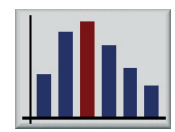
Schätzer von Spezienanzahlen (species abundance estimation)

Stefan Englert

Lehrstuhl für mathematische Statistik

22. Juli 2009

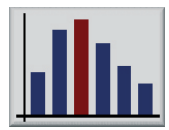


Notwendigkeit der Schätzung der Spezienanzahl

Angenommen eine Grundgesamtheit unterteilt sich in verschiedene Klassen (Spezien). In vielen Gelegenheiten ist weniger die relative Klassengröße als vielmehr die Anzahl der Klassen von Interesse.

Dieser Artenreichtum kann jedoch nur in Kollektiven in denen die Gesamtanzahl relativ klein ist durch Erfassung aller bestimmt werden.

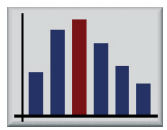
In allen anderen Fällen ist es notwendig die Spezienanzahl durch Schätzungen zu bestimmen.



Anwendungsmöglichkeiten

Insgesamt ergibt sich eine Vielzahl an Anwendungsmöglichkeiten für die Thematik, die hier stets als Schätzung der Spezienanzahlen bezeichnet wird.

- Schätzung bisher nicht beobachteter Pflanzen oder Tierspezies
- Anzahl der gleichen Einträge in einer sehr großen Datenbank
- Anzahl verschiedener Münzen oder Münzprägestätten einer Epoche
- Anzahl bisher nicht entdeckter Fehler einer Software
- Bisher nicht beobachtete Phänomene in der Astrologie
- Größe des Wortschatzes eines Autors

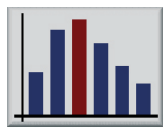


Problem des richtigen Schätzverfahrens

Bis zum heutigen Tag wurde eine Vielzahl an verschiedenen Schätzmethoden entwickelt. Jede dieser Schätzverfahren beansprucht für sich eine gewisse Art der Rechtfertigung und arbeitet in gewissen Situationen am Besten.

Ein Problem bei der Schätzung der Spezienanzahl ist also die Wahl des richtigen Schätzers.

Insbesondere trat genau dieses Problem in der Publikation *Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thougth* von Herrn Uwe Simon (Universität Graz, ehemalig Universität Würzburg) auf.

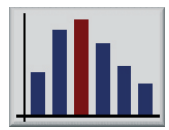


Konkreter Bezug zur Datenerhebung von Herrn Uwe Simon

In diesem Artikel wurden – vereinfacht dargestellt – drei verschiedene ribosomale (rDNA) Genregionen von vier verschiedenen Pilzarten nach Polymerasekettenreaktion kloniert und jeweils bestimmt, ob die klonierte Kolonie der „Konsensus“-Variante (Zusammensetzung des aufgenommenen Gens) oder einer polymorphen Variante (mit einer oder mehreren Punktmutationen) entsprach.

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.

Dazu wurden verschiedene Schätzverfahren angewandt, deren Ergebnisse sich jedoch sehr stark unterscheiden.

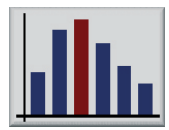


Zielsetzung

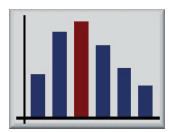
Ich habe mir mit meiner Diplomarbeit zum Ziel gesetzt, ein besseres Verständnis dafür zu vermitteln, wie ausgewählte Spezienanzahlschätzer bei den speziellen Verteilungen der Spezienhäufigkeiten, wie sie bei den Daten von Herrn Uwe Simon aufgetreten sind, arbeiten.

Die bisher entwickelten Spezienanzahlschätzverfahren sollen dabei sowohl theoretisch als auch im Rahmen einer Simulationsstudie untersucht werden und wurden gegebenenfalls auf die konkrete Situation angepasst.

Letztendlich soll es möglich sein eine Empfehlung abzugeben, welches Schätzverfahren in dieser Anwendung die besten Ergebnisse liefert.



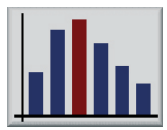
Inhaltsverzeichnis



Multiple Capture-Recapture Experimente

Um die Anzahl der verschiedenen Klassen oder Spezien einer Grundgesamtheit zu schätzen werden multiple Capture-Recapture Experimente betrachtet.

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 Spezies bestimmt und sowohl die zuvor schon einmal gemessenen als auch die neu festgestellten Spezien notiert. Dieses Verfahren wird nun bis zu einer festen Anzahl an Messdurchgängen n wiederholt.



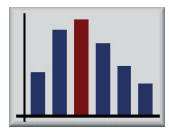
Datenerfassung

Letztendlich erhält man nach Durchführung des Experiments für jeden Messdurchgang die Information der Anwesenheit oder Abwesenheit der einzelnen Spezien.

Sei $j = 1, \ldots, S$ und $i = 1, \ldots, n$ 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 $2^{S \cdot n}$ solcher Matrizen.

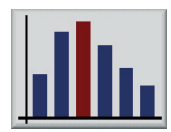


Beispiel

Sei $S = 5$ und $n = 3$. In drei Messdurchgängen seinen 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 nicht bekannt ist. Vielmehr kann nur der Teil der Matrix beobachtet werden, der aus den Spalten (Spezien) besteht, in denen mindestens eine Eins vorkommt.

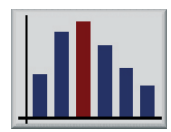


Beispiel

Sei $S = 5$ und $n = 3$. In drei Messdurchgängen seinen drei verschiedene Spezien wie folgt beobachtet worden:

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

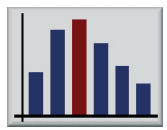
Ziel ist es nun aus dieser Matrix die Anzahl der Spezien zu schätzen, die in der Grundgesamtheit vorliegen.



Modelle

Wurde eine zufällige Stichprobe in dieser Weise gezogen, so haben sich bisher sieben verschiedene Grundmodelle zur Beschreibung der Daten herausgebildet.

Es bezeichne nun p_{ij} die Wahrscheinlichkeit, dass Spezies j bei der Untersuchung i gemessen wird.

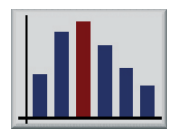


Modelle

Wurde eine zufällige Stichprobe in dieser Weise gezogen, so haben sich bisher sieben verschiedene Grundmodelle zur Beschreibung der Daten herausgebildet.

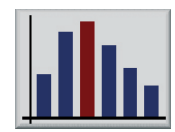
Es bezeichne nun p_{ij} die Wahrscheinlichkeit, dass Spezies j bei der Untersuchung i gemessen wird.

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix}$$



Modelle

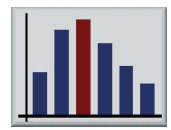
Modell M_0 Es gilt $\forall_{j=1}^{S} \forall_{i=1}^{n} p_{ij} = p$.



Modelle

Modell M_0 Es gilt $\forall_{j=1}^{S} \forall_{i=1}^{n} p_{ij} = p$.

Modell M_t Es gilt $\forall_{j=1}^{S} p_{ij} = p_i$. Für jeden Messdurchgang besitzten demnach alle Spezien die gleiche Erscheinungshäufigkeit. Diese kann jedoch mit der Zeit variieren.

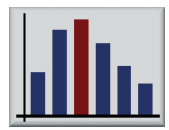


Modelle

Modell M_0 Es gilt $\forall_{j=1}^{S} \forall_{i=1}^{n} p_{ij} = p$.

Modell M_t Es gilt $\forall_{j=1}^{S} p_{ij} = p_i$. Für jeden Messdurchgang besitzten demnach alle Spezien die gleiche Erscheinungshäufigkeit. Diese kann jedoch mit der Zeit variieren.

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



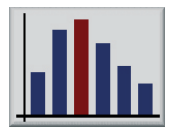
Modelle

Modell M_0 Es gilt $\forall_{j=1}^{S} \forall_{i=1}^{n} p_{ij} = p$.

Modell M_t Es gilt $\forall_{j=1}^{S} p_{ij} = p_i$. Für jeden Messdurchgang besitzten demnach alle Spezien die gleiche Erscheinungshäufigkeit. Diese kann jedoch mit der Zeit variieren.

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 $\forall_{i=1}^{n} p_{ij} = p_j$, d.h. die Spezienhäufigkeiten unterscheiden sich, sind jedoch unabhängig vom jeweiligen Messdurchgang.



Modelle

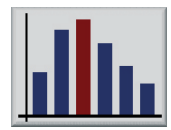
Modell M_0 Es gilt $\forall_{j=1}^{S} \forall_{i=1}^{n} p_{ij} = p$.

Modell M_t Es gilt $\forall_{j=1}^{S} p_{ij} = p_i$. Für jeden Messdurchgang besitzten demnach alle Spezien die gleiche Erscheinungshäufigkeit. Diese kann jedoch mit der Zeit variieren.

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 $\forall_{i=1}^{n} p_{ij} = p_j$, d.h. die Spezienhäufigkeiten unterscheiden sich, sind jedoch unabhängig vom jeweiligen Messdurchgang.

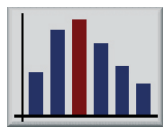
Insgesamt erhält man so die sieben Modelle M_0, M_t, M_b, M_h, M_{tb}, M_{th} und M_{tbh}.



Einschränkungen

Von mir wurden bei der Behandlung der Spezienanzahlschätzung folgende Einschränkungen vorgenommen:

- Es wird nur das Modell M_h betrachtet.
- Die Heterogenität in der Stichprobe ergibt sich nur aus der Variation der Spezien bezüglich ihrer Häufigkeit und nicht beispielsweise durch Clusterung der Spezien.
- Jeder Messdurchgang des Caputre-Recapture Experiments besteht aus genau einer Beobachtung.

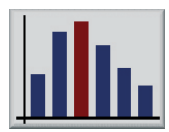


Einschränkungen

Von mir wurden bei der Behandlung der Spezienanzahlschätzung folgende Einschränkungen vorgenommen:

- Es wird nur das Modell M_h betrachtet.
- Die Heterogenität in der Stichprobe ergibt sich nur aus der Variation der Spezien bezüglich ihrer Häufigkeit und nicht beispielsweise durch Clusterung der Spezien.
- Jeder Messdurchgang des Caputre-Recapture Experiments besteht aus genau einer Beobachtung.

Die unter diesen Einschränkungen existierenden Schätzer der Spezienanzahlen lassen sich dabei wie folgt Kategorisieren:



Kategorisierung der Herangehensweise zur Spezienanzahlschätzung

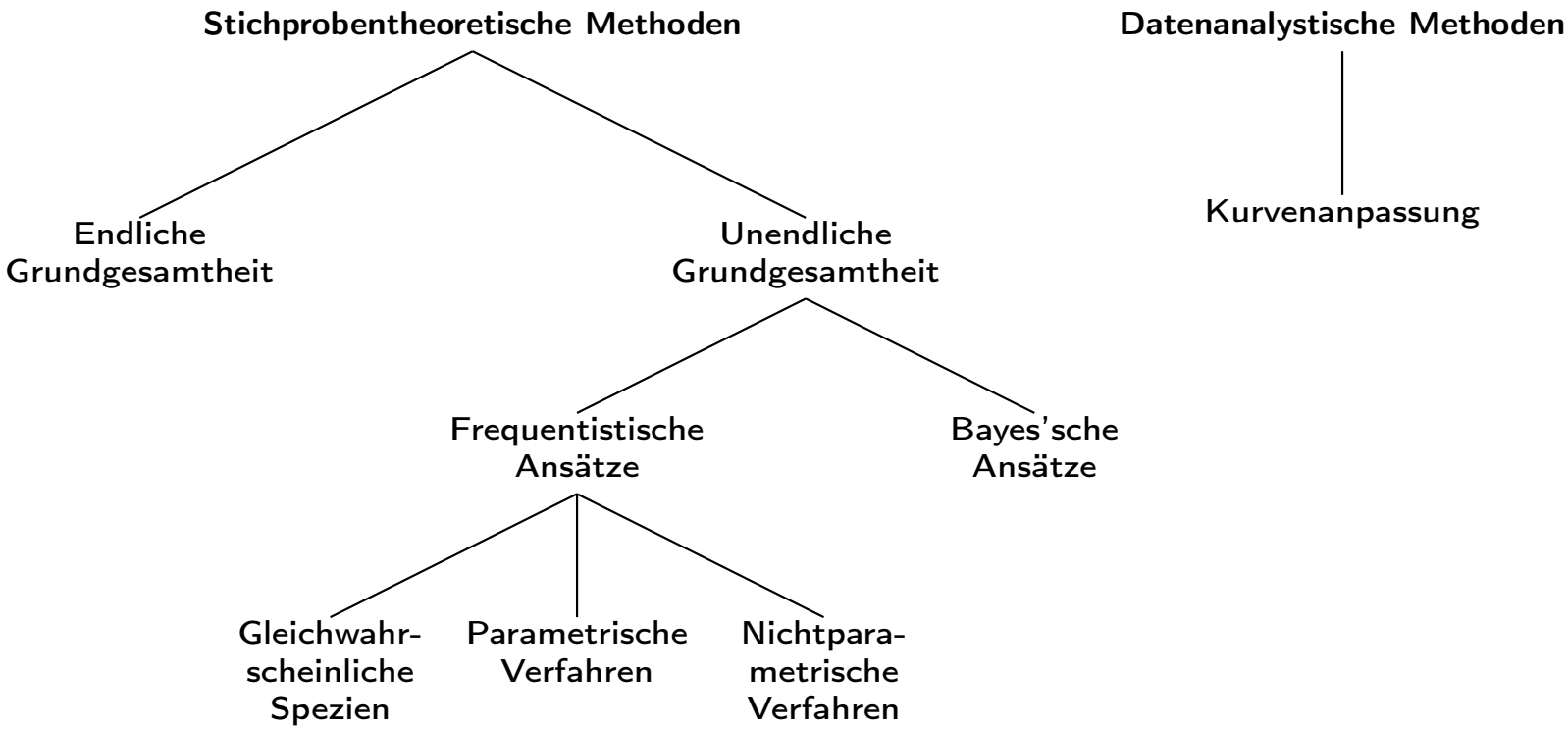
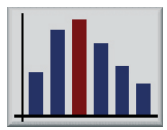


Abbildung: Kategorisierung der Herangehensweisen zur Spezienanzahlschätzung



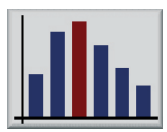
Anpassung auf die Daten von Herrn Uwe Simon

Alle bisherigen Einschränkungen und alle zukünftigen Anpassungen wurden in dieser Weise im Hinblick auf die Anwendung der Schätzverfahren auf die Daten von Herrn Uwe Simon getätigt.

Beispiel eines typischen Messergebnisses:

Pilzart	**Genregion**	**Anzahl Klonier-ungen**	**Spezien**	**Anteil Konsensus-variante**
Phoma exigua var. exigua	LSU	54	22	59%

Tabelle: Daten von Uwe K. Simon und Michael Weiß



Häufigkeitsverteilung

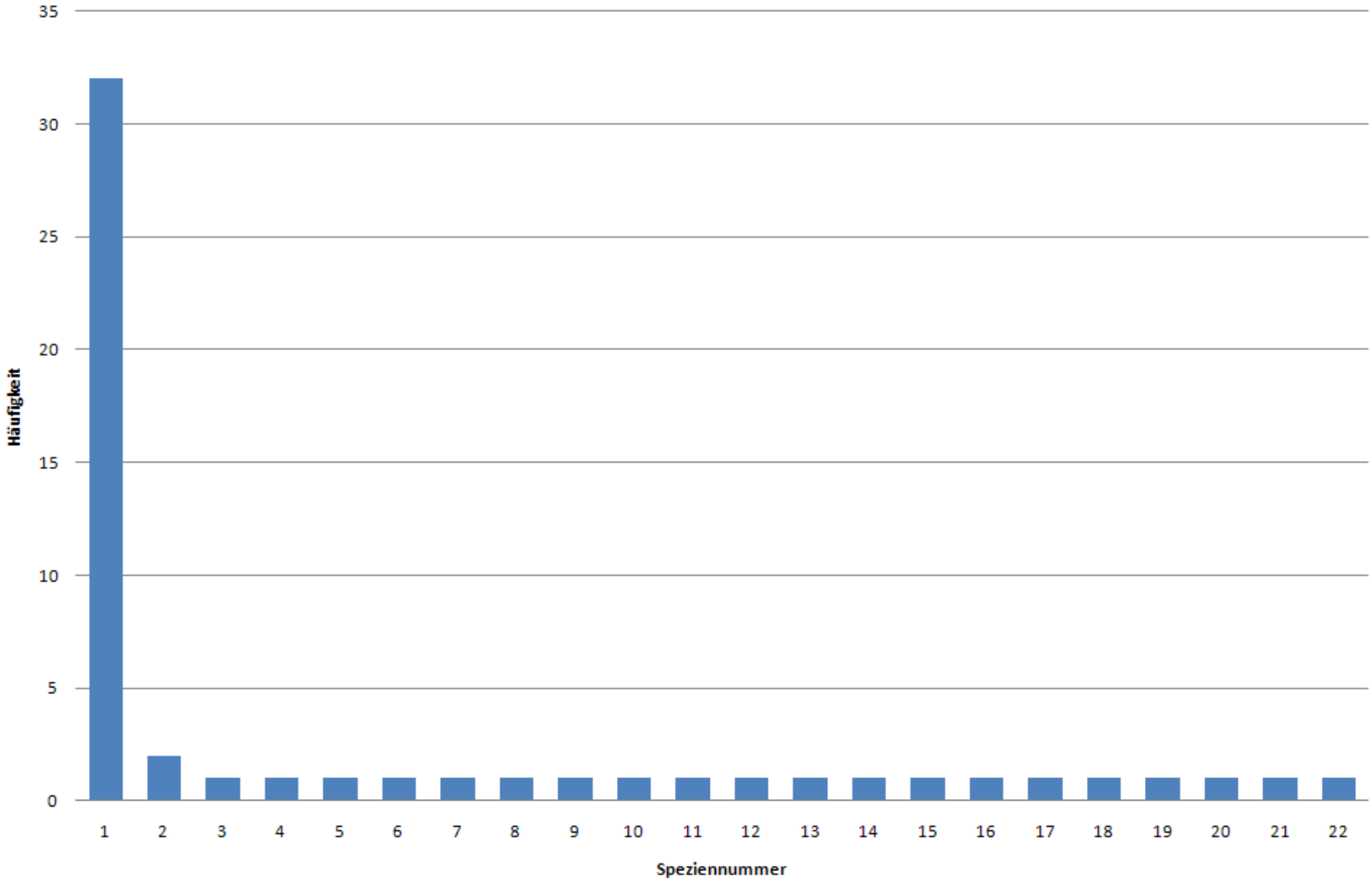
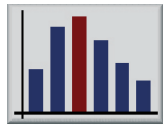


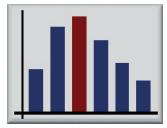
Abbildung: Häufigkeitsverteilung



Auswahl sinnvoller Schätzverfahren

- Da prinzipiell die Klonierungsvorgänge beliebig oft wiederholt werden können, muss die Grundgesamtheit als unendlich groß betrachtet werden.
- Da Bayes'sche Schätzverfahren stark von der angenommenen prior Verteilung abhängen und diese eigentlich durch Zusatzwissen festgelegt werden müsste, werden keine Bayes'schen Ansätze betrachtet.
- Aufgrund des sehr speziellen und stark schiefen Verlaufs der Häufigkeitsverteilung werden parametrische Schätzverfahren ehr schlechte Ergebnisse liefern.

Aus diesen Gründen werden im Folgenden nur nichtparametrische frequentistische Schätzverfahren, datenanalystische Methoden über Kurvenanpassung und aus historischen Gründen Schätzverfahren im Falle gleichwahrscheinlicher Spezien genauer betrachtet.



Auswahl sinnvoller Schätzverfahren

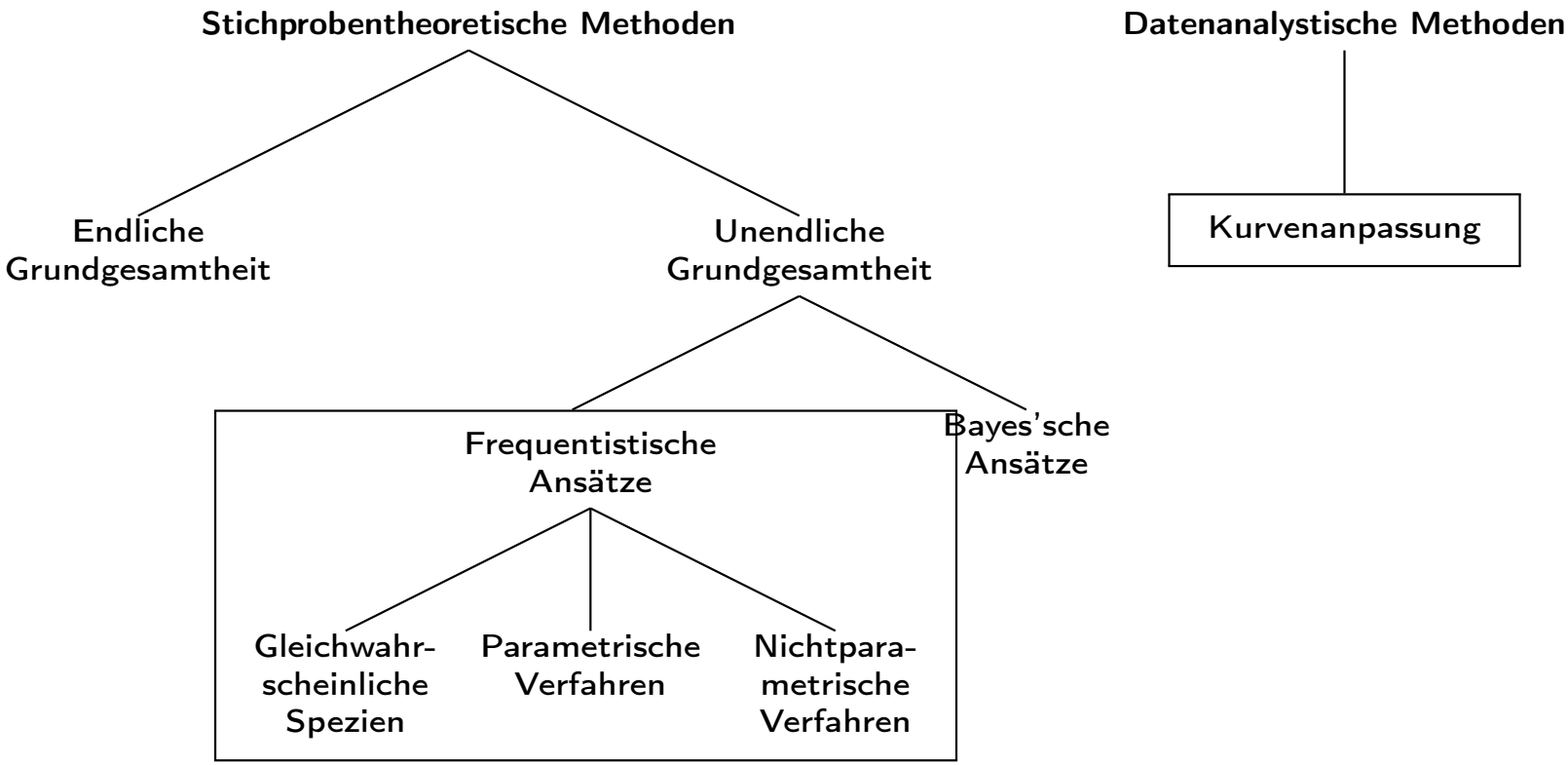
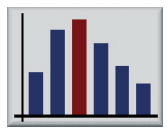
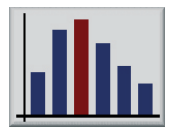


Abbildung: Kategorisierung der Herangehensweisen zur Spezienanzahlschätzung (betrachtete Verfahren sind hervorgehoben)



Bezeichnungen

S	Gesamtanzahl der Spezien im Kollektiv
n	Anzahl der Messdurchgänge
S_{obs}	In der Stichprobe gemessene Spezienanzahl
X_j	Anzahl der Individuen der Spezies j in der Stichprobe
F_1	Anzahl der nur einmal in der Stichprobe vorkommenden Spezien (Singletons)
F_2	Anzahl der nur genau zweimal in der Stichprobe vorkommenden Spezien (Dubletons)
F_i	Anzahl der nur genau i-Mal in der Stichprobe vorkommenden Spezien

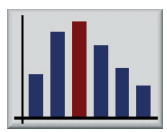


Schätzverfahren

Im Folgenden werden die Ideen verschiedener Schätzverfahren aufgezeigt und die dazu gehörigen Schätzer der Spezienanzahl angeben.

Einige Schätzer wurden dabei von mir speziell auf den hier betrachteten Fall angepasst.

Allen Schätzverfahren ist dabei gemeinsam, dass die Anzahl der in einer Stichprobe gefundenen Spezien immer als untere Schranke des Artenreichtums betrachtet werden muss.



Schätzer Darroch

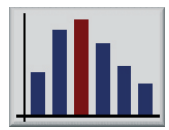
Das erste von mir betrachtete Schätzverfahren der Spezienanzahl ist das nach Darroch aus dem Jahr 1958.

Dieses Verfahren beschäftigt sich mit dem Modell M_0, der Gleichwahrscheinlichkeit aller Spezien. Ausgehend von diesem Modell und dem Artikel von Darroch begann meiner Einschätzung nach die Entwicklung der gesamten bisherigen Spezienschätztheorie.

Nimmt man 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.



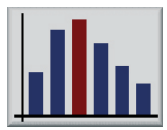
Schätzer Darroch

Die Wahrscheinlichkeit die gemessene Stichprobe zu erhalten ergibt sich dabei als Laplace Wahrscheinlichkeit zu

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

Da diese Wahrscheinlichkeit nur von der gesuchten Variablen S abhängt, lässt sich dessen Wert durch einen Maximum-Likelihood-Ansatz bezüglich dieser Variablen gewinnen. Der von Darroch vorgeschlagene Schätzer ist folglich die kleinste natürliche Lösung größer als S_{obs} von:

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



Schätzer Chao

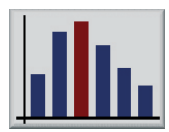
Der folgende Schätzer entwickelt von Anne Chao ist ein Schätzer für eine untere Schranke der Spezienanzahl.

Nimmt man an, dass die Erscheinungswahrscheinlichkeiten p der Spezien eine zufällige Größe mit Verteilungsfunktion F ist, so lässt sich zeigen, dass gilt:

$$E(F_i) = S \int_0^1 \binom{n}{i} p^i (1-p)^{n-i} \, dF(p)$$

Daraus ergibt sich zusammen mit der Cauchy-Schwarzschen Ungleichung für Wahrscheinlichkeitsintegrale

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

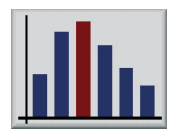


Schätzer Chao

Zusammen mit $S = S_{obs} + F_0$ erhält man nun einen Schätzer dadurch, dass man die entsprechenden $E(F_i)$ durch F_i ersetzt.

Es handelt sich um einen Schätzer für eine untere Grenze. Seine Fähigkeiten als Schätzer der Spezienanzahl sind aber ermutigend, insbesondere, wenn (S_{obs}, F_1, F_2) einen Großteil der Informationen enthält.

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

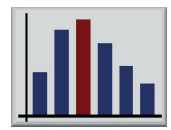


Bias im Equal Likely Fall

Im Falle des Equal Likely Falls wird die Ungleichung aus der der Schätzer hergeleitet wurde 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 die 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.



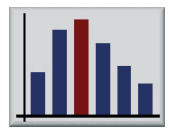
Biaskorrektur

Um diese Problematik zu umgehen wird in der Literatur eine Bias-Corrected Version des Schätzers von Chao vorgeschlagen.

Leider wird nicht dargestellt, wie diese Bias Korrektur vorgenommen wurden und deshalb habe ich eine Biaskorrektur entwickelt.

Sei nun $E(F_2) > 0$ angenommen. Nach dem zuvor Gezeigten gilt:

$$E(F_0) \overset{s.o.}{\geq} \frac{E(F_1)^2}{2\,E(F_2)} \quad \geq \quad \frac{E(F_1)^2}{2\,E(F_2)} \cdot P(F_2 \neq 0)$$



Biaskorrektur

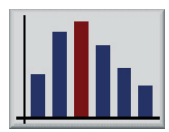
Ich habe nun einen Schätzer für

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

entwickelt und gezeigt, dass dieser erwartungstreu als Schätzer einer unteren Schranke der Spezienanzahl ist. Dabei erhält man bis auf einen Faktor $\frac{n-1}{n}$ genau die postulierte Bias-Corrected Version des Schätzers von Chao.

Ich empfehle, aufgrund meiner Herleitung, die Bias-Corrected Version des Schätzers in allen Gelegenheiten, da hier der Fall $F_2 = 0$ mit Wahrscheinlichkeit $P(F_2 = 0)$ berücksichtigt wird. Denn nur für $F_2 \neq 0$ kann S_{Chao} überhaupt berechnet werden.

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

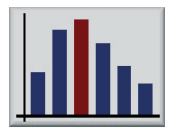


Schätzer über Sample Coverage

Bei der folgenden Herangehensweise handelt es sich um eine nichtparametrische Schätztechnik, die die Spezienhäufigkeit über den Sample Coverage C schätzt, das ist die Summe der Wahrscheinlichkeiten der beobachteten Spezien.

$$C = \sum_{j=1}^{S} p_j I[X_j > 0],$$

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



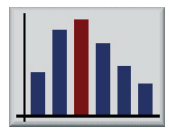
Schätzer über Sample Coverage

Durch eine Taylor-Entwicklung kann gezeigt werden, dass

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

ist, wobei γ der Coefficient of Variation ist und R ein Restterm ist. Vernachlässigt man diesen Term und ersetzt alle anderen Erwartungswerte geeignet durch entsprechende Schätzer, so erhält man einen Spezienanzahlschätzer über einen Sample Coverage Ansatz:

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



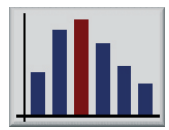
Modifikation

Der Schätzer über den Sample Coverage Ansatz wird so nur sehr schlechte Ergebnisse angewandt auf unsere Daten liefern, da die Heterogenitäten in den Spezienhäufigkeiten zu hoch sind. Ich habe ihn deshalb speziell auf diese Situation hin angepasst.

Es kann angenommen werden, dass in jedem vergleichbaren Experimentablauf die Konsensusvariante in jedem Fall in einer der n Durchführungen gemessen wird.

Es gilt also:

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



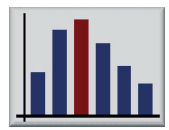
Modifikation

Statt nun, wie im Schätzer über Sample Coverage direkt einen Schätzer für S über S_{obs} zu entwickeln, schätzen wir S_{selten} und addieren den festen Wert 1 für die Konsensusvariante hinzu.

Dadurch verlieren wir einen Großteil der Heterogenität in unseren Daten und die verwendete Taylor-Approximation wird bedeutend besser. Im Gegensatz dazu haben wir aber nur noch eine Stichprobengröße von $n - m$, wobei m die Anzahl der Konsensusvarianten ist.

Analog zu vorher ergibt sich der Schätzer zu:

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

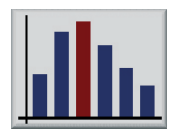


Jackknife

Der nächste betrachtete Schätzer verwendet die Jackknife Methode.

Diese wurde als allgemeine Methode zur Verringerung des Bias eines verzerrten Schätzers entwickelt.

Angenommen $\hat{\theta}$ ist eine relativ gute Approximation von θ. Um den Jackknife Schätzer zu erhalten führt man die folgenden Schritte durch:

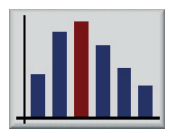


Allgemeine Jackknife Methode

1. Entferne eine der Beobachtungen. Dies sei x_i.
2. Berechne die Schätzung von θ basierend auf $(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ und bezeichne diese mit $\hat{\theta}_{-i}$.
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

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



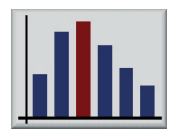
Schätzer Jackknife 1

Diese Schätzung ist als „first-order jackknife“ bekannt und hilft den Bias der Ordnung $1/n$ zu reduzieren.

In der Anwendung bei der Spezienanzahlschätzung wird $\theta := S$ geschätzt, basierend auf n unabhängigen Beobachtungen. Es wird angenommen, dass die beobachtete Spezienanzahl $\hat{\theta} := S_{obs}$ eine relativ gute Approximation ist.

Der nach der Jackknife Methode berechnete Schätzer ist damit:

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



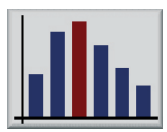
Verallgemeinerte Jackknife Methode

Schucany, Gray und Owen (1971) verallgemeinerten die Jackknife Methode um Bias höher Ordnungen zu entfernen.

Die Vorgehensweise ist dabei ähnlich, nur dass mehr als nur eine Beobachtung aus der Stichprobe entfernt wird.

Der berechnete Spezienschätzer für die Jackknife Methode mit Ordnung 2 ist dann:

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

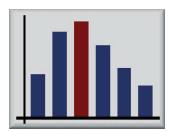


Datenanalystische Schätzverfahren

Im Gegensatz zu den vorherigen frequentistischen Verfahren wird im Folgenden ein Verfahren über Kurvenanpassung vorgestellt.

Eine *Species Accumulation Curve* (SAC) ist eine graphische Darstellung der gemessenen Spezien in irgendeiner Messeinheit und somit eine monoton steigende Funktion, die typischerweise asymptotisch verläuft. Sie kann dazu verwendet werden, die Spezienanzahl (als deren asymptotischen Wert) zu schätzen.

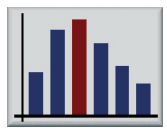
Genau in dieser Art und Weise habe ich die SAC verwenden. Diese wurde durch eine Michaelis Menten Funktion approximieren und deren asymptotischen Wert als Schätzwert der Spezienanzahl verwendet.



Simulationen

Um die Arbeitsweise der jeweilige Schätzer überprüfen zu können wurden Simulationen mit verschiedenen Verteilungen der Spezienhäufigkeiten durchgeführt. Zusätzlich wurde die verwendete Stichprobengröße und die Anzahl der Spezien variiert. In diesen Simulationen war also die tatsächliche Spezienanzahl bekannt und deshalb war es möglich zu beurteilen wie gut die Schätzverfahren in dieser Situation arbeiten.

Diese Simulationen sollen dabei helfen die Spezienschätzungen aus den Daten von Herrn Uwe Simon besser verstehen zu können. Die Verteilungen der Spezien wurden deshalb an die von Herrn Uwe Simon bestimmten Daten angepasst.



Häufigkeitsverteilung

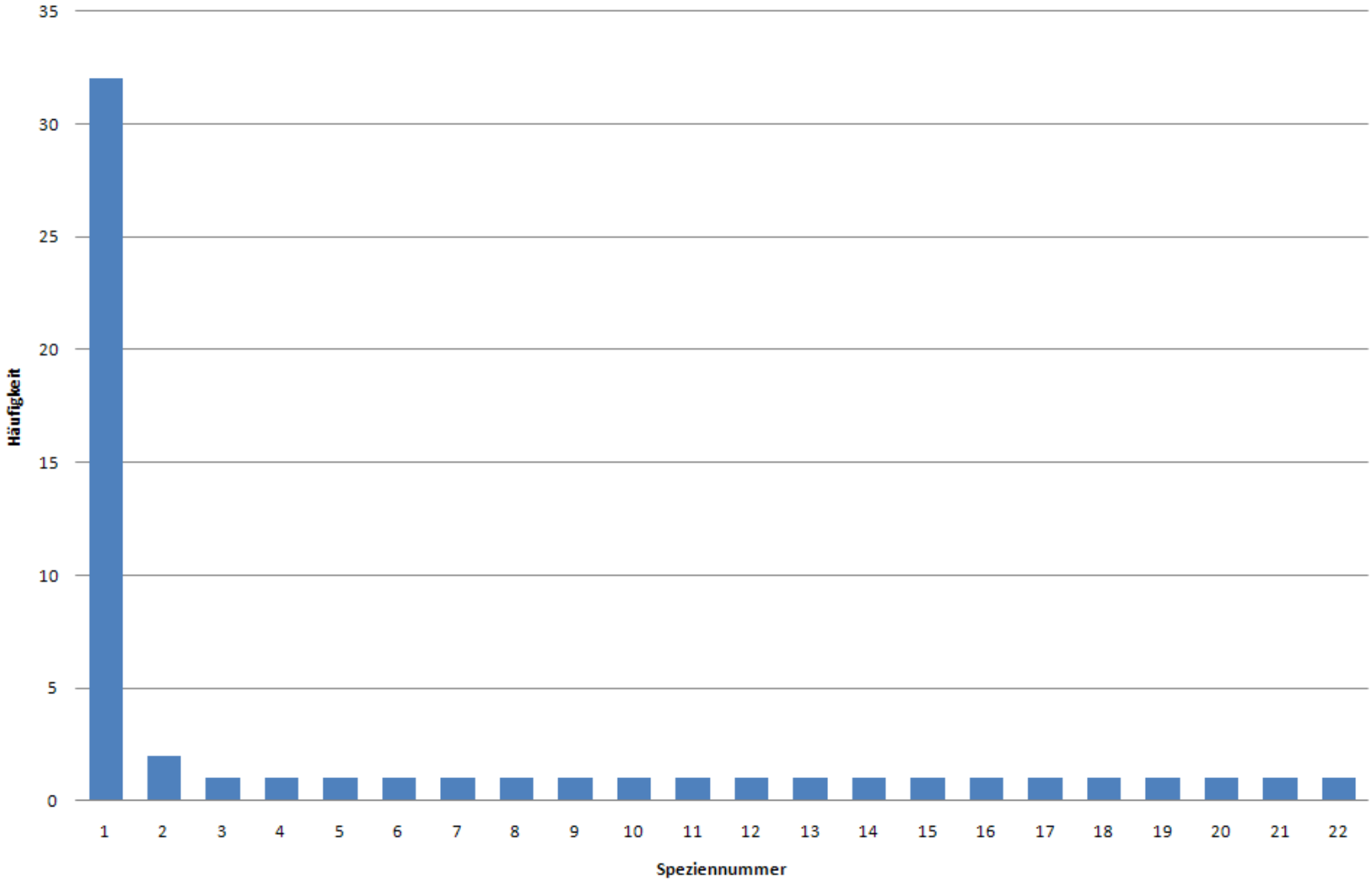
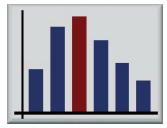


Abbildung: Häufigkeitsverteilung



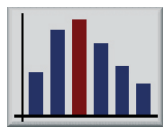
Angenommenen Verteilungen

Aufgrund der ziemlich auffälligen Verteilung der Daten von Herrn Uwe Simon wurden die Verteilungen so gewählt, dass eine sehr häufige Spezies (40, 60 und 80 %) und eine Vielzahl an seltenen Spezies vorkommen.

Den seltenen Spezien wurden dabei die folgenden parametrischen Verteilungen unterstellt:

- Gleichverteilung
- Truncated geometrische Verteilung mit sehr kleinem Koeffizienten ($\lambda = 0.01$)

In einer weiteren Untersuchung wurde auch kurz das Querschnittsverhalten, d.h. das Verhalten der Schätzer bei einer Variation des Glättungsparameters λ untersucht.

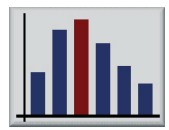


Simulationen

In den Simulationen wurden die vorgestellten Schätzer jeweils 1'000 Mal in den verschiedenen Kombinationen aus Stichprobengröße und tatsächlicher Spezienanzahl berechnet. Dabei ließen sich unter gewissen Umständen, z. B. wenn eine Stichprobe keine Spezies genau zwei Mal enthält ($F_2 = 0$) bestimmte Schätzer nicht berechnen.

Die Auswertung der Güte der Schätzer erfolgt über die Anzahl der nicht berechenbaren Schätzer („Fehlgeschlagen"), den Bias und den root mean squared error (RootMSE).

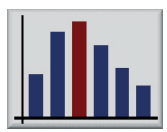
Die folgende Tabelle zeigt ein typisches Messergebnis für gleichverteilte seltene Spezien.



Simulationsergebnis für $S = 101$ und $n = 50, 100, 250$

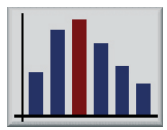
Schätzername	Darroch	Chao BC	Mod. Sample Coverage	Jackknife 1	Jackknife 2	Michaelis Menten
50						
Fehlgeschlagen	0	0	161	0	0	0
Mittelwert	22.02	86.74	119.92	35.40	49.77	35.12
RootMSE	79.12	54.45	69.42	65.89	52.1	66.81
100						
Fehlgeschlagen	0	0	1	0	0	0
Mittelwert	37.03	101.57	127.08	60.79	82.08	65.58
RootMSE	64.19	49.62	73.05	40.98	22.73	38.32
250						
Fehlgeschlagen	0	0	0	0	0	0
Mittelwert	66.41	101.11	105.96	101.3	119.81	115.92
RootMSE	34.92	16.29	16.56	8.22	23.35	19.32

Tabelle: Auszug aus den Simulationsergebnissen



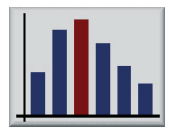
Zusammenfassung der Ergebnisse der Simulationsstudie

- Gleichverteilte seltene Spezien
  - Der Bias corrected Chao Schätzer zeigt durchwegs den kleinsten Bias und ist insbesondere bei größeren Spezienanzahlen den übrigen Schätzern sowohl im Bezug auf Bias als auch auf RootMSE überlegen.
  - Der modifizierte Schätzer über Sample Coverage neigt dazu die Spezienanzahl im Mittel sogar zu überschätzen.
  - Entspricht die Stichprobengröße in etwa der Spezienanzahl, so liefern die Methoden Jackknife 1 und Jackknife 2 sehr gute Ergebnisse.
  - Für größere Stichprobengrößen (ab 400) ist es nahezu aussichtslos die Spezienanzahl zu schätzen, da der relative RootMSE durchwegs über 0.5 liegt.



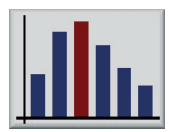
Zusammenfassung der Ergebnisse der Simulationsstudie

- Geometrisch verteilte seltene Spezien
  - Insbesondere in größeren Stichprobengrößen ist der modifizierte Schätzer über Sample Coverage den anderen bzgl. RootMSE überlegen.
  - Der Bias corrected Chao Schätzer zeigt stark negativen Bias und eignet sich nur als unterer Schrankenschätzer.
  - Mehr als 300 Spezien werden in keiner Konstellation geschätzt, da die Wahrscheinlichkeit für ein Auftreten einer solchen Spezies wohl schon zu gering ist.
  - Bei 80% seltenen Spezien ist unter einigen Konstellationen (mehr als 200 Spezien) die Schätzung in den betrachteten Situationen aussichtslos.



Zusammenfassung der Ergebnisse der Simulationsstudie

- Querschnittsuntersuchung
  - In diesem Fall zeigt der modifizierte Schätzer über Sample Coverage die besten Ergebnisse
  - Wie erwartet steigt der RootMSE mit zunehmendem λ stark an.
  - Im Bereich von $\lambda = 0$ bis 0.01 verbessert sich der modifizierte Schätzer über Sample Coverage sogar im Bezug auf RootMSE.



Empfehlungen für die Daten von Herrn Uwe Simon

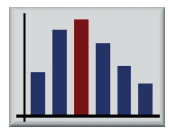
Aus den Simulationen können nur Empfehlungen unter zusätzlichen Annahmen abgeleitet werden.

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

Stichprobengröße $>$ erwartete Spezienanzahl : *Chao BC*, zuverlässige Schätzung mit geringem RootMSE.

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

Geringe Heterogenität in den seltenen Spezien : *Mod. Sample Coverage*, als Punktschätzer der Spezienanzahl. Dieser besitzt jedoch einen hohen RootMSE. Besser: *Chao BC* als Schätzer einer unteren Schranke.



Vielen Dank für Ihre Aufmerksamkeit!