | $n$ | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| **51** | 50 | *Fehlgeschlagen* | 0 | 0 | 43 | 0 | 26 | 0 | 0 | 0 |
| | | *Bias* | -31.66 | -33.52 | 12.89 | -3.43 | 16.51 | -20.50 | -10.13 | -22.34 |
| | | *RootMSE* | 31.87 | 33.63 | 41.17 | 25.76 | 42.86 | 21.15 | 13.03 | 23.70 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -20.83 | -22.37 | 4.81 | -0.46 | 5.72 | -4.61 | 6.00 | -4.80 |
| | | *RootMSE* | 21.13 | 22.57 | 20.70 | 16.79 | 18.07 | 7.53 | 11.56 | 9.33 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -9.44 | -10.44 | 1.09 | -0.52 | 1.09 | 5.47 | 8.89 | 8.36 |
| | | *RootMSE* | 9.81 | 10.78 | 7.50 | 6.57 | 6.09 | 7.42 | 12.55 | 10.06 |
| **101** | 50 | *Fehlgeschlagen* | 0 | 1 | 162 | 0 | 144 | 0 | 0 | 0 |
| | | *Bias* | -79.31 | -81.83 | 5.28 | -19.89 | 13.64 | -66.13 | -52.17 | -66.76 |
| | | *RootMSE* | 79.44 | 81.89 | 62.12 | 51.94 | 65.91 | 66.40 | 52.98 | 67.60 |
| **101** | 100 | *Fehlgeschlagen* | 0 | 0 | 3 | 0 | 1 | 0 | 0 | 0 |
| | | *Bias* | -64.88 | -67.49 | 14.01 | -5.96 | 15.84 | -41.84 | -21.63 | -38.17 |
| | | *RootMSE* | 65.07 | 67.60 | 69.34 | 44.26 | 62.69 | 42.55 | 24.89 | 40.46 |
| **101** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -43.77 | -45.90 | 0.78 | -3.98 | 1.83 | -11.55 | 8.76 | -0.77 |
| | | *RootMSE* | 44.04 | 46.09 | 24.33 | 21.66 | 21.03 | 14.27 | 16.54 | 12.95 |
| **401** | 50 | *Fehlgeschlagen* | 0 | 2 | 417 | 0 | 395 | 0 | 0 | 0 |
| | | *Bias* | -377.41 | -380.55 | -253.09 | -271.29 | -238.94 | -362.94 | -346.27 | -361.83 |
| | | *RootMSE* | 377.45 | 380.57 | 264.01 | 280.39 | 252.13 | 363.01 | 346.41 | 362.08 |
| **401** | 100 | *Fehlgeschlagen* | 0 | 0 | 35 | 0 | 21 | 0 | 0 | 0 |
| | | *Bias* | -359.71 | -363.64 | -131.08 | -201.33 | -123.56 | -331.51 | -302.47 | -320.94 |
| | | *RootMSE* | 359.76 | 363.67 | 217.24 | 233.76 | 214.55 | 331.63 | 302.76 | 321.51 |
| **401** | 200 | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -329.50 | -334.12 | -164.75 | -186.86 | -164.72 | -280.64 | -236.83 | -246.01 |
| | | *RootMSE* | 329.58 | 334.17 | 186.30 | 198.90 | 180.84 | 280.86 | 237.49 | 247.34 |

Tabelle 5.5: Simulationsergebnisse für geometrisch verteilte seltene Spezien und $p = 0.6$

| Schätzername | | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|---|
| $S$ | $n$ | | | | | | | | | |
| **26** | 50 | *Fehlgeschlagen* | 0 | 518 | 254 | 0 | 198 | 0 | 0 | 0 |
| | | *Bias* | -15.82 | -15.02 | 0.23 | -4.73 | 4.00 | -10.28 | -5.10 | -15.88 |
| | | *RootMSE* | 15.97 | 15.08 | 14.91 | 13.68 | 17.71 | 11.09 | 8.09 | 16.28 |
| **26** | 100 | *Fehlgeschlagen* | 0 | 699 | 17 | 0 | 1 | 0 | 0 | 0 |
| | | *Bias* | -10.28 | -8.63 | 5.69 | -0.32 | 6.61 | -2.31 | 3.13 | -9.61 |
| | | *RootMSE* | 10.53 | 8.73 | 18.05 | 10.86 | 18.24 | 4.86 | 7.50 | 10.34 |
| **26** | 200 | *Fehlgeschlagen* | 0 | 1000 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -4.01 | | 2.18 | 0.12 | 1.42 | 3.11 | 4.91 | -2.29 |
| | | *RootMSE* | 4.44 | | 8.32 | 5.13 | 4.81 | 4.72 | 8.11 | 3.76 |
| **51** | 50 | *Fehlgeschlagen* | 0 | 427 | 437 | 0 | 405 | 0 | 0 | 0 |
| | | *Bias* | -39.97 | -39.26 | -12.62 | -18.29 | -7.34 | -32.95 | -25.82 | -39.60 |
| | | *RootMSE* | 40.05 | 39.30 | 25.40 | 27.40 | 25.48 | 33.30 | 26.82 | 39.84 |
| **51** | 100 | *Fehlgeschlagen* | 0 | 364 | 54 | 0 | 29 | 0 | 0 | 0 |
| | | *Bias* | -32.60 | -31.84 | 11.20 | -3.63 | 15.97 | -20.40 | -9.95 | -30.41 |
| | | *RootMSE* | 32.73 | 31.91 | 39.59 | 26.43 | 42.99 | 21.16 | 13.10 | 30.93 |
| **51** | 200 | *Fehlgeschlagen* | 0 | 585 | 1 | 0 | 0 | 0 | 0 | 0 |
| | | *Bias* | -21.59 | -19.60 | 5.27 | -0.35 | 5.44 | -4.83 | 5.86 | -15.64 |
| | | *RootMSE* | 21.84 | 19.68 | 21.58 | 16.71 | 17.38 | 7.86 | 11.61 | 16.88 |
| **201** | 50 | *Fehlgeschlagen* | 0 | 320 | 738 | 0 | 734 | 0 | 0 | 0 |
| | | *Bias* | -189.20 | -188.85 | -148.65 | -152.54 | -143.56 | -180.95 | -172.14 | -188.35 |
| | | *RootMSE* | 189.22 | 188.86 | 151.07 | 155.12 | 146.39 | 181.03 | 172.32 | 188.42 |
| **201** | 100 | *Fehlgeschlagen* | 0 | 164 | 311 | 0 | 292 | 0 | 0 | 0 |
| | | *Bias* | -180.06 | -180.08 | -61.99 | -87.04 | -52.63 | -163.51 | -147.21 | -175.70 |
| | | *RootMSE* | 180.09 | 180.10 | 93.11 | 109.96 | 90.02 | 163.65 | 147.58 | 175.86 |
| **201** | 200 | *Fehlgeschlagen* | 0 | 96 | 12 | 0 | 4 | 0 | 0 | 0 |
| | | *Bias* | -163.84 | -164.02 | 7.98 | -45.12 | 11.46 | -134.07 | -107.22 | -148.98 |
| | | *RootMSE* | 163.91 | 164.07 | 141.22 | 93.14 | 132.21 | 134.40 | 108.15 | 149.51 |

Tabelle 5.6: Simulationsergebnisse für geometrisch verteilte seltene Spezien und $p = 0.8$

Es zeigt sich, dass der parametrische Schätzer nach Darroch in allen betrachteten Fällen einen stark negativen Bias besitzt. Dies war auch so zu erwarten, da dieser Schätzer die parametrische Grundannahme der gleichwahrscheinlichen Spezien verwendet und diese in unseren Simulationen nicht zutrifft. Jede Abweichung von dieser Grundannahme führt dazu, dass dieser Schätzer geringer ausfällt und somit die tatsächliche Spezienanzahl unterschätzt wird. Aufgrund der Darstellung

$$\mathrm{MSE}(\hat{S}) = \mathrm{Bias}^2(\hat{S}) + \mathrm{Var}(\hat{S}) \tag{5.1}$$

ist auch der RootMSE für diesen Schätzer groß. Insgesamt eignet sich der Schätzer von Darroch in dieser Situation nicht um den Artenreichtum zu schätzen.

Interessant ist jedoch, dass der Schätzer über die inverse Gauß-Verteilung (vgl. Abschnitt 4.2.2), der einen allgemeineren Ansatz verwendet als der Schätzer nach Darroch und die Verteilung der Zellhäufigkeiten aus den Daten heraus approximiert, sowohl bezüglich Bias als auch bezüglich RootMSE noch schlechtere Ergebnisse liefert als der Schätzer nach Darroch. Zusätzlich lässt sich dieser Schätzer dann, wenn eine häufige Spezies mit einem Anteil von 80% entsteht, sehr häufig nicht berechnen oder teilweise gar nicht. Dies mag daran liegen, dass die von uns simulierten Verteilungen sich schlecht durch eine inverse Gauß-Verteilung approximieren lassen.

Dieser Schätzer eignet sich also ebenso wenig wie der Schätzer von Darroch den Artenreichtum in den von uns betrachteten Situationen zu schätzen.

Die Simulationen bestätigen auch die in Abschnitt 4.3.1 erwähnten Bedenken über den Schätzer von Chao. Im Falle gleichwahrscheinlicher seltener Spezien zeigt der ursprünglich von Chao vorgeschlagene Schätzer stark positiven Bias. Dies ist umso gravierender, da dieser Schätzer eigentlich als Schätzer einer unteren Schranke der Spezienanzahl konstruiert wurde. So wird beispielsweise bei $p = 0.4$, $S = 101$ und $n = 50$ die tatsächliche Spezienanzahl im Mittel um ca. 25% überschätzt. Dies überträgt sich auch teilweise auf den Fall geometrisch verteilter seltener Spezien. Falls die tatsächliche Spezienanzahl nicht zu groß ist ($\leq 200$), besitzt der Schätzer nach Chao im Mittel einen positiven Bias. Gleichzeitig lässt er sich für größeren Artenreichtum oftmals nicht berechnen, wie beispielsweise bei $p = 0.8$, $S = 201$ und $n = 50$ in 812 bzw. 738 von 1000 Fällen.

Diese beiden Probleme wurden bereits im gleichen Abschnitt 4.3.1 angesprochen und durch eine Bias Korrektur beseitigt. Tabelle 5.1 bis 5.6 zeigt nun zum einen die in Satz 4.9 bewiesene Unverzerrtheit des unteren Schranken-Schätzers *Chao Bias Corrected* und gleichzeitig wird deutlich, dass die Eigenschaften dieses Schätzers als Punktschätzer des Artenreichtums im Falle gleichwahrscheinlicher seltener Spezien sehr ermutigend sind, da meistens nur ein zu vernachlässigender Bias vorliegt. Andererseits wird auch deutlich, dass eine Bias Korrektur nicht nur im Falle gleichwahrscheinlicher Spezien, wie von Frau Prof. Anne Chao gefordert, sondern auch in anderen Gelegenheiten notwendig ist. Die Simulationsergebnisse zeigen, dass diese sowohl im Falle gleichverteilter seltener Spezien als auch im Falle geometrisch verteilter seltener Spezien notwendig ist, da beispielsweise für $p = 0.6$, $S = 101$ und $n = 100$ der Artenreichtum um 14% überschätzt wird, während Chao Bias Corrected den Artenreichtum unter gleichen Bedingungen im Mittel um 6% unterschätzt, wie dies von einem Schätzer einer unteren Schranke des Artenreichtums auch zu erwarten wäre. Insgesamt untermauert dies nochmals unsere Empfehlung den Schätzer Chao Bias Corrected in allen Gelegenheiten, in denen der Artenreichtum geschätzt werden soll, anstelle des ursprünglich vorgeschlagenen Schätzers von Chao zu verwenden.

In den folgenden Auswertungen der Simulationsergebnisse werden wir deshalb nur noch den Schätzer *Chao Bias Corrected* anstelle des Schätzers nach Chao betrachten und aus oben genannten Gründen wird auch nicht weiter auf die Schätzer nach Darroch und Sichel (*Inverse Gaussian*) eingegangen.

Die Tabellen zeigen außerdem, dass das Verhalten der verbleibenden Schätzer *Chao Bias Corrected*, *Mod. Sample Coverage*, *Jackknife 1*, *Jackknife 2* und *Michaelis Menten* stark von den betrachteten Rahmenbedingungen bzw. den Zusatzannahmen, die man der Schätzung des Artenreichtums unterstellt, abhängen.

Betrachten wir zuerst die Simulationen, in denen den seltenen Spezien eine Gleichverteilung unterstellt wird. In diesen zeigt der Schätzer Chao Bias Corrected fast durchwegs den kleinsten Bias und ist insbesondere bei größerem Artenreichtum den übrigen Schätzern sowohl im Bezug auf Bias als auch auf RootMSE überlegen. Bei $S = 401$ und $n = 200$ liegt der Bias dieses Schätzers im Mittel 5 bzw. 30-Mal geringer als der der anderen betrachteten Verfahren, während er gleichzeitig den zweitkleinsten bzw. kleinsten RootMSE besitzt. Für kleinere Stichprobengrößen insbesondere im Bereich von 50 bis 70, der Anzahl der Durchführungen in den Experimenten von Uwe Simon, liegt jedoch der relative RootMSE (relativ zur tatsächlichen Spezienanzahl) teilweise bei 50%.

Generell ist zu beobachten, dass immer dann, wenn die Anzahl der Messdurchgänge deutlich unter der Spezienanzahl ($n < S/2$) liegt alle Schätzer einen großen RootMSE aufweisen. Dies ist insbesondere in der Konstellation $S = 401$ und $n = 50$ zu beobachten, bei der der modifizierte Ansatz über den Sample Coverage mit $\approx 40\%$ den geringsten relativen RootMSE besitzt. Gleiches Verhalten zeigt sich bei $S = 401$ und $n = 100$. Während hier das Verfahren über Sample Coverage und der Schätzer nach Chao einen vergleichsweise kleinen Bias verfügen besitzen die Jackknife-Methoden und der datenanalytische Schätzer nach Michaelis Menten einen vergleichsweise großen Bias bei einem annähernd gleichem RootMSE. Aus der Zerlegung des MSE aus Gleichung (5.1) kann damit geschlossen werden, dass erstere Verfahren eine wesentlich größere Streuung in den Schätzungen aufweisen.

Interessant ist jedoch das Verhalten der Schätzer immer dann, wenn die Spezienanzahl $S$ mit dem Stichprobenumfang $n$ übereinstimmt oder geringfügig kleiner ist. In dieser Situation liefern die Methoden Jackknife 1 und Jackknife 2 durchwegs bezüglich RootMSE die niedrigsten Werte und schätzen demnach die Spezienanzahl mit sehr guter Genauigkeit. Dabei ist unter diesen Beiden die Methode Jackknife 2 zu bevorzugen, die im Grunde aus einer Erweiterung des Verfahrens Jackknife 1 entstanden ist, wie in Abschnitt 4.3.4 dargestellt wurde. Generell liefern diese beiden Verfahren immer dann, wenn $S \leq n$ erfüllt ist zuverlässige Schätzungen des Artenreichtums.

Der Schätzer über den modifizierten Sample Coverage Ansatz liefert für gleichwahrscheinliche seltene Spezien oftmals einen stark positiven Bias und überschätzt somit im Mittel die tatsächlich vorliegende Spezienanzahl. Insbesondere dann, wenn die Stichprobengröße kleiner als die Speziengesamtanzahl ist, beispielsweise für $p = 0.6$, $S = 401$ und $n = 100$ oder 200.

Abschließend bleibt noch festzuhalten, dass für größere Spezienanzahlen ($S > 400$ für $p = 0.4$ und $p = 0.6$ bzw. $S > 200$ für $p = 0.8$) es unter den betrachteten Rahmenbedingungen nahezu aussichtslos ist die Spezienanzahl in den betrachteten Situationen mit einer vernünftigen Genauigkeit zu schätzten, da der relative RootMSE durchwegs über

0.14 für $p = 0.4$, 0.43 für $p = 0.6$ bzw. 0.51 für $p = 0.8$ liegt.

Betrachten wir nun die Simulationen mit zum Parameter $\lambda = 0.01$ geometrisch verteilten seltenen Spezien. Insgesamt ist das Verhalten aller Schätzer recht ähnlich, sodass die vorher gemachten Aussagen weitgehend auf diese Situation übertragen werden können. Interessant ist wieder, dass, wenn die Spezienanzahl $S$ mit dem Stichprobenumfang $n$ übereinstimmt oder geringfügig kleiner, ist die Methoden Jackknife 1 und Jackknife 2 den anderen bezüglich RootMSE deutlich überlegen sind, wobei auch hier wieder der Schätzer Jackknife 2 vorzuziehen ist. Auch ist wieder zu Beobachten dass diese beiden Verfahren immer dann, wenn $S \leq n$ erfüllt ist zuverlässige Schätzungen des Artenreichtums liefern.

Der modifizierte Schätzer über den Sample Coverage Ansatz neigt nun nicht mehr so stark dazu die tatsächliche Spezienanzahl zu überschätzen und liefert teilweise insbesondere für $S = 101$ und $n = 200$ nur einen Bias von 2% und somit bessere Ergebnisse bezüglich des Bias als der Bias corrected Chao Schätzer. Gleichzeitig liegt der RootMSE bei 12.24 bzw. 21.03. Dieser Schätzer kann also die Heterogenität in den seltenen Spezien mitberücksichtigen und eignet sich gut als Punktschätzer der Spezienanzahl, wenn die Stichprobengröße groß genug ist.

Der Schätzer Chao Bias Corrected liefert nun nur noch für kleine Spezienanzahlen ($S = 51$) den geringsten Bias und zeigt mit zunehmender Stichprobengröße stärker negativen Bias, was für einen Schätzer einer unteren Schranke der Spezienanzahl nicht weiter beunruhigend ist. Gleichzeitig ist es jedoch auch so, dass von keinem der Schätzer im Mittel mehr als 300 Spezien geschätzt wurden, auch wenn die tatsächliche Spezienanzahl deutlich höher lag.

Die Verfahren nach Jackknife und Michaelis Menten liefern insbesondere bei kleinen Stichproben ($n < S/2$) viel zu kleine Schätzungen des Artenreichtums, sodass der relative Bias und RootMSE auf durchschnittlich 50% ansteigt.

Die soeben gemachten Aussagen treffen auch auf die umfangreicheren Simulationen zu, die im Rahmen dieser Diplomarbeit gemacht wurden und sich als Anhang auf CD befinden.

Im Folgenden wollen wir nun noch das Querschnittsverhalten der Schätzer untersuchen. Dazu variieren wir für fest vorgegebene Stichprobengröße ($n = 250$) und Spezienanzahl ($S = 101$) den Parameter $\lambda$ der truncated geometrischen Verteilung der seltenen Spezien im Bereich von 0 bis 0.2. Dies soll es uns ermöglichen einschätzen zu können, wie die Schätzer auf Heterogenität in den seltenen Spezien reagieren. Für $\lambda = 0$ können wir dabei die Ergebnisse aus der Simulation für gleichwahrscheinliche seltene Spezien verwenden, da für $\lambda \to 0$ sich aus der truncated geometrischen Verteilung die Gleichverteilung ergibt.

Die Simulationsergebnisse dieses Querschnittsverhaltens sind in Tabelle 5.7 zusammengefasst.

| Schätzername | | Darroch | Inverse Gaussian | Chao | Chao Bias Corrected | Mod. Sample Coverage | Jackknife 1 | Jackknife 2 | Michaelis Menten |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda = 0$ | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -36.28 | -38.06 | 0.37 | -2.90 | 1.07 | -2.71 | 15.03 | 9.54 |
| | *RootMSE* | 36.60 | 38.30 | 17.20 | 16.00 | 15.13 | 8.77 | 20.45 | 15.52 |
| $\lambda = 0.001$ | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -34.73 | -36.68 | 3.35 | 0.05 | 4.72 | 0.04 | 18.50 | 14.39 |
| | *RootMSE* | 35.04 | 36.91 | 18.01 | 16.14 | 16.12 | 7.95 | 22.96 | 18.66 |
| $\lambda = 0.005$ | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -35.18 | -37.08 | 2.41 | -0.85 | 3.58 | -0.75 | 17.53 | 12.98 |
| | *RootMSE* | 35.51 | 37.34 | 18.14 | 16.47 | 15.98 | 8.37 | 22.42 | 17.91 |
| $\lambda = 0.01$ | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -36.45 | -38.21 | 0.74 | -2.62 | 1.16 | -2.91 | 14.99 | 8.97 |
| | *RootMSE* | 36.74 | 38.43 | 17.46 | 16.02 | 15.08 | 8.31 | 19.95 | 14.49 |
| $\lambda = 0.02$ | *Fehlgeschlagen* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -40.97 | -42.29 | -8.66 | -11.88 | -8.95 | -11.15 | 4.07 | -4.02 |
| | *RootMSE* | 41.22 | 42.50 | 19.24 | 19.50 | 17.12 | 13.79 | 14.14 | 11.55 |
| $\lambda = 0.05$ | *Fehlgeschlagen* | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -56.30 | -57.27 | -34.58 | -38.34 | -37.85 | -38.52 | -28.99 | -41.80 |
| | *RootMSE* | 56.40 | 57.37 | 37.84 | 40.35 | 39.31 | 39.07 | 30.92 | 42.27 |
| $\lambda = 0.10$ | *Fehlgeschlagen* | 0 | 975 | 5 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -71.12 | -66.52 | -59.02 | -63.01 | -62.20 | -62.68 | -58.01 | -67.71 |
| | *RootMSE* | 71.16 | 66.53 | 60.61 | 63.72 | 62.58 | 62.86 | 58.5 | 67.79 |
| $\lambda = 0.15$ | *Fehlgeschlagen* | 0 | 1000 | 35 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -78.17 | | -69.87 | -73.22 | -72.05 | -72.97 | -69.88 | -77.56 |
| | *RootMSE* | 78.20 | | 70.74 | 73.66 | 72.27 | 73.08 | 70.18 | 77.59 |
| $\lambda = 0.20$ | *Fehlgeschlagen* | 0 | 1000 | 77 | 0 | 0 | 0 | 0 | 0 |
| | *Bias* | -82.48 | | -76.81 | -79.62 | -77.95 | -79.00 | -76.78 | -82.94 |
| | *RootMSE* | 82.50 | | 77.23 | 79.81 | 78.11 | 79.08 | 76.96 | 82.96 |

Tabelle 5.7: Simulationsergebnisse für das Querschnittsverhalten der Schätzer

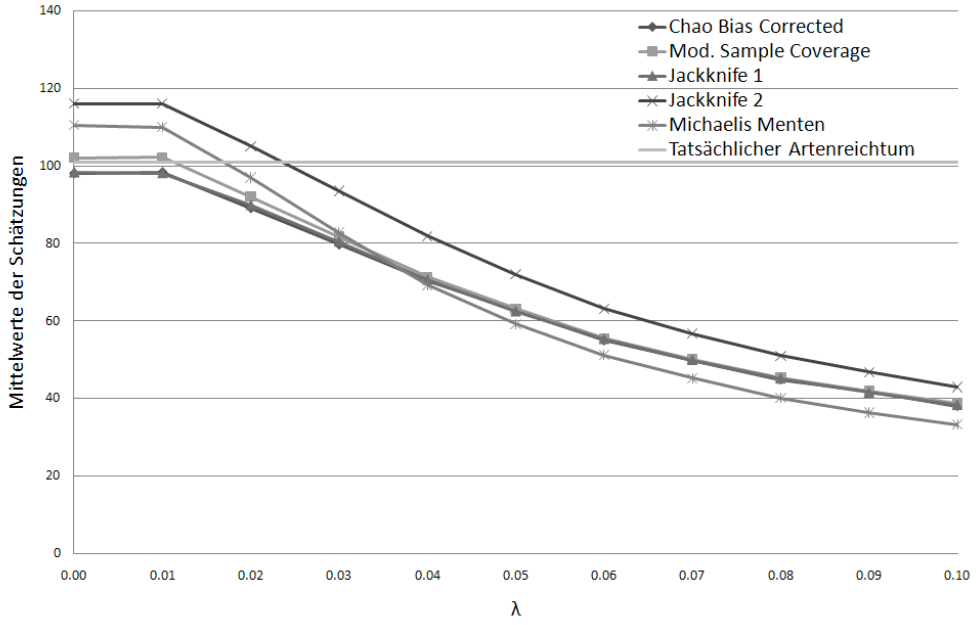


Abbildung 5.2: Mittelwerte der Schätzungen abgetragen gegen verwendetes $\lambda$ im Bereich von 0 bis 0.1 (Die durchgezogene Linie kennzeichnet die in der Simulation verwendete Spezienanzahl).

Wie bereits im Fall der gleichwahrscheinlichen Spezien festgestellt, wird ersichtlich, dass im Falle von $S = 101$ und $n = 250$ und $\lambda = 0$ (dies entspricht der Gleichverteilung) alle Schätzer bis auf der nach Darroch und Sichel die tatsächliche Spezienanzahl mit einer Abweichung von unter 20% (relativ zum Artenreichtum von 101) sowohl im Mittelwert als auch im RootMSE schätzen. Zusätzlich zeigt die Querschnittsanalyse insbesondere Abbildung 5.3, dass mit zunehmendem $\lambda$ der RootMSE aller Schätzer ansteigt.

Dabei wird deutlich, dass im Falle von $S = 101$ und $n = 250$, wenn also eine genügend große Stichprobengröße vorliegt, alle Schätzer weitgehend unverändert bei kleineren Abweichungen von der Gleichverteilung der seltenen Spezien bleiben, d.h. bei kleinen Werten von $\lambda$.

Im Bereich von $\lambda = 0$ bis $\lambda = 0.01$ beträgt die maximale Veränderung des RootMSE relativ zur Stichprobengröße lediglich 1.5% für alle Schätzer bis auf Michaelis Menten (4.2%) und Jackknife 2 (3.2%). Gleichzeitig bleibt auch der Bias der Schätzungen weitgehend konstant.

Für größere $\lambda$ steigt dann der RootMSE für alle Schätzer stark an. So verdoppelt sich der RootMSE für alle Schätzverfahren von $\lambda = 0.2$ auf $\lambda = 0.5$. Insgesamt ergeben sich aber keine klaren Präferenzen für einen der dargestellten Schätzer, der auch für größere Koeffizienten $\lambda$ oder über das gesamte Spektrum der verwendeten $\lambda$ den Artenreichtum zuverlässig schätzt. Während der Schätzer Jackknife 2 ab $\lambda = 0.03$ stets den kleinsten RootMSE besitzt überschätzt dieser für kleine Werte von $\lambda$ die tatsächliche Spezienanzahl deutlich und besitzt für $\lambda = 0.1$ oder 0 sogar den größten RootMSE unter den betrachteten Schätzverfahren. Alle weiteren Schätzer zeigen untereinander bezüglich Mittelwert (Bias) und RootMSE annähernd den gleichen Verlauf in Bezug auf $\lambda$, wie aus Abbildung 5.2 und 5.3 deutlich wird. Außerdem sieht man, dass ab $\lambda = 0.8$ alle aufgezeigten Schätzverfahren einen RootMSE relativ zur Spezienanzahl von über 50% aufweisen.

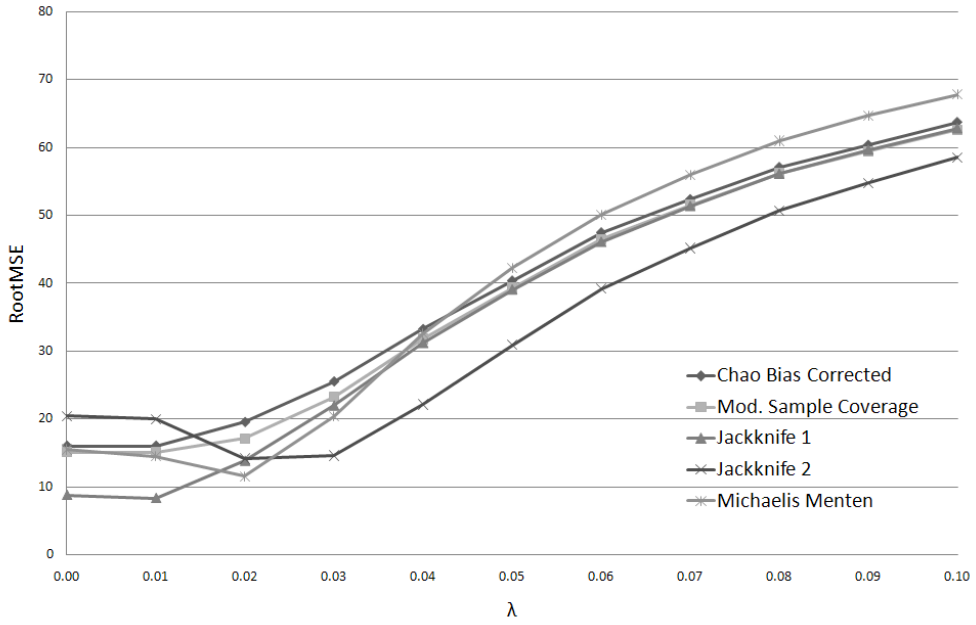


Abbildung 5.3: RootMSE der Schätzungen abgetragen gegen verwendetes $\lambda$ im Bereich von 0 bis 0.1.

# 6 Schätzung des Artenreichtums im konkreten Fall der Daten von Herrn Uwe Simon

In diesem Kapitel werden wir aus den theoretischen Betrachtungen aus Kapitel 4 zusammen mit den Ergebnissen der Simulationsstudie aus Kapitel 5 und Angaben aus der Literatur Schätzverfahren des Artenreichtums für den von uns konkret betrachteten Fall empfehlen und anschließend werden wir mithilfe dieser den Artenreichtum für die Daten von Uwe K. Simon und Michael Weiß aus Tabelle 1.1 schätzen.

Alle Empfehlungen für konkrete Schätzverfahren richten sich dabei nach den Werten für Bias, RootMSE und der erwarteten Anzahl der Fälle in denen die Schätzer nicht berechnet werden können.

## 6.1 Empfohlene Schätzverfahren

Die Simulationen haben gezeigt, dass parametrische Verfahren wie die nach Darroch und Sichel im Hinblick auf Bias und RootMSE ungeeignet sind in den von uns betrachteten Situationen den Artenreichtum zu schätzen. Wir können deshalb in dieser Situation nur nichtparametrische Verfahren und datenanalytische Methoden empfehlen. Durch nichtparametrische Herangehensweisen erreichen wir auch, dass wir keine Zusatzvoraussetzungen an die Verteilungen der Zellhäufigkeiten stellen müssen, um diese anwenden zu dürfen.

In der Onlinehilfe zu SPADE [CS03] wird empfohlen im Falle von heterogenen Zellhäufigkeiten den Schätzer nach Chao als einen Schätzer einer unteren Schranke der Spezienanzahl zu verwenden und den Schätzer aus dem modifizierten Sample Coverage Ansatz als Punktschätzer zu verwenden. Chao [Cha87] [Cha89] weist außerdem darauf hin, dass die Jackknife Schätzer oftmals die tatsächliche Spezienanzahl unterschätzen, insbesondere dann wenn alle Individuen nur ein oder zwei Mal gefangen werden. Dies wird auch durch Smith und van Belle [SvB84] bestätigt, der der Jackknife Methode zwar die Fähigkeit zuspricht den Bias zu verringern, wenn auch sie dazu neigt die tatsächliche Anzahl der Spezien zu unterschätzen, wenn es eine Vielzahl seltener Spezien gibt. Dies entspricht z. B. $n < S/2$ aus unseren Simulationen und konnte dort auch in dieser Weise bestätigt werden.

Huggins [Hug01] und Link [Lin03] haben allgemein gezeigt, dass die Größe einer Population bei Anwesenheit von Heterogenität (Modell $M_h$) im Allgemeinen nicht geschätzt werden kann. Selbst wenn die Stichprobengröße groß ist können zwei verschiedene Modelle, die die Daten jeweils sehr gut anpassen, deutlich verschiedene Schätzungen für die Spezienanzahl $S$ liefern. Demnach ist ohne zusätzliche Voraussetzungen eine Schätzung

einer unteren Schranke der Spezienanzahl oftmals das beste Resultat, das man erhalten kann.

Aus den Simulationen lassen sich deshalb nur unter gegebenen Zusatzvoraussetzungen oder Zusatzannahmen Empfehlungen aussprechen, die wir nun kurz zusammenfassen möchten. Wir nehmen wieder an, dass sich im Kollektiv eine Spezies befindet, die mit einer deutlich höheren Wahrscheinlichkeit bestimmt wird als die übrigen. Bei den Daten von Herrn Uwe Simon entpricht diese der Konsensusvariante.

Nehmen wir weiter an, dass die übrigen Spezien eine starke Heterogenität bezüglich ihrer Erscheinungswahrscheinlichkeiten aufweisen, so zeigen unsere Simulationsergebnisse insbesondere die Querschnittsuntersuchung, dass kein Schätzer in der Lage ist den Artenreichtum zuverlässig zu schätzen. Wir empfehlen deshalb den Schätzer Chao Bias Corrected als Schätzer einer unteren Schranke der Spezienanzahl. Dieser stellt die geringsten Anforderungen an die Zellwahrscheinlichkeiten. Es wird im Gegensatz zu den Jackknife Schätzverfahren nicht die Voraussetzung benötigt, dass der Erwartungswert der Schätzgröße eine Entwicklung in $\frac{1}{n}$ besitzt vgl. Gleichung (4.10). Es wird lediglich angenommen, dass die Erscheinungswahrscheinlichkeit $p$ eine zufällige Größe ist. Zudem wurde dieser direkt als Schätzer einer unteren Schranke entwickelt.

Wir sind außerdem der Meinung, dass aus Sicht des Anwenders eine untere Schranke des Artenreichtums oftmals sehr viel hilfreicher als ein größerer viel spekulativerer Wert für den gesamten Artenreichtum ist.

Können wir andererseits annehmen, dass die seltenen Spezien keine allzu große Heterogenität aufweisen, so können wir die aus der Simulationsstudie gewonnenen Ergebnisse verwenden und unter weiteren Annahmen an die Stichprobengröße, die wir nun aufzählen werden, die folgenden Schätzer empfehlen:

**Stichprobengröße $\approx$ erwartete Spezienanzahl** : *Jackknife 2*, wobei dieser eher noch die Spezienanzahl unterschätzt, aber einen geringen RootMSE aufweist.

**Stichprobengröße $>$ erwartete Spezienanzahl** : *Chao Bias Corrected* oder *Jackknife 1*, zuverlässige Schätzung mit geringem RootMSE.

**Stichprobengröße $<$ erwartete Spezienanzahl** : *Chao Bias Corrected*, als Punktschätzer, jedoch mit hohem RootMSE.

**Heterogenität in den seltenen Spezien** : *Mod. Sample Coverage*, als Punktschätzer der Spezienanzahl. Dieser besitzt jedoch einen hohen RootMSE und ist teilweise nicht berechenbar. Besser: *Chao Bias Corrected* als Schätzer einer unteren Schranke.

## 6.2 Ergebnisse der Schätzungen des Artenreichtums für die Daten von Uwe Simon

In diesem Abschnitt verwenden wir die Messergebnisse von Uwe K. Simon und Michael Weiß, die zusammengefasst in Tabelle 1.1 dargestellt sind, um daraus die entsprechenden Schätzungen des Artenreichtums nach den Formeln aus Kapitel 4 zu berechnen. In dieser

konkreten Situation handelt es sich bei dem Artenreichtum um eine Maßzahl für die Variabilität der entsprechenden Genregion.

Die Ergebnisse der empfohlenen Schätzer finden sich auf die zweite Nachkommastelle gerundet in Tabelle 6.1.

| **Pilzart** | **Genregion** | **Chao Bias Corrected** | **Mod. Sample Coverage** | **Jackknife 1** | **Jackknife 2** |
|---|---|---|---|---|---|
| Davidiella tassiana | SSU | 654.40 | Fehlgeschlagen | 72.28 | 106.84 |
| | ITS | 65.96 | Fehlgeschlagen | 22.79 | 33.38 |
| | LSU | 215.29 | 436.00 | 57.45 | 83.41 |
| Mycosphaerella punctiformis | SSU | 458.75 | Fehlgeschlagen | 60.50 | 89.50 |
| | ITS | 135.29 | Fehlgeschlagen | 32.77 | 48.31 |
| | LSU | 187.89 | Fehlgeschlagen | 38.65 | 56.96 |
| Phoma exigua var. exigua | SSU | 215.95 | 436.00 | 57.55 | 83.69 |
| | ITS | 37.50 | 57.47 | 17.86 | 25.57 |
| | LSU | 115.24 | 232.00 | 41.63 | 59.94 |
| Teratosphaeria microspora | SSU | 161.66 | 326.00 | 49.59 | 71.83 |
| | ITS | 55.18 | Fehlgeschlagen | 20.82 | 30.45 |
| | LSU | 90.58 | Fehlgeschlagen | 26.76 | 39.29 |

Tabelle 6.1: Schätzer des Artenreichtums aus den Daten von Uwe K. Simon und Michael Weiß

Aus Tabelle 6.1 wird sofort ersichtlich, dass die erhaltenen Schätzer des Artenreichtums, die durch die verschiedenen empfohlenen Schätzverfahren bestimmt wurden, stark unterschiedlich sind. Bei Pilzart *Davidiella tassiana* und Genregion *SSU* unterscheiden sich kleinster und größter Schätzwert sogar um den Faktor 9. Eigentlich wäre zu erwarten, dass sämtliche Verfahren, die schließlich alle die gleiche Größe $S$ schätzen, ähnliche Werte in dieser konkreten Situation liefern.

Besonders auffällig ist die Tatsache, dass der Schätzer *Chao Bias Corrected*, der als Schätzer einer unteren Schranke des Artenreichtums konstruiert wurde, im Vergleich zu den Schätzern über die Jackknife Methode deutlich höhere Werte annimmt. Gleichzeitig liefert der modifizierte Schätzer über den Sample Coverage Ansatz, falls er berechenbar ist, die größten Schätzwerte für den Artenreichtum.

Es lässt sich aus unseren bisherigen Erkenntnissen nicht folgern, ob der tatsächliche Artenreichtum eher durch die Werte der Verfahren *Chao Bias Corrected* und *Mod. Sample Coverage* approximiert wird und die Jackknife Methoden den Artenreichtum stark unterschätzen oder ob die Verfahren *Jackknife 1* und *Jackknife 2* den Artenreichtum besser schätzen und die Abweichungen der beiden anderen Schätzer auf die starke Streuung bzw. den positiven Bias dieser Schätzmethoden zurückzuführen ist. Diese beiden Phänomene konnten in den Simulationen nachgewiesen werden.

Insgesamt kann keine klare Präferenz für einen dieser Schätzer ausgesprochen werden, wenn man keine der im vorherigen Abschnitt aufgezählten Zusatzannahmen treffen möchte. Gleichzeitig betonen wir hier, dass sich durch die Entscheidung für ein Schätzverfahren mitunter schwerwiegende Konsequenzen ergeben können, da sich die erhaltenen Schätzwerte stark unterscheiden können.

# 7 Ausblick

Wir möchten nun noch kurz einen Ausblick geben, der über die Thematik dieser Diplomarbeit hinaus geht, und kurz darstellen, welche Forschungsschwerpunkte im Bereich der Schätzung des Artenreichtums (*species richness estimation*) in der Literatur beschrieben werden.

Der in dieser Arbeit betrachtete Fall, nämlich einer Beobachtung pro Messdurchgang, kann beispielsweise dahingehend verallgemeinert werden, dass mehr als ein Individuum pro Messdurchgang untersucht wird. Man nimmt dann an, dass die Anzahl der Individuen pro Messdurchgang selbst eine zufällige Größe ist. Diese Situation trifft beispielsweise zu, wenn ein Fischer mehrmals hintereinander eine zufällige Anzahl an Fischen aus einem Teich fängt.
Eine andere Verallgemeinerung wäre, dass die einzelnen Messdurchgänge Teilabschnitte eines größeren Flächenbereiches sind, die nacheinander untersucht werden.

Die momentane Forschung konzentriert sich bei der Beschreibung dieser Experimente nicht mehr alleine auf die Modelle gleicher Spezienwahrscheinlichkeiten (Modelle $M_0$ und $M_t$) oder auf Modelle nur mit Heterogenität innerhalb der Spezienerscheinungswahrscheinlichkeiten (Modell $M_h$), sondern es wird versucht die komplexeren Modelle, wie sie in Kapitel 2 genannt sind, zu beschreiben. Dabei werden beispielsweise bestehende Schätzer für das Modell $M_h$ um einen direkten Einfluss der Messdurchgänge erweitert (Modelle $M_{th}$ und $M_{tbh}$). So entwickelten Rivest und Baillareon [RB07] einen mit dem von Chao vergleichbaren Schätzer einer unteren Schranke der Spezienanzahl für das Modell $M_{th}$. Shen und He [SH08] ließen hingegen die Annahme des Ziehens mit Zurücklegen fallen und entwickelten Schätzer unter der Annahme des Ziehens ohne Zurücklegen.

Ebenso werden auch zunehmend erweiterte bzw. andere Fragestellungen im Rahmen der Schätzungen des Artenreichtums untersucht. Es werden nicht mehr nur feste Kollektive einzeln betrachtet und deren Artenreichtum bestimmt, sondern es werden verschiedene Kollektive vergleichend gegenübergestellt. So wird die Anzahl der Spezien untersucht, die man in einer weiteren bzw. größeren Stichprobe finden würde (Shen et al. [SCL03]) oder man interesiert sich für die gemeinsame Spezienanzahl zweier Grundgesamtheiten (Chao et al. [CSH06]).

Aber auch wenn sich die momentane Forschung schon weg von dem Modell bewegt, das einen expliziten Einfluss der Messdurchgänge ausschließt (Modell $M_h$), so hat unsere Arbeit doch gezeigt, dass in diesem Bereich noch kein universell einsetzbarer Schätzer existiert, der in allen Situationen, die das Modell $M_h$ erlaubt, sinnvoll angewendet werden kann, um den Artenreichtum zuverlässig zu schätzen.

Wir stimmen hier mit der Einschätzung von Bunge und Fritzpatrick [BF93] überein, dass es nichtsdestotrotz sein kann, dass für eine gewisse Gegebenheit einfach noch kein passendes Modell und somit kein passender Schätzer entwickelt wurde oder die beste

momentan vorhandene Methode keine zufriedenstellenden Ergebnisse liefert.

Es besteht also auch in diesem Bereich noch weiterer Handlungsbedarf zu dem diese Diplomarbeit ihren Beitrag geleistet hat. Wir konnten zeigen, wie Zusatzkenntnisse über die Verteilung der Zellhäufigkeiten dazu verwendet werden können einen bestehenden Schätzer so zu modifizieren, dass er in einer konkreten Anwendung sinnvoll angewendet werden kann. So konnte der ursprünglich entwickelte Schätzer über Sample Coverage erst nach einer geeigneten Modifikation sinnvoll auf den von uns konkret betrachteten Fall angewendet werden. Wir sind davon überzeugt, dass die Nutzung von individuellen Besonderheiten auch in anderen Situationen, in denen der Artenreichtum geschätzt werden soll, zu einer Verbesserung der bestehenden Schätzverfahren führen kann. Folglich sollte diese Möglichkeit bei jeder derartigen Anwendung genauer untersucht werden.

# Literaturverzeichnis


[Ber94] Berger, Elke: *Die Schätzverfahren bootstrap und jackknife.* Diplomarbeit an der Bayerischen Julius-Maximilians-Universität Würzburg, 1994.

[BF93] Bunge, John A. und Fritzpatrick, M.: *Estimating the Number of Species: A Review.* Journal of the American Statistical Association, 88(421):364–373, 1993.

[BFH95] Bunge, John A., Fritzpatrick, M. und Handley, John C.: *Comparison of three estimatiors of the number of species.* Journal of Applied Statistics, 22(1):45–59, 1995.

[BH91] Bunge, John A. und Handley, John C.: *Sampling to estimate the number of duplicates in a database.* Computational Statistics & Data Analysis, 11:65–74, 1991.

[BO78] Burnham, K. P. und Overton, W. S.: *Estimation of the size of a closed population when capture probabilities vary among animals.* Biometrika, 65:625–633, 1978.

[BRK87] Boender, C. G. E. und Rinnooy Kan, A. H. G.: *A multinomial Bayesian approch to the estimation of population and vocabulary size.* Biometrika, 74(4):849–856, 1987.

[Cas81] Castledine, B. J.: *A Bayesian analysis of multiple-recapture sampling for a cloased population.* Biometrika, 68(1):197–210, 1981.

[Cha84] Chao, Anne: *Nonparametric Estimation of the Number of Classes in a Population.* Scandinavian Journal of Statistics, 11:265–270, 1984.

[Cha87] Chao, Anne: *Estimating the Population Size for Capture-Recapture Data with Unequal Catchability.* Biometrics, 43:783–791, 1987.

[Cha89] Chao, Anne: *Estimating Population Size for Sparse Data in Capture-Recapture-Experiments.* Biometrics, 45:427–438, 1989.

[Cha05] Chao, Anne: *Species estimation and applications.* Encyclopedia of Statistical Sciences, 12:7907–7916, 2005. 2nd Edition.

[CL92] Chao, Anne und Lee, Shen Ming: *Estimating the Number of Classes via Sample Coverage.* Journal of the American Statistical Association, 87(417):210–217, 1992.

[CMC04] Colwell, Robert K., Mao, Chang Xuan und Chang, Jing: *Interpolaing, extrapolating, and comparing incidence-based species accumulation curves.* Ecology, 85(10):2717–2727, 2004.

[Col05] Colwell, R. K.: *EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application.* `http://purl.oclc.org/estimates`, 2005.

[CS03] Chao, Anne und Shen, Tsung Jen: *Program SPADE (Species Prediction And Diversity Estimation). Program and User's Guide.* `http://chao.stat.nthu.edu.tw`, 2003.

[CSH06] Chao, Anne, Shen, Tsung Jen und Hwang, Wen Han: *Application of Laplace's boundary-mode approximations to estimate species and shared species richness.* Australian & New Zealand Journal of Statistics, 48(2):117–128, 2006.

[CY93] Chao, Anne und Yang, Mark C. K.: *Stopping rules and estimation for recapture debugging with unequal failure rates.* Biometrika, 88:193–201, 1993.

[Dar58] Darroch, J. N.: *The Multiple-Recapture Census I. Estimation of a Closed Population.* Biometrika, 45:343–359, 1958.

[Est78] Esty, W. W.: *Confidence Intervals for the Coverage of Low Coverage Samples.* The Annals of Statistics, 10:190–196, 1978.

[Fal07] Falk, Michael: *Skript zur Vorlesung Mathematische Statistik.* `http://statistik.mathematik.uni-wuerzburg.de/~falk/downloads/mathematische_statistik.pdf`, 2007.

[Fal08] Falk, Michael: *Skript zur Vorlesung Stochastik I.* `http://statistik.mathematik.uni-wuerzburg.de/~falk/downloads/stochastik_I_08_09.pdf`, 2008.

[GT56] Good, I. J. und Toulmin, G. H.: *The Number Of New Species, And The Increse In Population Coverage, When A Sample Is Increased.* Biometrika, 43:45–63, 1956.

[HF83] Heltshe, James F. und Forrester, Nancy E.: *Estimating Species Richness Using the Jackknife Procedure.* Biometrics, 39:1–11, 1983.

[Hug01] Huggins, R.: *A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneoous capture probabilities.* Statistics an Probability Letters, 51:147–152, 2001.

[Irl05] Irle, Albrecht: *Wahrscheinlichkeitstheorie und Statistik.* Teubner, 2. Auflage, 2005.

[LC94] Lee, Shen Ming und Chao, Anne: *Estimating Population Size via Sample Coverage for Closed Capture-Recatpure Models.* Biometrics, 50:88–97, 1994.

[Lin03] Link, W. A.: *Nonidentifiability of population size from capture-recapture data with heterogeneoous detection probabilities.* Biometrics, 59:1123–1130, 2003.

[LJ84] Lewins, W. A. und Joanes, D. N.: *Bayesian Estimation of the Number of Species.* Biometrics, 40:323–328, 1984.

[MCC05] Mao, Chang Xuan, Colwell, Robert K. und Chang, Jing: *Estimating the Species Accumulation Curve Using Mixtures.* Biometrics, 61:433–441, 2005.

[McN73] McNeil, D.: *Estimating an author's vocabulary.* Journal of the American Statistical Association, 68:92–96, 1973.

[Pab06] Pabel, Helmut: *Skript zur Vorlesung Analysis II.* unpublished, 2006.

[Raa87] Raaijmakers, Jeroen G. W.: *Statistical Analysis of the Michaelis-Menten Equation.* Biometrics, 43:793–803, 1987.

[RB07] Rivest, Louis Paul und Baillareon, Sophie: *Applications and Extensions of Chao's Moment Estimator for the Size of a Closed Population.* Biometrics, 63:999–1006, 2007.

[Rob68] Robbins, H.: *Estimating the Total Probability of the Unobserved Outcomes of an Experiment.* Annals of Mathematical Statistics, 39:256–257, 1968.

[SCL03] Shen, Tsung Jen, Chao, Anne und Lin, Chin Feng: *Predicting the number of new species in further taxonomic sampling.* Ecology, 84(3):798–804, 2003.

[SGO71] Schucany, W. R., Gray, H. L. und Owen, D. B.: *On bias reduction in estimation.* Journal of the American Statistical Association, 66:524–533, 1971.

[SH08] Shen, Tsung Jen und He, Fangliang: *An incidence-based richness estimator for quadrats sampled without replacemant.* Ecology, 89(7):2052–2060, 2008.

[Sic86] Sichel, H. S.: *Parameter Estimation For A Word Frequency Distribution Based On Occupancy Theory.* Communications in Statistics – Theory and Methods, 15(3):935–949, 1986.

[Sic97] Sichel, H. S.: *Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution.* South African Statistical Journal, 31:13–37, 1997.

[SvB84] Smith, Eric P. und Belle, Gerald van: *Nonparametric Estimation of Species Richness.* Biometrics, 40:119–129, 1984.

[SW08] Simon, Uwe K. und Weiß, Michael: *Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thougth.* Molecular Biology and Evolution, 25(11):2251–2254, 2008.

[WC92] Wilson, Richard M. und Collins, Mark F.: *Capture-recapture estimation with samples of size one using frequency data.* Biometrika, 79(3):543–553, 1992.

[Wer06] Werner, Dirk: *Einführung in die höhere Analysis*. Springer, 2. Auflage, 2006.

[Zei] Zeitung: *DIE ZEIT*. Ausgabe, 12. Februar 2009.

# Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Hilfsmittel angefertigt habe.

Würzburg, den 13. Oktober 2009 Stefan Englert