Kollegium Spiritus Sanctus Brig

Maturitätsarbeit 2025/2026

Capture-Recapture-Verfahren

Schätzung unbekannter Populationsgrössen mit exemplarischer Anwendung auf erdnahe Asteroiden

Von:

Pfister Noémie Angelina, 5C

Eingereicht im Fachbereich Mathematik

Betreut durch:

Dr. Blumenthal Adrian

Brig, den 04.12.2025 Nummer 2025/76

Inhaltsverzeichnis

Kapitel 1

Einleitung

„The most problematic object is the one you don't know about."

Mit dieser Aussage weist die US-amerikanische Astronomin Amy Mainzer auf ein zentrales Problem der planetaren Verteidigung hin: erdnahe Asteroiden, die bisher unentdeckt geblieben sind und potenziell eine Gefahr für unseren Planeten darstellen könnten [47].

Im Sonnensystem wurden bislang fast 1.5 Millionen Asteroiden entdeckt und katalogisiert, von denen sich nur ein kleiner Bruchteil in Erdnähe befindet [39]. Die Beobachtung der erdnahen Asteroiden gestaltet sich jedoch oft schwierig, insbesondere bei kleinen und lichtschwachen Objekten. Daher ist die Erfassung noch sehr lückenhaft, was sich unter anderem daran zeigt, dass nahezu täglich neue erdnahe Asteroiden entdeckt werden [44]. Diese Unvollständigkeit ist besonders problematisch, da sich darunter auch Objekte befinden könnten, die eine Bedrohung darstellen. Doch wie lässt sich die Gesamtanzahl aller erdnahen Asteroiden abschätzen, wenn die Datenbasis unvollständig und verzerrt ist?

Ein vielversprechender Ansatz zur Lösung dieses Problems bieten die sogenannten Capture-Recapture-Verfahren. Ursprünglich wurden diese Verfahren zur Schätzung menschlicher Bevölkerungszahlen entwickelt, erlangten jedoch vor allem in der Ökologie, insbesondere zur Bestimmung von Tierpopulationen, grössere Bedeutung [26]. Inzwischen finden sie in vielen weiteren Bereichen Anwendung, beispielsweise in der Medizin [15], den Sozialwissenschaften [11] oder der Astronomie [9]. Solche Verfahren erlauben es unter bestimmten Annahmen, aus mehrfachen Beobachtungen Rückschlüsse auf die Grösse einer unbekannten Grundgesamtheit zu ziehen.

1.1 Ziele

Diese Arbeit befasst sich mit den Capture-Recapture-Verfahren. Dies sind statistische Methoden, durch welche die Grösse von Populationen abgeschätzt werden können. Besonders im Fokus stehen dabei der Lincoln-Petersen-Schätzer sowie der Chapman-Schätzer. Das erste konkrete Ziel besteht darin, diese beiden Schätzer detailliert zu erklären und ihre mathematischen Eigenschaften zu definieren. Des Weiteren sollen diese in verschiedenen Szenarien bezüglich ihrer Genauigkeit und Präzision mittels Simula-

tionen miteinander verglichen werden. Im letzten Teil dieser Arbeit wird das erlangte theoretische Wissen schliesslich auf eine praktische Anwendung übertragen, dies mit dem Ziel, die Anzahl der erdnahen Asteroiden abzuschätzen. Die resultierenden Schätzungen sollen in diesem Kontext analysiert und interpretiert werden, insbesondere hinsichtlich möglicher Fehlerquellen und der Aussagekraft der Ergebnisse.

1.2 Aufbau

Das zweite Kapitel bietet eine Einführung in die Capture-Recapture-Verfahren. Zunächst wird das grundlegende Konzept erläutert, gefolgt von einem kurzen historischen Überblick sowie einer Einteilung in verschiedene Arten von Capture-Recapture-Verfahren. Anschliessend werden die zugrunde liegenden Modellannahmen dargelegt und das Konzept der Fanghistorien eingeführt, anhand dessen die für die weitere Arbeit relevanten Grössen definiert werden. Im dritten Kapitel wird der Lincoln-Petersen-Schätzer vorgestellt. Dieser wird sowohl über Anteilsverhältnisse als auch unter Verwendung der Maximum-Likelihood-Methode hergeleitet. Zudem werden die Genauigkeit und Präzision dieses Schätzers näher untersucht. Das folgende Kapitel widmet sich dem Chapman-Schätzer, einem weiteren Schätzer aus den Capture-Recapture-Verfahren. Nach der Definition dieses weiterentwickelten Schätzers und der Darstellung seiner Vorteile wird ebenfalls auf seine Genauigkeit und Präzision eingegangen, wobei der Fokus besonders auf seiner Verzerrung liegt. Im fünften Kapitel werden die beiden Schätzer schliesslich mithilfe von Simulationen miteinander verglichen. Zunächst erfolgt der Vergleich unter der Annahme, dass alle Modellannahmen erfüllt sind, wobei insbesondere die Genauigkeit und Präzision der Schätzer berücksichtigt werden. Anschliessend wird die Robustheit der Schätzer im Hinblick auf potenzielle Verletzungen der Modellannahmen betrachtet. Das sechste und letzte Kapitel überträgt die theoretisch erarbeiteten Grundlagen letztendlich auf ein praktisches Beispiel: die Populationsschätzung erdnaher Asteroiden. Nach einer Einführung in die Thematik der erdnahen Asteroiden und deren Bedeutung folgt die Darstellung der Datengrundlage und der Methodik. Daraufhin werden die beiden vorgestellten Schätzverfahren auf die Daten angewendet, wobei die erhaltenen Resultate im Anschluss analysiert und interpretiert werden.

Im ersten Teil des Anhangs finden sich weiterführende Informationen zur hypergeometrischen Verteilung, einschliesslich der Herleitungen ihres Erwartungswerts und ihrer Varianz. Der zweite Teil befasst sich mit den Grundlagen der Maximum-Likelihood-Schätzung, die für die Herleitung des Lincoln-Petersen-Schätzers in Kapitel 3 Anwendung findet. Im letzten Teil des Anhangs sind schliesslich sämtliche Codes der Python-Programme enthalten, die einerseits für die Simulationen zum Vergleich der Schätzer in Kapitel 5 und andererseits für die Datenbeschaffung im Anwendungsbeispiel in Kapitel 6 verwendet werden.

Kapitel 2

Einführung in die Capture-Recapture-Verfahren

Dieses Kapitel führt in die Thematik der Capture-Recapture-Verfahren ein. Es beginnt mit einem Überblick über das Grundprinzip sowie einem kurzen historischen Rückblick, der mit einer Einteilung der Verfahren in verschiedene Arten kombiniert ist. Im Anschluss werden die Modellannahmen der in dieser Arbeit betrachteten Verfahren dargelegt. Ausserdem erfolgt eine Einführung in das Konzept der Fanghistorien, aus denen die im weiteren Verlauf verwendeten Grössen definiert werden.

2.1 Grundkonzept und historischer Hintergrund der Capture-Recapture-Verfahren

Capture-Recapture-Verfahren sind statistische Methoden zur Schätzung der Grösse einer Population, deren vollständige Erfassung nicht möglich oder zu aufwendig wäre [13]. Das Grundkonzept besteht darin, mindestens zwei Stichproben aus einer Population zu ziehen, wobei die erfassten Individuen jeweils eindeutig markiert werden. Zudem wird für jedes Individuum dokumentiert, ob es bereits in einer früheren Stichprobe erfasst wurde. Anhand der mehrfach markierten Individuen lässt sich die Gesamtzahl der Individuen in der Population abschätzen. Allgemein wird die Population umso grösser geschätzt, je kleiner der Anteil der wiedererkannten Individuen ist, da dies auf eine grössere Anzahl nicht erfasster Individuen deutet [25].

Die Ursprünge der Capture-Recapture-Verfahren reichen bis ins Jahr 1802 zurück, als der französische Mathematiker Pierre-Simon Laplace versuchte, die Grösse der französischen Bevölkerung anhand von Geburtenzahlen zu schätzen. Obwohl dies oft als erster dokumentierte Gebrauch einer Capture-Recapture-Idee gilt, lässt sich ein ähnliches Vorgehen bereits rund 200 Jahre früher nachweisen. Im frühen 17. Jahrhundert verwendete J. Graunt vergleichbare Methoden wie Laplace, um die Auswirkungen der Pest sowie die Grösse der englischen Bevölkerung zu schätzen. Beide griffen bei ihren Arbeiten auf einfache Verhältnisschätzungen zurück [26].

In der Literatur wird indes oft das Jahr 1896 als das Geburtsjahr der (modernen) Capture-Recapture-Verfahren angesehen, da die Methoden zu diesem Zeitpunkt erstmals Anwendung in der Ökologie fanden, was ihnen zu grösserer Bedeutung und Bekanntheit verhalf [13]. Der dänische Meeresbiologe C. G. J. Petersen markierte Fische mithilfe von Messingschildern und nutzte die Rückfänge als Grundlage für eine Schätzung der Populationsgrösse. Dieselbe Methode wurde 1930 vom US-amerikanischen Ornithologen F. C. Lincoln verwendet, um die Populationsgrösse von Wasservögeln in Nordamerika zu bestimmen. Obwohl beide Forscher auf den gleichen Prinzipien wie Graunt und Laplace aufbauten, wurde das Verfahren schliesslich nach ihnen benannt und ist heute als Lincoln-Petersen-Schätzer bekannt [8]. Kapitel 3 widmet sich diesem Schätzer, während in Kapitel 4 eine von D. G. Chapman vorgeschlagene Modifikation aus dem Jahre 1951 vorgestellt wird, die als Chapman-Schätzer bezeichnet wird.

Capture-Recapture-Verfahren lassen sich grundsätzlich in zwei Hauptkategorien unterteilen, nämlich in Verfahren für geschlossene und für offene Populationen [26]. Die ersten Verfahren, zu denen der Lincoln-Petersen-Schätzer zählt, wurden für geschlossene Populationen entwickelt, bei denen davon ausgegangen wird, dass die Individuen während des gesamten Untersuchungszeitraumes in ihrer Anzahl konstant bleiben. Der Lincoln-Petersen-Schätzer basiert auf lediglich zwei Stichproben und stellt eine einfache Methode zur Schätzung der Grösse geschlossener Populationen dar [30]. Eine wichtige Weiterentwicklung dieses Verfahrens fand 1938 durch Z. E. Schnabel statt, die es ermöglichte, die Anzahl der Stichproben beliebig zu wählen, was die Schätzungen flexibler und robuster macht [13]. Weitere Anpassungen erfolgten 1943 durch F. X. Schumacher und R. W. Eschmeyer, die das Verfahren als lineare Regression formulierten [8]. In den 1970er Jahren entwickelten Otis et al. zudem weiterführende Modelle, die die heterogenen Fangwahrscheinlichkeiten der Individuen berücksichtigen. Diese Unterschiede in den Fangwahrscheinlichkeiten können durch zeitliche Faktoren, unterschiedliche Verhaltensreaktionen auf das Fangen sowie durch die individuelle Heterogenität der Tiere erklärt werden [26].

Im Gegensatz dazu befassen sich Verfahren für offene Populationen mit der Schätzung von Populationsgrössen, bei denen die Zahl der Individuen während des Untersuchungszeitraumes aufgrund von Geburten, Todesfällen, Immigration und Emigration schwankt [13]. Bereits 1939 begann C. H. N. Jackson, Methoden zur Erfassung dieser dynamischen Populationen zu entwickeln, mit dem Ziel, die sich verändernde Populationsgrösse, Überlebensraten sowie die Zahl neu hinzukommender Individuen abzuschätzen [31]. Einen grossen Durchbruch erzielten R. M. Cormack, G. M. Jolly und G. A. F. Seber in den 1960er Jahren mit der Einführung der Maximum-Likelihood-Schätzung (vgl. Anhang B) für die Analyse von Capture-Recapture-Daten bei offenen Populationen. Daraus entstanden die sogenannten Cormack-Jolly-Seber (CJS) und Jolly-Seber (JS) Modelle [32]. Während das CJS-Modell ausschliesslich auf der Rückfangrate markierter Tiere basiert und lediglich Schätzungen zu Überlebensraten und Fangwahrscheinlichkeiten liefert, nutzt das JS-Modell auch die Verhältnisse zwischen markierten und nicht markierten Tieren, um so zusätzlich die Populationsgrösse zu schätzen [31]. Eine weitergehende Verallgemeinerung des JS-Modells bietet der Manly-Parr-Ansatz von B. F. J. Manly und M. J. Parr aus dem Jahre 1968, der darüber hinaus altersabhängige Faktoren berücksichtigt [31].

Ein weiteres Unterscheidungsmerkmal bei Capture-Recapture-Verfahren stellt die Behandlung der Zeit dar, wobei zwischen zeitlich diskreten und kontinuierlichen Modellen unterschieden wird [6]. In zeitlich diskreten Modellen erfolgen die Stichproben zu einer festgelegten Anzahl von Zeitpunkten. Die genauen Fangzeiten sowie die Reihenfolge der Individuen innerhalb einer einzelnen Stichprobe sind dabei unbekannt [26]. Zudem werden Wiederfänge innerhalb des Zeitraumes einer Stichprobe vernachlässigt [6]. Alle zuvor erwähnten Verfahren fallen in diese Kategorie.

Zeitlich kontinuierliche Modelle dagegen behandeln jeden Fangvorgang einzeln, wobei Fänge zu beliebigen Zeitpunkten stattfinden können und der exakte Zeitpunkt jedes Fangs dokumentiert wird. Solche Modelle können daher auch als Spezialfälle der zeitlich diskreten Modelle angesehen werden, bei denen jede Stichprobe nur ein einziges Individuum umfasst [26]. In der Praxis liefern kontinuierliche Verfahren eine detailliertere Datengrundlage, was in der Regel zu präziseren Ergebnissen führt. Allerdings ist die Anwendung dieser Modelle wesentlich aufwendiger. Typischerweise werden zeitlich kontinuierliche Verfahren in Studien über grosse Säugetiere wie Pottwale oder Grizzlybären sowie bei Insekten angewendet [6]. Ein frühes Beispiel hierfür ist eine Studie von C. C. Craig aus dem Jahre 1953 zur Erfassung von Schmetterlingen, bei der jeweils ein Schmetterling gefangen, mit einem Punkt Nagellack markiert und anschliessend sofort wieder freigelassen wurde [26]. Obwohl Craig das Verfahren damals nicht als zeitlich kontinuierlich bezeichnete, gilt es heute als eine frühe Anwendung dieses Ansatzes [6]. Auch die bei den Verfahren für geschlossene Populationen erwähnten, zeitlich diskreten Modelle von Otis et al. wurden inzwischen modifiziert, sodass diese auch in einer kontinuierlichen Form existieren [26].

In der vorliegenden Arbeit wird der Fokus auf zwei zeitlich diskrete Modelle für geschlossene Populationen gelegt, die auf jeweils zwei Stichproben basieren. Der Lincoln-Petersen-Schätzer wird in Kapitel 3 eingeführt, während Kapitel 4 sich mit dem Chapman-Schätzer beschäftigt. Diese grundlegenden Verfahren bilden das Fundament der Capture-Recapture-Verfahren, da auch weiterentwickelte und komplexere Modelle letztlich auf den Prinzipien dieser klassischen Ansätze aufbauen.

2.2 Modellannahmen

Alle Modelle und Verfahren, die in der statistischen Analyse Anwendung finden, basieren auf bestimmten Annahmen, die dazu dienen, komplexe Phänomene der realen Welt zu vereinfachen und damit besser analysierbar zu machen. Diese Annahmen stellen dabei eine Idealisierung der tatsächlichen Gegebenheiten dar [32]. Den in dieser Arbeit betrachteten Capture-Recapture-Verfahren liegen die folgenden Annahmen zugrunde:

- *Geschlossene Population.*

  Während des gesamten Untersuchungszeitraumes bleibt die Population konstant. Demographische Veränderungen in Form von Geburten, Todesfällen oder Migrationsbewegungen werden ausgeschlossen [13].

- *Homogene Fangwahrscheinlichkeiten.*

  Innerhalb der Population weisen alle Individuen die gleiche Wahrscheinlichkeit auf, in jede der verschiedenen Stichproben einbezogen zu werden [13]. Individuell beobachtbare Merkmale wie Geschlecht, Alter oder Gewicht sind dabei irrelevant [26].

- *Unabhängigkeit der Stichproben.*

  Die Erfassung eines Individuums in einer Stichprobe hat keinen Einfluss auf die Wahrscheinlichkeit, erneut in eine Stichprobe zu gelangen, da die Fangwahrscheinlichkeit als unabhängig von der Fanghistorie angenommen wird. Dementsprechend weisen markierte wie unmarkierte Individuen die gleiche Fangwahrscheinlichkeit auf [13]. Voraussetzung hierfür ist, dass zwischen den einzelnen Stichproben eine vollständige Durchmischung der Population erfolgt [29]. Diese Durchmischungsphase darf jedoch nicht zu lange andauern, da anderenfalls eine Veränderung der Population droht, was der Annahme einer geschlossenen Population widerspräche [1].

  Eine Abhängigkeit zwischen den Stichproben läge vor, wenn der Fang das Verhalten der Individuen beeinflusste, indem sie infolgedessen entweder eine erhöhte Scheu gegenüber den Fallen oder eine gesteigerte Anziehung zu diesen entwickelten. Im Englischen werden diese Verhaltensweisen als *trap-shy* beziehungsweise *trap-happy* bezeichnet [13].

- *Kein Markierungsverlust.*

  Die Markierungen bleiben über die Stichproben hinweg vollständig erhalten [8]. Somit ist eine eindeutige Reidentifikation von Individuen möglich, die zuvor bereits erfasst wurden [1].

2.3 Fanghistorien und deren Notation

Die Fanghistorie (engl. *Capture History*) stellt ein zentrales Konzept der Capture-Recapture-Verfahren dar und bildet die Grundlage für die Definition wesentlicher Variablen. Sie wird als binäre Sequenz dargestellt, bei der die einzelnen Einträge darüber Auskunft geben, ob ein bestimmtes Individuum zum jeweiligen Erhebungszeitpunkt der Stichproben gefangen wurde [26]. Im Falle eines Capture-Recapture-Verfahrens mit zwei Erhebungszeitpunkten kann die Fanghistorie durch das Zahlenpaar (x_1, x_2) dargestellt werden, wobei x_t den Fangstatus des Individuums zum Zeitpunkt $t = 1, 2$ angibt. Wurde das Individuum gefangen, so ist $x_t = 1$. Wurde es dagegen nicht gefangen, so ist $x_t = 0$ [30]. Insgesamt ergeben sich $2^2 = 4$ mögliche Fanghistorien: $(1, 1)$, $(1, 0)$ und $(0, 1)$ für mindestens einmal gefangene Individuen sowie $(0, 0)$ für solche, die zu keinem Zeitpunkt erfasst wurden [41].

Erhobene Daten können als Anzahl der Individuen mit einer bestimmten Fanghistorie zusammengefasst werden. Dabei bezeichnet $n_{i,j}$ die Anzahl der Individuen mit der Fanghistorie (i, j) für $i, j \in \{0, 1\}$ [30]. Die resultierenden Kombinationen sind in Tabelle 1 in Form einer 2×2-Kontingenztafel dargestellt [13].

		In der 2. Stichprobe?	
		Nein	Ja
In der 1. Stichprobe?	Nein	$n_{0,0}$	$n_{0,1}$
	Ja	$n_{1,0}$	$n_{1,1}$

Tabelle 1: Kontingenztafel der Häufigkeitsverteilung von Fanghistorien bei zwei Erhebungszeitpunkten.

Zur weiteren Verwendung im Rahmen dieser Arbeit werden folgende Grössen definiert, welche Werte im Bereich der natürlichen Zahlen $\mathbb{N}$ annehmen:

$$M := n_{1,0} + n_{1,1} \quad \text{Anzahl aller Individuen in der ersten Stichprobe,} \tag{2.1}$$

$$n := n_{0,1} + n_{1,1} \quad \text{Anzahl aller Individuen in der zweiten Stichprobe,} \tag{2.2}$$

$$m := n_{1,1} \quad \text{Anzahl markierter Individuen in der zweiten Stichprobe.} \tag{2.3}$$

Die Gesamtzahl der Individuen, die mindestens einmal gefangen wurden, wird mit $n_{obs} := n_{1,1}+n_{1,0}+n_{0,1}$ bezeichnet. Damit lässt sich die Gesamtpopulationsgrösse durch

$$N := \sum_{i=0}^{1}\sum_{j=0}^{1} n_{i,j} = n_{0,0} + n_{obs} \tag{2.4}$$

ausdrücken, wobei $n_{0,0}$, die Anzahl der zu keinem Zeitpunkt erfassten Individuen, unbekannt ist [30]. Folglich hängt die Schätzung der Gesamtpopulationsgrösse unmittelbar von der Schätzung des unbekannten Anteils $n_{0,0}$ ab [26]. Einen Überblick über die neu definierten Grössen bietet die Tabelle 2 [10].

		In der 2. Stichprobe?		
		Nein	Ja	
In der 1. Stichprobe?	Nein	$n_{0,0} = ?$	$n - m$	
	Ja	$M - m$	m	M
			n	$N = ?$

Tabelle 2: Übersicht über die neu eingeführten Grössen.

Kapitel 3

Der Lincoln-Petersen-Schätzer

In diesem Kapitel wird der Lincoln-Petersen-Schätzer eingeführt. Zunächst werden das zugrunde liegende mathematische Prinzip sowie die formale Definition erläutert. Anschliessend wird der Schätzer unter Verwendung der Maximum-Likelihood-Methode, deren theoretische Grundlagen dem Anhang B zu entnehmen sind, hergeleitet. Des Weiteren werden seine Genauigkeit und Präzision betrachtet, welche die Basis für das Kapitel 5 bilden.

3.1 Definition und mathematisches Grundprinzip des Lincoln-Petersen-Schätzers

Der Lincoln-Petersen-Schätzer stellt einen einfachen Verhältnis-Schätzer dar, der das Fundament der Capture-Recapture-Verfahren legt [13]. Auch weiterentwickelte und komplexere Methoden lassen sich letztlich auf diesen Ansatz zurückführen [8]. Den Ausgangspunkt dieses Verfahrens bildet eine erste Stichprobe (Capture), in der M Individuen aus einer geschlossenen Population mit unbekannter Grösse N gefangen und eindeutig markiert werden. Damit beträgt der Anteil markierter Individuen an der Gesamtpopulation $\frac{M}{N}$ [6]. Anschliessend werden die markierten Individuen wieder freigelassen und erhalten ausreichend Zeit, sich homogen unter den übrigen, nicht markierten Individuen zu verteilen [26]. Nach dieser Durchmischungsphase folgt eine zweite, unabhängige Stichprobe der Grösse n (Recapture). Dabei wird die Anzahl m der bereits markierten und erneut gefangenen Individuen erfasst. Der Anteil markierter Individuen in dieser zweiten Stichprobe beträgt somit $\frac{m}{n}$ [6]. Unter den getroffenen Annahmen homogener Fangwahrscheinlichkeiten und der Unabhängigkeit beider Stichproben kann davon ausgegangen werden, dass der Anteil markierter Individuen in der zweiten Stichprobe näherungsweise dem Anteil in der Gesamtpopulation entspricht [8]:

$$\frac{m}{n} \approx \frac{M}{N} \,. \tag{3.1}$$

Durch Umstellen dieser Näherung ergibt sich eine Schätzung der Gesamtpopulationsgrösse N,

$$\hat{N}_{LP} = \frac{Mn}{m} \,, \tag{3.2}$$

wobei $\hat{N}_{LP}$ als Lincoln-Petersen-Schätzer bezeichnet wird, der in der Literatur auch als Lincoln-Index bekannt ist [6].

Entscheidend ist die Feststellung, dass die Anzahl m der in der zweiten Stichprobe bereits markierten, erneut gefangenen Individuen einer hypergeometrischen Verteilung folgt, was formal durch $m \sim H(N, M, n)$ beschrieben wird [10]. Dies ist darauf zurückzuführen, dass in der zweiten Stichprobe n Individuen aus einer Gesamtpopulation der Grösse N gefangen werden, wobei M dieser N Individuen markiert sind. Da die gefangenen Individuen nicht vor Abschluss der zweiten Stichprobe wieder freigelassen werden, ändert sich die Wahrscheinlichkeit, ein markiertes Individuum zu fangen, mit jedem einzelnen Fang, was ein charakteristisches Merkmal der hypergeometrischen Verteilung darstellt [1]. Nähere Informationen zu dieser Verteilung finden sich im Anhang A.

3.2 Herleitung des Lincoln-Petersen-Schätzers mittels der Maximum-Likelihood-Methode

Neben der im Unterkapitel 3.1 eingeführten Definition des Lincoln-Petersen-Schätzers über Anteilsverhältnisse lässt sich dieser auch formal unter Verwendung der Maximum-Likelihood-Methode herleiten, deren theoretische Grundlagen dem Anhang B zu entnehmen sind. Gegeben seien die Werte M, n und m als feste Parameter. Die Gesamtpopulationsgrösse N sei hingegen unbekannt. Basierend auf der hypergeometrischen Verteilung ergibt sich die folgende Likelihood-Funktion in Abhängigkeit von $N \in \mathbb{N}_0$ mit $N \geq M + n - m$ [10]:

$$L(N) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \,. \tag{3.3}$$

Die Binomialkoeffizienten dieser Funktion sind für ganze Zahlen $0 \leq b \leq a$ folgendermassen definiert [37]:

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \,. \tag{3.4}$$

Hierbei bezeichnet das Ausrufezeichen die Fakultät, die für jede natürliche Zahl n als das Produkt aller natürlichen Zahlen von n bis 1 festgelegt ist: $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1$. Zudem gilt $0! = 1$ [40].

Die Likelihood-Funktion 3.3, deren Wertebereich im Intervall $[0, 1]$ liegt, beschreibt, wie plausibel verschiedene Werte der unbekannten Populationsgrösse N auf Grundlage der beobachteten Daten sind. Dabei wird angenommen, dass in einer Stichprobe der Grösse n genau m der insgesamt M markierten Individuen enthalten sind. Zur Schätzung der Gesamtpopulationsgrösse N stellt sich nun die Frage, für welchen Wert von N diese Funktion ihr Maximum erreicht (engl. *Maximum Likelihood Estimation*) [42]. Qualitativ kann festgestellt werden, dass $L(N)$ zunächst monoton wächst und dann wieder abfällt [10]. Da die Likelihood-Funktion jedoch keine stetige und differenzierbare Funktion von N ist, lässt sich das Maximum nicht durch Ableitung bestimmen. Stattdessen kann auf einen Quotientenvergleich zurückgegriffen werden, dessen Ansatz darin besteht, die Verhältnisse der Likelihoods für benachbarte Werte von N zu

untersuchen, um so das Maximum der Funktion zu identifizieren [49]. Dabei wird das Verhältnis

$$\begin{aligned} R(N) \;=\; \frac{L(N)}{L(N-1)} \;&=\; \frac{\binom{M}{m}\binom{N-M}{n-m}/\binom{N}{n}}{\binom{M}{m}\binom{N-M-1}{n-m}/\binom{N-1}{n}} \;=\; \frac{\binom{N-M}{n-m}\binom{N-1}{n}}{\binom{N-M-1}{n-m}\binom{N}{n}} \\ &= \frac{\frac{(N-M)!}{(n-m)!(N-M-n+m)!}\frac{(N-1)!}{n!(N-n-1)!}}{\frac{(N-M-1)!}{(n-m)!(N-M-n+m-1)!}\frac{N!}{n!(N-n)!}} \\ &= \frac{(N-M)!(N-1)!(N-M-n+m-1)!(N-n)!}{(N-M-1)!N!(N-M-n+m)!(N-n-1)!} \\ &= \frac{(N-M)(N-n)}{N(N-M-n+m)} \end{aligned} \tag{3.5}$$

betrachtet. Dieses Verhältnis $R(N)$ ist genau dann grösser als 1, wenn der Likelihood-Wert für N grösser ist als für $N-1$, was bedeutet, dass die Likelihood-Funktion an der Stelle N im Vergleich zu $N-1$ noch ansteigt. Der Zähler des Quotienten 3.5 muss in diesem Falle grösser sein als dessen Nenner [49]:

$$\begin{aligned} & (N-M)(N-n) \;>\; N(N-M-n+m) \\ \Leftrightarrow\quad & N^2-Nn-NM+Mn \;>\; N^2-NM-Nn+Nm \\ \Leftrightarrow\quad & Mn \;>\; Nm \\ \Leftrightarrow\quad & N \;<\; \frac{Mn}{m}\,. \end{aligned}$$

Daraus ergibt sich, dass $R(N) > 1$ für $N < \frac{Mn}{m}$, während umgekehrt $R(N) < 1$ für $N > \frac{Mn}{m}$ gilt [24]. Wäre der Parameterraum kontinuierlich auf $\mathbb{R}^+$, so würde die Likelihood-Funktion $L(N)$ ihr Maximum genau bei $N = \frac{Mn}{m}$ erreichen [13]. Da N definitionsgemäss jedoch nur ganzzahlige Werte annehmen kann, kommen stattdessen die beiden ganzzahligen Nachbarn von $\frac{Mn}{m}$ als Kandidaten für den Maximum-Likelihood-Schätzer (engl. *Maximum Likelihood Estimator*) in Betracht [24]:

$$\left\lfloor \frac{Mn}{m} \right\rfloor \quad \text{und} \quad \left\lfloor \frac{Mn}{m} \right\rfloor + 1\,. \tag{3.6}$$

Zumal $\lfloor \frac{Mn}{m} \rfloor + 1 > \frac{Mn}{m}$ gilt, folgt $R(\lfloor \frac{Mn}{m} \rfloor + 1) < 1$, was äquivalent ist zu $L(\lfloor \frac{Mn}{m} \rfloor + 1)/L(\lfloor \frac{Mn}{m} \rfloor) < 1$. Daraus resultiert, dass $L(\lfloor \frac{Mn}{m} \rfloor + 1) < L(\lfloor \frac{Mn}{m} \rfloor)$. Der Maximum-Likelihood-Schätzer der unbekannten Gesamtpopulationsgrösse N ist somit gegeben durch

$$\hat{N}_{ML} = \left\lfloor \frac{Mn}{m} \right\rfloor, \tag{3.7}$$

was dem abgerundeten Lincoln-Petersen-Schätzer entspricht [49].

3.3 Genauigkeit und Präzision des Lincoln-Petersen-Schätzers

Zur Bewertung der Güte eines Schätzers muss dieser sowohl in Bezug auf seine Genauigkeit als auch auf seine Präzision untersucht werden. Diese beiden Begriffe sind klar voneinander abzugrenzen und sollten nicht verwechselt werden. Die Genauigkeit beschreibt die Nähe des geschätzten Wertes zum wahren (in

der Regel unbekannten) Parameterwert und gibt somit Auskunft darüber, wie zutreffend die Schätzung im Mittel ist. Sie kann durch den Erwartungswert des Schätzers und seine systematische Verzerrung, auch als Bias bekannt, quantifiziert werden [28]. Die Präzision hingegen bezieht sich auf die Streuung der Schätzwerte und ist ein Mass für deren Reproduzierbarkeit. Eine präzise Schätzung ist durch eine hohe Konsistenz der Ergebnisse gekennzeichnet, unabhängig davon, ob diese dem wahren Wert nahe kommen. Sie kann durch die Varianz der Schätzwerte und die damit verbundene Standardabweichung beschrieben werden [35].

3.3.1 Erwartungswert und Verzerrung

Der Erwartungswert des Lincoln-Petersen-Schätzers ist ein gewichtetes Mittel über alle möglichen Werte, die $\hat{N}_{LP}$ annehmen kann. Dabei ergeben sich die Gewichte aus der hypergeometrischen Wahrscheinlichkeitsverteilung der Zufallsvariablen m [48]. Da M und n fixe, nicht-stochastische Terme sind, dürfen diese vor den Erwartungswert gezogen werden [40]:

$$\mathbb{E}[\hat{N}_{LP}] = \mathbb{E}\left[\frac{Mn}{m}\right] = Mn \cdot \mathbb{E}\left[\frac{1}{m}\right]. \tag{3.8}$$

Der Ausdruck $\frac{1}{m}$ stellt eine nichtlineare Transformation der Zufallsvariablen m dar. Für nichtlineare Funktionen f gilt im Allgemeinen, dass $\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$. Folglich ist $\mathbb{E}\left[\frac{1}{m}\right] \neq \frac{1}{\mathbb{E}[m]}$, was bedeutet, dass sich der Erwartungswert dieses Bruches nicht trivial bestimmen lässt [48]. Jedoch bietet die Taylorentwicklung zweiten Grades eine Möglichkeit, $\frac{1}{m}$ durch

$$\frac{1}{m} \approx \frac{1}{x_0} - \frac{1}{x_0^2}(m - x_0) + \frac{1}{x_0^3}(m - x_0)^2 \quad \text{mit} \quad x_0 = \mathbb{E}[m] = \frac{Mn}{N} \text{ (vgl. Anhang A)} \tag{3.9}$$

zu approximieren [10]. Infolge dieser Approximation erscheint die Zufallsvariable m nur noch in linearer und quadratischer Form, was die Berechnung des Erwartungswerts deutlich vereinfacht. Der Erwartungswert einer Summe von Zufallsvariablen ist gleich der Summe ihrer einzelnen Erwartungswerte, wobei nichtstochastische Terme, analog zum Vorgehen in Gleichung 3.8, vor den Erwartungswert gezogen werden dürfen. Damit lässt sich der Erwartungswert von $\frac{1}{m}$ wie folgt ausdrücken [40]:

$$\mathbb{E}\left[\frac{1}{m}\right] \approx \frac{1}{x_0} - \frac{1}{x_0^2} \cdot \mathbb{E}[m - x_0] + \frac{1}{x_0^3} \cdot \mathbb{E}\left[(m - x_0)^2\right]. \tag{3.10}$$

Da $\mathbb{E}[m - x_0] = \mathbb{E}[m] - x_0 = 0$, entfällt der lineare Term in Gleichung 3.10. Darüber hinaus entspricht $\mathbb{E}\left[(m - x_0)^2\right]$ definitionsgemäss $\mathrm{Var}(m)$ [48]. Somit vereinfacht sich der Ausdruck zu

$$\mathbb{E}\left[\frac{1}{m}\right] \approx \frac{1}{x_0} + \frac{\mathrm{Var}(m)}{x_0^3}. \tag{3.11}$$

Die erhaltene Näherung von $\mathbb{E}\left[\frac{1}{m}\right]$ kann nun in die Ausgangsgleichung 3.8 eingesetzt werden. Dabei wird $\mathrm{Var}(m)$ durch $\frac{Mn}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$ ersetzt, was der Varianz der hypergeometrisch verteilten Zufallsvariablen m entspricht [37]. Die Herleitung dieser Varianz findet sich im Anhang A. Des Weiteren ist der Entwicklungsort des Taylorpolynoms, x_0, wie in 3.9 angegeben, durch den Erwartungswert $\mathbb{E}[m] = \frac{Mn}{N}$ festgelegt,

der ebenfalls im Anhang A hergeleitet wird. Damit ergibt sich für den Erwartungswert des Lincoln-Petersen-Schätzers die folgende Approximation:

$$\begin{aligned}
\mathbb{E}[\hat{N}_{LP}] = Mn \cdot \mathbb{E}\left[\frac{1}{m}\right] &\approx Mn\left(\frac{1}{x_0} + \frac{\mathrm{Var}(m)}{x_0^3}\right) \\
&= Mn\left(\frac{1}{(\frac{Mn}{N})} + \frac{\left[\frac{Mn}{N}\cdot(1-\frac{M}{N})\cdot\frac{N-n}{N-1}\right]}{(\frac{Mn}{N})^3}\right) \\
&= Mn\left(\frac{N}{Mn} + \frac{\left[\frac{Mn(N-M)(N-n)}{N^2(N-1)}\right]}{(\frac{Mn}{N})^3}\right) \\
&= N + \frac{N(N-M)(N-n)}{Mn(N-1)}\,. \qquad (3.12)
\end{aligned}$$

Ausgehend von der hergeleiteten Näherung des Erwartungswerts lässt sich nun die Verzerrung des Lincoln-Petersen-Schätzers bestimmen. Diese ist definiert als die Differenz zwischen dem Erwartungswert des Schätzers und dem wahren Wert der zu schätzenden Grösse, in diesem Falle der tatsächlichen Gesamtpopulationsgrösse N [28]. Für den Schätzer $\hat{N}_{LP}$ ergibt sich somit

$$\text{Bias}[\hat{N}_{LP}] = \mathbb{E}[\hat{N}_{LP}] - N \approx \frac{N(N-M)(N-n)}{Mn(N-1)}\,. \qquad (3.13)$$

Unter der Voraussetzung, dass $N > M > 0$ und $N > n > 0$, ist dieser Ausdruck stets positiv. Dies impliziert, dass der Lincoln-Petersen-Schätzer eine positive Verzerrung aufweist, was bedeutet, dass die wahre Populationsgrösse im Mittel überschätzt wird [10]. Besonders ausgeprägt ist diese Verzerrung für kleine Stichprobengrössen M und n in Relation zur Gesamtpopulation N [31]. Umgekehrt, wenn $M \to N$ und $n \to N$, wächst der Nenner gegen N^3, während die Terme $N - M$ und $N - n$ im Zähler gegen null konvergieren. Die Verzerrung verschwindet damit im Grenzwert:

$$\lim_{M\to N,\, n\to N} \frac{N(N-M)(N-n)}{Mn(N-1)} = 0\,. \qquad (3.14)$$

Folglich ist der Lincoln-Petersen-Schätzer asymptotisch erwartungstreu [17].

3.3.2 Varianz und Standardabweichung

Die Varianz des Lincoln-Petersen-Schätzers misst die Streuung der geschätzten Populationsgrösse $\hat{N}$ um ihren Erwartungswert und spiegelt damit die Unsicherheit wider, die mit der Schätzung verbunden ist [48]. Da M und n fixe, nicht-stochastische Terme sind, dürfen diese vor die Varianz gezogen werden, wobei M und n quadriert werden [40]:

$$\mathrm{Var}(\hat{N}_{LP}) = \mathrm{Var}\left(\frac{Mn}{m}\right) = M^2n^2 \cdot \mathrm{Var}\left(\frac{1}{m}\right). \qquad (3.15)$$

Wie bereits bei der Herleitung der Approximation des Erwartungswerts des Lincoln-Petersen-Schätzers in Abschnitt 3.3.1 vermerkt, stellt der Term $\frac{1}{m}$ eine nichtlineare Transformation der Zufallsvariablen m dar,

weshalb dieser Term auch hier nicht durch eine elementare Methode bestimmt werden kann [48]. Jedoch kann erneut auf eine Taylorentwicklung zurückgegriffen werden, wobei in diesem Falle lediglich der erste Grad berücksichtigt wird. Dabei kann $\frac{1}{m}$ durch

$$\frac{1}{m} \approx \frac{1}{x_0} - \frac{1}{x_0^2}(m - x_0) \quad \text{mit} \quad x_0 = \mathbb{E}[m] = \frac{Mn}{N} \text{ (vgl. Anhang A)} \tag{3.16}$$

approximieren werden [10]. Die Zufallsvariable m erscheint infolgedessen nur noch in linearer Form. Demnach lässt sich die Varianz von $\frac{1}{m}$ wie folgt ausdrücken:

$$\mathrm{Var}\left(\frac{1}{m}\right) \approx \mathrm{Var}\left(\frac{1}{x_0} - \frac{1}{x_0^2} \cdot (m - x_0)\right). \tag{3.17}$$

Da $\frac{1}{x_0}$ eine Konstante darstellt, bleibt die Varianz durch den additiven konstanten Term unverändert, weshalb dieser wegfällt [48]. Im zweiten Term in der Varianz kann der konstante Faktor, analog zum Vorgehen in Gleichung 3.15, vor die Varianz gezogen werden, sodass in der Varianz nur noch der Ausdruck $m - x_0$ verbleibt [40]. Auch hier fällt die additive Konstante x_0 weg [48]. Der ganze Ausdruck vereinfacht sich somit zu

$$\mathrm{Var}\left(\frac{1}{m}\right) \approx \left(-\frac{1}{x_0^2}\right)^2 \cdot \mathrm{Var}(m - x_0) = \frac{1}{x_0^4} \cdot \mathrm{Var}(m).$$

Die erhaltene Näherung von $\mathrm{Var}\left(\frac{1}{m}\right)$ kann nun in die Ausgangsgleichung 3.15 eingesetzt werden. Dabei wird $\mathrm{Var}(m)$ durch $\frac{Mn}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$ ersetzt, was der Varianz der hypergeometrisch verteilten Zufallsvariablen m entspricht [37]. Zudem ist der Entwicklungsort des Taylorpolynoms, x_0, wie in 3.16 angegeben, durch den Erwartungswert $\mathbb{E}[m] = \frac{Mn}{N}$ festgelegt. Sowohl $\mathrm{Var}(m)$ als auch $\mathbb{E}[m]$ werden im Anhang A hergeleitet. Für die Varianz des Lincoln-Petersen-Schätzers ergibt sich damit die folgende Approximation:

$$\begin{aligned}
\mathrm{Var}(\hat{N}_{LP}) = M^2 n^2 \cdot \mathrm{Var}\left(\frac{1}{m}\right) &\approx M^2 n^2 \cdot \frac{1}{x_0^4} \cdot \mathrm{Var}(m) \\
&= M^2 n^2 \cdot \frac{1}{(\frac{Mn}{N})^4} \cdot \frac{Mn}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \frac{N-n}{N-1} \\
&= \frac{N^2(N-M)(N-n)}{Mn(N-1)}\,.
\end{aligned} \tag{3.18}$$

Ausgehend von der hergeleiteten Näherung der Varianz lässt sich nun die Standardabweichung des Lincoln-Petersen-Schätzers bestimmen. Die Standardabweichung, abgekürzt als SD (engl. *Standard Deviation*), ist als positive Quadratwurzel der Varianz definiert [28]. Für den Schätzer $\hat{N}_{LP}$ ergibt sich somit

$$\mathrm{SD}(\hat{N}_{LP}) = \sqrt{\mathrm{Var}(\hat{N}_{LP})} \approx \sqrt{\frac{N^2(N-M)(N-n)}{Mn(N-1)}}\,. \tag{3.19}$$

In der Praxis ist die Grösse der Gesamtpopulation N jedoch meist unbekannt, sodass für die Berechnung der Varianz sowie der Standardabweichung N durch $\hat{N}_{LP} = \frac{Mn}{m}$ ersetzt wird [31].

Kapitel 4

Der Chapman-Schätzer

Im vorliegenden Kapitel wird der Chapman-Schätzer als Weiterentwicklung des zuvor in Kapitel 3 vorgestellten Lincoln-Petersen-Schätzers eingeführt. Zunächst wird die Motivation für diese Modifikation dargelegt, gefolgt von der formalen Definition des Chapman-Schätzers. Analog zum Lincoln-Petersen-Schätzer werden anschliessend die Genauigkeit und Präzision des Chapman-Schätzers untersucht, was die Grundlage für den Vergleich beider Schätzer in Kapitel 5 bildet.

4.1 Motivation für die Modifikation des Lincoln-Petersen-Schätzers

Obwohl der Lincoln-Petersen-Schätzer aufgrund der intuitiven Herleitung über Anteilsverhältnisse und seiner einfachen Berechnungsweise leicht zu verstehen und anzuwenden ist, weist er zwei wesentliche Schwächen auf. Zum einen ist der Schätzer systematisch verzerrt (vgl. Abschnitt 3.3.1). Insbesondere bei kleinen Stichprobenumfängen in Relation zur Gesamtpopulation tendiert er dazu, die wahre Populationsgrösse zu überschätzen [31]. Zum anderen ergibt sich ein numerisches Problem, wenn in der zweiten Stichprobe kein einziges bereits markiertes Individuum erneut gefangen wird. In diesem Falle ist $m = 0$, was zu einer Division durch null führt und den Lincoln-Petersen-Schätzer unbrauchbar macht, da dieser unter solchen Umständen nicht definiert ist [4]. Eine vollständige Trennung der beiden Stichproben verhindert jegliche Rückschlüsse auf die Verbindung zwischen ihnen. Es lässt sich lediglich feststellen, dass mindestens $M + n$ Individuen existieren müssen, also mindestens so viele wie in den beiden Stichproben zusammen, wobei die tatsächliche Populationsgrösse potenziell erheblich darüber liegen kann [13].

Angesichts dieser beiden Schwächen wurde nach einem Schätzer gesucht, der sowohl rechnerisch stabil als auch möglichst unverzerrt ist. In einer im Jahre 1951 veröffentlichten Arbeit schlug der kanadisch-amerikanische Statistiker Douglas G. Chapman folgenden Schätzer vor [7]:

$$\hat{N}_C = \frac{(M+1)(n+1)}{(m+1)} - 1 \,. \tag{4.1}$$

Dieser heute als Chapman-Schätzer bekannte Ausdruck, im Folgenden mit $\hat{N}_C$ bezeichnet, stellt eine Modifikation des ursprünglichen Lincoln-Petersen-Schätzers dar [7]. Ein entscheidender Vorteil liegt darin, dass der Schätzer auch dann definiert bleibt, wenn $m = 0$ ist. Obwohl die Unsicherheit in einem solchen

Extremfall natürlich sehr hoch ist, liefert der Chapman-Schätzer dennoch einen zwar grossen, aber endlichen Schätzwert $\hat{N}_C = (M+1)(n+1) - 1$ und bleibt somit auch unter problematischen Bedingungen anwendbar, in denen der Lincoln-Petersen-Schätzer versagt [13].

4.2 Genauigkeit und Präzision des Chapman-Schätzers

Im folgenden Unterkapitel werden analog zur Analyse des Lincoln-Petersen-Schätzers die statistischen Eigenschaften des Chapman-Schätzers näher untersucht. Dabei liegt das Hauptaugenmerk auf dem Erwartungswert und der daraus resultierenden Verzerrung.

4.2.1 Erwartungswert und Verzerrung

Die Herleitung des Erwartungswerts des Chapman-Schätzers beginnt mit derselben Vorgehensweise wie derjenigen des Lincoln-Petersen-Schätzers in Abschnitt 3.3.1. Zunächst wird der Erwartungswert in einzelne Terme aufgeteilt, was aufgrund der Linearität des Erwartungswertoperators möglich ist. So kann der Erwartungswert einer Summe im Allgemeinen als Summenwert der Erwartungswerte umformuliert werden [37]. Im ersten Erwartungswert sind $(M+1)$ und $(n+1)$ fixe, nicht-stochastische Terme, weshalb diese vor den Erwartungswert gezogen werden dürfen [40]. Der zweite Erwartungswert dagegen bezieht sich auf eine Konstante, deren Erwartungswert definitionsgemäss dem Wert der Konstanten selbst entspricht [48]. Somit hängt der Erwartungswert des Chapman-Schätzers ausschliesslich vom Term $\frac{1}{m+1}$ ab, wobei m eine hypergeometrisch verteilte Zufallsvariable darstellt:

$$\mathbb{E}[\hat{N}_C] = \mathbb{E}\left[\frac{(M+1)(n+1)}{(m+1)} - 1\right] = (M+1)(n+1) \cdot \mathbb{E}\left[\frac{1}{m+1}\right] - 1\,. \tag{4.2}$$

Der Erwartungswert des Terms $\frac{1}{m+1}$ lässt sich gemäss der Definition des Erwartungswerts einer diskreten Zufallsvariablen als Summe der Produkte aus allen möglichen Werten und ihren jeweiligen Wahrscheinlichkeiten ausdrücken [28]. Die Wahrscheinlichkeiten ergeben sich dabei aus der hypergeometrischen Wahrscheinlichkeitsverteilung, die im Anhang A näher erläutert wird:

$$\mathbb{E}\left[\frac{1}{m+1}\right] = \sum_{k=0}^{n} \frac{1}{k+1} \cdot \mathbb{P}(m=k) = \sum_{k=0}^{n} \frac{1}{k+1} \cdot \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}\,. \tag{4.3}$$

Streng genommen läuft die Summe lediglich bis $\min(M,n)$, da m höchstens so gross wie die kleinere der beiden Stichproben M und n sein kann. Wird die obere Grenze jedoch mit n angegeben, so führt dies zum selben Resultat, da alle Terme mit $k > M$ aufgrund des Binomialkoeffizienten $\binom{M}{k} = 0$ wegfallen [27]. Des Weiteren lässt sich unter Verwendung der Fakultätsdarstellung des Binomialkoeffizienten zeigen, dass

$$\frac{1}{k+1}\binom{M}{k} = \frac{1}{k+1} \cdot \frac{M!}{k!(M-k)!} = \frac{1}{M+1} \cdot \frac{(M+1)!}{(k+1)!(M-k)!} = \frac{1}{M+1}\binom{M+1}{k+1}, \tag{4.4}$$

wodurch sich der Erwartungswert aus Gleichung 4.3 zu

$$\mathbb{E}\left[\frac{1}{m+1}\right] = \sum_{k=0}^{n} \frac{1}{M+1} \cdot \frac{\binom{M+1}{k+1}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{4.5}$$

umformen lässt [37]. Zur Vereinfachung der Summe wird der Summationsindex k verschoben, indem $r = k + 1$ gesetzt wird. Da k die Werte 0 bis n annimmt, läuft r folglich von 1 bis $n + 1$. In der Summe wird der alte Index k jeweils durch $r - 1$ ersetzt. Zudem lassen sich Konstanten, die nicht von der Laufvariablen r abhängen, ausserhalb der Summe faktorisiert [28]. Dadurch ergibt sich

$$\mathbb{E}\left[\frac{1}{m+1}\right] = \frac{1}{(M+1)\binom{N}{n}} \cdot \sum_{r=1}^{n+1} \binom{M+1}{r}\binom{N-M}{n-r+1}. \tag{4.6}$$

Um diese Summe weiter zu berechnen, wird die Vandermonde-Identität herangezogen, welche besagt, dass

$$\sum_{r=0}^{n+1} \binom{M+1}{r}\binom{N-M}{n-r+1} = \binom{N+1}{n+1} \tag{4.7}$$

für N, M, $m \in \mathbb{N}_0$ mit $N > M$ gilt. Ein Beweis dieser Identität findet sich in [21]. Da die Summe in Gleichung 4.6 jedoch erst bei $r = 1$ beginnt, muss der erste Term für $r = 0$ von der Summe in 4.7 subtrahiert werden. Somit erhält man

$$\begin{aligned}\sum_{r=1}^{n+1} \binom{M+1}{r}\binom{N-M}{n-r+1} &= \sum_{r=0}^{n+1} \binom{M+1}{r}\binom{N-M}{n-r+1} - \sum_{r=0}^{0} \binom{M+1}{r}\binom{N-M}{n-r+1} \\ &= \binom{N+1}{n+1} - \binom{N-M}{n+1},\end{aligned} \tag{4.8}$$

wobei der Binomialkoeffizient $\binom{M+1}{r}$ für $r = 0$ den Wert 1 ergibt [37]. Wird dieses Ergebnis in Gleichung 4.6 eingesetzt, führt dies auf den folgenden Ausdruck:

$$\mathbb{E}\left[\frac{1}{m+1}\right] = \frac{1}{M+1} \cdot \frac{\binom{N+1}{n+1} - \binom{N-M}{n+1}}{\binom{N}{n}}. \tag{4.9}$$

In einem nächsten Schritt werden die beiden Quotienten im zweiten Term der Gleichung 4.9 separat betrachtet. Der erste Quotient kann mithilfe der Fakultätsdarstellung der Binomialkoeffizienten zu

$$\frac{\binom{N+1}{n+1}}{\binom{N}{n}} = \frac{\frac{(N+1)!}{(n+1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{N+1}{n+1} \tag{4.10}$$

umgeformt werden [28]. Analog ergibt sich für den zweiten Quotienten

$$\frac{\binom{N-M}{n+1}}{\binom{N}{n}} = \frac{\frac{(N-M)!}{(n+1)!(N-M-n-1)!}}{\frac{N!}{n!(N-n)!}} = \frac{(N-M)!(N-n)!}{N!(n+1)(N-M-n-1)!}. \tag{4.11}$$

Setzt man die Resultate der beiden Quotienten 4.10 und 4.11 in Gleichung 4.9 ein und klammert anschliessend $\frac{N+1}{n+1}$ aus, so erhält man

$$\begin{aligned}\mathbb{E}\left[\frac{1}{m+1}\right] &= \frac{1}{M+1}\left(\frac{N+1}{n+1} - \frac{(N-M)!(N-n)!}{N!(n+1)(N-M-n-1)!}\right) \\ &= \frac{N+1}{(M+1)(n+1)}\left(1 - \frac{(N-M)!(N-n)!}{(N+1)!(N-M-n-1)!}\right).\end{aligned} \tag{4.12}$$

Dieses Resultat für $\mathbb{E}\left[\frac{1}{m+1}\right]$ kann nun in die Ausgangsgleichung 4.2 eingesetzt werden, was schliesslich zum Erwartungswert des Chapman-Schätzers führt [7]:

$$\begin{aligned}\mathbb{E}[\hat{N}_C] &= (M+1)(n+1)\cdot\mathbb{E}\left[\frac{1}{m+1}\right]-1\\ &= (N+1)\left(1-\frac{(N-M)!(N-n)!}{(N+1)!(N-M-n-1)!}\right)-1\\ &= N-\frac{(N-M)!(N-n)!}{N!(N-M-n-1)!}\,.\end{aligned} \tag{4.13}$$

Ausgehend von diesem hergeleiteten Erwartungswert lässt sich die Verzerrung des Chapman-Schätzers bestimmen. Für den Schätzer $\hat{N}_C$ ergibt sich

$$\text{Bias}[\hat{N}_C] = \mathbb{E}[\hat{N}_C] - N = -\frac{(N-M)!(N-n)!}{N!(N-M-n-1)!}\,. \tag{4.14}$$

Gemäss der Literatur ist diese Verzerrung jedoch vernachlässigbar, sofern $m \geq 7$ gilt, also mindestens 7 bereits markierte Individuen in der zweiten Stichprobe erneut gefangen werden. Unter Verwendung einer binomialen Näherung kann ausserdem gezeigt werden, dass Bias$[\hat{N}_C] = 0$, wenn $M+n \geq N$ [20]. Damit wird die Schwäche des Lincoln-Petersen-Schätzers, die Tendenz zur Überschätzung der Populationsgrösse, in den allermeisten praktischen Fällen korrigiert [26].

4.2.2 Varianz und Standardabweichung

Da die Herleitung der Varianz des Chapman-Schätzers den Rahmen dieser Arbeit sprengen würde, wird sie im Folgenden ohne Beweis eingeführt. In der Literatur wird die Varianz üblicherweise durch die folgende Näherungsformel beschrieben [13]:

$$\text{Var}(\hat{N}_C) \approx N^2\left[\frac{N}{Mn}+2\left(\frac{N}{Mn}\right)^2+6\left(\frac{N}{Mn}\right)^3\right]. \tag{4.15}$$

Dieser Ausdruck wurde von Douglas G. Chapman in einer im Jahre 1951 veröffentlichten Arbeit (vgl. [7]) hergeleitet. Die Definition der Varianz

$$\text{Var}(X) = \mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2-2X\,\mathbb{E}[X]+\mathbb{E}[X]^2\right] = \mathbb{E}[X^2]-\mathbb{E}[X]^2 \tag{4.16}$$

bildet die Grundlage der Herleitung [37]. Einerseits greift Chapman auf den Erwartungswert $\mathbb{E}\left[\frac{1}{m+1}\right]$ zurück, der dem in Gleichung 4.12 dargestellten Ausdruck entspricht. Andererseits approximiert er den Erwartungswert des Terms $\frac{1}{(m+1)^2}$ mithilfe einer Reihenentwicklung, deren Herleitung auf eine von Frederick F. Stephan im Jahre 1945 veröffentlichte Arbeit zurückgeht (vgl. [22]).

Ausgehend von der approximierten Varianz lässt sich die Standardabweichung des Chapman-Schätzers bestimmen, die der positiven Quadratwurzel der Varianz entspricht [28]. Für den Schätzer $\hat{N}_C$ ergibt sich somit

$$\text{SD}(\hat{N}_C) = \sqrt{\text{Var}(\hat{N}_C)} \approx \sqrt{N^2\left[\frac{N}{Mn}+2\left(\frac{N}{Mn}\right)^2+6\left(\frac{N}{Mn}\right)^3\right]}\,. \tag{4.17}$$

Da in der Praxis die Grösse der Gesamtpopulation N in der Regel unbekannt ist, wird für die Berechnung der Varianz sowie der Standardabweichung N meist durch $\hat{N}_C = \frac{(M+1)(n+1)}{(m+1)} - 1$ ersetzt [31].

Kapitel 5

Vergleich der Schätzverfahren: Lincoln-Petersen versus Chapman

In diesem Kapitel werden der Lincoln-Petersen-Schätzer und der Chapman-Schätzer mithilfe von Simulationen miteinander verglichen. Der in Anhang C präsentierte Python-Code dient dabei stets als Grundlage. Während im ersten Teil die Modellannahmen aus dem Unterkapitel 2.2 noch vollständig eingehalten werden, werden diese im zweiten Teil gezielt gebrochen, um die Robustheit der Schätzverfahren zu testen.

5.1 Vergleich der Schätzverfahren im Kontext erfüllter Modellannahmen

Simulationen sind wertvolle Instrumente zur Untersuchung des Verhaltens statistischer Methoden unter kontrollierten Bedingungen [13]. Sie beruhen auf dem Gesetz der grossen Zahlen, welches besagt, dass die Ergebnisse einer zufälligen Simulation mit zunehmender Wiederholungszahl gegen den Erwartungswert konvergieren [37]. Im ersten Teil des Kapitels werden dabei alle Modellannahmen als erfüllt betrachtet.

5.1.1 Simulation 1: Populationsschätzung bei konstanten Parametern

In der ersten Simulation wird die wahre Populationsgrösse $N = 500$ als unbekannt angenommen. In jedem Simulationsdurchlauf werden zunächst $M = 100$ Individuen zufällig aus der Population gefangen und markiert (Capture). Im Anschluss erfolgt eine zweite, unabhängige Stichprobe, die ebenfalls $n = 100$ Individuen enthält (Recapture). Dabei wird die Anzahl m der bereits markierten und erneut gefangenen Individuen erfasst. Dieses Vorgehen wird insgesamt $5'000$ mal wiederholt, wobei der Lincoln-Petersen-Schätzer sowie der Chapman-Schätzer für jeden Durchlauf auf Basis der beobachteten Werte M, n und m je einen Schätzwert $\hat{N}$ für die Populationsgrösse liefern. Die Verteilung dieser jeweils $5'000$ Schätzwerte ist in den Boxplots in Abbildung 1 dargestellt. Zudem sind die wichtigsten statistischen Kennwerte der Schätzwerte $\hat{N}$, gerundet auf zwei Dezimalstellen, der Tabelle 3 zu entnehmen.

Schätzer	Mittelwert (Verzerrung)	Median	IQR	Spannweite	$\mathrm{Var}(\hat{N})$
Lincoln-Petersen	$517.53\ (17.53)$	500	101.01	$[294.12; 1250]$	$10'433.02$
Chapman	$500.05\ (0.05)$	484.76	93.37	$[290.46; 1'132.44]$	$8'683.97$

Tabelle 3: Statistische Kennwerte der Schätzwerte $\hat{N}$ aus $5'000$ Simulationen ($M = n = 100$, $N = 500$).

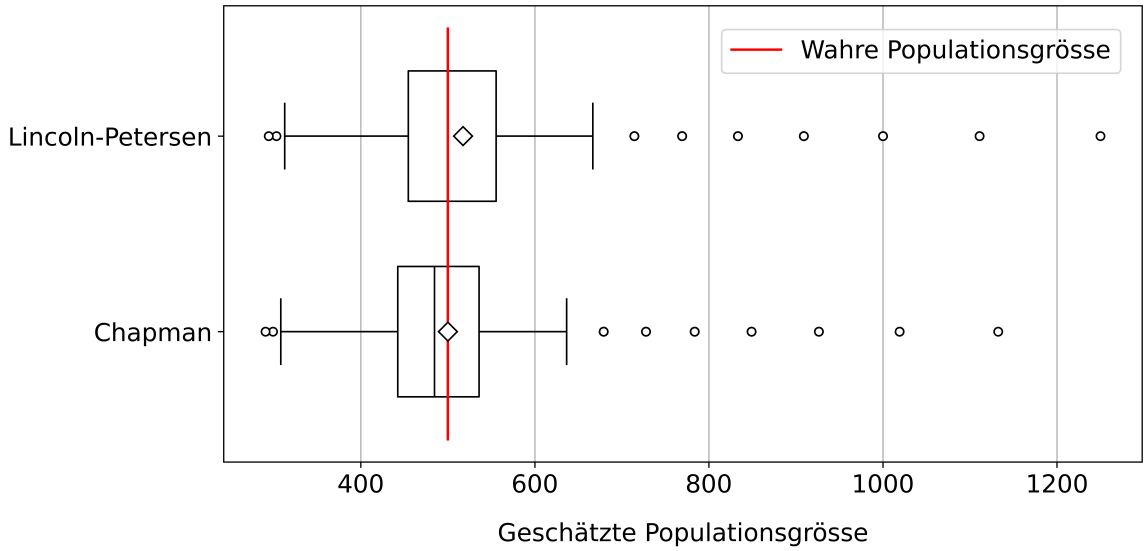

Abbildung 1: Boxplots der Schätzwerte $\hat{N}$ aus $5'000$ Simulationen ($M = n = 100$, $N = 500$).

Die Mittelwerte der Schätzwerte, die in den Boxplots durch Rhomben hervorgehoben werden, zeigen interessante Unterschiede zwischen den beiden Schätzverfahren. Der Lincoln-Petersen-Schätzer liefert einen Mittelwert von 517.53, was eine positive Verzerrung im Vergleich zur wahren Populationsgrösse von $N = 500$ widerspiegelt. Diese Verzerrung ist zu erwarten, da sie mit der theoretischen Herleitung des Erwartungswerts und der Verzerrung in Abschnitt 3.3.1 übereinstimmt. Berechnet man den Erwartungswert des Lincoln-Petersen-Schätzers mit der hergeleiteten approximativen Formel 3.12, so ergibt sich für die gewählten Parameter $M = n = 100$ ein Wert von

$$\mathbb{E}[\hat{N}_{LP}] \approx N + \frac{N(N-M)(N-n)}{Mn(N-1)} = 500 + \frac{500 \cdot (500-100)(500-100)}{100 \cdot 100 \cdot (500-1)} \approx 516.03\,,$$

der dem in der Simulation erhaltenen Mittelwert von 517.53 sehr nahe kommt. Im Gegensatz dazu liegt der Mittelwert des Chapman-Schätzers bei 500.05, was eine nahezu vernachlässigbare Verzerrung im Vergleich zur wahren Populationsgrösse darstellt. Bei den gewählten Parametern $M = 100$ und $n = 100$ ist der Wert von m in der Regel ≥ 7. Unter diesen Bedingungen wird die Verzerrung des Chapman-Schätzers gemäss der Literatur als vernachlässigbar angesehen [20]. Der Median der Schätzwerte zeigt beim Lincoln-Petersen-Schätzer mit genau 500 eine deutlich bessere Übereinstimmung mit der wahren Populationsgrösse als beim Chpaman-Schätzer, dessen Median bei 484.76 liegt. Bei näherer Betrachtung der Boxplots wird ersichtlich, dass der Chapman-Schätzer in der Mehrheit der Fälle die wahre Populationsgrösse unterschätzt. Jedoch finden sich einzelne Ausreisser mit deutlich höheren Schätzwerten, die die Verzerrung im Mittelwert wieder ausgleichen.

Betrachtet man die Streuung der Schätzwerte, so erkennt man, dass sowohl der Interquartilsabstand (IQR) als auch die Spannweite beim Chapman-Schätzer kleiner sind als beim Lincoln-Petersen-Schätzer. Die kleinsten Schätzwerte beider Verfahren sind nahezu gleich, wobei der Lincoln-Petersen-Schätzer einen minimalen Wert von 294.12 und der Chapman-Schätzer einen Wert von 290.46 erreicht. Allerdings weist der Lincoln-Petersen-Schätzer mit $1'250$ eine deutlich grössere Streuung nach oben auf als der Chapman-

Schätzer, dessen maximaler Wert bei $1'132.44$ liegt. Auch die Varianz der Schätzwerte ist beim Lincoln-Petersen-Schätzer mit $10'433.02$ grösser als beim Chapman-Schätzer, dessen Varianz $8'683.97$ beträgt. Nichtsdestotrotz liefern beide Schätzverfahren in einzelnen Fällen Schätzwerte, die mehr als doppelt so gross sind wie die wahre Populationsgrösse. Dies wird insbesondere in den Boxplots ersichtlich, die Ausreisser zeigen, die besonders nach oben hin ausgeprägt sind.

Auf Grundlage der Ergebnisse dieser ersten Simulation ist der Chapman-Schätzer dem Lincoln-Petersen-Schätzer vorzuziehen. Auch wenn der Median des Lincoln-Petersen-Schätzers näher an der wahren Populationsgrösse liegt, zeigt dieser eine positive Verzerrung. Der Chapman-Schätzer hingegen weist eine vernachlässigbare Verzerrung auf und liefert zudem stabilere Schätzungen mit geringerer Streuung, was ihn insgesamt zuverlässiger macht. Bei beiden Schätzern finden sich jedoch Ausreisser, insbesondere nach oben, was bei der Anwendung deren beachtet werden sollte.

5.1.2 Simulation 2: Populationsschätzung bei variierenden Parametern

In der zweiten Simulation wird die wahre Populationsgrösse $N = 500$ erneut als unbekannt angenommen. Im Vergleich zur ersten Simulation variieren hier jedoch die beiden Parameter M und n, wobei diese in allen Durchläufen jeweils denselben Wert annehmen ($M = n$). Für jeden betrachteten Fall werden zunächst M Individuen zufällig aus der Population gefangen und markiert (Capture). Anschliessend erfolgt eine zweite, unabhängige Stichprobe der Grösse n, die gleich viele Individuen umfasst wie die erste (Recapture). Anhand der Anzahl m der bereits markierten und erneut gefangenen Individuen liefern der Lincoln-Petersen- sowie der Chapman-Schätzer jeweils einen Schätzwert $\hat{N}$ für die Populationsgrösse. Dabei wird die Simulation für jeden betrachteten Fall insgesamt $5'000$ mal wiederholt. Die resultierenden Mittelwerte und Standardabweichungen der geschätzten Populationsgrössen sind in den Fehlerbalkendiagrammen in Abbildung 2 dargestellt.

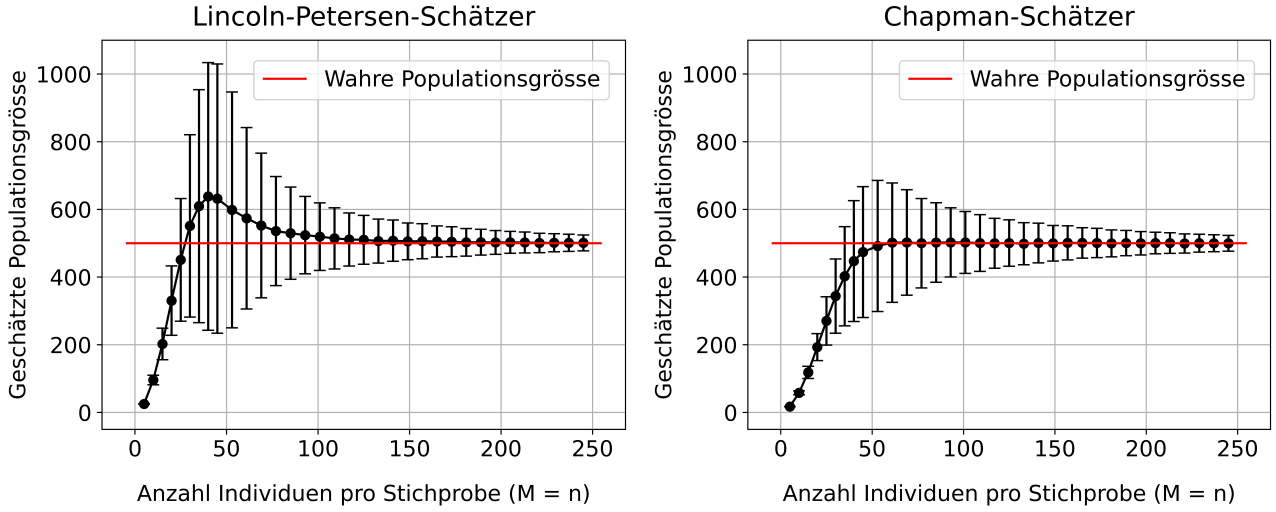

Abbildung 2: Fehlerbalkendiagramme der Schätzwerte $\hat{N}$ bei variierenden Parametern ($M = n$, $N = 500$).

Im Fehlerbalkendiagramm des Lincoln-Petersen-Schätzers zeigt sich, dass dieser insbesondere bei kleinen Stichprobenumfängen eine deutliche Verzerrung aufweist. Für kleine Werte von $M = n$, bei Stichprobengrössen unter etwa 25 Individuen, wird die wahre Populationsgrösse $N = 500$ systematisch unterschätzt. Diese Verzerrung entsteht dadurch, dass der Zähler des Lincoln-Petersen-Schätzers, Mn, bei solch kleinen Stichprobengrössen die wahre Populationsgrösse N gar nicht oder nur in begrenztem Masse überschreiten kann. Zudem wird dieser Zähler durch die Anzahl m der erneut gefangenen Individuen dividiert, was den Schätzwert weiter verringert, selbst wenn m sehr klein ist. Simulationsdurchläufe mit $m = 0$ wurden aus der Analyse ausgeschlossen, da der Lincoln-Petersen-Schätzer in diesen Fällen aufgrund der Division durch null keinen Schätzwert liefert. Auch beim Chapman-Schätzer fehlen diese Fälle, obwohl dieser unter den genannten Bedingungen einen gültigen Schätzwert liefern könnte.

Im Bereich zwischen $M = n \approx 25$ und $M = n \approx 75$ tritt hingegen eine Überschätzung der Populationsgrösse durch den Lincoln-Petersen-Schätzer auf. Diese Überschätzung entspricht der typischen positiven Verzerrung, die in Abschnitt 3.3.1 hergeleitet wird [10]. Erst ab einem Stichprobenumfang von etwa $M = n \approx 100$ nähern sich die Mittelwerte der Schätzwerte $\hat{N}_{LP}$ der tatsächlichen Populationgrösse an und stabilisieren sich. Diese Beobachtung bestätigt die in Gleichung 3.14 dargelegte Eigenschaft des Lincoln-Petersen-Schätzers, dass dieser asymptotisch erwartungstreu ist [17]. Darüber hinaus nimmt die Standardabweichung mit wachsendem Stichprobenumfang ab, was sich an den immer kleiner werdenden Fehlerbalken im Diagramm ablesen lässt. Betrachtet man den Ausdruck 3.19 der hergeleiteten approximativen Standardabweichung, so erkennt man, dass der Zähler dieses Terms gegen null strebt, wenn $M \to N$ und $n \to N$, was zur Folge hat, dass der gesamte Term gegen null konvergiert:

$$\lim_{M \to N,\, n \to N} \mathrm{SD}(\hat{N}_{LP}) = \lim_{M \to N,\, n \to N} \sqrt{\frac{N^2(N-M)(N-n)}{Mn(N-1)}} = 0\,.$$

Dies impliziert eine zunehmende Präzision des Lincoln-Petersen-Schätzers im Grenzwert.

Aus dem Fehlerbalkendiagramm des Chapman-Schätzers geht hervor, dass dieser im Vergleich zum Lincoln-Petersen-Schätzer eine deutlich geringere Verzerrung aufweist. Zwar unterschätzt auch der Chapman-Schätzer bei kleinen Stichprobengrössen unter etwa 50 Individuen, aus demselben Grund wie der Lincoln-Petersen-Schätzer, die Populationsgrösse, jedoch nähern sich die Mittelwerte der Schätzwerte $\hat{N}_C$ bereits ab einer Stichprobengrösse von $M = n \approx 50$ sehr genau der wahren Populationsgrösse an. In der zugrunde liegenden Simulation entspricht diese Stichprobengrösse von 50 Individuen 10% der Gesamtpopulation. Bei grösseren Populationen gelingt diese Annäherung jedoch bereits bei anteilsmässig kleineren Stichprobengrössen. Eine Tendenz zur Überschätzung ist beim Chapman-Schätzer zu keinem Zeitpunkt erkennbar. Dies bestätigt die in Abschnitt 4.2.1 dargelegte Eigenschaft des Chapman-Schätzers, dass dessen Verzerrung für $M + n \geq N$ komplett entfällt beziehungsweise für $m \geq 7$ vernachlässigbar wird [20]. Der Chapman-Schätzer korrigiert daher die systematische positive Verzerrung des Lincoln-Petersen-Schätzers. Darüber hinaus ist die Streuung der Schätzwerte besonders im Bereich von $M = n \approx 25$ bis $M = n \approx 75$ geringer als beim Lincoln-Petersen-Schätzer, während sich ausserhalb

dieses Bereichs ein vergleichbares Verhalten beider Schätzer hinsichtlich der Präzision zeigt. So nimmt mit zunehmender Stichprobengrösse auch die Standardabweichung des Chapman-Schätzers weiter ab, was an den immer kleiner werdenden Fehlerbalken zu erkennen ist.

Obwohl beide Schätzer mit zunehmender Stichprobengrösse gegen die wahre Populationsgrösse konvergieren, was deren Konsistenz unterstreicht, erweist sich der Chapman-Schätzer auf Grundlage der Ergebnisse dieser zweiten Simulation erneut als der bevorzugte Schätzer. Während sich die Schätzwerte beider Verfahren ab einer Stichprobengrösse von $M = n \approx 50$ hinsichtlich der Genauigkeit und Präzision kaum unterscheiden, bietet der Chapman-Schätzer bei kleinen Stichprobengrössen klare Vorteile. Er korrigiert die positive Verzerrung des Lincoln-Petersen-Schätzers und liefert somit genauere Schätzwerte, wobei der zusätzliche Rechenaufwand im Vergleich zum Lincoln-Petersen-Schätzer vernachlässigbar bleibt.

5.2 Vergleich der Schätzverfahren bei Verletzungen der Modellannahmen

Sowohl der Lincoln-Petersen-Schätzer als auch der Chapman-Schätzer basieren auf idealisierten Modellannahmen, die im Unterkapitel 2.2 definiert werden. In der praktischen Anwendung sind diese Annahmen jedoch äusserst selten vollständig erfüllt. Daher wird in den folgenden Abschnitten untersucht, wie gezielte Verletzungen der Modellannahmen die Schätzwerte beider Verfahren beeinflussen, was eine Überprüfung ihrer Robustheit ermöglicht.

Verletzung der Annahme einer geschlossenen Population

Die Annahme einer geschlossenen Population wird verletzt, wenn sich diese während des Untersuchungszeitraumes, beispielsweise in einer Tierpopulation durch Geburten, Todesfälle oder Migrationsbewegungen, verändert [13]. Um beurteilen zu können, wie sich solche demographischen Veränderungen auf die Schätzwerte auswirken, muss zwischen der Populationsgrösse zum Zeitpunkt der ersten Stichprobe N_1 und jener zum Zeitpunkt der zweiten Stichprobe N_2 unterschieden werden [10].

Sind markierte und unmarkierte Individuen gleichermassen von Abwanderung oder Mortalität betroffen, nimmt die Gesamtzahl der Individuen ab, während der Anteil markierter Individuen unverändert bleibt. In diesem Falle bleibt die Schätzung von N_1 unverzerrt. Wird jedoch N_2 geschätzt, kommt es zu einer systematischen Überschätzung [10]. Die Schätzer berücksichtigen weiterhin die ursprüngliche Anzahl markierter Individuen M, obwohl zum Zeitpunkt der zweiten Stichprobe tatsächlich weniger markierte Individuen in der Population vorhanden sind. Da M sowohl im Zähler des Lincoln-Petersen-Schätzers als auch des Chapman-Schätzers zu finden ist, führt dies zu einer Überschätzung der wahren Populationsgrösse N_2 [32]. Der Chapman-Schätzer liefert im Mittel einen etwas genaueren Schätzwert als der Lincoln-Petersen-Schätzer, jedoch sind die Verzerrungen beider Verfahren im Allgemeinen vergleichbar, was sich auch in Simulationen zeigt.

Kommen durch Zuwanderung oder Geburten neue Individuen hinzu, sind diese stets unmarkiert, wodurch der Anteil markierter Individuen in der Population sinkt, während die Anzahl markierter Individuen M konstant bleibt. Unter diesen Umständen hat die demographische Veränderung keinen Einfluss auf die Schätzung von N_2, da der Anteil markierter Individuen zum Zeitpunkt der zweiten Stichprobe korrekt erfasst wird [10]. Im Vergleich zum Anfangszustand werden jedoch aufgrund des verringerten Anteils markierter Individuen weniger bereits markierte Individuen erneut gefangen, sodass m kleiner ausfällt. Da m im Nenner beider Schätzer vorzufinden ist, führt dies zu einer Überschätzung von N_1 [26]. Auch hier sind die Schätzwerte des Chapman-Schätzers im Durchschnitt etwas genauer, obwohl die Verzerrungen in einem vergleichbaren Mass wie beim Lincoln-Petersen-Schätzer auftreten.

In realen Populationen treten Geburten, Todesfälle sowie Migrationsbewegungen meist gleichzeitig auf. Da Zuwanderung und Geburten den Anteil markierter Individuen verringern und damit zu einer Überschätzung von N_1 führen, während Abwanderung und Mortalität die Anzahl markierter Tiere zwischen den Stichproben reduzieren und dadurch eine Überschätzung von N_2 bewirken, ergibt sich insgesamt eine Tendenz zur Überschätzung der Populationsgrösse zu beiden Erhebungszeitpunkten [35]. Dabei reagieren der Lincoln-Petersen-Schätzer und der Chapman-Schätzer nahezu identisch auf diese demographischen Veränderungen.

Verletzung der Annahme homogener Fangwahrscheinlichkeiten

Weisen die Individuen innerhalb der Population unterschiedliche Wahrscheinlichkeiten auf, in jede der verschiedenen Stichproben einbezogen zu werden, wird die Annahme homogener Fangwahrscheinlichkeiten verletzt [13]. Solche Unterschiede lassen sich typischerweise auf individuell beobachtbare Merkmale wie Geschlecht, Alter oder Gewicht zurückführen [26]. Sind einige Individuen leichter zu fangen als andere, erhöht sich deren Wahrscheinlichkeit, sowohl in der ersten als auch in der zweiten Stichprobe erfasst zu werden. Dadurch vergrössert sich die Überschneidung der beiden Stichproben, und die Anzahl der bereits markierten, erneut gefangenen Individuen m nimmt zu. Da m beim Lincoln-Petersen-Schätzer als auch beim Chapman-Schätzer im Nenner steht, führt ein grösserer Wert von m zu einer Unterschätzung der wahren Populationsgrösse N [10].

Simulationsergebnisse bestätigen diesen Zusammenhang. Auch bei normalverteilten individuellen Fangwahrscheinlichkeiten ergibt sich eine systematische, wenn auch insgesamt nur geringe Unterschätzung von N. Da der Lincoln-Petersen-Schätzer im Mittel etwas höhere Schätzwerte liefert als der Chapman-Schätzer, liegt dieser tendenziell näher an der tatsächlichen Populationsgrösse. Die Ausprägung der Verzerrung ist jedoch bei beiden Schätzverfahren vergleichbar.

Verletzung der Annahme unabhängiger Stichproben

Der Fang eines Individuums beeinflusst in realen Populationen oft dessen Wahrscheinlichkeit, in einer späteren Stichprobe erneut gefangen zu werden, wodurch die Annahme unabhängiger Stichproben ver-

letzt wird. Individuen können infolge des ersten Fangs entweder eine erhöhte Scheu gegenüber den Fallen (trap-happy) oder eine gesteigerte Anziehung zu diesen (trap-shy) entwickeln. Diese Verhaltensänderungen führen dazu, dass markierte und unmarkierte Individuen in der zweiten Stichprobe unterschiedliche Fangwahrscheinlichkeiten aufweisen [13]. Eine weitere Ursache für eine Abhängigkeit kann eine zu kurze Durchmischungsphase zwischen den beiden Stichproben sein. Wenn die markierten Individuen nicht genügend Zeit erhalten, sich vollständig unter den nicht markierten zu verteilen, gelangen diese mit erhöhter Wahrscheinlichkeit erneut in die zweite Stichprobe [29].

Zeigen gefangene Individuen eine erhöhte Scheu gegenüber den Fallen, werden diese in der zweiten Stichprobe seltener gefangen. Die Überschneidung zwischen den beiden Stichproben fällt damit geringer aus als unter der Annahme unabhängiger Stichproben. Folglich sinkt die Anzahl der bereits markierten, erneut gefangenen Individuen m, was bei beiden Schätzverfahren zu einer Überschätzung der wahren Populationsgrösse führt [30]. Treten dagegen eine gesteigerte Anziehung zu den Fallen oder eine unzureichende Durchmischung auf, gelangen markierte Individuen häufiger in die zweite Stichprobe, wodurch sich die Überschneidung der Stichproben vergrössert und damit m erhöht, was eine Unterschätzung der tatsächlichen Populationsgrösse zur Folge hat [10].

Zur Veranschaulichung eines solchen Szenarios wird im Folgenden eine Simulation durchgeführt, in welcher die wahre Populationsgrösse $N = 500$ als unbekannt angenommen wird. In jedem Simulationsdurchlauf werden zunächst $M = 100$ Individuen zufällig aus der Population gefangen und markiert (Capture). Im Anschluss erfolgt eine zweite Stichprobe, für deren Auswahl ein Zufallsgenerator verwendet wird, der potenzielle Kandidaten bestimmt. Wurde das vom Zufallsgenerator bestimmte Individuum bereits in der ersten Stichprobe markiert, so wird es mit einer Wahrscheinlichkeit von 100% in die zweite Stichprobe aufgenommen. Handelt es sich dagegen um ein unmarkiertes Individuum, gelangt es lediglich mit einer Wahrscheinlichkeit von 50% in die zweite Stichprobe. Wird ein Individuum nicht aufgenommen oder befindet es sich bereits in der zweiten Stichprobe, wählt der Zufallsgenerator ein neues Individuum aus der Population aus, wobei dieser Prozess so lange wiederholt wird, bis die zweite Stichprobe ebenfalls $n = 100$ Individuen enthält (Recapture). Die Anzahl m der bereits markierten und erneut gefangenen Individuen wird dabei kontinuierlich erfasst. Dieses Vorgehen wird insgesamt $5'000$ mal wiederholt, wobei der Lincoln-Petersen-Schätzer sowie der Chapman-Schätzer für jeden Durchlauf auf Basis der beobachteten Werte M, n und m je einen Schätzwert $\hat{N}$ für die Populationsgrösse liefern. Die Verteilung dieser jeweils $5'000$ Schätzwerte ist in den Boxplots in Abbildung 3 dargestellt.

Die Simulation bestätigt die zu erwartende deutliche Unterschätzung der Populationsgrösse. Im Mittel liefert der Lincoln-Petersen-Schätzer einen Schätzwert von 323.71, während der Chapman-Schätzer durchschnittlich von 318.70 Individuen ausgeht. Diese systematische Verzerrung ist damit zu begründen, dass die Überschneidung der beiden Stichproben aufgrund der erhöhten Fangwahrscheinlichkeit markierter Individuen künstlich vergrössert wird. Dadurch stiegt die Anzahl der Wiederfänge m, die bei beiden Schätzern im Nenner steht, was zu einer Unterschätzung der wahren Populationsgrösse führt. Der Lincoln-

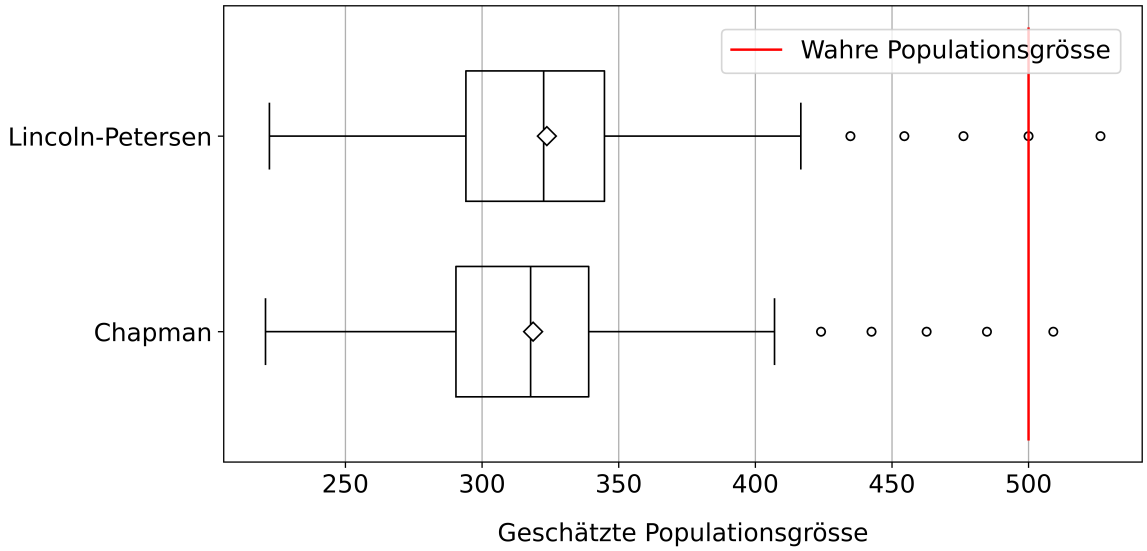

Abbildung 3: Boxplots der Schätzwerte $\hat{N}$ aus $5'000$ Sim. mit abh. Stichproben ($M = n = 100$, $N = 500$).

Petersen-Schätzer liegt zwar etwas näher am wahren Wert, jedoch ist dies lediglich darauf zurückzuführen, dass er im Durchschnitt geringfügig höhere Schätzwerte liefert als der Chapman-Schätzer. Insgesamt wird die Population in diesem Falle um mehr als 45% unterschätzt. Auffällig ist zudem der kleinere Interquartilsabstand, der beim Lincoln-Petersen-Schätzer 51 sowie beim Chapman-Schätzer 49 beträgt und damit deutlich unter denjenigen der Simulation 1 mit unabhängigen Stichproben liegt (vgl. Abschnitt 5.1.1). Die erhöhte Präzision entsteht, weil durch die Abhängigkeit der Stichproben die Werte von m weniger zufälligen Schwankungen unterliegen und stärker stabilisiert werden. Jedoch stellt diese höhere Präzision keine Verbesserung der Schätzung dar, da die Schätzwerte trotz geringerer Streuung deutlich verzerrt sind.

Verletzung der Annahme keines Markierungsverlusts

Während des Untersuchungszeitraumes können beispielsweise in einer Tierpopulation Markierungen verloren gehen, indem Bänder abfallen oder Fellmarkierungen verblassen [8]. Nimmt man an, dass ein Anteil $\lambda \in]0, 1[$ der ursprünglich markierten Individuen M seine Markierung verliert, verbleiben zum Zeitpunkt der zweiten Stichprobe noch $M^* = M(1 - \lambda)$ sichtbar markierte Individuen in der Population [32]. Da der Anteil der wiedergefangenen Individuen in der zweiten Stichprobe gemäss der Herleitung des Lincoln-Petersen-Schätzers über Anteilsverhältnisse ungefähr dem Anteil markierter Individuen in der Gesamtpopulation entspricht (vgl. Unterkapitel 3.1), sinkt die Anzahl der bereits markierten, erneut gefangenen Individuen entsprechend auf $m^* = m(1 - \lambda)$ [26]. Wird m^* anstelle von m im Lincoln-Petersen-Schätzer eingesetzt, ergibt sich

$$\hat{N}^*_{LP} = \frac{Mn}{m^*} = \frac{Mn}{m(1-\lambda)} = \frac{\hat{N}_{LP}}{(1-\lambda)} \, . \tag{5.1}$$

Da der Term $1-\lambda$ im Nenner des Ausdrucks 5.1 stets kleiner als 1 ist, wird die wahre Populationsgrösse N somit systematisch überschätzt, wobei die Verzerrung mit steigendem λ kontinuierlich zunimmt [10]. Ein Markierungsverlust von 10% $(\lambda = 0.1)$ führt bereits zu einer Überschätzung um rund 11.1%. Betrachtet

man analog den Chapman-Schätzer, erhält man für diesen entsprechend

$$\hat{N}_C^* = \frac{(M+1)(n+1)}{(m^*+1)} - 1 = \frac{(M+1)(n+1)}{(m(1-\lambda)+1)} - 1 \,. \tag{5.2}$$

Da hier der Nenner zusätzlich um 1 erhöht wird, fällt die Verzerrung bei kleinen Wiederfangzahlen etwas geringer aus als beim Lincoln-Petersen-Schätzer. Dennoch führt auch bei diesem Schätzer ein Markierungsverlust zu einer systematischen Überschätzung der tatsächlichen Populationsgrösse. Simulationsergebnisse zeigen, dass bei $\lambda = 0.1$ die Überschätzung des Chapman-Schätzers im Durchschnitt rund 8.8% beträgt.

Kapitel 6

Anwendungsbeispiel: Populationsschätzung erdnaher Asteroiden

Dieses Kapitel widmet sich einem praktischen Beispiel: der Populationsschätzung erdnaher Asteroiden. Nach einer Einführung in die Thematik folgt die Darstellung der Datengrundlage und der Methodik. Anschliessend werden die beiden in Kapitel 3 und 4 vorgestellten Schätzverfahren auf die Daten angewendet, wobei die erhaltenen Resultate im Anschluss analysiert und interpretiert werden.

6.1 Definition und Relevanz erdnaher Asteroiden

Asteroiden sind unregelmässig geformte, meist aus verschiedenen Gesteinen und Metallen bestehende Kleinkörper, die sich auf Umlaufbahnen um die Sonne bewegen. Sie erfüllen jedoch nicht die Kriterien eines Zwergplaneten gemäss der Internationalen Astronomischen Union (IAU), da sie im Gegensatz zu diesen nicht über eine ausreichende Masse verfügen, um durch ihre Eigengravitation eine hydrostatisch ausgeglichene, annähernd sphärische Form auszubilden [34]. Stattdessen sind Asteroiden, ebenso wie Kometen, Überreste des frühen Sonnensystems, die entscheidende Informationen zur Rekonstruktion seiner Entstehung vor etwa 4.6 Milliarden Jahren in sich tragen [33]. Im Unterschied zu Kometen sind Asteroiden jedoch inerte Körper. Selbst wenn sie sich der Sonne nähern, bilden sie keine sichtbare Koma aus, da sie nur geringe Mengen flüchtiger Bestandteile enthalten, die in Sonnennähe verdampfen könnten [43]. In ihrer Grösse variieren Asteroiden erheblich, von einigen Dutzend Metern bis hin zu Durchmessern von mehreren Kilometern [43]. Kleinere Fragmente mit einem Durchmesser unter etwa zehn Metern werden aber in der Regel als Meteoroiden bezeichnet [34].

Die Mehrzahl der bislang fast 1.5 Millionen bekannten Asteroiden umkreist die Sonne im sogenannten Asteroidengürtel, einem Bereich zwischen den Umlaufbahnen der Planeten Mars und Jupiter. Gravitative Störungen des Gasriesen Jupiter verhinderten dort die Akkretion planetaren Materials und damit die Bildung eines weiteren Planeten [39]. Einige Asteroiden bewegen sich jedoch auf Bahnen, die sie der Sonne viel näher bringen als üblich. Nähert sich ein Asteroid der Sonne bis auf 1.3 Astronomische Einheiten, was

rund 194 Millionen Kilometern entspricht, so wird dieser als erdnaher Asteroid (engl. *Near-Earth Asteroid*, kurz NEA) klassifiziert. Eine Astronomische Einheit, abgekürzt im Deutschen durch AE, bezeichnet dabei den mittleren Abstand zwischen Erde und Sonne, der etwa 150 Millionen Kilometer beträgt [43].

Erdnahe Asteroiden werden anhand ihrer orbitalen Eigenschaften, basierend auf der Aphel-Distanz (Q), der Perihel-Distanz (q) sowie der grossen Halbachse (a), in vier Hauptkategorien unterteilt [33]. Während der Aphel den Punkt der Umlaufbahn eines Himmelskörpers bezeichnet, der am weitesten von der Sonne entfernt ist, markiert der Perihel den Punkt, an dem der Himmelskörper der Sonne am nächsten kommt [36]. Diese beiden Entfernungen sowie die grosse Halbachse, die dem mittleren Abstand der Erde zur Sonne entspricht, sind entscheidend, da sie die potenzielle Nähe eines Asteroiden zur Erde bestimmen [34]. Die Klassifikation in die vier Kategorien *Amor*, *Apollo*, *Aten* und *Atira*, benannt nach den erdnahen Asteroiden *1221 Amor*, *1862 Apollo*, *2062 Aten* und *163693 Atira*, erfolgt gemäss den in Tabelle 4 dargestellten Definitionen [45].

Amor	Apollo	Aten	Atira
$a > 1.0\,\mathrm{AE}$	$a > 1.0\,\mathrm{AE}$	$a < 1.0\,\mathrm{AE}$	$a < 1.0\,\mathrm{AE}$
$1.017\,\mathrm{AE} < q < 1.3\,\mathrm{AE}$	$q < 1.017\,\mathrm{AE}$	$Q > 0.983\,\mathrm{AE}$	$Q < 0.983\,\mathrm{AE}$

Tabelle 4: Definitionen der NEA-Kategorien [45].

Die Werte $1.017\,\mathrm{AE}$ und $0.983\,\mathrm{AE}$ entsprechen der Aphel- beziehungsweise der Perihel-Distanz der Erde zur Sonne, jeweils auf drei Nachkommastellen gerundet [34].

Amor-Asteroiden nähern sich der Erdbahn an, verlaufen jedoch immer ausserhalb dieser, aber noch innerhalb der Marsbahn. Ihre Perihel-Distanz liegt zwischen dem Aphel der Erde und dem Perihel des Mars. Daher stellen sie für die Erde keine unmittelbare Kollisionsgefahr dar, sind aber dennoch von Interesse für die Überwachung, falls sich ihre Umlaufbahnen zukünftig ändern sollten [34]. Apollo- und Aten-Asteroiden kreuzen dagegen die Erdbahn. Aten-Asteroiden, die häufigste Gruppe, haben kleinere Halbachsen als die Erde, während Apollo-Asteroiden grössere Halbachsen aufweisen. Apollo-Asteroiden kommen in etwa gleich häufig vor wie Amor-Asteroiden [36]. Die kleinste Gruppe bilden die Atira-Asteroiden, die auch als Inner Earth Objects (IEOs) bezeichnet werden. Ihre Umlaufbahnen verlaufen vollständig innerhalb der Erdbahn, weshalb sie von der Erde aus immer in Nähe der Sonne stehen, was sie schwer zu entdecken und beobachten macht [45].

Die Untersuchung erdnaher Asteroiden ist sowohl aus wissenschaftlicher als auch aus wirtschaftlicher sowie sicherheitstechnischer Perspektive von grossem Interesse. Aus wissenschaftlicher Sicht stellen Asteroiden Überreste der frühen Entwicklungsphasen des Sonnensystems dar und enthalten weitgehend unverarbeitetes Material, das auf der Erde infolge geologischer Prozesse wie Akkretion, Plattentektonik, Vulkanismus oder Metamorphose verloren gegangen ist. Somit liefern sie entscheidende Informationen zur Rekonstruktion der Entstehung des Sonnensystems [33]. Auch aus einer wirtschaftlichen Sicht betrachtet gewinnen NEAs zunehmend an Bedeutung. Aufgrund ihrer relativen Nähe zur Erde gelten sie als

potenziell gut erreichbare Quellen für die Gewinnung wertvoller Ressourcen, insbesondere im Hinblick auf zukünftige interplanetare Missionen sowie die langfristige industrielle Nutzung des Weltraumes [43]. Die begrenzten Vorkommen seltener Erdmetalle und halbleitender Elemente auf der Erde könnten durch die Rohstoffreserven von Asteroiden ergänzt oder sogar ersetzt werden. So machen Erdmetalle einen beträchtlichen Teil der Masse von NEAs aus. Schätzungen zufolge könnte ein einziger metallischer Asteroid mit einem Durchmesser von nur einem Kilometer bereits eine doppelt so grosse Menge an Platinmetallen enthalten wie sie bislang auf der Erde gefördert wurde [14]. Neben den wirtschaftlichen Chancen bieten erdnahe Asteroiden aber auch sicherheitstechnische Herausforderungen. Selbst kleinere NEAs können aufgrund ihrer hohen Relativgeschwindigkeit im Falle einer Kollision mit der Erde verheerende Folgen haben. Der weithin akzeptierte Zusammenhang zwischen dem Einschlag eines Asteroiden oder Kometen mit einem Durchmesser von etwa zehn Kilometern und dem Massensterben von Lebewesen an der Kreide-Tertiär-Grenze verdeutlicht das enorme Zerstörungspotenzial solcher Ereignisse. Die dabei freigesetzte Energie übersteigt jene aller bekannten Naturphänomene wie Erdbeben, Vulkaneruptionen oder Tsunamis bei Weitem. Aus diesem Grunde ist eine kontinuierliche Beobachtung und Katalogisierung erdnaher Asteroiden von zentraler Bedeutung, um potenzielle Kollisionsgefahren frühzeitig zu erkennen und gegebenenfalls präventive Massnahmen zum Schutz der Erde ergreifen zu können [33].

6.2 Datengrundlage und methodisches Vorgehen

Die zur Populationsschätzung erdnaher Asteroiden verwendeten Daten stammen von der Near-Earth Object Dynamics Site (NEODyS) [44]. Dabei handelt es sich um eine öffentlich zugängliche Online-Datenbank, die Beobachtungsdaten sowie Informationen zu Orbitalparametern erdnaher Asteroiden enthält. Entwickelt wurde NEODyS im Jahre 1999 von einer Forschungsgruppe des Fachbereichs Mathematik der Universität Pisa und wird seither kontinuierlich betrieben und täglich aktualisiert. Seit 2011 ist auch die Europäische Weltraumorganisation ESA an dem Projekt beteiligt. Neben der wissenschaftlichen Zusammenarbeit trägt diese zur Finanzierung bei [44].

Ein wesentlicher Vorteil, alle Daten aus einer einzigen Quelle zu beziehen, liegt in der einheitlichen Namensgebung der erfassten NEAs. Da die Benennung von Asteroiden in einem zweistufigen Prozess erfolgt, kann es bei der Verwendung unterschiedlicher Datenbanken zu Inkonsistenzen kommen: Während ein Objekt in der einen Datenbank noch unter seiner provisorischen Bezeichnung, einer alphanumerischen Kombination, registriert ist, wird es in einer anderen möglicherweise bereits unter seinem endgültigen Namen aufgeführt [45]. Um solche Verwechslungen auszuschliessen, wird in dieser Arbeit ausschliesslich auf Daten von NEODyS zurückgegriffen.

Da nahezu täglich erdnahe Asteroiden in der Datenbank ergänzt werden, wird der 1. Juni 2025 als Stichtag für die Datenbeschaffung festgelegt. Zu diesem Zeitpunkt sind Beobachtungsdaten von insgesamt 2616 Observatorien verfügbar. Jedes Observatorium hat dabei in einer eigenen Tabelle festgehalten, welche erdnahen Asteroiden es bislang zu welchen Zeitpunkten beobachtet hat. Mithilfe eines Python-

Programms, das im Anhang C zu finden ist, werden aus diesen Tabellen die Namen aller erfassten NEAs extrahiert. Zumal in dieser Arbeit diskrete Capture-Recapture-Verfahren verwendet werden, werden Mehrfachsichtungen desselben Asteroiden durch ein und dasselbe Observatorium jeweils zusammengefasst. Das Ziel besteht darin, für jedes Observatorium die Anzahl unterschiedlicher, mindestens einmal erfasster NEAs zu bestimmen. Nach dieser Datendeduplikation ist ein Vergleich der Observatorien hinsichtlich der Anzahl beobachteter erdnaher Asteroiden möglich. Da sowohl die Genauigkeit als auch die Präzision mit zunehmender Stichprobengrösse steigt (vgl. Unterkapitel 5.1), werden für die weiterführende Analyse nur diejenigen Observatorien, welche die meisten unterschiedlichen NEAs erfasst haben, berücksichtigt. Die drei Observatorien mit den umfangreichsten Beobachtungsdaten befinden sich alle in den USA und sind

- das Mount Lemmon Survey in Arizona (G96) mit 24'194 erfassten NEAs,
- das Pan-STARRS1 auf Hawaii (F51) mit 20'160 erfassten NEAs sowie
- das Astronomical Research Observatory in Illinois (H21) mit 19'017 erfassten NEAs.

Im weiteren Verlauf der Arbeit werden diese Observatorien der Übersichtlichkeit halber entsprechend ihrer offiziellen Sternwarten-Codes als G96, F51 und H21 bezeichnet.

In einem nächsten Schritt werden mit einem weiteren Python-Programm, dessen Code ebenfalls im Anhang C aufgeführt ist, die spezifischen Daten der einzelnen NEAs gesammelt, die von mindestens einem der drei ausgewählten Observatorien gesichtet wurden. Neben der Information, ob das jeweilige Objekt als potenziell gefährlich eingestuft wird (Einschlagswahrscheinlichkeit bis zum Jahre 2100 grösser als 10^{-11} [44]), werden auch die absolute Helligkeit sowie verschiedene Orbitalparameter, wie beispielsweise die grosse Halbachse oder die Aphel- und Periheldistanz, erfasst. Die gesammelten Daten erlauben es, eine Kategorisierung der NEAs gemäss den in Tabelle 4 definierten Kriterien vorzunehmen. Darüber hinaus wird analysiert, welche dieser Objekte von mehreren der drei Observatorien registriert wurden. Alle gesammelten Daten werden schliesslich in einer Excel-Datei gespeichert, wobei für jedes der drei Observatorien sowie für die jeweiligen Überschneidungen separate Tabellen erstellt werden. Eine Übersicht über die erfassten Daten bietet die Tabelle 5, in der die Anzahl der Objekte pro NEA-Kategorie dargestellt wird, die von den Observatorien G96, F51 und H21 sowie von jeweils zwei dieser Observatorien beobachtet wurden. Ausserdem enthält die Tabelle die Gesamtzahlen für jedes Observatorium und die entsprechenden Überschneidungen. In den Zellen ist zudem in Klammern vermerkt, wie viele der jeweiligen NEAs als potenziell gefährlich eingestuft werden.

	|G96|	|F51|	|H21|	|G96 ∩ F51|	|G96 ∩ H21|	|F51 ∩ H21|
Amor	8'610 (25)	8'743 (19)	7'586 (13)	5'505 (4)	5'289 (4)	5'383 (9)
Apollo	13'700 (995)	10'136 (511)	10'049 (503)	6'161 (210)	6'874 (316)	5'510 (177)
Aten	1'869 (143)	1'268 (69)	1'377 (89)	733 (24)	916 (57)	669 (26)
Atira	15 (0)	13 (0)	5 (0)	7 (0)	5 (0)	1 (0)
Total	24'194 (1'163)	20'160 (599)	19'017 (605)	12'406 (238)	13'084 (377)	11'563 (212)

Tabelle 5: Datenübersicht über die ausgewählten Observatorien und deren Überschneidungen, Anzahl der davon als potenziell gefährlich eingestuften NEAs jeweils in Klammern angegeben.

6.3 Anwendung der Schätzverfahren

Auf die in Tabelle 5 dargestellten Daten werden im Folgenden der Lincoln-Petersen-Schätzer (Kapitel 3) sowie der Chapman-Schätzer (Kapitel 4) angewendet. Beide Verfahren liefern für jedes Paar von Observatorien und deren Schnittmengen jeweils einen Schätzwert der Gesamtpopulation erdnaher Asteroiden sowie der Anzahl der Objekte, die eine vollständige Risiko-Liste umfassen würde. Zudem werden zu allen Schätzwerten die entsprechenden Standardabweichungen auf Grundlage der in den Abschnitten 3.3.2 und 4.2.2 eingeführten Approximationen berechnet. Dabei wird in den jeweiligen Formeln N durch $\hat{N}_{LP}$ beziehungsweise $\hat{N}_C$ ersetzt. Die auf ganze Zahlen gerundeten Ergebnisse sind in den Tabellen 6 und 7 dargestellt.

Datengrundlage	Lincoln-Petersen-Schätzer		Chapman-Schätzer	
	$\hat{N}_{LP}$	$\mathrm{SD}(\hat{N}_{LP})$	$\hat{N}_C$	$\mathrm{SD}(\hat{N}_C)$
G96 und F51	39'316	153	39'315	352
G96 und H21	35'165	116	35'164	307
F51 und H21	33'156	126	33'156	308

Tabelle 6: Populationsschätzungen erdnaher Asteroiden einschliesslich ihrer Standardabweichungen.

Datengrundlage	Lincoln-Petersen-Schätzer		Chapman-Schätzer	
	$\hat{N}_{LP}$	$\mathrm{SD}(\hat{N}_{LP})$	$\hat{N}_C$	$\mathrm{SD}(\hat{N}_C)$
G96 und F51	2'927	131	2'921	189
G96 und H21	1'866	48	1'865	96
F51 und H21	1'709	76	1'706	118

Tabelle 7: Schätzungen der Grösse der Risiko-Liste einschliesslich ihrer Standardabweichungen.

Wie aus den Tabellen 6 und 7 hervorgeht, liefern der Lincoln-Petersen-Schätzer und der Chapman-Schätzer für gleiche Datengrundlagen nahezu identische Schätzwerte. Sowohl bei der Schätzung der Gesamtpopulation als auch der Grösse der Risiko-Liste liegen die absoluten Abweichungen zwischen den beiden Verfahren im einstelligen Bereich und sind in Anbetracht der Grössenordnung der geschätzten Werte vernachlässigbar. Abhängig davon, welche Observatorien als Datengrundlage verwendet werden, treten jedoch leichte Schwankungen in den Schätzwerten auf. Im Mittel schätzt der Lincoln-Petersen-Schätzer die Gesamtpopulation erdnaher Asteroiden auf $\hat{N}_{LP,\,pop} \approx 35'879$, wovon schätzungsweise $\hat{N}_{LP,\,risk} \approx 2'167$ Objekte Teil der Risiko-Liste sind. Der Chapman-Schätzer geht im Durchschnitt von $\hat{N}_{C,\,pop} \approx 35'878$ NEAs aus, wobei er die vollständige Risiko-Liste auf $\hat{N}_{C,\,risk} \approx 2'164$ Objekte schätzt.

Die approximativen Standardabweichungen des Chapman-Schätzers sind systematisch grösser als jene des Lincoln-Petersen-Schätzers. Bei der Schätzung der Gesamtpopulation erdnaher Asteroiden liegen sie mehr als doppelt so hoch, betragen jedoch nichtsdestotrotz weniger als 1% des jeweiligen Schätzwerts. Für die Schätzung der Anzahl Objekte auf der vollständigen Risiko-Liste ist der relative Anteil der Standardabweichung bei beiden Schätzverfahren leicht höher. Während die Standardabweichung des

Chapman-Schätzers bei jeweils etwas über 5% des entsprechenden Schätzwerts liegt, überschreitet die Standardabweichung des Lincoln-Petersen-Schätzers diese Grenze in keinem der drei betrachteten Fälle.

6.4 Analyse und Interpretation der Ergebnisse

Die berechneten Schätzwerte aus dem Unterkapitel 6.3 zeigen insgesamt eine hohe Konsistenz sowie relativ niedrige Standardabweichungen, was Merkmale einer zuverlässigen Schätzung sind. Jedoch bleibt es zu überprüfen, ob die Modellannahmen aus dem Unterkapitel 2.2, die den angewendeten Schätzverfahren zugrunde liegen, tatsächlich erfüllt sind. Im Folgenden werden diese getroffenen Annahmen systematisch überprüft.

Annahme einer geschlossenen Population

Grundsätzlich können erdnahe Asteroiden mit der Erde kollidieren, jedoch sind solche Einschläge relativ selten. Statistisch gesehen trifft etwa alle 10 Jahre ein NEA mit einem Durchmesser von $10\,\mathrm{m}$ auf die Erde. Bei Objekten von $50\,\mathrm{m}$ Durchmesser ist im Mittel einmal in 1'000 Jahren ein Einschlag zu erwarten, während eine Kollision eines NEAs von $150\,\mathrm{m}$ Durchmesser nur noch etwa alle 10'000 Jahre auftritt [39]. Allgemein nimmt die Häufigkeit von Einschlägen mit zunehmender Grösse der erdnahen Asteroiden ab. Dies ist darauf zurückzuführen, dass kleinere NEAs deutlich häufiger auftreten, während grössere Objekte, insbesondere solche im Kilometerbereich, vergleichsweise selten sind [43]. Selbst wenn es aber gelegentlich zu einem Einschlag kommt, bleibt die Zahl dieser Ereignisse im Verhältnis zur gesamten bekannten Population erdnaher Asteroiden vernachlässigbar. Bis Oktober 2025 sind etwas mehr als 39'700 NEAs katalogisiert. Angesichts dieser grossen Zahl kann die Annahme, dass es sich bei den erdnahen Asteroiden um eine geschlossene Population handelt, als zutreffend angesehen werden.

Annahme homogener Fangwahrscheinlichkeiten

Erdnahe Asteroiden stellen eine ausgesprochen heterogene Objektgruppe dar. Allein in ihrer Grösse variieren NEAs erheblich. Ihre Durchmesser reichen von einigen Dutzend Metern bis hin zu mehreren Kilometern, wobei die Anzahl der Objekte mit zunehmender Grösse stark abnimmt [43]. Ebenso sind die Orbitalparameter, wie beispielsweise die grosse Halbachse, die Exzentrität oder Periheldistanz, für alle erdnahen Asteroiden unterschiedlich, selbst innerhalb derselben NEA-Kategorie [2]. Auch die absolute Helligkeit H, angegeben in Magnituden (mag), weist eine erhebliche Spannbreite auf. Sie beschreibt die scheinbare Helligkeit eines Objekts, wenn sich dieses sowohl $1\,\mathrm{AE}$ von der Sonne als auch $1\,\mathrm{AE}$ vom Beobachter entfernt in Opposition befindet. Je kleiner der Wert von H, desto heller erscheint das Objekt [12]. Diese Vielfalt individueller Eigenschaften führt dazu, dass nicht alle erdnahen Asteroiden mit derselben Wahrscheinlichkeit erfasst werden. Die Annahme homogener Fangwahrscheinlichkeiten ist daher verletzt.

Auf Grundlage der individuell erfassten Informationen zu den NEAs, darunter verschiedene Orbitalparameter sowie die absolute Helligkeit, lässt sich die Heterogenität der Fangwahrscheinlichkeit belegen. Angaben zu den Durchmessern liegen in der NEODyS-Datenbank keine vor, da diese besonders für kleinere

NEAs meist unbekannt sind oder lediglich indirekt über die absolute Helligkeit und angenommene Albedowerte abgeschätzt werden können [45]. Eine Auswertung der verfügbaren Parameter zeigt, dass sich die Unterschiede in den Fangwahrscheinlichkeiten insbesondere in der absoluten Helligkeit widerspiegeln. In Abbildung 4 zeigt das linke Histogramm die Verteilung der absoluten Helligkeiten der von den einzelnen Observatorien erfassten NEAs in absoluten Zahlen, während rechts die Verteilung der Objekte dargestellt ist, die von jeweils zwei Observatorien registriert wurden.

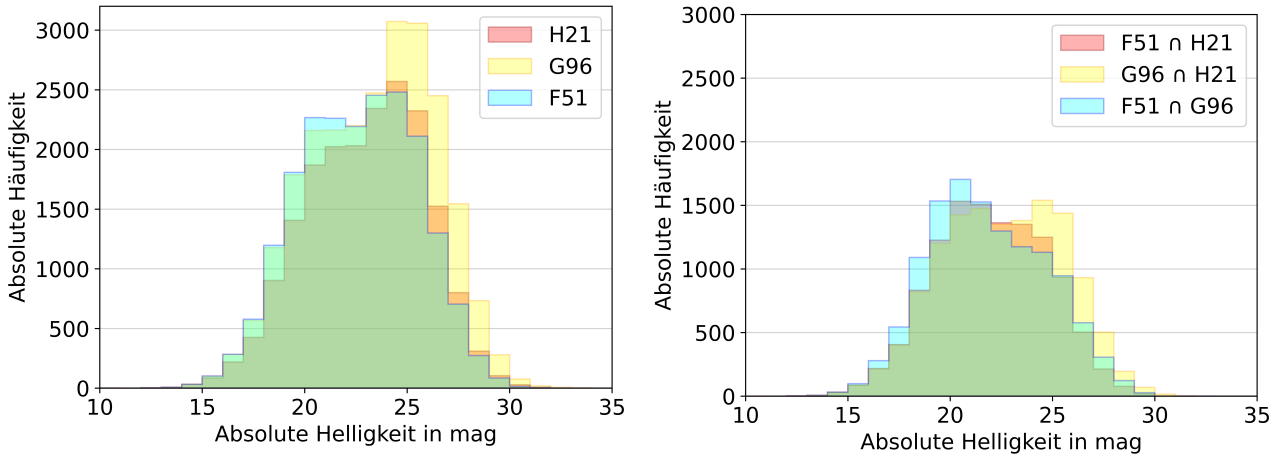

Abbildung 4: Histogramme der Verteilungen der absoluten Helligkeiten der NEAs.
Links: erfasst durch die einzelnen Observatorien. *Rechts*: erfasst durch jeweils zwei Observatorien.

Bereits ein Vergleich dieser beiden Histogramme legt nahe, dass die wiedergefangenen erdnahen Asteroiden im Mittel eine etwas geringere absolute Helligkeit aufweisen und somit tendenziell heller sind. Zur Überprüfung dieser Hypothese werden die NEAs nach Helligkeitsintervallen gruppiert, wobei anschliessend in allen Intervallen für jedes Observatorium untersucht wird, welcher prozentuale Anteil der dort erfassten Objekte auch von den beiden anderen Observatorien registriert wurde. So lassen sich alle sechs möglichen Vergleichspaare zwischen den drei Observatorien analysieren. Die prozentualen Anteile entsprechen dabei approximierten Recapture-Wahrscheinlichkeiten. Unter der Annahme zeitlich konstanter Beobachtungsbedingungen entsprechen diese zudem näherungsweise der Capture-Wahrscheinlichkeit. Die resultierenden Werte sind in Abbildung 5 dargestellt, wobei die schwarze Linie die Mittelwerte der approximierten Fangwahrscheinlichkeiten markiert. Wie deutlich zu erkennen ist, nimmt die Fangwahrscheinlichkeit mit zunehmender absoluter Helligkeit stetig ab. Lichtschwächere NEAs werden somit mit deutlich geringerer Wahrscheinlichkeit entdeckt, was mit den Erwartungen übereinstimmt. Jedoch ist zu beachten, dass die dargestellten Fangwahrscheinlichkeiten positiv verzerrt sind, da diese ausschliesslich auf Objekten basieren, die

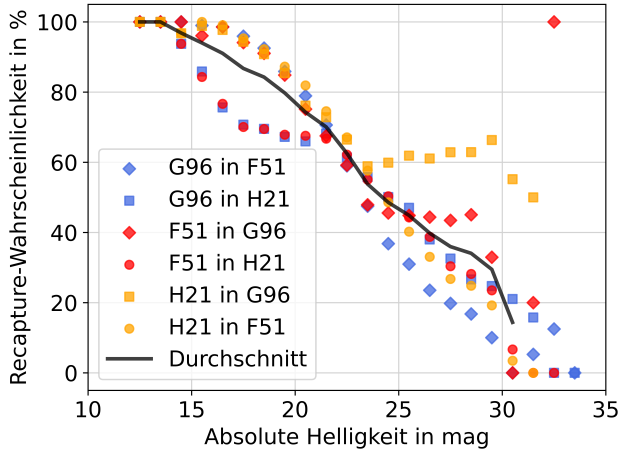

Abbildung 5: Modellierung der Fangwahrscheinlichkeiten anhand doppelt erfasster NEAs.

mindestens einmal entdeckt wurden. Nicht erfasste NEAs bleiben dabei unberücksichtigt, wodurch die tatsächlichen Fangwahrscheinlichkeiten vermutlich überschätzt werden.

Annahme unabhängiger Stichproben

Da es sich bei erdnahen Asteroiden um unbelebte Objekte handelt, kann die Erfassung dieser keine Verhaltensänderungen hervorrufen, wie sie in klassischen Capture-Recapture-Studien mit Tieren auftreten können [26]. Jedoch besteht das Problem darin, dass es zwischen den Stichproben nicht zu einer vollständigen Durchmischung kommt. Die Datenerhebung erfolgt stets durch die gleichen drei Observatorien, die alle in den USA lokalisiert sind [44]. Dadurch überschneiden sich ihre beobachteten Himmelsausschnitte erheblich, sodass NEAs in diesen Bereichen häufig von mehreren Observatorien gemeinsam erfasst werden, während Objekte ausserhalb dieser Regionen unberücksichtigt bleiben. Des Weiteren arbeiten alle drei Observatorien im sichtbaren Wellenlängenbereich des elektromagnetischen Spektrums. Das Observatorium F51 kann mit einer maximalen Wellenlänge von rund $963\,\mathrm{nm}$ zwar bis ins nahe Infrarot vordringen, jedoch operiert auch dieses primär im sichtbaren Wellenlängenbereich [5]. Somit bestehen sowohl hinsichtlich der beobachteten Himmelsregion als auch der genutzten Wellenlängenbereiche Ähnlichkeiten zwischen den Datensätzen, was zu einer deutlichen positiven Abhängigkeit der Stichproben führt. Die Überschneidungen der erfassten NEAs fällt damit viel grösser aus als es bei statistisch unabhängigen Stichproben der Fall wäre.

Die Abhängigkeit zwischen den Stichproben lässt sich auch statistisch untersuchen, indem die erfassten Daten der erdnahen Asteroiden mittels eines Chi-Quadrat-Tests analysiert werden. Die Grundidee besteht darin, die tatsächlich beobachteten Häufigkeiten mit denjenigen zu vergleichen, die man unter der Annahme vollständiger Unabhängigkeit erwarten würde. Je stärker diese voneinander abweichen, desto unwahrscheinlicher ist eine Unabhängigkeit [40]. Im Folgenden wird der Chi-Quadrat-Test exemplarisch verwendet, um die Daten der Observatorien G96 und F51 auf Abhängigkeit zu überprüfen. Dazu werden drei disjunkte Mengen gebildet: NEAs, die ausschliesslich von G96 erfasst wurden, solche, die ausschliesslich von F51 registriert wurden, sowie jene, die von beiden Observatorien gesichtet wurden. Alle Objekte dieser drei Mengen werden anschliessend gemäss den in Tabelle 4 definierten Kriterien in die vier NEA-Kategorien eingeteilt. Zudem wird für jede Menge die Spaltensumme, für jede NEA-Kategorie die Zeilensumme sowie die Gesamtsumme aller in den Test einbezogenen NEAs berechnet. Die entsprechenden Werte sind in Tabelle 8 dargestellt.

	$\lvert G96 \setminus (F51 \cup H21)\rvert$	$\lvert F51 \setminus (G96 \cup H21)\rvert$	$\lvert G96 \cap F51\rvert$	Total
Amor	3’105	3’238	5’505	11’848
Apollo	7’539	3’975	6’161	17’675
Aten	1’136	535	733	2’404
Atira	8	6	7	21
Total	11’788	7’754	12’406	31’948

Tabelle 8: Kontingenztabelle der gesammelten Daten.

Unter der Nullhypothese H_0 geht man davon aus, dass die Verteilung über die vier NEA-Kategorien für alle drei Mengen gleich ist, sofern die Daten der Observatorien G96 und F51 unabhängig voneinander erhoben wurden [28]. Auf dieser Grundlage wird in einem nächsten Schritt für jede Zelle (i, j) der Kontingenztabelle 8 die jeweilige erwartete Häufigkeiten $E_{i,j}$ mit der folgenden Formel bestimmt [37]:

$$E_{i,j} = \frac{(\text{Zeilensumme}_i) \cdot (\text{Spaltensumme}_j)}{\text{Gesamtsumme aller Objekte}} \quad \text{mit} \quad i \in \{1,2,3,4\} \quad \text{und} \quad j \in \{1,2,3\}. \tag{6.1}$$

Die auf diese Weise berechneten erwarteten Häufigkeiten, jeweils auf zwei Dezimalstellen gerundet, finden sich in Tabelle 9.

	|G96 \ (F51 ∪ H21)|	|F51 \ (G96 ∪ H21)|	|G96 ∩ F51|
Amor	4'371.61	2'875.59	4'600.80
Apollo	6'521.63	4'289.84	6'863.53
Aten	887.01	583.47	933.52
Atira	7.75	5.10	8.15

Tabelle 9: Erwartete Häufigkeiten unter der Annahme der Unabhängigkeit.

Anschliessend wird untersucht, wie stark die tatsächlichen Beobachtungen von den erwarteten Häufigkeiten abweichen. Dazu berechnet man den sogenannten Chi-Quadrat-Wert χ^2, der folgendermassen definiert ist [40]:

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \, . \tag{6.2}$$

Dabei bezeichnet $O_{i,j}$ die beobachtete Häufigkeit der Zelle (i, j), während $E_{i,j}$ der erwarteten Häufigkeit derselben entspricht [40]. Je stärker die Abhängigkeit ist, desto grösser wird der resultierende Chi-Quadrat-Wert [28], der in diesem Falle $\chi^2_{beobachtet} \approx 961.39$ entspricht. Unter der Nullhypothese H_0 folgt dieser Wert einer Chi-Quadrat-Verteilung, deren Form von der Anzahl der Freiheitsgrade abhängt. Die Freiheitsgrade lassen sich durch $(\text{Anzahl Zeilen} - 1) \cdot (\text{Anzahl Spalten} - 1) = (4-1) \cdot (3-1) = 3 \cdot 2 = 6$ bestimmen [40]. Der p-Wert, der durch $p = P(\chi^2_6 \geq \chi^2_{beobachtet})$ gegeben ist, gibt nun an, wie wahrscheinlich es ist, unter der Annahme der Unabhängigkeit einen mindestens so grossen Chi-Quadrat-Wert zu erhalten, wenn die Nullhypothese H_0 wahr ist [37]. In diesem Falle ergibt sich ein p-Wert, der praktisch gleich Null ist und somit deutlich unter den gebräuchlichen Signifikanzniveaus wie 0.05, 0.01 oder 0.001 liegt. Dies bedeutet, dass die beobachtete Abweichung äusserst unwahrscheinlich ist, wenn tatsächlich eine Unabhängigkeit bestünde. Die Nullhypothese H_0 muss daher klar verworfen werden, da statistisch gesehen eine signifikante Abhängigkeit zwischen den Beobachtungsdaten der Observatorien G96 und F51 besteht [37]. Mit derselben Methode lässt sich auch zwischen den Observatorien G96 und H21 sowie zwischen F51 und H21 eine signifikante Abhängigkeit zeigen.

Annahme keines Markierungsverlusts

Obwohl die erfassten erdnahen Asteroiden nicht mit physischen Markierungen wie Bändchen oder Chips versehen werden können, wie es in klassischen Capture-Recapture-Studien bei Tieren üblich ist [26], erfolgt ihre Markierung unmittelbar nach der Entdeckung durch eine systematische Katalogisierung. Jeder entdeckte erdnahe Asteroid wird einheitlich erfasst und erhält eine Identifikation, die zunächst in Form einer provisorischen alphanumerischen Bezeichnung vergeben wird, bevor ihm ein endgültiger Name zugewiesen wird [45]. Da alle Daten aus derselben Datenbank stammen, ist gewährleistet, dass sowohl vorläufige als auch endgültige Namen einheitlich verwendet werden. Zudem werden sämtliche aufgezeichneten Orbitalparameter sowie andere spezifische Daten des Objekts, wie die absolute Helligkeit, festgehalten und in der öffentlich zugänglichen NEODyS-Datenbank gespeichert [44]. Diese Vorgehensweise ermöglicht eine Reidentifikation bereits bekannter Asteroiden. Die Modellannahme einer fehlerfreien Markierung kann in diesem Falle somit als gegeben betrachtet werden.

Fazit und Vergleich mit Schätzwerten aus der Literatur

Von den im Unterkapitel 2.2 beschriebenen Modellannahmen werden im Anwendungsbeispiel zwei deutlich verletzt: die Annahme homogener Fangwahrscheinlichkeiten sowie die Annahme unabhängiger Stichproben. Eine Verletzung dieser Annahmen führt sowohl beim Lincoln-Petersen-Schätzer als auch beim Chapman-Schätzer zu einer systematischen Unterschätzung (vgl. Unterkapitel 5.2). Daher liegt es nahe, dass die tatsächliche Anzahl erdnaher Asteroiden deutlich höher liegt und auch die vollständige Risiko-Liste wesentlich mehr NEAs umfasst, als in Abschnitt 6.3 geschätzt wird.

Bis Oktober 2025 sind etwas mehr als $39'700$ erdnahe Asteroiden katalogisiert, von denen rund $2'100$ auf der Risiko-Liste stehen [44]. Vergleicht man diese Zahlen mit den in Abschnitt 6.3 berechneten Schätzwerten, die im Mittel auf Hunderter gerundet $39'900$ NEAs für die Gesamtpopulation sowie $2'200$ NEAs für die vollständige Risiko-Liste ergeben, so zeigt sich, dass diese in derselben Grössenordnung liegen wie die bis dahin bekannten Werte. Jedoch ist die Erfassung der NEAs noch längst nicht vollständig, was sich unter anderem daran zeigt, dass nahezu täglich neue erdnahe Asteroiden entdeckt werden [44]. Die tatsächliche Anzahl dürfte somit deutlich höher liegen. Dies bestätigt die theoretische Erwartung, dass die beiden angewandten Schätzverfahren auf Grundlage der verwendeten Daten die wahre Anzahl erdnaher Asteroiden systematisch unterschätzen.

Auch Ergebnisse aus der Fachliteratur, die auf deutlich komplexeren, bias-korrigierten Modellen beruhen, stützen diese Vermutung. In der Regel werden solche Modelle auf die Population erdnaher Objekte (engl. *Near-Earth Objects*, kurz NEOs) angewendet, die all jene Objekte umfassen, die eine Perihel-Distanz von weniger als $1.3\,\mathrm{AE}$ aufweisen [34]. Neben erdnahen Asteroiden zählen hierzu unter anderem auch erdnahe Kometen, jedoch machen die erdnahen Asteroiden den mit Abstand grössten Anteil aus, der an allen bis Oktober 2025 katalogisierten NEOs etwa 99.7% beträgt [45]. Die Ergebnisse dieser Modelle lassen sich aus diesem Grunde gut auf die NEAs übertragen. Im Allgemeinen beruhen diese Modelle

auf der Annahme, dass die grössten und hellsten NEOs, deren Grösse und absolute Helligkeit in einem exponentiellen Zusammenhang stehen, aufgrund ihrer guten Sichtbarkeit grösstenteils entdeckt sind [12]. Erdnahe Objekte mit einem Durchmesser von über $5\,\mathrm{km}$ gelten als vollständig erfasst [33], während die Erfassungsrate für Objekte im Kilometerbereich bei über 95% liegt [19]. Mathematisch wird die NEO-Population als Funktion in Abhängigkeit von vier Parametern beschrieben: der grossen Halbachse a, der Exzentrizität e, der Inklination i sowie der absoluten Helligkeit H. Zwischen den beobachteten und den tatsächlich existierenden erdnahen Objekten besteht der folgende Zusammenhang [12]:

$$n(a, e, i, H) = \varepsilon(a, e, i, H) \cdot N(a, e, i, H). \tag{6.3}$$

Dabei bezeichnet $n(a, e, i, H)$ die Anzahl der beobachteten Objekte, $N(a, e, i, H)$ die tatsächlich existierende NEO-Population und $\varepsilon(a, e, i, H)$ die Entdeckungswahrscheinlichkeit, dass ein Objekt mit gegebenen Orbitaleigenschaften a, e und i sowie einer absoluten Helligkeit H von den eingesetzten Beobachtungsstandorten erfasst wird. Durch Umformung dieser Gleichung lässt sich aus den Beobachtungsdaten auf die wahre, bias-korrigierte Verteilung der NEOs schliessen [12]. Des Weiteren wird die Gesamtpopulation als Summe verschiedener Quellregionen s modelliert [2]:

$$N(a, e, i, H) = \sum_{s} N_s(H) \cdot R_s(a, e, i). \tag{6.4}$$

Quellregionen sind Bereiche im Sonnensystem, aus denen die NEOs ursprünglich stammen, etwa der Asteroidengürtel, die transneptunische Region oder die Oortsche Wolke. Innerhalb dieser Regionen können Resonanzen oder nahe Begegnungen mit Planeten die Umlaufbahnen von Objekten so verändern, dass diese schliesslich die Kriterien für eine Klassifikation als NEO erfüllen [2]. Die Modellierung typischer Orbitalparameter (a, e, i) für jede Quelle s erfolgt mithilfe numerischer Bahnintegration, bei denen die Bewegungen zahlreicher hypothetischer Objekte über Millionen von Jahren simuliert werden, um zu analysieren, wie sich ihre Umlaufbahnen, beispielsweise durch die Gravitationskräfte der Planeten, im Laufe der Zeit verändern [12]. Die Parameter der Helligkeitsverteilungen $N_s(H)$ sowie die relativen Anteile der verschiedenen Quellregionen werden anschliessend mittels einer Maximum-Likelihood-Schätzung (vgl. Anhang B) bestimmt. Dabei werden die Parameter so angepasst, dass die durch das Modell vorhergesagte Verteilung der NEOs mit den Beobachtungsdaten bestmöglich übereinstimmt [2].

Die Ergebnisse solcher Modelle zeigen, dass die Anzahl der erdnahen Objekte und damit auch der erdnahen Asteroiden mit zunehmender absoluter Helligkeit ansteigt, was bedeutet, dass kleinere und lichtschwächere Objekte wesentlich zahlreicher existieren als grössere und hellere [33]. Granvik et al. (2018) schätzen beispielsweise, dass es $802^{+48}_{-42} \cdot 10^3$ NEOs mit einer absoluten Helligkeit von $H < 25$ gibt [12], während Bottke et al. (2002) für $H < 22$ von etwa $24'500$ NEOs ausgehen [2]. Für Objekte mit einem Durchmesser von weniger als $100\,\mathrm{m}$, was ungefähr einer Helligkeit von $H > 23.2$ entspricht [2], schätzt S. J. Stuart (2003) rund $85'000$ NEOs [33]. Der im Unterkapitel 6.3 erhaltene mittlere Schätzwert von etwa $35'900$ NEAs bezieht sich auf eine Helligkeitsspanne von 9.5 bis $33.6\,\mathrm{mag}$, liegt jedoch deutlich unter den in der Literatur genannten Werten, die sich teilweise sogar auf kleinere Helligkeitsbereiche beschränken. Dies bestätigt erneut, dass die in dieser Arbeit angewandten Schätzverfahren die tatsächliche

NEA-Populationsgrösse bedeutend unterschätzen. Obwohl sich in der Literatur keine direkten Angaben zu Schätzungen der vollständigen Risiko-Liste finden beziehungsweise zur Anzahl erdnaher Asteroiden, die bis zum Jahre 2100 eine Einschlagswahrscheinlichkeit von grösser als 10^{-11} aufweisen, ist davon auszugehen, dass auch die im Unterkapitel 6.3 ermittelte mittlere Schätzung von rund $2'200$ NEAs deutlich zu niedrig angesetzt ist.

Insgesamt ist die Aussagekraft der in dieser Arbeit erzielten Schätzwerte daher gering. Zwar liegen die berechneten Werte zufälligerweise nahe an den derzeit bekannten Zahlen, jedoch spiegeln sie weder die tatsächliche NEA-Populationsgrösse, noch die Grösse der vollständigen Risiko-Liste wider. Die niedrigen berechneten Standardabweichungen sind in diesem Falle kein Hinweis auf eine besonders hohe Präzision, sondern ebenfalls eine Folge der verletzten Modellannahmen.

Kapitel 7

Schlussfolgerung

In diesem Kapitel werden zunächst die wichtigsten Erkenntnisse dieser Arbeit zusammengefasst. Anschliessend wird ein Ausblick auf mögliche Erweiterungen gegeben.

7.1 Erkenntnisse

Capture-Recapture-Verfahren umfassen eine ganze Familie statistischer Methoden, die darauf abzielen, unter bestimmten Annahmen aus mehrfachen Beobachtungen Rückschlüsse auf die Grösse einer unbekannten Grundgesamtheit zu ziehen. Trotz ihrer heutigen Vielfalt lassen sich alle Verfahren auf dasselbe Grundprinzip, die Verhältnisschätzung des Lincoln-Petersen-Schätzers, zurückführen. Bedenkt man, dass dieser Schätzer auf lediglich zwei Stichproben basiert und sehr einfach anzuwenden ist, so ist umso bemerkenswerter, wie genaue und präzise Schätzwerte er unter erfüllten Modellannahmen liefert, sofern eine Überschneidung zwischen den Stichproben besteht. Nichtsdestotrotz ist der Lincoln-Petersen-Schätzer positiv verzerrt, auch wenn er asymptotisch erwartungstreu und konsistent ist. Der Chapman-Schätzer als Weiterentwicklung korrigiert diese Verzerrung weitgehend und liefert selbst bei fehlender Überschneidung der beiden Stichproben noch einen Schätzwert. Besonders bei kleinen Stichprobengrössen, bei denen die Überschätzungstendenz des Lincoln-Petersen-Schätzers am ausgeprägtesten ist, liefert der Chapman-Schätzer sehr zuverlässige Schätzwerte, und dies sogar mit geringerer Standardabweichung, wobei der zusätzliche Rechenaufwand vernachlässigbar bleibt. Dennoch können bei beiden Schätzverfahren einzelne Ausreisser auftreten, insbesondere nach oben. Untersucht man die beiden Schätzer unter gezielten Verletzungen der Modellannahmen, so verhalten sich diese im Allgemeinen sehr ähnlich. Grundsätzlich werden die Schätzwerte umso schlechter, je stärker die Annahmen verletzt sind, wobei bereits kleinere Verletzungen deutliche Verzerrungen hervorrufen können.

Letztere Problematik wird auch im Anwendungsbeispiel zur Populationsschätzung erdnaher Asteroiden sichtbar. Zwar zeigen die Schätzwerte des Lincoln-Petersen-Schätzers und des Chapman-Schätzers für die Populationsgrösse erdnaher Asteroiden (NEAs) insgesamt eine hohe Konsistenz sowie relativ niedrige Standardabweichungen, was eigentlich Merkmale einer zuverlässigen Schätzung sind, jedoch sind wesentliche Modellannahmen klar verletzt. Die Fangwahrscheinlichkeiten sind heterogen, was mit der Vielfalt individueller Eigenschaften der NEAs zu begründen ist, etwa ihren unterschiedlichen Grössen, Orbitalpa-

rametern und absoluten Helligkeiten. Eine Modellierung der Fangwahrscheinlichkeit in Abhängigkeit der absoluten Helligkeit verdeutlicht diese Heterogenität, auch wenn die resultierenden Fangwahrscheinlichkeiten in diesem Falle positiv verzerrt sind, da die Datengrundlage mindestens einmal erfasste Objekte bilden. Des Weiteren besteht eine positive Abhängigkeit zwischen den Stichproben, da die Datenerhebung der NEAs stets durch die gleichen drei Observatorien erfolgt, die alle überwiegend im sichtbaren Wellenlängenbereich arbeiten und aufgrund ihrer geographischen Lokalisation ähnliche Himmelsregionen abdecken. Die Abhängigkeit zwischen den Daten der einzelnen Observatorien kann unter Verwendung von Chi-Quadrat-Tests gezeigt werden. Sowohl heterogene Fangwahrscheinlichkeiten als auch eine positive Abhängigkeit der Stichproben führen zu einer Unterschätzung der wahren Populationsgrösse erdnaher Asteroiden, wobei ein Vergleich mit den aktuell katalogisierten erdnahen Asteroiden sowie mit der Fachliteratur diese Annahme bestätigt. Die Aussagekraft der in diesem Anwendungsbeispiel erzielten Schätzwerte ist daher gering und führt zu keinem wirklichen Gewinn neuer Erkenntnisse. Für verlässlichere Resultate müsste auf komplexere Modelle zurückgegriffen werden, die insbesondere die Heterogenität der Fangwahrscheinlichkeiten sowie die positive Abhängigkeit der Stichproben berücksichtigen.

7.2 Ausblick

Aufgrund der Vielzahl an bestehenden Capture-Recapture-Verfahren bietet sich ein breites Feld an Möglichkeiten zur Weiterführung dieser Arbeit. In Kapitel 2 werden einige dieser Verfahren im historischen Überblick kurz angesprochen, jedoch finden sich in der Literatur noch zahlreiche weitere Methoden, wobei auch diese gezielt miteinander verglichen werden könnten. Eine mögliche Erweiterung bestünde darin, den Vergleich der Schätzverfahren noch zu vertiefen, indem beispielsweise in Simulationen mehrere Modellannahmen gleichzeitig verletzt werden. Darüber hinaus könnten Methoden zur Modellauswahl herangezogen werden, um komplexere Modelle systematisch zu bewerten. Hierbei bieten sich die Akaike-Gewichte sowie das Akaike Information Criterion (AIC) an, die den Vergleich unterschiedlicher Modelle aufgrund ihrer Güte und Komplexität ermöglichen [26]. Im Kontext des Anwendungsbeispiels der Populationsschätzung erdnaher Asteroiden wäre zudem eine genauere Betrachtung statistischer Tests sinnvoll, insbesondere des im Unterkapitel 6.4 verwendeten Chi-Quadrat-Tests und der zugehörigen Verteilung. Ausserdem würde sich die Anwendung komplexerer Modelle anbieten, die die Heterogenität der Fangwahrscheinlichkeiten und die Abhängigkeit der Stichproben berücksichtigen. Dabei könnte das Modell M_{bh} nach Otis et al. (1978) hinzugezogen werden, das sowohl Verhaltensänderungen (*behavioral responses*) als auch unterschiedliche Fangwahrscheinlichkeiten (*heterogeneity*) modelliert [18]. Da die in der NEODyS-Datenbank erfassten Beobachtungen stets mit Zeitstempeln versehen sind, könnte des Weiteren auch die zeitlich kontinuierliche Variante M_{bh}^{*} dieses Modells angewendet werden [26]. Schliesslich könnten zusätzlich Bayes'sche Modelle eingeführt werden, die die unbekannten Parameter nicht als feste Werte, sondern als Zufallsgrössen mit entsprechenden Wahrscheinlichkeitsverteilungen behandeln [32]. Besonders bei der Analyse der Fangwahrscheinlichkeiten erdnaher Asteroiden wäre dieser Ansatz von Vorteil, da auf diese Art und Weise zusätzliche Einflussgrössen wie die absolute Helligkeit oder verschiedene Orbitalparameter gezielt in die Modellierung einbezogen werden könnten.

Anhang A

Die hypergeometrische Wahrscheinlichkeitsverteilung

Dieser Teil des Anhangs enthält nähere Informationen zur hypergeometrischen Verteilung. Darüber hinaus werden ihr Erwartungswert und ihre Varianz hergeleitet, welche in Kapitel 3 und 4 für die Bestimmung des Erwartungswerts sowie der Varianz des Lincoln-Petersen-Schätzers beziehungsweise des Chapman-Schätzers verwendet werden.

A.1 Grundlagen der hypergeometrischen Verteilung

Die hypergeometrische Wahrscheinlichkeitsverteilung, häufig kurz als hypergeometrische Verteilung bezeichnet, stellt eine diskrete Wahrscheinlichkeitsverteilung dar, die ein Zufallsexperiment mit genau zwei möglichen Ausgängen beschreibt: „Erfolg“ oder „Misserfolg“ [28]. Das zugrunde liegende Zufallsexperiment lässt sich anschaulich durch ein Urnenmodell darstellen. Gegeben sei eine Urne mit insgesamt N Kugeln, von denen M markiert sind beziehungsweise als Erfolge gelten. Aus dieser Urne werden zufällig n Kugeln gezogen, wobei dies im Gegensatz zur Binomialverteilung ohne Zurücklegen erfolgt. Die Zusammensetzung der Grundgesamtheit ändert sich somit nach jeder Ziehung, was bedeutet, dass die einzelnen Ziehungen nicht unabhängig voneinander sind [37]. Die hypergeometrische Wahrscheinlichkeitsverteilung ist somit durch die folgenden drei Parameter charakterisiert [27]:

$$N \qquad \text{Gesamtzahl der Elemente,} \tag{A.1}$$

$$M \qquad \text{Anzahl der Erfolge in der Grundgesamtheit,} \tag{A.2}$$

$$n \qquad \text{Stichprobengrösse.} \tag{A.3}$$

Die Wahrscheinlichkeit, genau m Erfolge in der Stichprobe von n gezogenen Kugeln zu erhalten, ergibt sich aus dem Verhältnis der Anzahl der günstigen Ziehungen zur Gesamtzahl aller möglichen Ziehungen [1]. Die Anzahl der Möglichkeiten, genau m Erfolge und damit $n - m$ Misserfolge zu ziehen, beträgt $\binom{M}{m}\binom{N-M}{n-m}$, während es insgesamt $\binom{N}{n}$ Möglichkeiten gibt, n aus N Kugeln auszuwählen [37]. Daraus ergibt sich die Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung

$$P(X = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} , \tag{A.4}$$

wobei $\max(0, n-(N-M)) \leq m \leq \min(n, M)$ gilt. Formal wird dies durch $X \sim H(N, M, n)$ ausgedrückt [28].

Ist die Gesamtzahl der Elemente N sehr gross, sowohl in absoluter Zahl als auch im Verhältnis zu M und n, so bleibt die Erfolgswahrscheinlichkeit während des gesamten Zufallsexperiments in etwa konstant. In diesem Falle kann die hypergeometrische Verteilung näherungsweise durch die einfacher zu berechnende Binomialverteilung beschrieben werden, deren Wahrscheinlichkeitsfunktion folgendermassen definiert ist [27]:

$$P(X = m) = \binom{n}{m} \cdot p^m \cdot (1-p)^{n-m} \quad \text{mit} \quad p = \frac{M}{N} \,. \tag{A.5}$$

In der Praxis besteht die Konvention, dass die Binomialverteilung als hinreichend genaue Approximation der hypergeometrischen Verteilung angesehen wird, wenn $\frac{n}{N} < 0.05$, was bedeutet, dass weniger als fünf Prozent der Gesamtzahl der Elemente gezogen werden [1].

A.2 Herleitung des Erwartungswerts der hypergeometrischen Verteilung

Die folgende Herleitung des Erwartungswerts der hypergeometrischen Verteilung basiert auf dem in [37] verwendeten Ansatz. Der Erwartungswert einer diskreten Zufallsvariablen X lässt sich definitionsgemäss als Summe der Produkte aus allen möglichen Werten und ihren jeweiligen Wahrscheinlichkeiten ausdrücken [48]. Da X einer hypergeometrischen Verteilung folgt, ergibt sich somit

$$\mathbb{E}[X] \;=\; \sum_{m=1}^{n} m \cdot \mathbb{P}(X = m) = \sum_{m=1}^{n} m \cdot \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} \,. \tag{A.6}$$

Zur Vereinfachung dieses Ausdrucks wird eine kombinatorische Identität herangezogen [37]. Unter Verwendung der Fakultätsdarstellung des Binomialkoeffizienten kann für $1 \leq b \leq a$ gezeigt werden, dass

$$\begin{aligned} b\binom{a}{b} &= b \cdot \frac{a!}{b!(a-b)!} \\ &= \frac{a(a-1)!}{(b-1)!(a-1-(b-1))!} \\ &= a\binom{a-1}{b-1}, \end{aligned} \tag{A.7}$$

woraus unmittelbar folgt, dass

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1}. \tag{A.8}$$

Wird diese Identität auf die in Ausdruck A.6 enthaltenen Binomialkoeffizienten $\binom{M}{m}$ und $\binom{N}{n}$ angewendet, entsteht ein konstanter Vorfaktor, der unabhängig von der Laufvariablen m ist und daher aus der Summe ausgeklammert werden kann [28]. Dadurch ergibt sich

$$\mathbb{E}[X] \;=\; \sum_{m=1}^{n} \frac{M}{\frac{N}{n}} \cdot \frac{\binom{M-1}{m-1}\binom{N-M}{n-m}}{\binom{N-1}{n-1}} = \frac{Mn}{N} \cdot \sum_{m=1}^{n} \frac{\binom{M-1}{m-1}\binom{N-M}{n-m}}{\binom{N-1}{n-1}} \,. \tag{A.9}$$

Der verbleibende Summenausdruck entspricht dabei der Summe aller Wahrscheinlichkeiten einer hypergeometrischen Verteilung mit den Parametern $N' = N - 1$, $M' = M - 1$ und $n' = n - 1$. Da die Summe aller Wahrscheinlichkeiten einer Wahrscheinlichkeitsverteilung den Wert 1 annimmt [40], ist der Erwartungswert der hypergeometrischen Verteilung folglich gegeben durch

$$\mathbb{E}[X] = \frac{Mn}{N} . \tag{A.10}$$

Damit entspricht der Erwartungswert dem Produkt aus dem Stichprobenumfang n und dem Anteil der Erfolge in der Grundgesamtheit $\frac{M}{N}$.

A.3 Herleitung der Varianz der hypergeometrischen Verteilung

Ausgehend vom in Abschnitt A.2 hergeleiteten Erwartungswert der hypergeometrischen Verteilung bietet es sich an, die folgende Definition der Varianz heranzuziehen, um diese für die hypergeometrische Verteilung zu bestimmen [37]:

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}\left[X(X-1)\right] + \mathbb{E}[X] - \mathbb{E}[X]^2 . \tag{A.11}$$

Dabei wird der Ausdruck X^2 durch $X(X-1)+X$ ersetzt, da sich der Erwartungswert von $X(X-1)+X$ in diesem Falle einfacher bestimmen lässt. Der Erwartungswert $\mathbb{E}[X]$ ist bereits bekannt und entspricht dem in Gleichung A.10 angegebenen Ausdruck. Damit verbleibt lediglich der Erwartungswert $\mathbb{E}[X(X-1)]$ als unbekannte Grösse. Definitionsgemäss lässt sich dieser als Summe der Produkte aus allen möglichen Werten und ihren jeweiligen Wahrscheinlichkeiten ausdrücken, wobei X einer hypergeometrischen Verteilung folgt [28]:

$$\mathbb{E}\left[X(X-1)\right] = \sum_{m=2}^{n} m(m-1)\cdot\mathbb{P}(X=m) = \sum_{m=2}^{n} m(m-1)\cdot\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}} . \tag{A.12}$$

Zur Vereinfachung dieses Ausdrucks wird die in Abschnitt A.2 eingeführte Identität verwendet, die besagt, dass $b\binom{a}{b} = a\binom{a-1}{b-a} \Leftrightarrow \binom{a}{b} = \frac{a}{b}\binom{a-1}{b-a}$ [37]. Diese Identität wird zunächst auf $m\binom{M}{m}$ sowie $\binom{N}{n}$ und anschliessend erneut auf $(M-1)\binom{M-1}{m-1}$ sowie $\binom{N-1}{n-1}$ angewendet, wodurch ein konstanter Vorfaktor entsteht, der unabhängig von der Laufvariablen m ist und daher aus der Summe ausgeklammert werden kann [28]. Dadurch ergibt sich

$$\begin{aligned}
\mathbb{E}\left[X(X-1)\right] &= \sum_{m=2}^{n} M(m-1)\cdot\frac{\binom{M-1}{m-1}\binom{N-M}{n-m}}{\frac{N}{n}\binom{N-1}{n-1}} \\
&= \sum_{m=2}^{n} M(M-1)\cdot\frac{\binom{M-2}{m-2}\binom{N-M}{n-m}}{\frac{N}{n}\cdot\frac{N-1}{n-1}\binom{N-2}{n-2}} \\
&= \frac{n(n-1)M(M-1)}{N(N-1)}\cdot\sum_{m=2}^{n}\frac{\binom{M-2}{m-2}\binom{N-M}{n-m}}{\binom{N-2}{n-2}} .
\end{aligned} \tag{A.13}$$

Der verbleibende Summenausdruck entspricht dabei der Summe aller Wahrscheinlichkeiten einer hypergeometrischen Verteilung mit den Parametern $N' = N - 2$, $M' = M - 2$ und $n' = n - 2$. Da die

Summe aller Wahrscheinlichkeiten einer Wahrscheinlichkeitsverteilung den Wert 1 annimmt [40], ist der Erwartungswert von $X(X-1)$ folglich gegeben durch

$$\mathbb{E}\left[X(X-1)\right] = \frac{n(n-1)M(M-1)}{N(N-1)}\,. \tag{A.14}$$

Setzt man diesen gemeinsam mit dem bekannten Erwartungswert $\mathbb{E}[X] = \frac{Mn}{N}$ in die Ausgangsgleichung A.11 ein, so lässt sich die Varianz der hypergeometrischen Verteilung bestimmen. Um den Ausdruck für die Varianz übersichtlicher zu gestalten, wird zunächst $\frac{Mn}{N}$ ausgeklammert, und die verbleibenden Brüche innerhalb der Klammer werden auf einen gemeinsamen Nenner gebracht. Der Zähler vereinfacht sich dabei zu $(N-M)(N-n)$, woraus schliesslich der in der Literatur am häufigsten angegebene Ausdruck für die Varianz der hypergeometrischen Verteilung resultiert [28]:

$$\begin{aligned}
\mathrm{Var}(X) &= \frac{n(n-1)M(M-1)}{N(N-1)} + \frac{Mn}{N} - \left(\frac{Mn}{N}\right)^2 \\
&= \frac{Mn}{N}\left(\frac{(n-1)(M-1)}{(N-1)} + 1 - \frac{Mn}{N}\right) \\
&= \frac{Mn}{N}\left(\frac{N(n-1)(M-1) + N(N-1) - Mn(N-1)}{N(N-1)}\right) \\
&= \frac{Mn}{N^2(N-1)}\Big(NMn - Nn - NM + N + N^2 - N - NMn + Mn\Big) \\
&= \frac{Mn(N-M)(N-n)}{N^2(N-1)} \;=\; \frac{Mn}{N}\cdot\left(1-\frac{M}{N}\right)\cdot\frac{N-n}{N-1}\,.
\end{aligned} \tag{A.15}$$

Die Varianz der hypergeometrischen Verteilung entspricht somit der Varianz der Binomialverteilung, multipliziert mit dem Korrekturfaktor $\frac{N-n}{N-1}$ [27].

Anhang B

Die Maximum-Likelihood-Methode

In diesem Teil des Anhangs werden die theoretischen Grundlagen der Maximum-Likelihood-Methode erläutert, die im Unterkapitel 3.2 zur Herleitung des Lincoln-Petersen-Schätzers Anwendung finden.

Die Maximum-Likelihood-Methode ist ein statistisches Verfahren zur Parameterschätzung [35]. Häufig wird sie R. A. Fisher zugeschrieben, der diese im Jahre 1912 in seinen ersten statistischen Arbeiten einführte. Allerdings lässt sich die Methode bis ins 18. Jahrhundert zurückverfolgen, wo sie unter anderem bereits von J. H. Lambert und D. Bernoulli verwendet wurde [28].

Gegeben sei eine Zufallsstichprobe $x = (x_1, x_2, \ldots, x_n)$ mit Umfang n. Jedes Stichprobenelement x_i kann dabei im Allgemeinen einen ganzen Satz von Variablen darstellen. Es wird angenommen, dass jedes x_i einer Wahrscheinlichkeitsdichte $f(x|\theta)$ folgt, die durch die Parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ bestimmt wird [49]. Zudem wird vorausgesetzt, dass die Stichprobenelemente unabhängig und identisch verteilt sind [37]. Die Wahrscheinlichkeit, die gesamte Stichprobe zu erhalten, ergibt sich daher als Produkt der Wahrscheinlichkeiten der einzelnen Stichprobenelemente [28]:

$$P(x_1, x_2, \ldots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \ldots \cdot f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{B.1}$$

Für feste Werte der Stichprobenelemente $x_1, x_2, \ldots, x_n$ ist diese gemeinsame Wahrscheinlichkeitsdichtefunktion eine Funktion der Parameter θ [28]. Diese Funktion,

$$L(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta), \tag{B.2}$$

wird als Likelihood-Funktion bezeichnet. Im Unterschied zu Wahrscheinlichkeitsverteilungen ist diese jedoch nicht normiert. Die Likelihood-Funktion beschreibt, wie plausibel es ist, die Stichprobenelemente $x_1, x_2, \ldots, x_n$ für unterschiedliche Sätze von Parametern θ zu erhalten [37]. Das Ziel der Maximum-Likelihood-Methode besteht nun darin, den Satz von Parametern θ zu finden, der die Likelihood-Funktion maximiert. Dies ist gleichbedeutend mit der Maximierung der Wahrscheinlichkeit, die beobachteten Daten unter den gegebenen Parametern zu erhalten [49]. Der gesuchte Satz von Parametern, bezeichnet durch $\hat{\theta}_{ML}$, wird dabei als Maximum-Likelihood-Schätzer (engl. *Maximum Likelihood Estimator*) bezeichnet [28].

Formal ausgedrückt wird der Maximum-Likelihood-Schätzer durch den folgenden Ausdruck definiert [49]:

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta). \tag{B.3}$$

Die Konsistenz stellt ein zentrales Element von Maximum-Likelihood-Schätzern dar. Mit zunehmender Stichprobengrösse n konvergiert $\hat{\theta}_{ML}$ gegen den wahren Satz von Parametern [49]. In der Praxis ist es jedoch oft umständlich, die Likelihood-Funktion direkt zu maximieren, da das Produkt vieler Wahrscheinlichkeiten zu sehr kleinen, aber nie negativen Zahlen führen kann. Aus numerischen Gründen wird daher häufig der Logarithmus der Likelihood-Funktion verwendet [40]. Diese Funktion,

$$\ell(\theta|x_1, x_2, \ldots, x_n) = \ln L(\theta|x_1, x_2, \ldots, x_n), \tag{B.4}$$

auch als Log-Likelihood-Funktion bekannt, hat den Vorteil, die einzelnen Funktionswerte auf eine besser handhabbare Skala zu verschieben. Zudem wandelt der Logarithmus das Produkt der Wahrscheinlichkeiten in eine Summe um:

$$\ell(\theta|x_1, x_2, \ldots, x_n) = \ln \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta). \tag{B.5}$$

Summen lassen sich in der Regel leichter maximieren als Produkte, was die Berechnung deutlich vereinfacht [28]. Der Logarithmus verändert dabei nicht die Lage der Maxima, da es sich um eine monotone Transformation handelt. Der Maximum-Likelihood-Schätzer der Log-Likelihood-Funktion ist somit gleich dem Maximum-Likelihood-Schätzer der Likelihood-Funktion [40]. Allerdings kann die Likelihood-Funktion, insbesondere bei grossen Datensätzen oder komplexeren Modellen, viele lokale Maxima aufweisen, was die Identifizierung des globalen Maximums zu einer rechnerischen Herausforderung macht [49].

Anhang C

Programmcodes

In diesem Teil des Anhangs werden die Codes der Python-Programme vorgestellt, die in Kapitel 5 für die Simulationen zum Vergleich des Lincoln-Petersen-Schätzers mit dem Chapman-Schätzer sowie in Kapitel 6 für die Datenbeschaffung zur Populationsschätzung erdnaher Asteroiden verwendet werden.

C.1 Flexibler Populationsschätzungs-Simulator

Der folgende Abschnitt präsentiert den Code des Python-Programms, der als Grundlage für sämtliche Simulationen in Kapitel 5 dient. Mit den damit erzeugten Daten können je nach Bedarf Visualisierungen, wie beispielsweise Boxplots, erstellt oder ausgewählte Kennzahlen numerisch ausgegeben werden.

```
import random
import numpy as np

# Listen zur Speicherung der Schätzwerte
N_LP_geschaetzt = []  # Schätzwerte des Lincoln-Petersen-Schätzers
N_C_geschaetzt = []   # Schätzwerte des Chapman-Schätzers

# Anzahl Durchläufe der Simulation ohne Resultat (Division durch null)
kein_resultat = 0

# Gesamtpopulationsgrösse (N)
populationsgroesse = 500
# Anzahl Individuen der ersten Stichprobe bzw. Anzahl markierter Individuen der Gesamtpopulation (M)
M = round(populationsgroesse/5)
# Anzahl Individuen der zweiten Stichprobe (n)
n = round(populationsgroesse/5)

# Populationsveränderungen (Zuwanderung: pbew > 0, Abwanderung: pbew < 0)
pbew = 0

# Fangwahrscheinlichkeiten der einzelnen Individuen
"""
mean = 0.5
std_dev = 0.1
def fangw():
  return min(1, max(0, np.random.normal(loc = mean, scale = std_dev)))
"""
def fangw():
  return random.uniform(0.5, 0.5)

# Verhaltensreaktionen auf das Fangen (trap-shy: trb > 1, trap-happy: trb < 1)
trb = 1

# Anzahl Durchläufe der Simulation
d = 5000
```

```
# erste Stichprobe (Capture)
def capture():
  # zufällig ausgewählte Individuen, die potenziell gefangen werden können
  potcapture = random.sample(range(0, len(population)), 1)

  # wenn generierte Zufallszahl kleiner ist als Fangwahrscheinlichkeit, gilt das Individuum als gefangen
  if population[potcapture[0]] <= 1:
    if random.random() < population[potcapture[0]]:
      population[potcapture[0]] += 1 # gefangen, erhöht Fangwahrscheinlichkeit um eins
    else:
      capture()  # nicht gefangen
  else:
    capture()    # nicht gefangen

# zweite Stichprobe (Recapture)
def recapture():
  global m
  # zufällig ausgewählte Individuen, die potenziell gefangen werden können
  potrecapture = random.sample(range(0, len(population)), 1)

  # ist die Fangwahrscheinlichkeit > 1, so wurde das Individuum bereits in der ersten Stichprobe erfasst
  if population[potrecapture[0]] > 1:
    # wenn generierte Zufallszahl kleiner ist als Fangwahrscheinlichkeit, gilt das Individuum als gefangen
    if random.random() < (population[potrecapture[0]] - trb):
      m += 1  # erhöht Zähler der erneut gefangenen Individuen um eins
      population[potrecapture[0]] = 0  # gefangen, setzt Fangwahrscheinlichkeit auf null: verhindert erneuten Fang
    else:
      recapture()  # nicht gefangen
  else:
    # wenn generierte Zufallszahl kleiner ist als Fangwahrscheinlichkeit, gilt das Individuum als gefangen
    if random.random() < population[potrecapture[0]]:
      population[potrecapture[0]] = 0  # gefangen, setzt Fangwahrscheinlichkeit auf null: verhindert erneuten Fang
    else:
      recapture()  # nicht gefangen

# für jeden Durchlauf der Simulation wird for-Schlaufe einmal ausgeführt
for i in range(d):
  # generiert die Fangwahrscheinlichkeiten für sämtliche Individuen der Population
  population = [fangw() for x in range(populationsgroesse)]

  # führt Funktion capture() M mal aus
  for i in range(M):
    capture()

  # überprüft, ob es zu Populationsveränderungen kommt
  if pbew != 0:
    if pbew > 0:  # Zuwanderung
      for i in range(pbew):
        population.append(fangw())  # generiert die Fangwahrscheinlichkeiten der neu hinzukommenden Individuen
    else:  # Abwanderung
      for i in range(abs(pbew)):
        idx = random.randrange(len(population))
        population.pop(idx)  # löscht Individuen der Population nach dem Zufallsprinzip

  m = 0  # setzt Zähler der erneut gefangenen Individuen auf null
  # führt die Funktion recapture() n mal aus
  for i in range(n):
    recapture()

  if m == 0:
    kein_resultat += 1  # erhöht Zähler der Anzahl Durchläufe ohne Resultat um eins, da m = 0
  else:
    N_LP = n*M/m  # berechnet den Schätzwert des Lincoln-Petersen-Schätzers
    N_LP_geschaetzt.append(N_LP)
    N_C = (n+1)*(M+1)/(m+1)-1  # berechnet den Schätzwert des Chapman-Schätzers
    N_C_geschaetzt.append(N_C)
```

C.2 Datenbeschaffung für die Populationsschätzung erdnaher Asteroiden

In diesem Unterkapitel werden die Codes der Python-Programme aus dem Unterkapitel 6.2 präsentiert. Mit dem ersten Programm werden die Namen sowie die Anzahl eindeutiger NEAs pro Observatorium ermittelt.

```
from bs4 import BeautifulSoup
from urllib.request import urlopen

# Listen zur Speicherung von URLs
observatories = []  # URLs der Observatorien
observations = []   # URLs der Tabellen mit den Beobachtungsdaten

# öffnet Hauptseite mit Liste aller Observatorien und sammelt alle Links
with urlopen("https://newton.spacedys.com/neodys/index.php?pc=2.0&o=") as response:
    soup = BeautifulSoup(response, "html.parser")
    for anchor in soup.find_all("a"):  # extrahiert alle <a>-Tags (Links)
        observatories.append("https://newton.spacedys.com/neodys/" + anchor.get("href", "/"))

# entfernt die ersten 24 Links, da diese von der Menüleiste stammen
observatories = observatories[24:]

# sammelt für jedes Observatorium Link zu den Tabellen mit den Beobachtungsdaten
for i in range(len(observatories)):  # Range muss ggf. manuell in kleinere Intervalle unterteilt werden
    with urlopen(observatories[i]) as response:
        soup = BeautifulSoup(response, "html.parser")
        for anchor in soup.find_all("a"):
            o = anchor.get("href", "/")
            # Links, die mit "&ab=0" enden, führen jewils zum Anfang der Tabellen
            if o.endswith("&ab=0"):
                observations.append("https://newton.spacedys.com/neodys/" + anchor.get("href", "/"))

# wertet Beobachtungsdaten aus und extrahiert eindeutige NEAs
for i in range(len(observations)):
    # extrahiert die letzten drei Zeichen der URL vor "&ab=0", was Observatoriums-Code entspricht
    name = observations[i]
    name = name[:-5][-3:]
    print("Observatorium: " + name)

    # schreibt Namen des Observatoriums in die Datei "ausgabe.txt"
    with open("ausgabe.txt", "a", encoding="utf-8") as datei:
        datei.write("Observatorium: " + name + "\n")

    objects = []  # Liste eindeutiger NEAs
    count = 0     # Zähler für die Anzahl eindeutiger NEAs

    j = 0
    while True:
        # erzeugt URL für die aktuelle Seite der Tabellen mit Beobachtungsdaten
        obs = observations[i]
        obs = obs[:-1] + f"{j}"
        print(obs)

        with urlopen(obs) as response:
            soup = BeautifulSoup(response, "html.parser")
            rows = soup.find_all("tr")  # sucht alle Tabellenzeilen auf der Seite

            if not rows:
                break  # wenn keine Zeilen vorhanden sind, wurde das Ende der Tabelle erreicht

            first_column = []  # Liste zur Speicherung der Inhalte der ersten Spalte (Namen der NEAs)

            # geht jede Zeile durch und extrahiert jeweils den Text der ersten Zelle
            for row in rows:
                cells = row.find_all("td")
                if cells:
                    first_column.append(cells[0].get_text(strip=True))
```

```
            # da jedes NEA doppelt aufgeführt ist, wird nur die erste Hälfte der Liste verwendet
            half = len(first_column) // 2
            for x in first_column[:half]:
                if x not in objects:
                    objects.append(x)
                    count += 1  # erhöht Zähler eindeutiger NEAs um eins
            j += 1  # navigiert zur nächsten Seite, auf der die Tabelle fortgesetzt wird

    # schreibt Namen und Anzahl der NEAs in Ausgabedatei
    with open("ausgabe.txt", "a", encoding="utf-8") as datei:
        for object in objects:
            datei.write(object + "\n")
        datei.write("Anzahl Elemente: " + str(count) + "\n\n")

    print("Anzahl Elemente: " + str(count) + "\n")
```

Das zweite Programm dient der Sammlung spezifischer Daten zu den einzelnen NEAs. Neben der Information, ob das jeweilige Objekt als potenziell gefährlich eingestuft wird, werden auch die absolute Helligkeit sowie verschiedene Orbitalparameter, wie beispielsweise die grosse Halbachse oder die Aphel- und Periheldistanz, erfasst.

```
import requests
from bs4 import BeautifulSoup

# Funktion zur Prüfung, ob String eine Gleitkommazahl (float) ist
def number(x):
    try:
        float(x)
        return True
    except ValueError:
        return False

# Liste mit Namen erdnaher Asteroiden, zu denen Daten gesammelt werden sollen
list = ["2014YC35", "2009OZ4", "2018BY6", "2010AF3", ...]  # Liste der Übersicht halber gekürzt

# Basis-URL der NEODyS-Seite zur Abfrage von Objektinformationen
url = "https://newton.spacedys.com/neodys/index.php?pc=1.1.0&n="

# iteriert über alle NEAs in der Liste
for i in range(len(list)):
    link = url + list[i]  # setzt vollständiges URL für das jeweilige Objekt zusammen
    response = requests.get(link)  # sendet HTTP-Anfrage an die Seite

    data = []  # Liste zur Speicherung extrahierter numerischer Daten
    risk = 0   # Risiko-Indikator, zählt Anzahl gefundener Links zur Risiko-Liste

    # kontrolliert, ob Seite erfolgreich geladen wurde
    if response.status_code == 200:
        soup = BeautifulSoup(response.text, "html.parser")

        # überprüft, ob NEA aktuell auf der Risiko-Liste steht
        for anchor in soup.find_all("a"):
            o = anchor.get("href", "/")
            if o.endswith("pc=4.1"):  # Link zur Risiko-Liste endet auf "pc=4.1"
                risk += 1

        # extrahiert alle Tabellen auf der Seite
        tables = soup.find_all("table")

        # greift auf die erste Tabelle zu, welche Daten zu Orbitalparametern enthält
        first_table = tables[0]
        rows = first_table.find_all("tr")
        first_row = rows[0]  # gesuchte Daten befinden sich in der ersten Tabellenzeile
        columns = first_row.find_all("td")

        # öffnet Ausgabedatei, um Daten fortlaufend zu speichern
        with open("datenausgabe.txt", "a", encoding="utf-8") as file:
            file.write("\n" + list[i] + ", ")  # schreibt Name des NEAs in die Datei
```

```
            # überprüft jede Spalte auf numerische Inhalte und speichert diese
            for c in columns:
                c_text = c.text.strip()
                if number(c_text):  # nur Zahlen werden gespeichert
                    data.append(c_text)

            # entfernt zusätzliche Werte, die nicht bei allen NEAs aufgeführt sind
            if len(data) == 20:
                for d in data:
                    file.write(d + ", ")
            elif len(data) == 21:
                del data[12]  # entfernt einen zusätzlichen Wert
                for d in data:
                    file.write(d + ", ")
            elif len(data) == 22:
                del data[12]  # entfernt zwei zusätzliche Werte
                del data[12]
                for d in data:
                    file.write(d + ", ")
            else:
                print("Error D")  # gibt eine Fehlermeldung bei unerwarteter Datenstruktur aus

            if risk == 2:  # zwei Links zur Risiko-Liste wurden gefunden
                file.write("1")  # NEA wird auf der Risiko-Liste geführt
            else:
                file.write("0")  # NEA wird nicht auf der Risiko-Liste geführt

        # zeigt Fortschritt in der Konsole an
        print(list[i])

    else:
        # gibt eine Fehlermeldung aus, falls Seite nicht erfolgreich geladen werden konnte
        print("Error L")
```

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Artikel

[1] Adrian Blumenthal, „IX Induktive Statistik," *Fachschaft Mathematik, Kollegium Spiritus Sanctus Brig*, Feb. 2025.

[2] Bottke et al., „Debiased Orbital and Absolute Magnitude Distribution of the Near-Earth Objects," *Icarus*, Jg. 156, Nr. 2, S. 399–433, 2002.

[3] Lionel Briand, Khaled Emam, Bernd Freimut und Oliver Laitenberger, „A Comprehensive Evaluation of Capture-Recapture Models for Estimating Software Defect Content," *Software Engineering, IEEE Transactions on*, Jg. 26, S. 518–540, Jul. 2000.

[4] Sarah Brittain und Dankmar Boehning, „Estimators in capture–recapture studies with two sources," *AStA Advances in Statistical Analysis*, Jg. 93, S. 23–47, 2008.

[5] Chambers et al., „The Pan-STARRS1 Surveys," *arXiv e-prints*, Dez. 2019.

[6] Anne Chao, „An overview of Closed capture-recapture models," *Journal of Agricultural, Biological, and Environmental Statistics*, Jg. 6, S. 158–175, Jun. 2001.

[7] Douglas G. Chapman, „Some properties of the hypergeometric distribution with applications to zoological sample censuses," *Statistics*, Jg. 1, Nr. 7, S. 131–160, 1951.

[8] Kyle Dettloff, „Assessment of bias and precision among simple closed population mark-recapture estimators," *Fisheries Research*, Jg. 265, S. 106 756, Mai 2023.

[9] Anke van Dyk, „Capturing transients: an application of biostatistics to astronomy," *Department of Astronomy, University of Cape Town*, Nov. 2021.

[10] Joachim Engel, „Markieren - Einfangen - Schätzen: Wie viele wilde Tiere?" *Stochastik in der Schule*, Nr. 2, S. 17–24, 2000.

[11] Thomas Gautschi und Dominik Hangartner, „Die Untersuchung verborgener Populationen: Eine Capture-Recapture-Studie mit Heroinabhängigen," *Zeitschrift für Soziologie*, Jg. 39, Okt. 2010.

[12] Granvik et al., „Debiased orbit and absolute-magnitude distributions for near-Earth objects," *Icarus*, Jg. 312, S. 181–207, 2018.

[13] Laura Gruber, „Capture-Recapture Methods: A Comparison of different Estimators and Confidence Intervals," *Institute of Applied Statistics, Johannes Kepler University Linz*, 2023.

[14] Zaki Hasnain, Christopher Lamb und Shane Ross, „Capturing near-Earth asteroids around Earth," *Acta Astronautica*, Jg. 81, S. 523–531, Dez. 2012.

[15] Lima Ramos et al., „A Review of Capture-recapture Methods and Its Possibilities in Ophthalmology and Vision Sciences (Ophthalmic Epidemiology)," *Ophthalmic epidemiology*, Mai 2020.

[16] Naaheed Mukadam, Louise Marston, Katie Flanagan, Etuini Ma'u, Gary Cheung und Dankmar Boehning, „Estimating undiagnosed dementia in England using capture recapture techniques," *BMC Geriatrics*, Jg. 25, Jan. 2005.

[17] Stephen O. Nyangoma, „On the Capture-Recapture Methods of Estimating Population Size," *Department of Mathematics, University of Nairobi*, Jul. 1991.

[18] Otis et al., „Statistical Inference from Capture Data on Closed Animal Populations," *Wildlife Monographs*, Jg. 62, S. 1–135, 1978.

[19] Quanzhi Ye et al., „A Twilight Search for Atiras, Vatiras, and Co-orbital Asteroids: Preliminary Results," *The Astronomical Journal*, Jg. 159, Nr. 2, Jan. 2020.

[20] George A. F. Seber, „The Effects of Trap Response on Tag Recapture Estimates," *Biometrics*, Jg. 26, Nr. 1, S. 13–22, 1970.

[21] Michael Spivey, „The Chu-Vandermonde Identity via Leibniz's Identity for Derivatives," *The College Mathematics Journal*, Jg. 47, Nr. 3, S. 219–220, 2016.

[22] Frederick F. Stephan, „The Expected Value and Variance of the Reciprocal and Other Negative Powers of a Positive Bernoullian Variate," *Annals of Mathematical Statistics*, Jg. 16, S. 50–61, 1945.

[23] Janet T. Wittes, „331. Note: On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate," *Biometrics*, Jg. 28, Nr. 2, S. 592–597, 1972.

[24] Hanwen Zhang, „A Note About Maximum Likelihood Estimator in Hypergeometric Distribution," *Revista Comunicaciones en Estadistica*, Jg. 2, Aug. 2009.

[25] Xiao Zhang, „Confidence Intervals for Population Size in a Capture-Recapture Problem.," *East Tennessee State University*, Aug. 2007.

Bücher

[26] Steven C. Amstrup, Trent L. McDonald und Bryan F. J. Manly, *Handbook of Capture-Recapture Analysis*. Princeton University Press, Okt. 2005.

[27] Karl Bury, *Statistical Distributions in Engineering*. Cambridge University Press, 1999.

[28] Yadolah Dodge, *The Concise Encyclopedia of Statistics*. Springer New York, Feb. 2008.

[29] William R. Heyer, Maureen A. Donnelly und Roy W. McDiarmid, *Measuring and Monitoring Biological Diversity: Standard Methods for Amphibians*, W. Ronald Heyer, Hrsg. Smithsonian Institution Press, 1994, Kap. 8, S. 183–205.

[30] Ruth King und Rachel McCrea, *Integrated Population Biology and Modeling, Part B*, Arni S. R. Srinivasa Rao und Calyampudi Radhakrishna Rao, Hrsg. Elsevier, 2019, Bd. 40, Kap. 2, S. 33–83.

[31] Charles J. Krebs, *Ecological Methodology*, 2. Aufl. Benjamin Cummings, 1998, Kap. 2.

[32] Rachel S. McCrea und Byron J. T. Morgan, *Analysis of Capture-Recapture Data*. Chapman und Hall/CRC, 2016.

[33] Lucy A. McFadden und Richard P. Binzel, *Encyclopedia of the Solar System*, Second Edition, Lucy A. McFadden, Paul R. Weissman und Torrence V. Johnson, Hrsg. Academic Press, 2007, Kap. 14, S. 283–300.

[34] Muriel Gargaud et al., *Encyclopedia of Astrobiology*, First Edition. Springer Science & Business Media, 2011, Jan. 2015.

[35] Larkin Powell und George Gale, *Estimation of Parameters for Animal Populations: a primer for the rest of us*. Caught Napping Publications, Aug. 2015.

[36] Ian Ridpath, *A Dictionary of Astronomy*, Second Edition. Oxford University Press, 2012.

[37] Rinaldo B. Schinazi, *Probability with Statistical Applications*. Birkhäuser Cham, Feb. 2022.

Internetquellen

[38] Minor Planet Center. „IAU MPC. “Adresse: https://www.minor_planetcenter.net/mpcops/documentation/provisional-designation-definition/.

[39] ESA’s Near-Earth Object Coordination Centre. „NEOCC. “Adresse: https://neo.ssa.esa.int/home.

[40] Reinhard Furrer. „(A Gentle) Introduction to Statistic. “Adresse: https://user.math.uzh.ch/vorlesungen/sta120/fs20/web/script_sta120.pdf#page38.

[41] James Barry Grand. „Lecture 06 – Multinomials and CMR Models for closed population. “Adresse: https://webhome.auburn.edu/~grandjb/wildpop/lectures/lect_06.pdf.

[42] Zakhar Kabluchko. „Skript zur Vorlesung Mathematische Statistik. “Adresse: https://www.uni-muenster.de/imperia/md/content/Stochastik/skript_statistik.pdf.

[43] Jet Propulsion Laboratory. „Keeping an Eye on Space Rocks. “Adresse: https://www.jpl.nasa.gov/keeping-an-eye-on-space-rocks/.

[44] Milani Comparetti et al. „Near Earth Objects - Dynamic Site. “Adresse: https://newton.spacedys.com/neodys/index.php?pc=0.

[45] NASA’s Center for Near-Earth Object Studies. „CNEOS. “Adresse: https://cneos.jpl.nasa.gov.

[46] Pfeifer et al. „Grundzüge der Statistischen Ökologie. “Adresse: https://uol.de/f/5/inst/mathe/personen/dietmar.pfeifer/Monographien/StatOek.pdf.

[47] Brandon Specktor. „’Planet killer’ asteroids are hiding in the sun’s glare. Can we stop them in time? “Adresse: https://www.livescience.com/space/asteroids/the-sun-is-blinding-us-to-thousands-of-potentially-lethal-asteroids-can-scientists-spot-them-before-its-too-late.

[48] Prapun Suksompong. „9 Expectation and Variance. “Adresse: https://www2.siit.tu.ac.th/prapun/ecs315_2014_1/ECS315%209%20u3.pdf.

[49] Joseph C. Watkins. „Topic 15: Maximum Likelihood Estimation. “Adresse: https://math.arizona.edu/~jwatkins/o-mle.pdf.

Erklärung zu generativer KI und KI-gestützten Technologien im Schreibprozess

Während der Vorbereitung dieser Arbeit benutzte der Autor ChatGPT 4o-mini (Gratisversion), um die Lesbarkeit und den Lesefluss des Textes zu verbessern, ohne den Inhalt zu beeinflussen, um Übersetzungen vorzunehmen und als Schreibhilfe in LaTeX. Nach der Nutzung dieses Tools hat der Autor den Inhalt überprüft und bearbeitet und übernimmt die volle Verantwortung für den Inhalt der veröffentlichten Arbeit.

Authentizitätserklärung

Ich bezeuge mit meiner Unterschrift, dass ich diese Arbeit selbstständig verfasst und erlaubte Hilfen von Drittpersonen korrekt veranschaulicht habe. Informationen, die ich wörtlich oder sinngemäss von anderen Publikationen oder Quellen, unter anderem aus dem Internet oder mithilfe künstlicher Intelligenz (KI), verfasst habe, sind klar, wiederauffindbar zitiert und machen nur einen geringen Anteil der Arbeit aus. Die in meiner Arbeit benützten Hilfsmittel entsprechen der Wahrheit, sind vollständig aufgelistet und auf diese Weise noch nicht veröffentlicht worden. Mir ist bewusst, dass ich unter Umständen zu den von mir gebrauchten Hilfsmitteln Stellung nehmen muss.

Ort und Datum: _____

Unterschrift der Schülerin: _____