Skewsamp: Fallzahlabschätzung bei schiefen Verteilungen in R

Johannes Brachem
johannes.brachem@stud.uni-goettingen.de

Dominik Strache
dominik.strache@stud.uni-goettingen.de

15. December 2021

Georg-August-Universität Göttingen

Inhaltsverzeichnis

1. Einleitung

Über viele Disziplinen hinweg ist eine der wichtigsten Methoden zum wissenschaftlichen Erkenntnisgewinn der Vergleich einer Experimental- mit einer Kontrollgruppe in einem kontrollierten, randomisierten Experiment. Über diesen Vergleich kann bspw. die Wirksamkeit eines Impfstoffs ermittelt werden: Wie häufig treten Infektionen mit einem Erreger bei Geimpften auf (Experimentalgruppe), verglichen mit Personen, die ein Placebo-Präparat erhalten (Kontrollgruppe)? In der Regel werden anhand der in einem Experiment gesammelten Daten statistische Tests durchgeführt. Damit soll ermittelt werden, ob etwaige Gruppenunterschiede statistisch signifikant sind, d.h. ob sie so groß sind, dass sie mit hoher Wahrscheinlichkeit nicht zufällig zustandegekommen sind. Je mehr Datenpunkte zur Verfügung stehen, desto sicherer werden statistische Schätzwerte und damit auch die Ergebnisse statistischer Tests. Das Risiko falsch positiver Ergebnisse (Typ-I Fehler) wird in der Regel über das α-Fehler Niveau kontrolliert und bei 5% konstant gehalten, die Anzahl der Datenpunkte wirkt sich deshalb in der Praxis in erster Linie auf das Risiko falsch negativer Ergebnisse (Typ-II Fehler) aus.

Mehr Daten sind also aus rein statistischer Sicht besser als weniger Daten - allerdings muss in den meisten Fällen bei der Datenerhebung eine Abwägung getroffen werden. In dieser Abwägung steht auf der einen Seite die Genauigkeit des statistischen Tests. Auf der anderen Seite stehen praktische und ethische Gründe dafür, die geringstmögliche sinnvolle Anzahl von Datenpunkten zu erheben. Das sind häufig finanzielle oder zeitliche Erwägungen, da Datenerhebungen mit Kosten und Aufwand verbunden und die verfügbaren Ressourcen begrenzt sind. Dazu gehört aber, bspw. in klinischen Studien, auch die Verantwortung von Forschenden, Versuchsteilnehmende nicht unnötig einem Risiko auszusetzen. Lässt sich bspw. die Wirkung eines neuen Medikaments bereits mit 100 Versuchsteilnehmenden gut untersuchen, so sollten nicht unnötigerweise 150 Personen in eine Studie aufgenommen und damit dem Risiko von Nebenwirkungen ausgesetzt werden.

Um solche Abwägungen formal zu unterfüttern ist es gute Praxis, als Teil der Versuchsplanung eine *a priori*-Power-Analyse durchzuführen. Damit wird die Anzahl von Datenpunkten ermittelt, die für das Erreichen einer bestimmten *Power*, auch *Teststärke* genannt, erforderlich ist (Lakens, 2021). Die Power eines Tests ist im Kontext des Experiments die Wahrscheinlichkeit, dass ein real bestehender Gruppenunterschied tatsächlich gefunden wird. Eine hohe Power ist gleichbedeutet mit einem niedrigen Risiko für einen Typ-II-Fehler. Die im Einzelfall angewandte Methodik zur Fallzahlabschätzung ist abhängig von den untersuchten Daten und dem geplanten statistischen Test.

Die Durchführung einer a prior Power-Analyse kann auch für Fachleute eine komplexe Aufgabe sein, doch für viele häufig anzutreffende Konstellationen gibt es gut

anwendbare und gut dokumentierte Hilfsmittel, bspw. das Programm G*Power (Faul et al., 2007) oder das R-Paket {pwr} (Champely, 2020). Nicht abgedeckt sind in den breit verfügbaren Hilfsmitteln allerdings Fälle, in denen die zugrundeliegenden Daten eines Gruppenvergleichs potentiell *schiefen* Verteilungen, wie z.B. der Gamma-Verteilung, folgen. Solche Verteilungen findet man bspw., wenn Messwerte ausschließlich positiv, aber mit hoher Wahrscheinlich sehr niedrig sind, z.B. wenn man das Gewicht von leichten Gegenständen misst. Die Verteilung der Daten muss bei der Auswahl des statistischen Tests und damit auch bei der Fallzahlabschätzung berücksichtigt werden, um das Risiko falscher Schlussfolgerungen und ineffizienter Experimentdesigns minimal zu halten.

Im Fokus dieses Berichts stehen zwei Ansätze zur Fallzahlabschätzung bei schiefen Verteilungen: Ein parametrischer Ansatz für generalisierte lineare Modelle (GLM) (Cundill & Alexander, 2015) und ein nichtparametrischer Ansatz für den Wilcoxon-Mann-Whitney-Test (Chakraborti et al., 2006). Wir stellen in Abschnitt 2 die beiden Ansätze vor und beschreiben anschließend in Abschnitt 3 unser R-Paket {skewsamp}, mit dem wir beide Ansätze in einem getesteten und dokumentierten Open-Source-Paket mit nutzerfreundlichem Interface einfacher zugänglich machen. In Abschnitt 4 stellen wir Simulationsstudien vor, in denen wir unter Verwendung unseres Pakets die jeweiligen Originalbefunde zur Evaluation der beiden Ansätze replizieren und ihre Robustheit näher auf die Probe stellen. Schließlich ziehen wir in Abschnitt 5 ein Fazit.

2. Theorie

In diesem Abschnitt geben wir zunächst einen Überblick über die relevanten Konzepte und stellen dann die einzelnen Verfahren zur Fallzahlabschätzung näher vor.

2.1. Relevante Konzepte

Von zentraler Bedeutung in der Fallzahlabschätzung sind die Effektstärke δ, die Wahrscheinlichkeit eines Typ-I-Fehlers (falsch positiv) α, die Power κ und natürlich die Fallzahl oder Stichprobengröße N[1]. Über die Effektstärke wird der Unterschied zwischen Kontroll- und Experimentalgruppe quantifiziert. Auf sie sind die Hypothesen des statistischen Tests gerichtet, im Standardfall die Nullhypothese $H_0 : \delta = 0$ (kein Unterschied) und die Alternativhypothese $H_1 : \delta \neq 0$ (Unterschied ist nicht 0). Für den eigentlichen Test wird aufgrund der verfügbaren Daten eine Teststatistik $T_{\hat{\delta}}$ für die geschätzte Effektstärke $\hat{\delta}$ berechnet, deren Verteilung unter Annahme der H_0 bekannt ist. Liegt die Teststatistik innerhalb eines zuvor definierten Ablehnungsbereichs B, wird

[1] Um Unklarheiten zu vermeiden, bezeichnen wir in dieser Arbeit die Fallzahl in einer einzelnen Gruppe als n und die Fallzahl für die Studie insgesamt als N.

die Nullhypothese verworfen. Das α-Niveau ist definiert als die Wahrscheinlichkeit, die H_0 fälschlicherweise zu verwerfen, formal:

$$\alpha = P(T_{\hat{\delta}} \in B \mid H_0).$$

In ähnlicher Weise ist die Wahrscheinlichkeit eines Typ-II-Fehlers β definiert als die Wahrscheinlichkeit, die H_0 fälschlicherweise *nicht* zu verwerfen:

$$\beta = P(T_{\hat{\delta}} \notin B \mid H_1).$$

Die Power κ lässt sich direkt daraus ableiten als die Wahrscheinlichkeit, die H_0 korrekterweise zu verwerfen:

$$\begin{aligned} \kappa &= 1 - \beta \\ &= P(T_{\hat{\delta}} \in B \mid H_1). \end{aligned}$$

Die Effektstärke, Fallzahl, Power und das Alpha-Niveau hängen bei statistischen Tests so zusammen, dass jeweils eine der Größen aus den drei übrigen berechnet werden kann. Zur Fallzahlabschätzung müssen Forschende deshalb festlegen, welche Typ-I-Fehler-Wahrscheinlichkeit α sie tolerieren möchten und mit welcher Wahrscheinlichkeit κ sie eine richtig-positive Testentscheidung anstreben, wenn die wahre Effektstärke mindestens δ ist:

$$N = f(\alpha, \kappa, \delta)$$

Die genaue Form dieser Berechnung unterscheidet sich je nach der Form der Teststatistik, dabei finden sich aber immer gewisse Regelmäßigkeiten: (1) Je höher das tolerierte Risiko von falsch-positiven Befunden α gewählt wird, desto weniger Datenpunkte werden benötigt, (2) je geringer die angestrebte Power κ, desto weniger Datenpunkte werden benötigt und (3) je größer die Effektstärke δ, für die die angegebene Power angestrebt wird, desto weniger Datenpunkte werden benötigt.

2.2. Fallzahlabschätzung in generalisierten linearen Modellen

Ein mächtiges Instrument zur Analyse schief verteilter Daten ist das generalisierte lineare Modell (GLM), das die Aufstellung von Regressionsmodellen für Messwerte aus Verteilungen der Exponentialfamilie erlaubt. Dabei wird der Erwartungswert der Messwerte $E(y_i) = \mu_i$, $i = 1, \ldots, N$ über eine *link*-Funktion g mit einem linearen Maß η_i in Beziehung gesetzt, so dass $\eta_i = g(\mu_i)$. Mithilfe dieser link-Funktion wird ein lineares Regressionsmodell aufgestellt, das im Falle eines einfachen Vergleichs einer Kontroll- mit einer Experimentalgruppe mit einer einzigen Prädiktorvariable x_i auskommt, die

als Indikatorvariable die Gruppenzugehörigkeit anzeigt:

$$\eta_i = \beta_0 + \beta_1 x_i.$$

Effektstärke

Üblicherweise wird die Kontrollgruppe als $x_i = 0$ kodiert, so dass der Gruppenunterschied in der transformierten Größe η_i über β_1 gegeben ist:

$$\beta_1 = g(\mu_1) - g(\mu_0), \tag{1}$$

wobei μ_1 den Mittelwert in der Experimental- und μ_0 den Mittelwert in der Kontrollgruppe bezeichnet. Der Regressionskoeffizient β_1 ist somit ein Maß der Effektstärke und kann häufig in gut interpretierbare Größen transformiert werden. So kann für den häufig verwendeten log-Link das Maß $\delta = 1 - \frac{\mu_1}{\mu_0}$ genutzt werden. Hier gibt δ an, um welchen Anteil die Messgröße in der Experimentalgruppe im Vergleich zur Kontrollgruppe zurückgegangen (bei $\delta > 0$) oder gestiegen (bei $\delta < 0$) ist. Eine Effektsärke von $\delta = 0.3$ bedeutet bspw., dass ein Messwert aus der Experimentalgruppe im Mittel um 30% kleiner ist, als ein Messwert aus der Kontrollgruppe. Cundill & Alexander (2015) bezeichnen dieses Maß als Wirksamkeit (*efficacy*), da sich ihre Arbeit in erster Linie auf die Wirksamkeit einer in der Experimentalgruppe angewendeten Intervention bezieht.

Der t-Test für die Nullhypothese $H_0 : \beta_1 = 0$ ist zugleich ein Test für die $H_0 : \delta = 0$. Diese Beziehung kommt wie folgt zustande. Bei Einsetzen der log-Linkfunktion in Gleichung 1 ergibt sich

$$\beta_1 = log(\mu_1) - log(\mu_0) = log\left(\frac{\mu_1}{\mu_0}\right)$$

und somit $e^{\beta_1} = \frac{\mu_1}{\mu_0}$. Durch einsetzen in die Formel für δ erhalten wir

$$\delta = 1 - e^{\beta_1}.$$

Ist also $\beta_1 = 0$, dann wird $e^{\beta_1} = 1$ und somit $\delta = 0$. Eine analoge Beziehung gilt für den logit-Link im Fall binomial-verteilter Daten. Die test-Statistik für $H_0 : \beta_1 = 0$ ist $\hat{\beta}_1 / \sqrt{\widehat{Var}(\hat{\beta}_1)}$ und folgt unter der Annahme der Nullhypothese einer t-Verteilung und asymptotisch einer Standard-Normalverteilung.

Fallzahlabschätzung

Cundill & Alexander (2015) stellen auf der Grundlage des generalisierten linearen Modells und aufbauend auf der Arbeit von Lachin (1981) zwei leicht unterschiedliche

allgemeine Formeln zur Fallzahlabschätzung zur Verfügung. Von Lachin übernehmen die Autoren die Annahme, dass die *Teststatistik* - nicht die Daten - normalverteilt ist, und dass die Varianz der Teststatistik unter Geltung von H_0 und H_1 gleich ist. Wir arbeiten mit der von den Autoren empfohlenen Variante:

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}\frac{V(\mu_1)}{(d\mu/d\eta|_{\mu=\mu_1})^2} + \frac{1}{Q_0}\frac{V(\mu_0)}{(d\mu/d\eta|_{\mu=\mu_0})^2}}}{g(\mu_0) - g(\mu_1)}. \tag{2}$$

Q_0 und Q_1 bezeichnen die relative Größe der Kontroll- bzw. Experimentalgruppe gemessen an der gesamten Stichprobe. Analog bezeichnen μ_0 und μ_1 die Erwartungswerte beider Gruppen. Z_a ist das a-Quantil der Standardnormalverteilung, $V(\mu)$ die Varianz in Abhängigkeit des Erwartungswerts. Unter Berücksichtigung von Gleichung 1 kann der Nenner auch als $-\beta_1$ geschrieben werden. So wird sichtbar, dass das α-Niveau und die Power $1-\beta$ über die Z-Quantile und die Effektstärke durch die Differenz im Nenner in Gleichung 2 einfließen. Die ermittelte Fallzahl N ist hier die Gesamtgröße der Stichprobe. Da Gleichung 2 eine flexible Festlegung der gewünschten Größenverhältnisse zwischen den beiden Gruppen ermöglicht, muss zur Bestimmung der Größe der einzelnen Gruppen die jeweilige relative Größe Q_0, bzw. Q_1 mit der Gesamtgröße mutipliziert werden.

Die Autoren stellen ausformulierte Gleichungen für vier Verteilungen aus der Exponentialfamilie zur Verfügung, im Einzelnen sind das die Poisson-, die Binomial-, die negative Binomial- und die Gamma-Verteilung. Wir gehen exemplarisch näher auf die Herleitung der Fallzahlabschätzung für die Poisson-Verteilung aus Gleichung 2 ein und geben dann die übrigen drei Varianten an.

Poisson-verteilte Daten

Als Link-Funktion g wird der Logarithmus genutzt, um die Positivität des Erwartungswerts zu gewährleisten, d.h. $\eta_j = log(\mu_j)$ mit $j = 0, 1$. Da bei der Poisson-Verteilung die Varianz dem Erwartungswert entspricht, gilt $V(\mu) = \mu$, sodass in der allgemeinen Formel für beide Gruppen jeweils $V(\mu_j)$ durch μ_j ersetzt werden kann. Der Ableitungsterm $d\mu/d\eta$ lässt sich mit Substitution folgendermaßen umformen:

$$\frac{d}{d\,\eta}\mu = \frac{d}{d\,\eta}e^{\eta} = e^{\eta} = \mu$$

Somit ergibt sich für die allgemeinen Terme im Zähler von Gleichung 2

$$\frac{V(\mu_j)}{(d\mu/d\eta\,|_{\mu=\mu_j})^2} = \frac{\mu_j}{\mu_j^2} = \frac{1}{\mu_j}$$

und insgesamt folgende Formel zur GLM-basierten Fallzahlabschätzung für poisson-verteilte Daten:

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{Q_1\mu_1} + \frac{1}{Q_0\mu_0}}}{log(\mu_0) - log(\mu_1)}.$$

Negativ-binomial, gamma- und binomial-verteile Daten

Die negative Binomialverteilung hat als Generalisierung der Poisson-Verteilung einen zusätzlichen Dispersionsparameter k, der in der Fallzahlabschätzung in $V(\mu) = \mu + \frac{\mu^2}{k}$ zum Tragen kommt. Mit dem üblichen log-Link bleibt es bei $d\mu/d\eta = \mu$, so dass die Fallzahlabschätzung insgesamt durch

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{Q_1}(\frac{1}{\mu_1} + \frac{1}{k_1}) + \frac{1}{Q_0}(\frac{1}{\mu_0} + \frac{1}{k_0})}}{log(\mu_0) - log(\mu_1)}$$

ausgedrückt werden kann.

Für die Gammaverteilung verwenden Cundill & Alexander (2015) ebenfalls den log-Link, so dass weiterhein $d\mu/d\eta = \mu$ gilt. Die Autoren wählen außerdem eine Parametrisierung der Gamma-Verteilung auf Basis des Erwartungswerts mit Formparameter κ und Skalenparameter $\frac{\mu}{\kappa}$. Da so $V(\mu) = \mu^2/\kappa$, ergibt sich als Fallzahlabschätzung:

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{Q_1\kappa_1} + \frac{1}{Q_0\kappa_0}}}{log(\mu_0) - log(\mu_1)}.$$

Für die Binomialverteilung mit der Anzahl an Versuchen d und Erfolgswahrscheinlichkeit p ergibt sich bei Verwendung des kanonischen logit-Links

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta})\sqrt{\frac{1}{Q_1 p_1(1-p_1)} + \frac{1}{Q_0 p_0(1-p_0)}}}{\sqrt{d}\,(logit(p_0) - logit(p_1))},$$

da $d\mu/d\eta = \mu(1-\mu)$.

2.3. Fallzahlabschätzung für den Wilcoxon-Mann-Whitney-Test

Ein Nachteil des GLM-basierten Ansatzes ist, dass die Verteilung der untersuchten Daten und bestimmte Parameter bekannt sein müssen, bspw. der Form-Parameter im Fall der Gamma-Verteilung. Der Wilcoxon-Mann-Whitney Test (WMW-Test) bietet die Möglichkeit einer anderen Herangehensweise an das Problem schief verteilter Daten durch den Verzicht auf die Annahme konkreter theoretischer Verteilungen. Erforderliche Annahmen für den zwei-Stichproben WMW-Test sind, dass die untersuchten Variablen

ordinal skaliert sind, d.h. ihrer Größe nach geordnet werden können, und dass die beiden untersuchten Stichproben voneinander unabhängig sind. In einer allgemeinen Formulierung wird im WMW-Test folgende Nullhypothese über die unabhängigen Zufallsvariablen X und Y untersucht:

$$H_0 : P(X < Y) = P(X > Y),$$

d.h. es ist gleich wahrscheinlich, dass eine Realisation von X größer oder kleiner ist als eine Realisation von Y. Wird die Nullhypothese, verworfen kann angenommen werden, dass Realisationen einer der beiden Zufallsvariablen tendenziell größer sind, als Realisationen der anderen. Im Experimentalkontext könnte das bedeuten, dass Beobachtungen aus der Experimentalgruppe tendenziell höhere (oder niedrigere) Werte aufweisen, als Beobachtungen aus der Kontrollgruppe. Die Teststatistik wird wie folgt gebildet. Nehmen wir zwei unabhängige Stichproben $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ an. Dann wird zunächst jeder Wert x_i mit allen Werten y_j verglichen, indem wir zählen, für wie viele Werte $x_i > y_i$ gilt. Dazuaddiert wird die halbe Anzahl von Gleichheit beider Werte:

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I(x_i > y_i) + 0.5 \sum_{i=1}^{m} \sum_{j=1}^{n} I(x_i = y_i).$$

In der Formel ist I die Indikatorfunktion. Die Summe U geteilt durch die Anzahl der Vergleiche mn ist die Teststatistik und asymptotsich normalverteilt.

Fallzahlabschätzung nach Noether

Noether (1987) liefert eine Vorgehensweise zur Fallzahlabschätzung auf Basis der Wahrscheinlichkeit $p = P(X < Y)$ als Effektstärke:

$$N = \frac{[Z_\alpha + Z_\beta]^2}{12c(1-c)(p-0.5)^2}, \tag{3}$$

wobei Z_α und Z_β die jeweiligen Quantile der Standardnormalverteilung für die Typ-I-Fehlerrate α und die Typ-II-Fehlerrate β bezeichnen. Die Konstante c gibt über $n_X = cN$ an, wie groß die X-Stichprobe gemessen an der Gesamtstichprobe ist. Bei Annahme gleicher Stichprobengrößen in beiden Gruppen eines Experiments vereinfacht sich daher der Nenner der Gleichung zu $3(p-0.5)^2$. Zu beachten ist, dass Noether hier die Fallzahl für einen *einseitigen* Test abschätzt. Die Quantile der Standardnormalverteilung sind Teil der Berechnung, weil die Test-Statistik des WMW-Tests - nicht also die untersuchten Variablen - asymptotisch normalverteilt ist. Noether vereinfacht seine Berechnung, indem er auf die Berücksichtigung gleicher Werte in der Bildung der Teststatistik verzichtet.

Mithilfe von Gleichung 3 kann die Fallzahl für den WMW-Test abgeschätzt werden,

wenn eine sinnvolle Zielgröße p definiert werden kann. Für viele Forschungsfragen mit Gruppenvergleichen ist es allerdings nicht nur wichtig, *ob* Beobachtungen aus einer Gruppe dazu tendieren, größer zu sein als Beobachtungen aus einer anderen Gruppe, sondern auch welches *Ausmaß* Größenunterschiede tendenziell haben. Eine solche Fragestellung kann ohne Zusatzannahmen kaum über das Maß p ausgedrückt werden. Deshalb wird für den WMW-Test in manchen Fällen die Zusatzannahme getroffen, dass die Verteilungsfunktionen der Daten in der Kontroll- und der Experimentalgruppe die gleiche Form haben und sich nur über eine Verschiebung δ unterscheiden. Bezeichnen wir die der Kontrollgruppe zugrundeliegende Verteilunsgfunktion als F_X und die der Experimentalgruppe zurgrundeliegende Verteilungsfunktion als F_Y, dann kann die Annahme formal als $F_Y(x) = F_X(x - \delta)$ ausgedrückt werden. Der Test fällt dann in das *location shift* Paradigma und mit der als $H_0 : \delta = 0$ formulierbaren Nullhypothese ergibt sich ein Test auf Gleichheit der Mediane.

Location shift Fallzahlabschätzung nach Chakraborti et al.

Chakraborti et al. (2006) präsentieren und testen drei Vorgehensweisen zur Fallzahlabschätzung für den WMW-Test im *location shift* Paradigma. Sie stützen sich dabei auf vorherige Arbeit von Hamilton & Collings (1991). Wir beziehen uns ausschließlich auf den von Chakraborti et al. (2006) so betitelten NECDF-Schätzer (Noether Empirical Cumulative Density Function), da dieser Schätzer im Bericht der Autoren die größte Genauigkeit und die beste Performance zeigte.

Wir fassen das Vorgehen zunächst kurz zusammen und beschreiben dann Teilschritte in größerem Detail. Das Verfahren erfordert die Verfügbarkeit von Vorinformationen über die zu erwartenden Daten aus jeder der zu untersuchenden Versuchsbedingungen. Chakraborti et al. (2006) nehmen als konkretes Szenario für solche Vorinformationen an, dass Daten aus einem Pilot-Experiment mit kleiner Stichprobengröße, also zwei kleinen Pilot-Stichproben zur Verfügung stehen. Dieses Szenario ist von substanzieller praktischer Relevanz, da die Pilotierung mit kleinen Stichproben ein wichtiger und häufiger Bestandteil der Versuchsplanung ist. Wir bezeichnen die Stichprobe aus der Kontrollgruppe als $X_1, \ldots, X_{m_X}$ und die Stichprobe aus der Experimentalgruppe als $Y_1, \ldots, Y_{m_Y}$[2].

Für jede von diesen zwei Pilot-Stichproben werden nun zwei geglättete empirische Verteilungsfunktionen (Empirical Cumulative Density Function, ECDF) konstruiert, d.h. $G_X(x)$ und $G_X(x - \delta)$, sowie $G_Y(y)$ und $G_Y(y - \delta)$. Abbildung 1 zeigt Beispiele solcher ECDFs, die exakte Konstruktion beschreiben wir in Anhang A. Da diese Funktionen

[2] Chakraborti et al. (2006) arbeiten zur Vereinfachung der Notation mit gleichen Stichprobengrößen in den Pilotstichproben. Wir machen hier lediglich deutlich, dass unterschiedlich große Stichproben ebenso möglich sind.

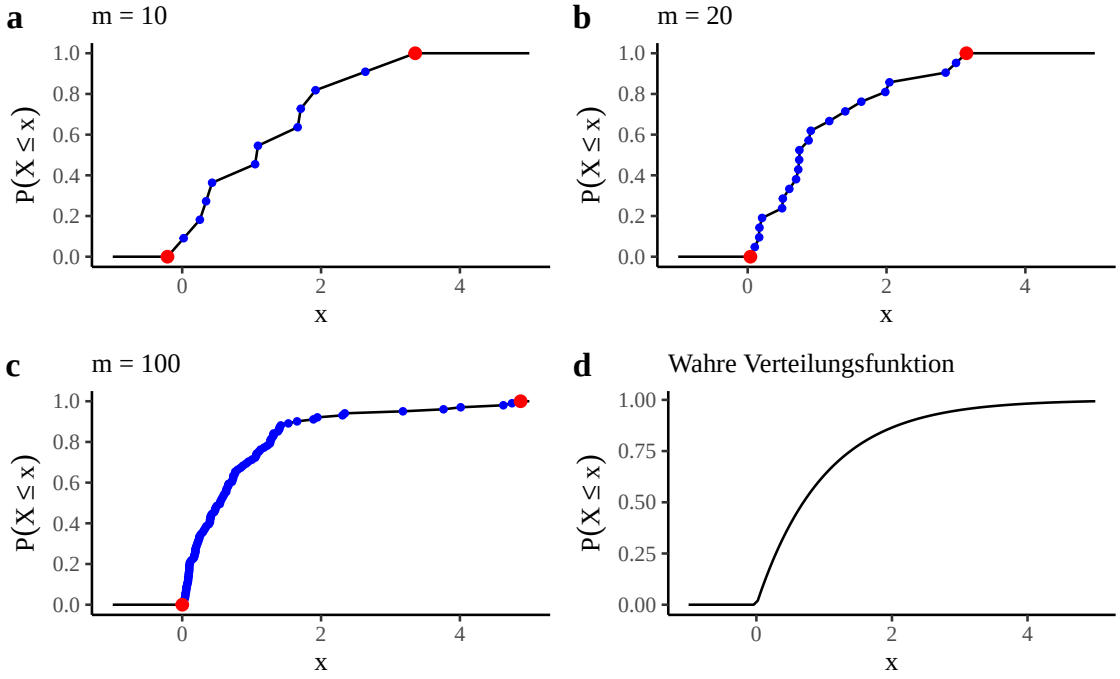

Abbildung 1 Empirische Verteilungsfunktionen auf Basis unterschiedlich großer Stichproben der Größe m aus einer Exponentialverteilung mit $\lambda = 1$ (a-c) und wahre Verteilungsfunktion (d). Die blauen Punkte sind beobachtete Datenpunkte aus den Stichproben, die roten Punkte sind extrapolierte Endpunkte.

abschnittsweise linear sind, kann eine Schätzung von p über Integration analytisch bestimmt werden. Die Schätzung für p auf Basis der X-Stichprobe ist

$$\hat{p}_X = \int G_X(x) dG_X(x - \delta).$$

Auf Basis der Y-Stichprobe wird analog $\hat{p}_Y$ geschätzt. Die Herleitung zur Implementierung dieses Integrals beschreiben wir in Anhang A. Durch Einsetzen der beiden Schätzwerte in Gleichung 3 erhalten wir zwei Schätzungen $\hat{N}_Y$ und $\hat{N}_X$ für die erforderliche Fallzahl und bilden das gewichtete Mittel

$$\hat{N}_{NECDF} = \frac{m_X \times \hat{N}_X + m_Y \times \hat{N}_Y}{m_X + m_Y} \tag{4}$$

als finalen Schätzwert für die erforderliche Gesamtanzahl an Beobachtungen Da Chakraborti et al. (2006) in Gleichung 3 gleiche Gruppengrößen im Zielexperiment, also $c = 0.5$, annehmen und den Term zusätzlich durch Zwei teilen, erhalten sie einen Schätzer für die Größe *einer* Teilstichprobe. Wir bleiben hier bei der allgemeineren Formulierung, die die Spezifikation unterschiedlich großer Teilstichproben erlaubt.

Obergrenze durch Resampling

In einer Simulationsstudie ermitteln Chakraborti et al. (2006) für Daten aus vier verschiedenen Verteilungen die Genauigkeit ihres Verfahrens, indem sie jeweils 200 Paare von Pilotstichproben mit jeweils gleicher Größe $m = 10$ oder $m = 20$ aus einer bekannten Verteilung ziehen und den $\hat{N}_{ECDF}$-Schätzer ermitteln. Der Schätzer liefert in den von den Autoren berichteten Fällen im Mittel gute Fallzahlabschätzungen, weist allerdings aufgrund der Abhängigkeit von kleinen, zufällig gezogenen Pilot-Stichproben eine große Varianz auf. Die Varianz der Schätzung wird erwartbar mit steigender Größe der Pilotstichproben geringer und die Genauigkeit steigt. Das Erheben großer Pilotstichproben zur Verbesserung der Schätzung konterkariert allerdings in gewisser Weise den eigentlichen Zweck der Fallzahlabschätzung, wie auch Chakraborti et al. (2006) anmerken. Um auch angesichts kleiner Pilotstichproben das Risiko einer Unterschätzung der Fallzahl zu verringern, stellen die Autoren eine Resampling-Methode zur Abschätzung einer Obergrenze vor. Das Vorgehen ist dabei wie folgt:

1. Aus der Pilot-Stichprobe $X_1, \ldots, X_{n_X}$ wird wie gehabt die ECDF G_X konstruiert.
2. Aus G_X werden N_{sim} weitere Pilot-Stichproben der gleichen Größe gezogen.
3. Für jede dieser simulierten Pilot-Stichproben wird jeweils erneut die empirische Verteilungsfunktion $G_{Xi}, i = 1, \ldots, N_{sim}$ aufgestellt.
4. Anhand dieser ECDFs werden N_{sim} Schätzungen $\hat{p}_{Xi}$ errechnet und zur Fallzahlabschätzung in Gleichung 3 eingesetzt, um die Schätzungen $\hat{N}_{Xi}$ zu erhalten.
5. Die Schritte 1-4 werden analog mit der zweiten Pilot-Stichprobe durchgeführt, so dass N_{sim} Fallzahlabschätzungen $\hat{N}_{NECDF,i}$ nach Gleichung 4 gebildet werden können.

Als Obergrenze der benötigten Fallzahl verwenden die Autoren das 90%-Quantil der so erhaltenen Verteilung von N_{sim} Fallzahlabschätzungen. Wir bezeichnen diese Schätzung der Obergrenze im Folgenden als $\hat{N}^{(90)}$.

2.4. Vergleich beider Verfahren

Die beiden vorgestellten Verfahren zur Fallzahlabschätzung bei schiefen Verteilungen verfolgen grundsätzlich andere Ansätze mit unterschiedlichen Vor- und Nachteilen.

Bei der generalisierten Regression zeigen die Autoren in ihrer Simulationsstudie, dass die von ihnen berechneten Fallzahlen sehr genau zu der vorher festgelegten Power führen. Bei Kenntnis über die exakten Verteilungen ist die generalisierte Regression somit Fallzahlabschätzungen unter Normalverteilungsannahme klar überlegen.

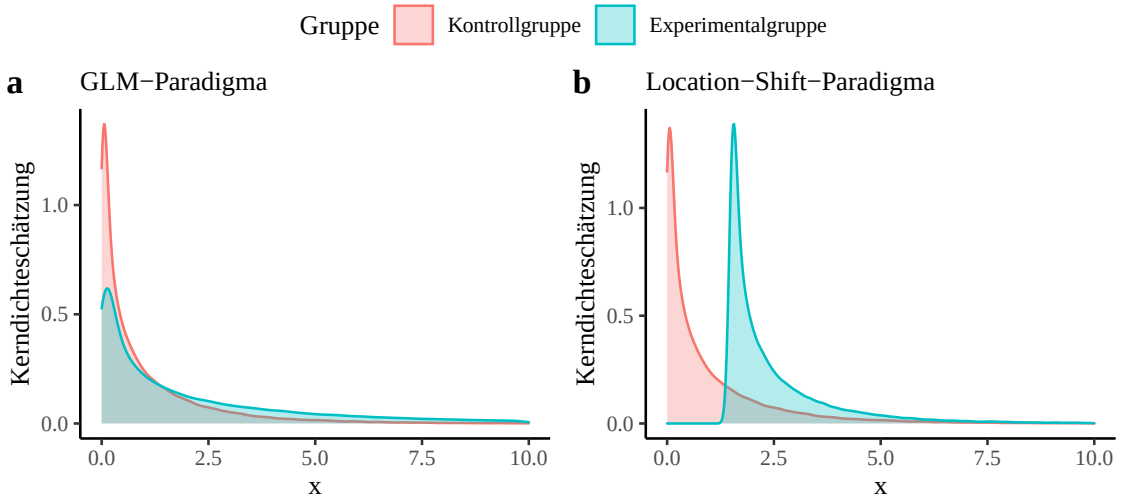

Abbildung 2 Gegenüberstellung der parametrischen Modellierung und nichtparametrischen *location shift* Modellierung des Vergleichs einer Kontroll- und einer Experimentalgruppe. Ausgangspunkt ist in beiden Fällen in der Kontrollgruppe eine Gamma-Verteilung mit Formparameter $\kappa = 0.5$ und Skalenparameter $\theta = \frac{\mu_0}{\kappa}$, wobei $\mu_0 = 1$. In beiden Grafiken beträgt der Mittelwertsunterschied zwischen Kontroll- und Experimentalgruppe $\mu_1 - \mu_0 = 1.5$. In Teilabbildung **a** werden die Daten der Experimentalgruppe aus einer Gamma-Verteilung mit gleichem Formparameter und anderem Mittelwert $\mu_1 = 2.5$ gezogen. In Teilabbildung **b** werden die Daten der Experimentalgruppe aus derselben Verteilung gezogen wie die Daten der Kontrollgruppe und anschließend um 1.5 verschoben.

Für die Anwendung des GLM-Verfahrens müssen Forschende im Vorhinein definieren, mit welcher Verteilung sie ihre Daten modellieren möchten. Das ist nur dann möglich, wenn Informationen verfügbar sind, die eine solche Verteilungsannahme erlauben. Der nichtparametrische *location shift* Ansatz scheint in dieser Hinsicht flexibler zu sein: Die Notwendigkeit, vorab Informationen über die Daten einzuholen, ist durch die Verwendung von Pilot-Stichproben ein integraler Bestandteil des Verfahrens. Problematisch ist dabei allerdings, dass eine starke Abhängigkeit von den Pilotstichproben entsteht und die Fallzahlabschätzung nach Chakraborti et al. (2006) in der Folge eine große Varianz aufweist. So entsteht für Forschende Unsicherheit darüber, ob die ermittelte Fallzahl tatsächlich geeignet ist, die angestrebte Power zu realisieren.

Der parametrische Ansatz weist dieses Problem nicht auf. Wie die Simulationen von Cundill & Alexander (2015) zeigen, wird die angestrebte Power für die untersuchten Szenarien erreicht.

Der *location shift* Ansatz ist zudem nicht so flexibel, wie das Etikett der Freiheit von einer spezifischen theoretischen Verteilungsannahme womöglich suggeriert. In der Tat ist die Annahme $F_X(x) = F_Y(x - \delta)$ ebenfalls eine starke Verteilungsannahme, die unter anderem auch Gleichheit der Varianzen in beiden Gruppen erfordert. Gerade bei schief verteilten Daten ist eine Verletzung dieser Annahme in der Praxis nicht unwahrscheinlich. Im GLM-basierten Ansatz kann eine veränderte Verteilunsform modelliert werden; gleichzeitig kann die Situation einer bloßen Verschiebung bei schiefen Ver-

teilungen im GLM-Ansatz weniger gut abgebildet werden. In Abbildung 2 sind die unterschiedlichen Verteilungsannahmen illustriert. Aufgrund der Anwendung unter deutlich verschiedenen Annahmen können die Verfahren unserer Ansicht nach nicht als direkte Alternativen füreinander betrachtet werden. Forschende sollten sich zur Fallzahlabschätzung allerdings damit befassen, welches Paradigma für die jeweilige Fragestellung sinnvoll ist, und was die Stärken und Schwächen der beiden Verfahren sind.

3. R-Paket skewsamp

Unser R-Paket {`skewsamp`} macht beide Verfahren zur Fallzahlabschätzung einfacher und sicherer als bisher zugänglich. Cundill & Alexander (2015) stellen ein R-Skript frei zur Nutzung zur Verfügung, das grundsätzlich die Anwendung ihres Verfahrens ermöglicht[3]. Der Code ist allerdings nicht durch automatische Tests abgesichert und abgesehen von kurzen Instruktionen undokumentiert. In unseren Testläufen unter R 4.1 funktionierte außerdem die Anwendung der Funktion auf binomial-verteilte Daten nicht. Chakraborti et al. (2006) haben ihr Verfahren in der proprietären Programmiersprache Mathematica umgesetzt. Die Funktionalität ist nicht direkt verfügbar; die Autoren erklären allerdings ihre Bereitschaft, sie auf Anfrage zur Verfügung zu stellen.

Unser R-Paket ist mit automatischen Unit-Tests ausgestattet und die Funktionen sind dokumentiert, so dass die Anwendung sicher und verständlich ist. Als detailliertere Dokumentation dient auch dieser Bericht. In Anhang **??** ist die PDF-Version der Paketdokumentation verfügbar. Mit 56 automatischen Tests erreichen wir eine Code-Abdeckung von ca. 92%. Dieser Indikator allein ist natürlich kein hinreichendes Kriterium, um die korrekte Funktionsweise des Codes zu garantieren. Durch unsere Simulationsstudien (siehe nächster Abschnitt) ist die Evidenz für die Korrektheit unserer Implementierungen allerdings bestechend.

Durch die Implementierung in der kostenfreien Open-Source-Programmiersprache R besteht potentiell eine sehr gute Verfügbarkeit für Forschende. Das Paket ist nicht von dritten R-Paketen abhängig und modular über tendenziell kurze Funktionen mit hoher Kohäsion aufgebaut. Durch diese Design-Entscheidungen erreichen wir eine hohe Wartbarkeit. Das Paket kann über die Datei anhand des Codes in Codeblock 1 installiert werden:

In R kann über das Kürzel `?name` auf die Dokumentation einzelner Funktionen zugegriffen werden, bspw. `?n_gamma`. Im Folgenden geben wir einen Überblick über die

[3] https://static-content.springer.com/esm/art%3A10.1186%2Fs12874-015-0023-0/MediaObjects/12874_2015_23_MOESM3_ESM.zip

implementierten Funktionen und erklären beispielhaft die Anwendung.

3.1. Fallzahlabschätzung in generalisierten linearen Modellen

Für die Fallzahlabschätzung in generalisierten linearen Modellen haben wir das Verfahren für alle vier von Cundill & Alexander (2015) vorgestellten Verteilungen implementiert; Tabelle 1 zeigt eine Übersicht der Funktionsnamen. In Codeblock 2 ist ein Beispiel-Input und -Output für das Beispiel der Gamma-Funktion dargestellt. Um Missverständnisse bei der Anwendung zu vermeiden haben wir ein besonderes Augenmerk auf die Darstellung wichtiger Merkmale im Output gelegt. So zeigt der Output standardmäßig die Gesamtgröße der Stichprobe und die Größe der einzelnen Gruppen, die eingegebene Effektstärke und Art der Effektstärke, das verwendete α-Niveau und die Ziel-Power an. Auch die Richtung des Tests (ein-, oder zweiseitig) wird explizit angegeben. Aufgrund dieses ausführlichen Outputs haben wir uns dafür entschieden, für größere Nutzerfreundlichkeit default-Werte für das α-Niveau (0.05) und die Power (90%) zu definieren. Die Funktionen `n_gamma` und `n_negbinom` erlauben die Festlegung des Form-, bzw. Dispersionsparameters getrennt für beide Gruppen, da auch die zugrundeliegende Formeln die Trennung erlauben. Da in der Anwendung in generalisierten linearen Modellen allerdings in der Regel eine über beide Gruppen konstante Form, bzw. Dispersion angenommen wird, akzeptiert die Funktion auch die Spezifikation der Parameter nur für die Kontrollgruppe und nimmt in diesem Fall konstante Werte in beiden Gruppen an.

Fortgeschrittene Nutzer können die abstraktere Funktion `n_glm` nutzen, um eine Fallzahlabschätzung für eine beliebige Verteilung der Exponentialfamilie zu erhalten.

3.2. Fallzahlabschätzung für den Wilcoxon-Mann-Whitney-Test

Für die Fallzahlabschätzung im Wilcoxon-Mann-Whitney Test unter Location-Shift Paradigma stellt unser Paket über die Funktion `n_locshift` den NECDF-Schätzer für die erforderliche Fallzahl auf Basis zweier Pilot-Stichproben zur Verfügung. Codeblock 3 zeigt einen Beispielinput und -output. Das Paket erlaubt außerdem die Ermittlung einer geschätzten Obergrenze der Fallzahl anhand des von Chakraborti et al. (2006) vorgestellten Resampling-Verfahrens in der Funktion `n_locshift_bound`. Die Funktion gibt standardmäßig das 90%-Quantil der per Resampling ermittelten Schätzwerte

Codeblock 1. Installation von skewsamp aus GitHub.

```
# install.packages("devtools") # the package devtools is required
devtools::install_github("https://github.com/jobrachem/skewsamp")
```

Tabelle 1 Übersicht über implementierte, an Nutzer gerichtete Funktionen im Paket {skewsamp}.

Funktion	Beschreibung
Fallzahlabschätzung für generalisierte lineare Modelle	
n_poisson	Fallzahlabschätzung für poisson-verteilte Daten
n_negbinom	Fallzahlabschätzung für negativ-binomial-verteilte Daten
n_gamma	Fallzahlabschätzung für gamma-verteilte Daten
n_binom	Fallzahlabschätzung für binomial-verteilte Daten
n_glm	Generelle Implementierung der Fallzahlabschätzung für generalisierte lineare Modelle für fortgeschrittene Nutzerinnen und Nutzer
Fallzahlabschätzung für Location Shift im WMW-Test	
n_locshift	NECDF-Fallzahlabschätzung für Location Shift im WMW-Test
n_locshift_bound	Aus Resampling geschätzte Obergrenze für die erforderliche Fallzahl
resample_n_locshift	Vektor aus NECDF-Fallzahlabschätzungen basierend auf Resamples aus Pilot-Stichproben
Empirische Verteilungsfunktion	
pemp	Empirische Verteilungsfunktion
demp	Empirische Dichtefunktion
qemp	Empirische Quantilsfunktion
remp	Funktion zum ziehen zufälliger Werte aus der empirischen Verteilungsfunktion

Codeblock 2. Beispiel-Input und -Output der {skewsamp}-Funktion n_gamma.

```
>skewsamp::n_gamma(mean0 = 8.46, effect = 0.3, shape0 = 0.639, alpha = 0.05,
↪  power = 0.9)

Estimated sample size for group difference.
Generalized Regression, Gamma Distribution, link: log

N (total)           517.02
n0 (Group 0)        258.51
n1 (Group 1)        258.51

Effect size         0.3
Effect type         1 - (mean1 / mean0)
Type I error        0.05
Target power        0.9
Two-sided           TRUE

Call: skewsamp::n_gamma(mean0 = 8.46, effect = 0.3, shape0 = 0.639, alpha =
↪  0.05, power = 0.9)
```

Codeblock 3. Beispiel-Input und -Output der {skewsamp}-Funktion n_locshift. Da die Fallzahlabschätzung abhängig von den zufällig gezogenen Pilotstichproben ist, können die Ergebnisse bei Wiederholung anders aufallen.

```
> skewsamp::n_locshift(rexp(10), rexp(10), delta = 0.3, alpha = 0.05, power
↪  = 0.9)

Estimated sample size for group difference.
Wilcoxon-Mann-Whitney Test, location shift

N (total)         75.94
n0 (Group 0)      37.97
n1 (Group 1)      37.97

Effect size       0.3
Effect type       location shift
Type I error      0.05
Target power      0.9
Two-sided         FALSE

Call: skewsamp::n_locshift(s1 = rexp(10), s2 = rexp(10), delta = 0.3, alpha
↪  = 0.05, power = 0.9)
```

aus, über den Parameter q kann aber auch ein anderes Quantil ausgegeben werden. Zusätzlich kann auch über die Funktion resample_n_locshift der volle Resampling-Vektor ausgegeben werden. Diese Funktionen erfordern jeweils die Eingabe von zwei Pilot-Stichproben und die Eingabe des Location Shifts δ.

Neben den reinen Funktionen zur Fallzahlabschätzung stellen wir in diesem Teil des Pakets auch die Funktionen rund um die empirische Verteilungsfunktion zur Verfügung, die wir die Umsetzung des Verfahrens implementiert haben. Das umfasst die Verteiluns-, Dichte-, und Quantilsfunktion, sowie eine Funktion zum ziehen zufälliger Werte aus der empirischen Verteilungsfunktion. Die zugehörigen theoretischen Details sind in Anhang A.1 dargestellt. Eine Übersicht der Funktionen ist in Tabelle 1 enthalten.

4. Simulationsstudien

Wir führen aus zwei Gründen Simulationsstudien durch. Zunächst möchten wir die Ergebnisse der Originalarbeiten von Cundill & Alexander (2015) und Chakraborti et al. (2006) mit unseren R-Implementationen ihrer Verfahren replizieren und damit verifizieren, dass unsere Implementationen korrekt funktionieren. Dann möchten wir über die Originalarbeiten hinausgehen und überprüfen, wie robust die beiden Verfahren in weiteren Szenarien sind. Da erste Pilot-Versuche hier auf eine größere Variation unter verschiedenen Bedingungen beim NECDF-Schätzer hindeuteten, während die Variation bei Cundill & Alexander sehr beschränkt zu Tage trat, haben wir die weitergehende Simulation für NECDF deutlich detaillierter gestaltet. Im Folgenden beschreiben wir

das Design der Simulationsstudien näher. Für alle Simulationen arbeiten wir mit einer angestrebten Power von 90%, einer Typ-I-Fehlerrate von $\alpha = 0.05$ und gleichen Gruppengrößen in der Kontroll- und Experimentalgruppe. Die Daten und der R-Code für unsere Simulationen stehen über Anhang C zur Verfügung.

4.1. Simulation 1: Replikation der GLM-Fallzahlabschätzung

Wir beschreiben in den folgenden Abschnitten die Designs und Ergebnisse unserer Replikationen der Simulationsstudien von Cundill & Alexander (2015).

Negative Binomialverteilung

Wie Cundill & Alexander (2015) nehmen wir das Beispiel von Brooker et al. (2005) als Ausgangspunkt für die Untersuchung der negativen Binomialverteilung. In dem Beispiel geht es um die Wirksamkeit eines Impfstoffs gegen Infektionen mit Hakenwürmern. Der Erfolg des Mittels wird über die Anzahl gefundener Eier in Stuhlproben überprüft. Zur Modellierung solcher Zähldaten ist die negative Binomialverteilung eine angemessene Wahl. Der Mittelwert in der Kontrollgruppe beträgt $\mu_0 = 71.4$ gefundene Eier pro Probe.

Das Effektstärkemaß ist die Wirksamkeit $\delta = 1 - \frac{\mu_1}{\mu_0}$, sie gibt an, um welchen Anteil die Zählung in der Experimentalgruppe im Vergleich zur Kontrollgruppe zurückgegangen (bei $\delta > 0$) oder gestiegen (bei $\delta < 0$) ist. Eine Wirksamkeit von 30% bedeutet dementsprechend, dass in der Experimentalgruppe 30% weniger Eier pro Probe gefunden wurden, als in der Kontrollgruppe.

Zur Datengenerierung nutzen wir die Mittelwerts-Parametrisierung der negativen Binomialverteilung mit Mittelwert μ und Dispersionsparameter k:

$$y \sim NB(\mu, k)$$

Wie in der Originalarbeit nehmen wir an, dass die Kontroll- und Experimentalgruppe sich im Mittelwert unterscheiden, wobei der Dispersionsparameter k über beide Gruppen konstant bleibt. Aus dem Mittelwert in der Kontrollgruppe und der Effektstärke δ ergibt sich ein Mittelwert von $\mu_1 = \mu_0(1 - \delta)$ in der Experimentalgruppe.

Das Vorgehen für die Simulation ist wie folgt:

1. Wir errechnen die erforderliche Anzahl von Beobachtungen N durch Eingabe der Parameter in die Funktion `n_negbinom`. Als link-Funktion nutzen wir den natürlichen Logarithmus. In der Vergleichsbedingung nutzen nutzen wir wie

die Originalautoren den *identity*-Link, womit die Fallzahlabschätzung in der Vergleichsbedingung der verbreiteten Variante unter Annahme normalverteilter Daten entspricht.

2. Anschließend ziehen wir $n = N/2$ Beobachtungen jeweils aus $NB(\mu_0, k)$ und $NB(\mu_1, k)$. Die Fallzahl n runden wir dabei auf die nächsthöhere Ganzzahl auf.

3. Wir wenden ein generalisiertes lineares Modell für negativ-binomial-verteilte abhängige Variablen mit log-Link auf die Daten an. Die Modellgleichung lautet

$$ln(E(y_i)) = \beta_0 + \beta_1 x_i,$$

mit $i = 1, \ldots, N$. Dabei zeigt die Indikatorvariable x_i die Gruppenzugehörigkeit an; der Wert 0 repräsentiert die Kontrollgruppe. Der t-Test für die $H_0 : \beta_1 = 0$ kann wie in Abschnitt 3.1 beschrieben als Test für $H_0 : \delta = 0$ genutzt werden. In Schritt 3 prüfen wir daher nach der Modellierung jeweils, ob diese H_0 beim Standard-α von 0.05 verworfen werden kann.

Die Schritte 1-3 werden $N_{sim} = 10000$ Mal wiederholt. Wir errechnen die beobachtete Power als die Anzahl der signifikanten Testergebnisse geteilt durch N_{sim}.

Zunächst halten wir die Effektstärke konstant bei $\delta = 1 - \frac{50}{71.4} \approx 0.3$ und variieren k. Dazu bilden wir eine Reihe aus 20 Werten mit gleichmäßigen Abständen zwischen 0.1 und 10. Anschließend halten wir k konstant bei $k = 0.33$ und variieren die Effektstärke δ. Dazu bilden wir eine Reihe aus 20 Werten mit gleichmäßigen Abständen zwischen 0.3 und 0.7. Für jeden der 20 k-Werte und jeden der 20 δ-Werte führen wir also 10 000 Wiederholungen der Schritte 1-3 jeweils einmal mit log-Link und einmal mit identity-Link in der Fallzahlabschätzung durch.

Ergebnisse und Diskussion In Abbildung 3 sind die Ergebnisse dargestellt, Teilabbildung *a* zeigt die Ergebnisse für die Variation von k und Teilabbildung *b* für die Variation von δ. Das Muster entspricht in beiden Fällen exakt dem Bericht von Cundill & Alexander (2015). Die Verwendung des identity-Links führt zu höheren geschätzten Fallzahlen als die Verwendung des log-Links. Der Unterschied ist sehr klein für $\delta \approx 0.3$ und nimmt mit steigendem δ zu. Während der log-Link insgesamt zu einer stabilen Power von ca. 90% führt, nähert sich die Power bei Verwendung des log-Links mit steigender Effektstärke 100% an. Der identity-Link führt also nicht zu einer geringeren Teststärke, sondern zu einer ineffizienteren Verwendung von Ressourcen, indem bis zu 60% mehr Datenpunkte erhoben werden, als eigentlich für die angestrebte Power notwendig wären ($\delta = 0.7$ und $k = 0.33$, Abbildung 3 b). Die Variation des Dispersionsparameters k hat keinen klar erkennbaren bedeutsamen Einfluss.

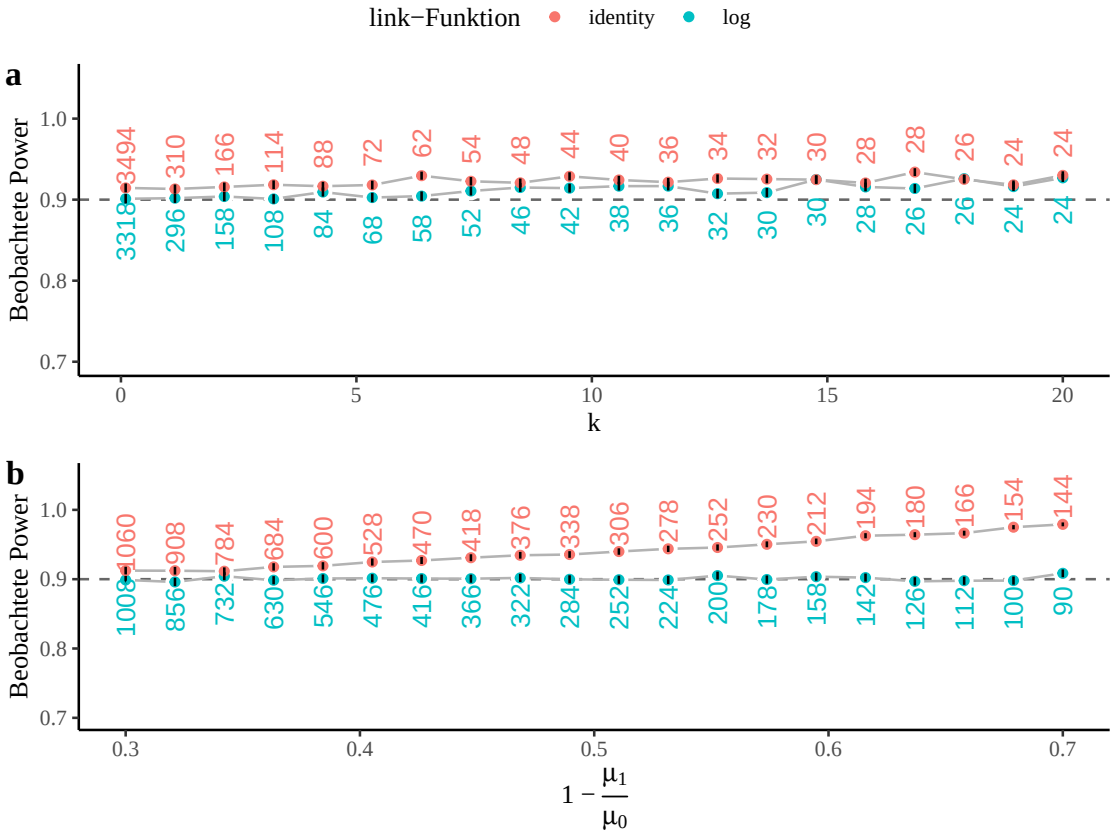

Abbildung 3 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für negativ-binomial verteilte Daten. Teilabbildung **a** zeigt Ergebnisse bei konstanter Effektstärke von $\delta = 1 - \frac{50}{71.4}$ und Variation des Dispersionsparameters k. Teilabbildung **b** zeigt Ergebnisse bei konstantem $k = 0.3$, $\mu_0 = 71.4$ und Variation der Effektstärke. Die Zahlen ober- und unterhalb der Punkte geben die verwendete Anzahl von Datenpunkten an. Die vertikalen schwarzen Linien innerhalb der Punkte zeigen 95% Wald-Konfidenzintervalle für die beobachtete Power und sind ein Maß für den Monte-Carlo-Fehler.

Gamma-Verteilung

Für die Gamma-Verteilung arbeiten Cundill & Alexander (2015) mit einem Datensatz zur Konzentration des Insektizids Deltamethrin in Hängematten (Rodríguez et al., 2009). Der Mittelwert liegt bei $\mu_0 = 8.46\ mg/m^2$, mit Formparameter (*shape*) $\kappa = 0.639$. Als Effektstärkemaß kommt abermals der relative Mittelwertsunterschied δ zum Einsatz. Zur Datengenerierung nutzen wir die Mittelwerts- Parametrisierung der Gamma-Verteilung mit Formparameter κ und Skalenparameter $\frac{\mu}{\kappa}$:

$$y \sim Gamma(\kappa, \frac{\mu}{\kappa})$$

Zur Fallzahlabschätzung nutzen wir unsere Funktion `n_gamma` mit log-Link oder identity-Link und zur Analyse eine Gamma-Regression mit entsprechender Linkfunktion. Das Design der Simulation entspricht ansonsten der Vorgehensweise zur negativen Binomialverteilung mit Variation von δ. Der Formparameter κ ist fixiert bei $\kappa = 0.639$.

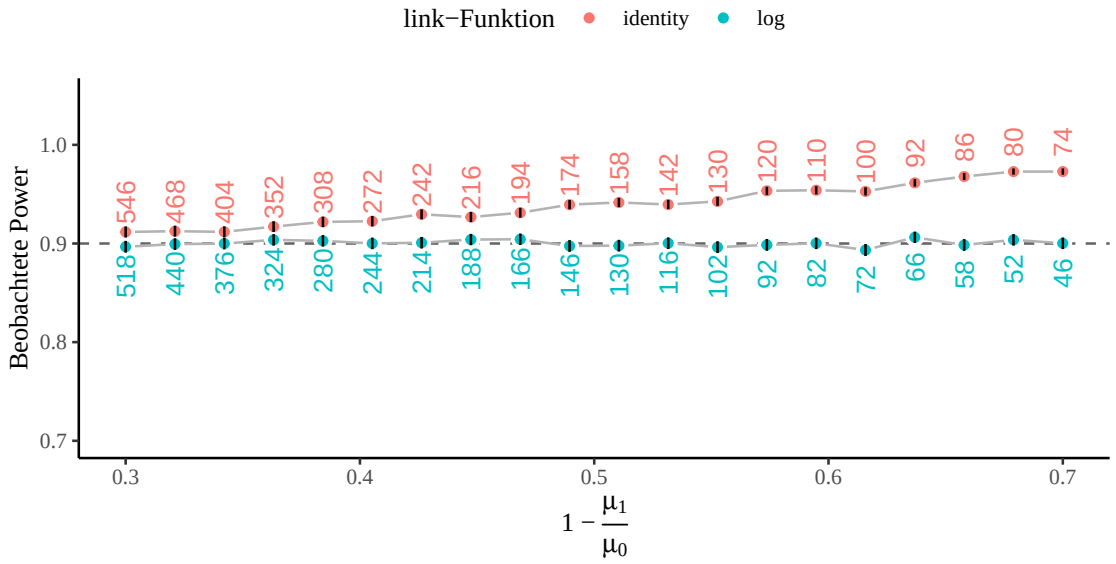

Abbildung 4 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für gamma-verteilte Daten, entsprechend Abbildung 3. Die Abbildung zeigt Ergebnisse bei konstantem Formparameter $\kappa = 0.639$, $\mu_0 = 8.46$ und Variation der Effektstärke.

Ergebnisse und Diskussion Die Ergebnisse sind in Abbildung 4 dargestellt und entsprechen den von Cundill & Alexander (2015) vorgestellten Befunden. Mit der GLM-basierten Fallzahlabschätzung wird die angestrebte Power genau erreicht. Wie schon bei der negativen Binomialverteilung führt die Verwendung des identity-Links mit steigender Effektstärke zu einer größeren Überschätzung der notwendigen Fallzahl für die angestrebte Power.

Poisson- und Binomialverteilung

Auch für diese Verteilungen können wir die Ergebnisse der Originalautoren replizieren. Wir fassen die Ergebnisse hier knapp zusammen. Die Unterschiede zwischen der Verwendung von identity-Link und log- (Poisson), bzw. logit-Link (Binomial) sind für diese beiden Verteilungen kleiner als für die Gamma- und die negative Binomialverteilung. Die jeweils kanonischen link-Funktionen zeigen eine leichte Tendenz zu konservativen Fallzahlschätzungen mit steigenden Effektstärken, d.h. die ermittelten Fallzahlen führen zu leicht höherer beobachteter Power im Vergleich zum angestrebten Wert. Die zugehörigen Abbildungen B.1 und B.2 sind im Anhang B verfügbar.

Diskussion

Mit diesen Ergebnissen können wir einerseits die korrekte Funktionsweise unserer Implementierung verifizieren und andererseits die Ergebnisse der Originalautoren erfolgreich

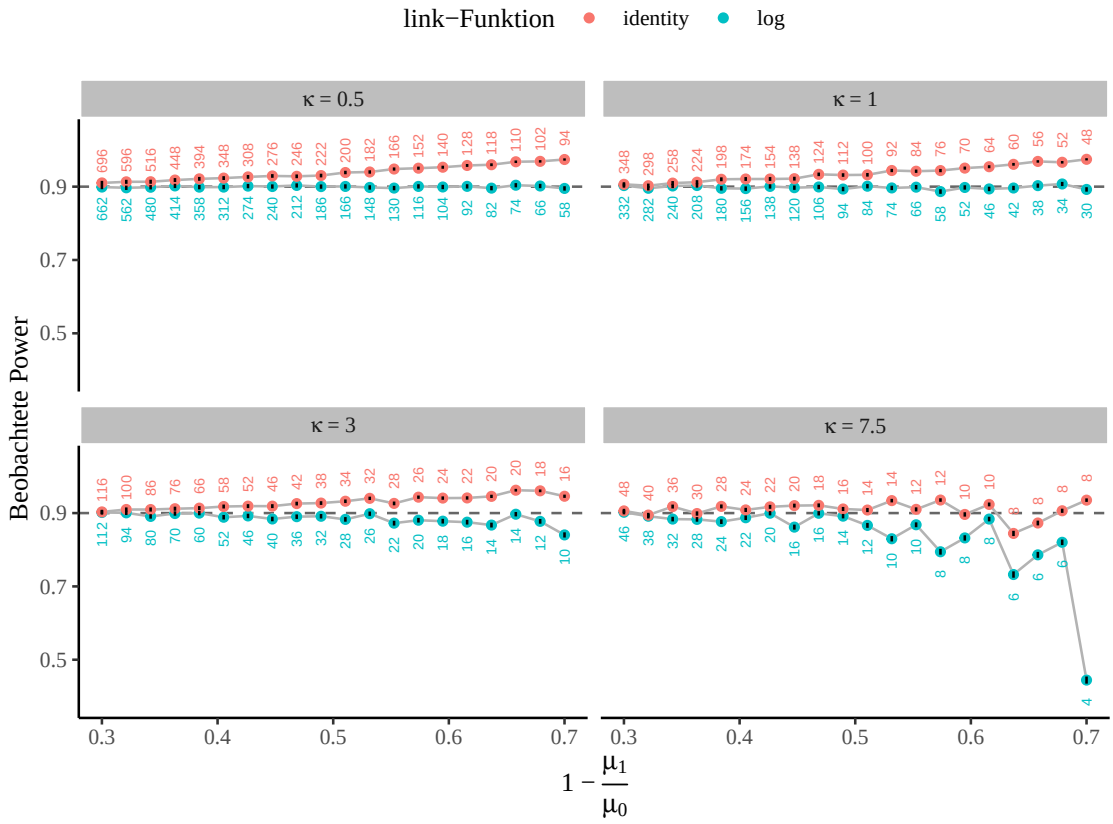

Abbildung 5 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für gamma-verteilte Daten, entsprechend Abbildung 3. Die Abbildung zeigt Ergebnisse bei verschiedenen Formparametern κ und Variation der Effektstärke.

replizieren. Für alle untersuchten Verteilungen ist die GLM-basierte Fallzahlabschätzung mit log-Link geeignet, mindestens die angestrebte Power zu erreichen. Die Verwendung des identity-Links kann dagegen bei der negativen Binomialverteilung und der Gammaverteilung zu einer substantiellen Überschätzung der notwendigen Fallzahl und damit zu einer ineffizienten Verwendung der verfügbaren Ressourcen führen.

4.2. Simulation 2: Zusätzliche Simulationen zur GLM-Fallzahlabschätzung

Für die Gamma-Verteilung führen wir eine Reihe von zusätzlichen Simulationen mit Variation des Formparameters κ und des Skalenparameter θ durch. Für die Datengenerierung verwenden wir in diesem Fall die klassische Parametrisierung der Gamma-Verteilung mit Form- und Skalenparameter:

$$y \sim Gamma(\kappa, \theta).$$

In diesem Fall ist $\mu_0 = \kappa\theta$ und, wie gehabt, $\mu_1 = \mu_0(1-\delta)$. Wir verwenden $\kappa \in \{0.5, 1, 2, 3, 7.5\}$ und $\theta \in \{0.5, 1, 2, 3, 4, 5, 6\}$ und untersuchen alle 35 Kombinationen dieser Werte. Für jedes dieser 35 Szenarien führen wir die gleiche Prozedur durch wie im Replikationsteil zur Gamma-Verteilung.

Ergebnisse und Diskussion Die Variation des Skalenparameters hat keinen Einfluss auf die Ergebnisse. Für eine prägnantere Darstellung platzieren wir die zugehörige Abbildung B.3 im Anhang. In Abbildung 5 sind die Ergebnisse für unterschiedliche Formparameter dargestellt. Bei der Variation des Formparameters zeigte sich größtenteils das gleiche Muster wie das von Cundill & Alexander (2015) berichtete. Im Extremfall, bei $\kappa = 7.5$ verändert sich das Muster allerdings deutlich, wie in der Abbildung sichtbar wird. Mit solch einem großen Formparameter sinkt die errechnete Fallzahl unter log-Link bei großer Effektstärke von $\delta = 0.7$ auf $N = 4$, was in unseren Simulationen nicht einmal für eine Power von 50% reicht. Hier funktioniert der identity-Link, mit dem $N = 8$ errechnet wird, besser. In Ansätzen zeigt sich eine Entwicklung in diese Richtung schon bei $\kappa = 3$, wenngleich deutlich weniger gravierend. In diesem Fall liegt die erzielte Power bei Verwendung des log-Links und $\delta = 0.7$ bei 84%.

Insgesamt demonstrieren die Ergebnisse die Robustheit der GLM-basierten Fallzahlabschätzung, die nur im Extremfall bei äußerst niedrigen Fallzahlen ungenau zu werden scheint.

4.3. Simulation 3: Replikation der NECDF-Fallzahlabschätzung

Zur Replikation der Ergebnisse von Chakraborti et al. (2006) nutzen wir im Grunde das gleiche Design wie die Originalautoren. Wir erhöhen allerdings die Anzahl der Simulationsdurchgänge und erweitern das Design um eine zusätzliche Simulation, in der wir überprüfen, welche Power wir mit der geschätzten Fallzahl praktisch erreichen. Wie die Originalautoren führen wir die Simulation zunächst mit der einfachen geschätzten Fallzahl durch, bevor wir die konservativere Resampling-Schätzung ebenfalls überprüfen.

Fallzahlabschätzung

Das Vorgehen zur Fallzahlabschätzung ist wie folgt:

1. Wir ziehen zwei Pilot-Stichproben der Größe m aus einer Wahrscheinlichkeitsverteilung.

2. Wir schätzen die erforderliche Fallzahl für eine Power von 90% bei $\alpha = 0.05$ mit der von uns implementierten Funktion `n_locshift()` (einfache Fallzahlabschätzung), bzw. `n_locshift_bound()` (Fallzahlabschätzung mit Resampling für Obergrenze der Schätzung).

3. Wir wiederholen die Schritte 1 und 2 $N_{sim} = 10000$ Mal (einfache Fallzahlabschätzung), bzw. $M_{sim} = 500$ Mal (mit Resampling), so dass wir ebenso viele Fallzahlschätzungen erhalten. Die geringere Anzahl an Wiederholungen für die Resampling-Schätzung ist notwendig, um den Rechenaufwand für uns in einem beherrschbaren Rahmen zu halten, da in dieser Bedingungen für jeden einzelnen Simulationsdurchgang schon 1000 Resampling-Schritte anfallen (jeweils 500 für jede der beiden Pilot-Stichproben).

So erhalten wir eine Verteilung der geschätzten Fallzahlen $\hat{N}_{NECDF,i}$, $i = 1, \ldots, N_{sim}$ und eine Verteilung der geschätzten Obergrenzen $\hat{N}_i^{(90)}$, $i = 1, \ldots, M_{sim}$. Dieses Prozedere führen wir für vier Wahrscheinlichkeitsverteilungen und zwei Stichprobengrößen ($m = 10$ und $m = 20$) der Pilotstudien durch. Wie in der Originalarbeit verwenden wir unterschiedliche Werte für die Effektstärke, die Lageverschiebung δ, für unterschiedliche Verteilungen. Die untersuchten Verteilungen sind die Normalverteilung ($\delta = 0.5$), die Gleichverteilung ($\delta = 0.2$), die Exponentialverteilung ($\delta = 0.35$), und die logistische Verteilung ($\delta = 0.8$). Da Chakraborti et al. (2006) keine Details über die von ihnen verwendeten Parameter der Verteilungen berichten, arbeiten wir mit Standardfällen: Für die Normalverteilung arbeiten wir mit $\mu = 0$ und $\sigma = 1$, für die Exponentialverteilung mit $\lambda = 1$, für die Gleichverteilung mit $a = 0$ und $b = 1$ und für die logistische Verteilung mit $\mu = 0$ und $s = 1$.

Empirische Power

Um die Qualität des Verfahrens noch genauer zu untersuchen, führen wir zusätzlich zur reinen Replikation des Vorgehens von Chakraborti et al. (2006) weitere Simulationen zur Bestimmung der tatsächlich erreichten Power durch. Das Vorgehen dabei ist wie folgt:

1. Wir ziehen zwei Stichproben der Größe $n = N/2$ aus der gleichen Wahrscheinlichkeitsverteilung, aus der die ursprünglichen Pilot-Stichproben gezogen wurden. Die Stichprobengrößen werden aufgerundet auf die nächsthöhere Ganzzahl.

2. Wir wenden die Lageverschiebung δ auf eine der beiden Stichproben an, indem wir von jedem Wert in der Stichprobe δ subtrahieren.

3. Wir wenden einen einseitigen Wilcoxon-Mann-Whitney Test auf die Daten an und prüfen, ob die H_0 verworfen werden kann.

4. Wir wiederholen die Schritte 1-3 10000 Mal.

Als Schätzung der tatsächlich erreichten Power teilen wir die Anzahl signifikanter Testergebnisse durch die Anzahl der Wiederholungen. Für N in Schritt 1 setzen wir folgende Werte ein:

1. Den Mittelwert aus den N_{sim} zuvor ermittelten Fallzahlabschätzungen, also

$$\bar{\hat{N}}_{NECDF} = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \hat{N}_{NECDF,i}.$$

2. Die $10\%-, 20\%-, \ldots, 80\%-, 90\%-$Quantile der Werte $\hat{N}_{NECDF,1}, \ldots, \hat{N}_{NECDF,N_{sim}}$. Dadurch ermitteln wir die Auswirkungen der Unsicherheit von $\hat{N}_{NECDF}$ als Folge seiner Abhängigkeit von den Pilot-Stichproben.

3. Den Mittelwert aus den M_{sim} zuvor ermittelten Obergrenzen der Fallzahlabschätungen, also

$$\bar{\hat{N}}^{(90)} = \frac{1}{M_{sim}} \sum_{i=1}^{M_{sim}} \hat{N}_i^{(90)}.$$

Ergebnisse und Diskussion

Einfache Fallzahlabschätzung Die Ergebnisse vergleichen wir in Abbildung 6 mit den von Chakraborti et al. (2006) berichteten Werten. Um eine gute Vergleichbarkeit zu gewährleisten, verwenden wir im Bericht wie die Originalautoren die Fallzahl für jeweils *eine* Gruppe. Unsere Ergebnisse sind praktisch identisch mit denen der Originalautoren, abgesehen von den bei einer Simulationsstudie erwartbaren leichten numerischen Unterschieden. In Abbildung 7a ist zusätzlich die Verteilung der Fallzahl-Schätzungen für $m = 20$ in Histogrammen dargestellt. Dort ist die Unsicherheit der Schätzung für alle untersuchten Verteilungen deutlich erkennbar.

In Tabelle 2 sind die Ergebnisse der Power-Simulation zusammengefasst. Auf Grundlage der mittleren Fallzahlschätzung werden in allen untersuchten Szenarien Werte erreicht, die sehr nah an der angestrebten Power von 90% liegen. Die größte Abweichung ist -2.2% im Falle der Normalverteilung mit $m = 10$. In vier der acht Szenarien beinhaltet das 95%-Konfidenzintervall den Zielwert, in jeweils zwei Fällen liegen beide Grenzen des Intervalls ober- bzw. unterhalb des Zielwerts. Im Mittel scheint die NECDF-Methode in den untersuchten Szenarien also recht brauchbare Schätzungen zu liefern. Eine Zahl von vier von acht Konfidenzintervallen, die den angestrebten Wert nicht enthalten, weist allerdings auf eine etwas zu unsichere Schätzung hin. Eine klare Tendenz zur Unter- oder Überschätzung ist auf Grundlage dieser Daten nicht erkennbar.

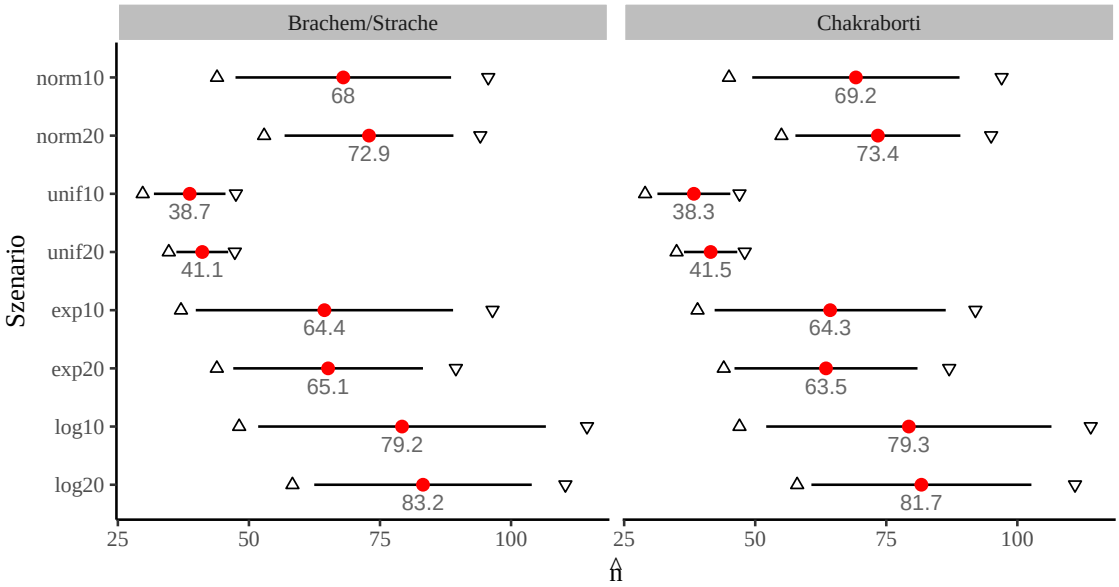

Abbildung 6 Vergleich der Simulationsergebnisse zwischen unserer Replikation (links) und Originalarbeit (rechts). Die rechte Abbildung basiert auf den NECDF-Daten aus Tabelle 4 in Chakraborti et al. (2006). Der rote Punkt, inkl. Annotierug ist der Mittelwert aus 10 000 (Replikation), bzw. 200 (Original) Simulationsdurchläufen. Die Dreieke zeigen das 10.- und das 90. Perzentil der geschätzten Fallzahlen. Die horizontale Linie zeigt eine Standardabweichung. Die Szenariennamen auf der y-Achse enthalten die Größe der verwendeten Pilot-Stichproben. Dargestellt ist jeweils die Fallzahlabschätzung für die Größe *einer* Gruppe.

Weiteren Aufschluss bieten die ebenfalls in Tabelle 2 dargestellten erreichten Power-Werte für das 10. und 90. Perzentils der geschätzten Fallzahlen. Die Verwendung des 90. Perzentils führte in allen Szenarien zu einer höheren als der angestrebten Power, der Unterschied zur Zielpower betrug zwischen 3.2% (Gleichverteilung, $m = 10$) und 7.3% (Exponentialverteilung, $m = 10$). Entsprechend führte die Verwendung des 10%-Quantils zu einer geringeren als der angestrebten Power. Hier fielen die erreichten Power-Werte durchweg deutlich geringer als aus als die angestrebten. Die größte Abweichung betrugt -18.5% (logistische Verteilung, $m = 10$) und nur in einem Fall fiel die Abweichung mit -5.6% (Gleichverteilung, $m = 20$) kleiner als 10 Prozentpunkte aus. In Abbildung B.4 in Anhang B sind die Ergebnisse für die übrigen Perzentile visualisiert.

Obergrenze durch Resampling Tabelle 3 zeigt die Ergebnisse des zweiten Teils der Replikation. Hier wird mittels Resampling eine Obergrenze für die benötigte Fallzahl geschätzt. Auch in diesem Fall entsprechen unsere Ergebnisse der Originalarbeit, abgesehen von kleinen simulationsbedingten numerischen Abweichungen. In Abbildung 7b ist die Verteilung der geschätzten Obergrenzen über die 500 Simulationsdurchläufe für $m = 20$ dargestellt. Es ist klar erkennbar, dass die Obergrenze die Gefahr einer Unterschätzung der notwendigen Fallzahl in allen untersuchten Szenarien im Vergleich zur einfachen Schätzung deutlich verringert. Die Varianz der Obergrenze bleibt dennoch substantiell, was in erster Linie eine Folge der Abhängigkeit von der ursprünglich

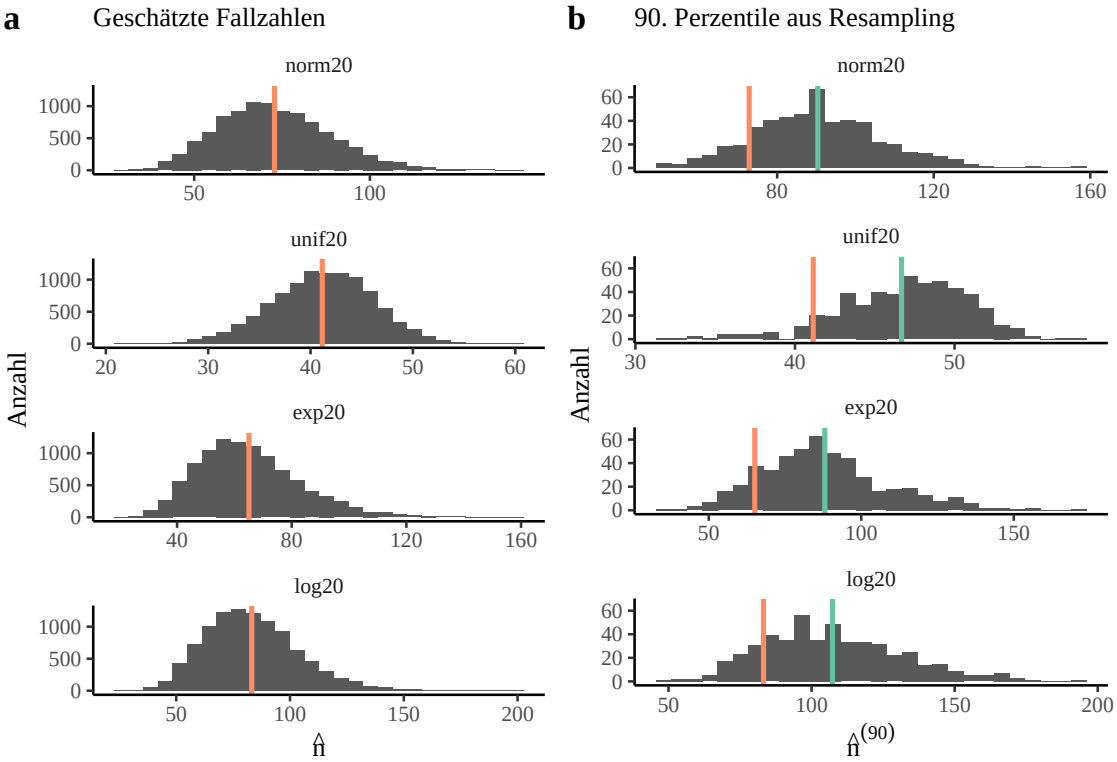

Abbildung 7 Verteilung von NECDF-Fallzahlabschätzungen und Obergrenzen für die Größe *einer* Gruppe in den Replikationsszenarien mit $m = 20$. Teilabbildung **a** zeigt die Verteilung der einzelnen Schätzungen in 10 000 Simulationsdurchgängen. Teilabbildung **b** zeigt die Verteilung von durch Resampling geschätzten Obergrenzen in 500 Simulationsdurchgängen. Die rötliche vertikale Linie zeigt in beiden Teilabbildungen den Mittelwert der 10 000 einfachen Fallzahlabschätzungen. Die grüne Linie in b) zeigt den Mittelwert der 500 Schätzungen der Obergrenze.

gezogengen Pilot-Stichprobe sein dürfte.

Diskussion Die Übereinstimmung unserer Ergebnisse mit denen von Chakraborti et al. (2006) deuten wir als Evidenz für die Korrektheit unserer Implementierung. Die Ergebnisse verifizieren außerdem die Originalbefunde: In den untersuchten Szenarien führt das NECDF-Verfahren im Mittel zu einer Power, die der angestrebten Power sehr nahe kommt. Deutlich wird aber auch, wie schon Chakraborti et al. (2006) berichten, dass die Leistung des Verfahrens stark schwankt, da sie von den verwendeten Pilot-Stichproben abhängt. So ist die Gefahr groß, dass die tatsächlich erforderliche Fallzahl in der Praxis unter- oder überschätzt wird. Durch die Einbeziehung einer durch Resampling aus den Pilotstichproben geschätzten Obergrenze für diese Fallzahl kann zwar das Risiko einer Unterschätzung vermindert werden, dies geschieht allerdings auf Kosten einer im Mittel substantiellen Überschätzung der Fallzahl.

Tabelle 2 Übersicht über die Ergebnisse unserer Replikation von Chakraborti et al. (2006). Die Tabelle zeigt die mittlere geschätzte Fallzahl sowie das 10. und das 90. Perzentil aus 10 000 Simulationsdurchgängen neben dem jeweils in weiteren 10 000 Simulationsdurchgängen beobachteten Anteil signifikanter WMW-Tests in Prozent in der Spalte *Power*. Für die Power-Schätzungen zeigen wir 95% Wald-Konfidenzintervalle als Maß für den Monte Carlo Fehler.

			Mittelwert			10. Perzentil			90. Perzentil		
Vert.	δ	m	n	Power	95%-KI	n	Power	95%-KI	n	Power	95%-KI
norm	0.50	10	68	87.8	[87.2, 88.5]	44	73.1	[72.2, 73.9]	96	95.7	[95.3, 96.1]
norm	0.50	20	73	89.8	[89.3, 90.4]	53	79.9	[79.1, 80.7]	95	95.8	[95.4, 96.2]
unif	0.20	10	39	88.1	[87.5, 88.7]	30	79.5	[78.7, 80.3]	48	93.2	[92.7, 93.7]
unif	0.20	20	42	90.4	[89.8, 90.9]	35	84.4	[83.6, 85.1]	48	93.3	[92.8, 93.8]
exp	0.35	10	65	90.7	[90.1, 91.3]	38	72.4	[71.5, 73.3]	97	97.3	[97.0, 97.6]
exp	0.35	20	66	90.7	[90.1, 91.2]	44	78.4	[77.6, 79.2]	90	96.7	[96.3, 97.0]
log	0.80	10	80	89.4	[88.8, 90.0]	49	71.5	[70.6, 72.3]	115	96.5	[96.1, 96.9]
log	0.80	20	84	90.2	[89.6, 90.8]	59	79.7	[78.9, 80.5]	111	96.4	[96.0, 96.7]

Tabelle 3 Übersicht über die Ergebnisse unserer Replikation zur Obergrenze der Fallzahlschätzung, $\bar{\hat{n}}^{(90)} = \bar{\hat{N}}^{(90)}/2$. Die Spalte $n_{CHAK}^{(90)}$ zeigt den in der Originalarbeit berichteten Wert. Die Spalte *SD* zeigt die Standardabweichung der Obergrenzen in unseren 500 Simulationsdurchläufen. Die Spalte *Power* zeigt den Anteil signifikanter WMW-Tests in Prozent aus 10 000 Simulationsdurchläufen. Die Spalte *95%-KI* zeigt 95%-Wald-Konfidenzintervalle für diese Power-Schätzung als Maß für den Monte Carlo Fehler.

Vert.	δ	m	$n_{CHAK}^{(90)}$	$\bar{\hat{n}}^{(90)}$	SD	Power	95%-KI
norm	0.50	10	87.4	85.7	22.4	94.1	[93.7, 94.6]
norm	0.50	20	89.5	90.4	16.3	94.7	[94.2, 95.1]
unif	0.20	10	45.6	45.4	6.1	92.5	[91.9, 93.0]
unif	0.20	20	46.3	46.7	4.1	93.0	[92.5, 93.5]
exp	0.35	10	95.3	93.1	30.6	97.0	[96.7, 97.4]
exp	0.35	20	86.5	88.1	21.0	96.4	[96.1, 96.8]
log	0.80	10	104.4	103.1	35.4	95.1	[94.7, 95.5]
log	0.80	20	105.5	107.3	24.3	95.8	[95.4, 96.2]

Tabelle 4 Variationen der Verteilungsparameter zur Generierung von Pilot-Stichproben und Test-Stichproben in den zusätzlichen Simulationen zum NECDF-Verfahren (Simulation 4).

Vert.	Parameter	Variationen
Gamma	δ	0.1, 0.5, 1, 1.5, 2
Gamma	θ	0.5, 1, 2, 3, 6
Gamma	κ	0.5, 1, 2, 3, 7.5
Norm	δ	0.5, 0.1, 1, 1.5, 2
Norm	σ	1, 2, 3, 4, 5
Exp	δ	0.1, 0.35, 0.5, 1, 1.5, 2
Exp	λ	0.2, 1, 2, 3, 4, 5

4.4. Simulation 4: Zusätzliche Simulationen zur NECDF

Das Ziel unserer zusätzlichen Simulationen für das NECDF-Verfahren war, die Robustheit des Verfahrens in einer größeren Bandbreite von Szenarien auf die Probe zu stellen, als dies in der Originalarbeit der Fall war. Dazu haben wir Simulationen auf Basis der Normal- und der Exponentialverteilung durchgeführt, bei denen wir die Standardabweichung σ bzw. die Rate λ variieren. Wir nehmen außerdem Szenarien auf Grundlage der Gamma-Verteilung auf, da die Gamma-Verteilung in der Praxis häufig zur Modellierung schiefer Daten eingesetzt wird. Für die Gamma-Verteilung untersuchen wir eine Reihe von Kombinationen verschiedener Werte für den Formparameter k und den Skalenparameter θ.Alle Szenarien führen wir außerdem für verschieden große Werte der Lageverschiebung δ durch.

Das Vorgehen ist analog zum Vorgehen bei der Replikationsstudie. Unsere zentrale Zielgröße ist die erreichte Power. Für die Power- Simulation verwenden wir als Fallzahl in jedem Szenario den auf die nächsthöhere Ganzzahl gerundeten Mittelwert aus 10000 Fallzahlschätzungen, die jeweils auf unterschiedlichen Pilot-Stichproben beruhen. Wir führen die Simulationen wie in der Originalarbeit jeweils einmal für Pilot-Stichproben der Größe $m = 10$ und $m = 20$ durch. Die Simulationen erlauben uns so Erkenntnisse darüber, ob das NECDF-Verfahren unter unterschiedlichen Bedingungen im Mittel zu Fallzahlabschätzungen führt, die geeignet sind, die angestrebte Power zu erreichen. Tabelle 4 zeigt eine Übersicht der untersuchten Szenarien.

Ergebnisse Abbildung 8 zeigt eine Zusammenfassung der Ergebnisse. Da das Muster der Ergebnisse bei $m = 10$ und $m = 20$ im Wesentlichen identisch ist, zeigen wir in der Abbildung nur die Ergebnisse für $m = 20$. Wir beschränken die Abbildung außerdem auf die Darstellung von jeweils drei Stufen der Variation von δ und, im Falle der Gamma-Verteilung, eine Abbildung mit niedrigem Skalenparameter (Teilabbildung c, links) und hohem Skalenparameter (Teilabbildung c, rechts). Diese Beschränkungen verbessern die Übersichtlichkeit und zeigen die wesentlichen Muster.

In der Abbildung wird deutlich, dass die beobachtete Power stark von der Größe der Lageverschiebung δ und den Parametern der Zugrundeliegenden Verteilung abhängt. Die beobachtete Power ist dabei umso höher, je größer die Lageverschiebung δ ist, teilweise wird die angestrebte Power von 90% sogar überschritten. Bei normalverteilten Daten führt eine größere Standardabweichung zu geringerer beobachteter Power, bei der Exponentialverteilung wirkt sich dagegen eine geringe Rate tendenziell negativ auf die beobachtete Power aus. Im Falle der Gamma-Verteilung sind sowohl hohe Werte des Form-, als auch hohe Werte des Skalenparameters mit geringerer beobachteter Power assoziiert. Zu besonders schwerwiegenden Unterschreitungen der angestrebten Power

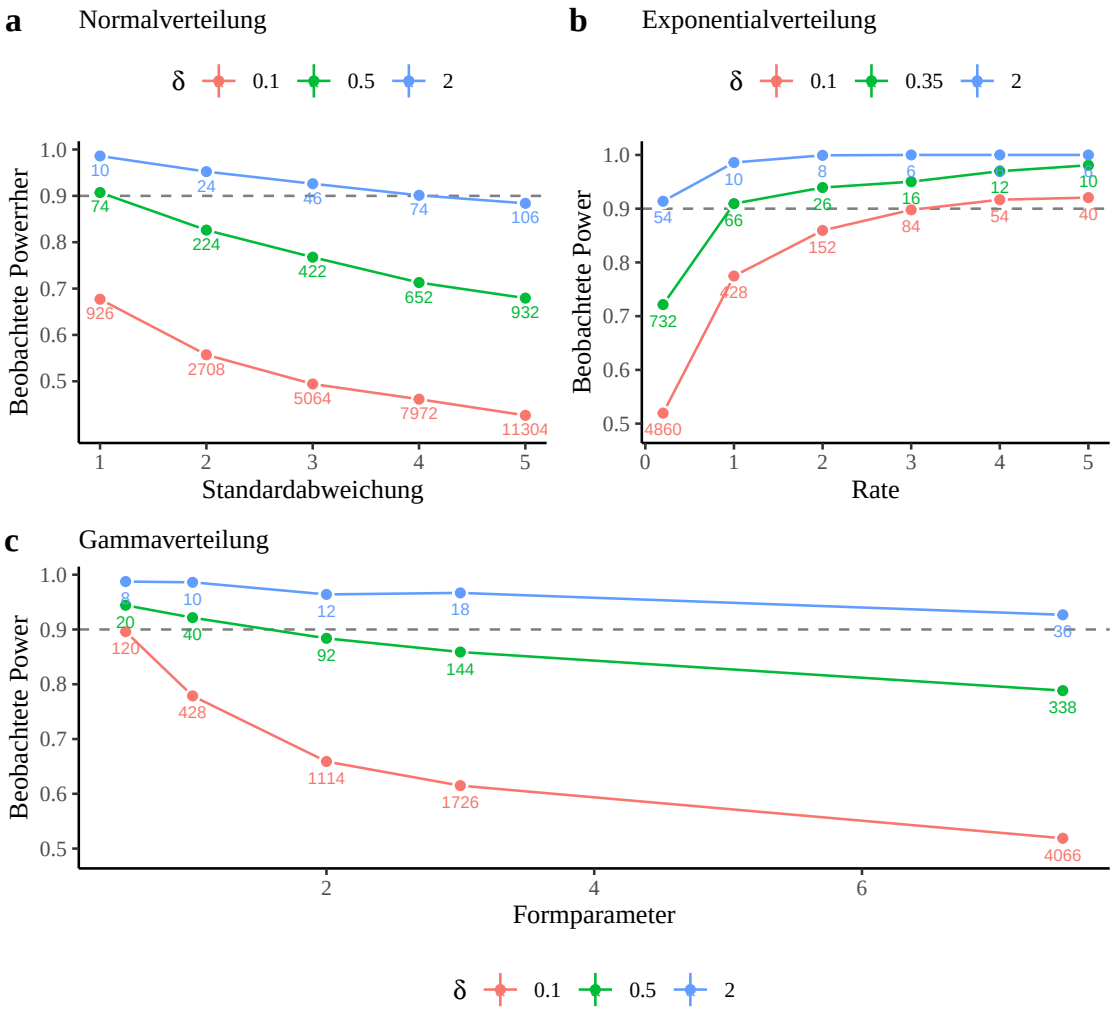

Abbildung 8 Beobachtete Power in zusätzlichen Simulationen unter Verschiedenen Bedingungen in der Normalverteilung (**a**), der Exponentialverteilung (**b**) und der Gammaverteilung (**c**) mit Skalenparameter $\theta = 1$. Unter den Punkten ist die jeweils verwendete Fallzahl für *eine* Gruppe angegeben.

kommt es bei Kombinationen aus geringer Effektstärke, hohem Form- und hohem Skalenparameter. So wird im Fall von $\mathcal{G}(k = 7.5, \theta = 1)$ und $\delta = 0.1$ mit der geschätzten Fallzahl von 4066 Beobachtungen in einer Gruppe tatsächlich eine Power von ca. 52% erreicht.

Diskussion In den zusätzlichen Simulationen zum NECDF-Verfahren zeigt sich eine problematische Abhängigkeit der Leistung von den konkreten Eigenschaften der untersuchten Daten. Insgesamt wecken diese Ergebnisse Zweifel an der Robustheit des NECDF-Verfahrens. Auffällig ist in diesem Zusammenhang, dass Chakraborti et al. (2006) im Falle der Normalverteilung und Exponentialverteilung jeweils genau solche Werte für δ wählen, die bei der jeweiligen Verteilung mit einer guten Erreichung der angestrebten Power einhergehen. Ihre Wahl von δ begründen die Autoren nicht näher.

5. Fazit

In diesem Bericht stellen wir zwei unterschiedliche Ansätze zur Fallzahlabschätzung bei schiefen Verteilungen vor, den GLM-basierten Ansatz nach Cundill & Alexander (2015) und den nichtparametrischen Ansatz für das *location shift* Paradigma im WMW-Test nach Chakraborti et al. (2006). Wir stellen für beide Ansätze eine einfach zu nutzende, dokumentierte und getestete Implementierung in der Open-Source-Programmiersprache R zur Verfügung. Wir haben außerdem die Simulationsstudien repliziert, mit denen die jeweiligen Originalautoren die Leistung ihrer Verfahren überprüfen. Dabei konnten wir in beiden Fällen die Berichte der Originalautoren anhand unserer Ergebnisse bestätigen und somit zugleich Evidenz für die Korrektheit unserer Implementierungen sammeln.

Die Replikationsstudien haben wir um eigene, zusätzliche Simulationen erweitert, um die Robustheit der Verfahren stärker zu testen. Dabei zeigten sich vor allem beim NECDF-Verfahren in der Beobachtung teils eklantante Unterschreitungen der angestrebten Power. Die Leistung des Verfahrens wies eine starke Abhängigkeit von den Parametern der zur Datengenerierung genutzten Verteilungen auf. Für die praktische Anwendung stellt das ein erhebliches Problem dar, da Forschende gerade bei der Planung einer Datenerhebung kaum abschätzen können, wie ihre Daten die Fallzahlabschätzung beeinflussen. Vor diesem Hintergrund können wir die Anwendung des NECDF- Verfahrens mit kleinen Pilotstichproben nicht ohne Weiteres empfehlen. Die entsprechenden Funktionen in {`skewsamp`} enthalten deshalb Hinweise zur Warnung. In jedem Fall sollte die abgeschätzte Obergrenze der benötigten Fallzahl in Entscheidungen auf Grundlage des NECDF-Verfahrens miteinbezogen werden.

Das GLM-basierte Verfahren zeigte sich dagegen robust in verschiedenen Szenarien, kann allerdings in Randfällen, in denen die ermittelte Fallzahl sehr gering ist, zu einer Unterschätzung der notwendigen Stichprobengröße führen. Die Anwendung dieses Verfahrens können wir empfehlen. Durch das Paket {`skewsamp`} ist das Verfahren einfach zugänglich.

Literatur

Brooker, S., Bethony, J. M., Rodrigues, L. C., Alexander, N., Geiger, S. M., & Hotez, P. J. (2005). Epidemiologic, immunologic and practical considerations in developing and evaluating a human hookworm vaccine. *Expert Review of Vaccines*, *4*(1), 35–50. https://doi.org/10.1586/14760584.4.1.35

Chakraborti, S., Hong, B., & Van De Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. *Technometrics*, *48*(1), 88–94. https://doi.org/10.1198/004017005000000193

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr

Cundill, B., & Alexander, N. D. E. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, *15*(1), 1–9. https://doi.org/10.1186/s12874-015-0023-0

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Hamilton, M. A., & Collings, B. J. (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics*, *33*(3), 327–337. https://doi.org/10.1080/00401706.1991.10484838

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, *2*(2), 93–113. https://doi.org/10.1016/0197-2456(81)90001-5

Lakens, D. (2021). *Sample Size Justification*. https://doi.org/https://doi.org/10.31234/osf.io/9d3yf

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, *82*(398), 645–647. https://doi.org/10.2307/2289477

Rodríguez, M., Pérez, L., Caicedo, J. C., Prieto, G., Arroyo, J. A., Kaur, H., Suárez-Mutis, M., De La Hoz, F., Lines, J., & Alexander, N. (2009). Composition and biting activity of anopheles (diptera: culicidae) in the Amazon region of Colombia. *Journal of Medical Entomology*, *46*(2), 307–315. https://doi.org/10.1603/033.046.0215

A. Zusätzliche Details zur Theorie

A.1. Empirische Verteilungsfunktion

Wir übernehmen die Parametriesierung der emprischen Verteilunsgfunktion, wie Chakraborti et al. (2006) sie berichten. Für eine Stichprobe $X_1, \ldots, X_m$ erstellen wir die nach Größe geordnete Reihe $X_{(1)}, \ldots, X_{(m)}$ und bilden künstliche Endpunkte[4]

$$\begin{aligned} X_{(0)} &= 2X_{(1)} - X_{(2)} \\ X_{(m+1)} &= 2X_{(m)} - X_{(m-1)}. \end{aligned}$$

Anhand der so erweiterten Reihe $X_{(0)}, \ldots, X_{(m+1)}$ ist die linear geglättete empirische Verteilungsfunktion dann gegeben durch

$$G_X(x) = \begin{cases} 0, & \text{wenn } x \leq X_{(0)} \\ \frac{i}{m+1} + \frac{x - X_{(i)}}{(m+1)(X_{(i+1)} - X_{(i)})}, & \text{wenn } X_{(i)} \leq x < X_{(i+1)} \\ 1, & \text{wenn } x \geq X_{(m+1)}, \end{cases}$$

mit $i = 0, 1, \ldots, m$. Daraus ergibt sich die empirische „Dichtefunktion“ als

$$g_X(x) = \begin{cases} \frac{1}{(m+1)(X_{(i+1)} - X_{(i)})}, & \text{wenn } X_{(i)} \leq x < X_{(i+1)} \\ 0, & \text{sonst} \end{cases}.$$

Außerdem leiten wir die empirische Quantilsfunktion als Inverse der empirischen Verteilungsfunktion her:

$$G_X^{-1}(p) = \begin{cases} (p(m+1) - i)\,(X_{(i+1)} - X_{(i)}) + X_{(i)}, & \text{wenn } 0 \leq p < 1 \\ X_{(m+1)}, & \text{wenn } p = 1, \end{cases}$$

wobei i so gewählt wird, dass $G_X(X_{(i)}) \leq p < G_X(X_{(i+1)})$ erfüllt ist. Die Quantilsfunktion nutzen wir zur Ziehung zufälliger Werte aus der empirischen Verteilungsfunktion in der Funktion `skewsamp::remp()`, indem wir $p \sim U(0, 1)$ ziehen und in G_X^{-1} einsetzen.

A.2. Herleitung der Berechnung von p

Mithilfe der empirischen Verteilungsfunktion G_X erfolgt die Berechnung der Wahrscheinlichkeit. Diese Ausführungen gehen vom Bericht der Autoren Chakraborti et

[4] Chakraborti et al. (2006) berichten irrtümlicherweise $X_{(0)} = 2X_{(2)} - X_{(1)}$, wodurch keine Erweiterung “nach unten” erreicht wird, da $2X_{(2)} - X_{(1)} > X_{(1)}$. Die hier dargestellte Variante erzeugt eine solche Erweiterung und entspricht auch der Darstellung in Hamilton & Collings (1991).

al. (2006) aus, stellen aber zusätzlich die einzelnen Schritte und das Endprodukt, eine geschlossene Formel für $\hat{p}_X$, dar. Die Wahrscheinlichkeit, dass eine zufällige Realisation von X kleiner ist als eine zufällige Realisation von Y lässt sich unter der *location shift*-Annahme von $F_X(y) = F_Y(y - \delta)$ wie folgt ausdrücken:

$$p = P(X < Y) = \int F_X(y)\, dF_Y(y) = \int F_X(y)\, dF_X(y - \delta).$$

Nun wird die Wahrscheinlichkeit auf Basis der empirischen Verteilungsfunktion G_X geschätzt:

$$\begin{aligned}\hat{p}_X &= \int G_X(x) dG_X(x - \delta) \\ &= \int G_X(x + \delta) dG_X(x) \\ &= \int G_X(x + \delta) g_X(x) dx\end{aligned}$$

Dieses Integral lässt sich wie folgt berechnen: Sei $Z = \{Z_{(1)}, ..., Z_{(2m+3)}\}$ die $(2m + 3)$-elementige und aufsteigend der Größe nach geordnete Menge bestehend aus $X_{(0)}, ..., X_{(m+1)}$ und $X_{(1)} - \delta, ..., X_{(m+1)} - \delta$. Dann gilt:

$$\begin{aligned}\hat{p}_X &= \int G_X(x + \delta) g_X(x) dx \\ &= \sum_{i=1}^{2m+2} \int_{Z_{(i)}}^{Z_{(i+1)}} G_X(x + \delta) g_X(x) dx.\end{aligned}$$

Diese Umformung ist möglich, da sowohl für $x < X_{(0)}$ als auch für $x > X_{(m+1)}$ gilt, dass $g_X(x) = 0$ ist. Somit haben diese Bereiche keinen Einfluss auf den Wert des Integrals. Aus dem gleichen Grund braucht der Wert $X_{(0)} - \delta$ nicht in die Menge Z aufgenommen werden. Nun nutzen wir aus, dass im Bereich zwischen $Z_{(i)}$ und $Z_{(i+1)}$ die Dichte $g_X(x)$ konstant bei $g(Z_{(i)})$ liegt und somit aus dem Integral nach vorne gezogen werden kann. Dadurch ergibt sich

$$\hat{p}_X = \sum_{i=1}^{2m+2} g_X(Z_{(i)}) \int_{Z_{(i)}}^{Z_{(i+1)}} G_X(x + \delta)\, dx.$$

Im nächsten Schritt nutzen wir aus, dass die Funktion $G_X(x + \delta)$ zwischen $Z_{(i)}$ und $Z_{(i+1)}$ linear ansteigt. Das Integral einer linearen Funktion auf einem abgeschlossenen Intervall kann mithilfe der Sehnentrapezformel berechnet werden, so dass sich schließlich Gleichung 5 ergibt:

$$\hat{p}_X = \sum_{i=1}^{2m+2} g_X(Z_{(i)}) \times (Z_{(i+1)} - Z_{(i)}) \times \frac{1}{2} \left[G_X(Z_{(i+1)} + \delta) + G_X(Z_{(i)} + \delta) \right]. \quad (5)$$

B. Zusätzliche Ergebnisse

B.1. GLM-basiertes Verfahren

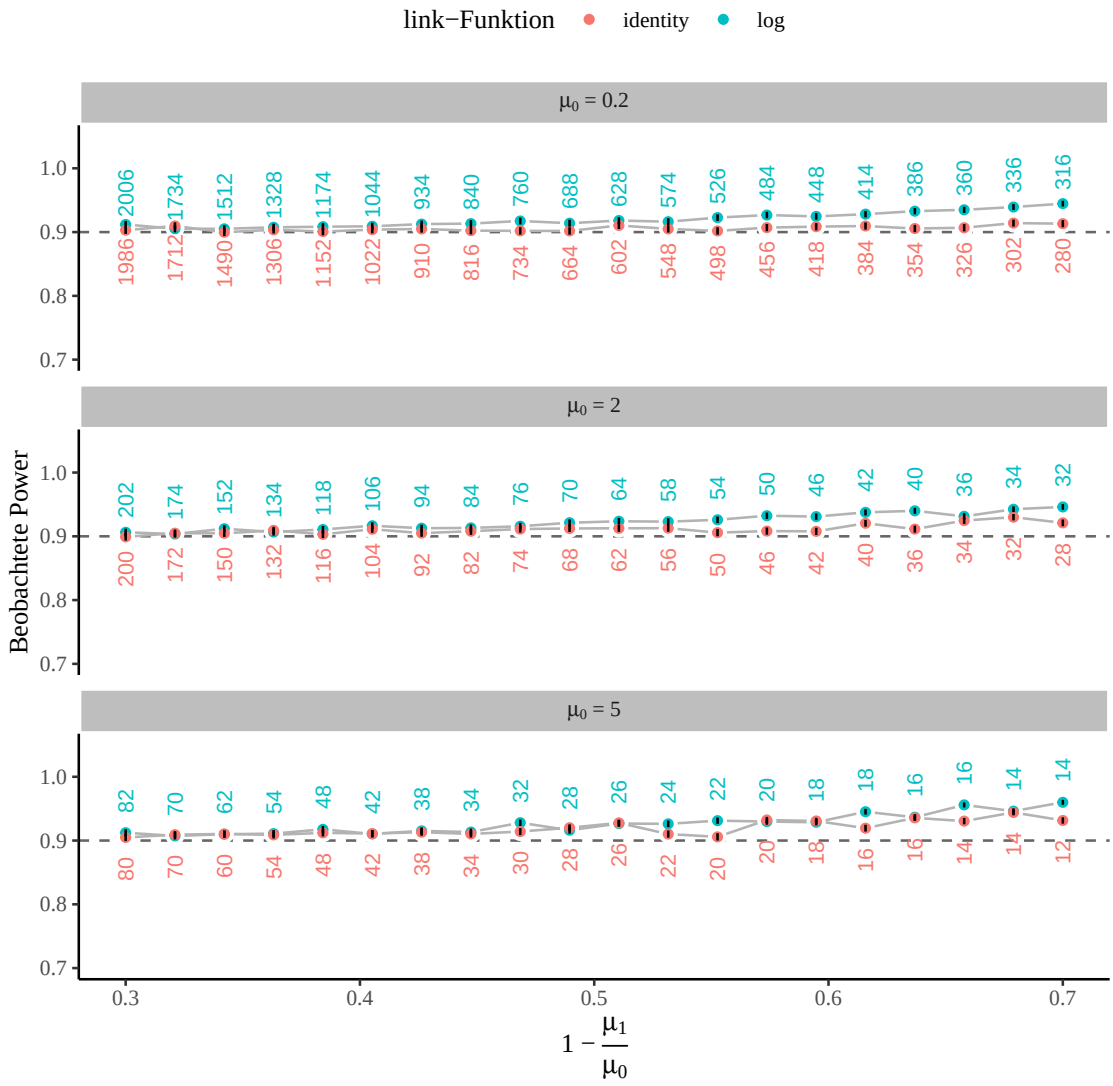

Abbildung B.1 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für poisson-verteilte Daten, entsprechend Abbildung 3. Die Abbildung zeigt Ergebnisse bei verschiedenen Werten von μ_0 und Variation der Effektstärke.

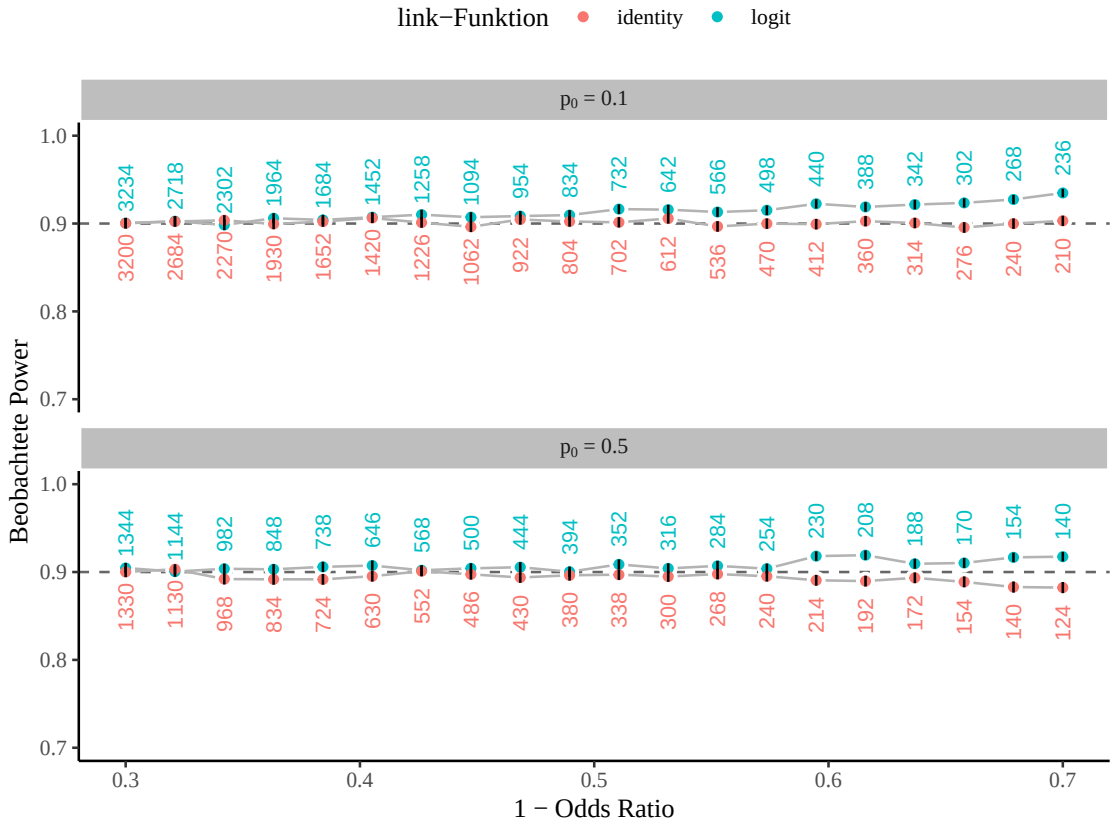

Abbildung B.2 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für binomial-verteilte Daten, entsprechend Abbildung 3. Die Abbildung zeigt Ergebnisse bei verschiedenen Wahrscheinlichkeiten p_0 und Variation der Effektstärke Odds Ratio.

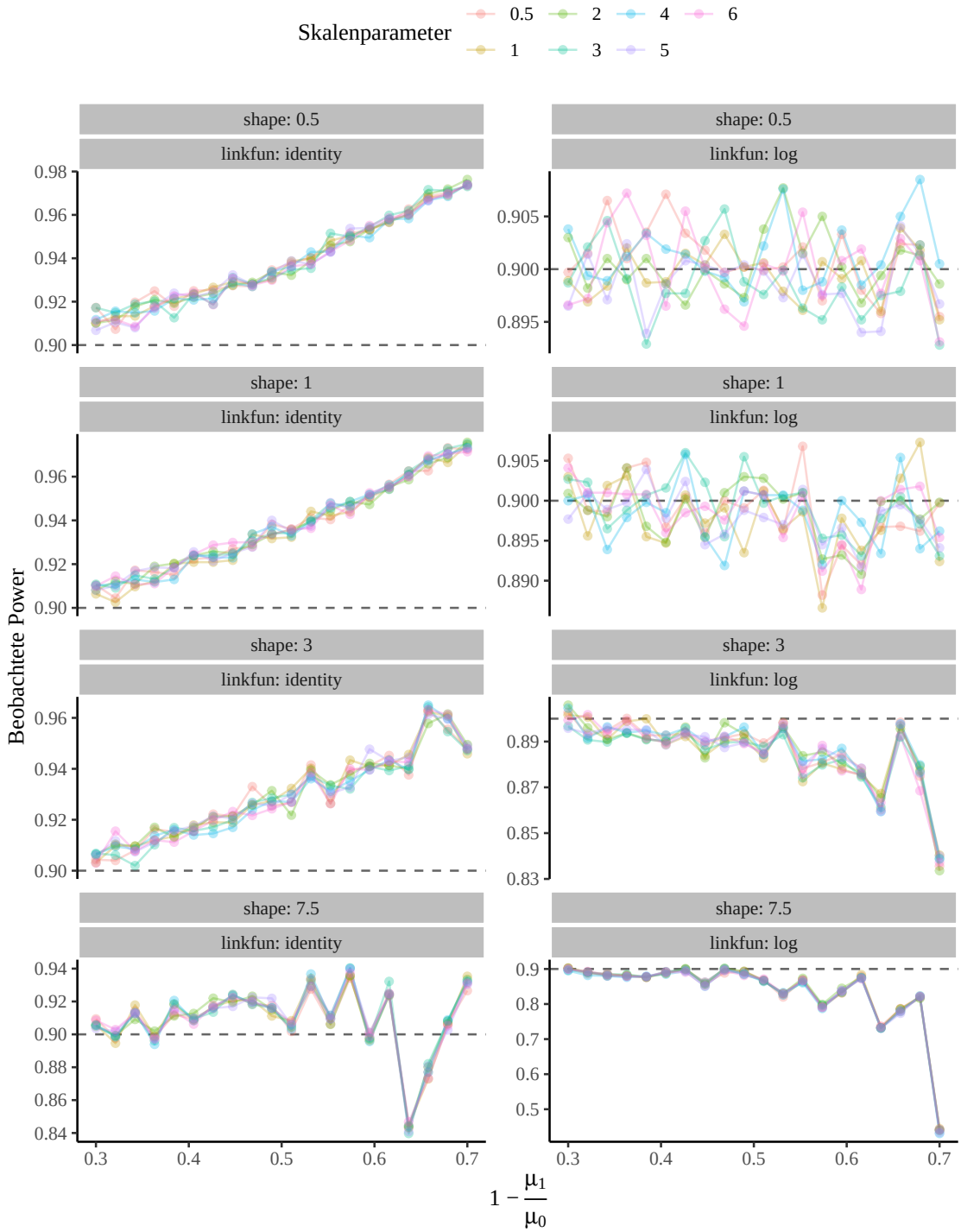

Abbildung B.3 Beobachtete Power bei Verwendung der Fallzahlabschätzung auf Basis jeweils des identity- und des log-Links für gamma-verteilte Daten. Die Abbildung zeigt Ergebnisse bei verschiedenen Skalenparametern θ und Variation der Effektstärke. Anzumerken ist, dass in den einzelnen Teilabbildungen die Skalierung der y-Achse automatisch gewählt ist, so dass sich die y-Achse bspw. bei *shape: 0.5, linkfun: log* (oben rechts) nur von 0.895 bis 0.905 erstreckt. Die Variation des Skalenparameters hat keinen erkennbaren Einfluss.

B.2. NECDF-Verfahren

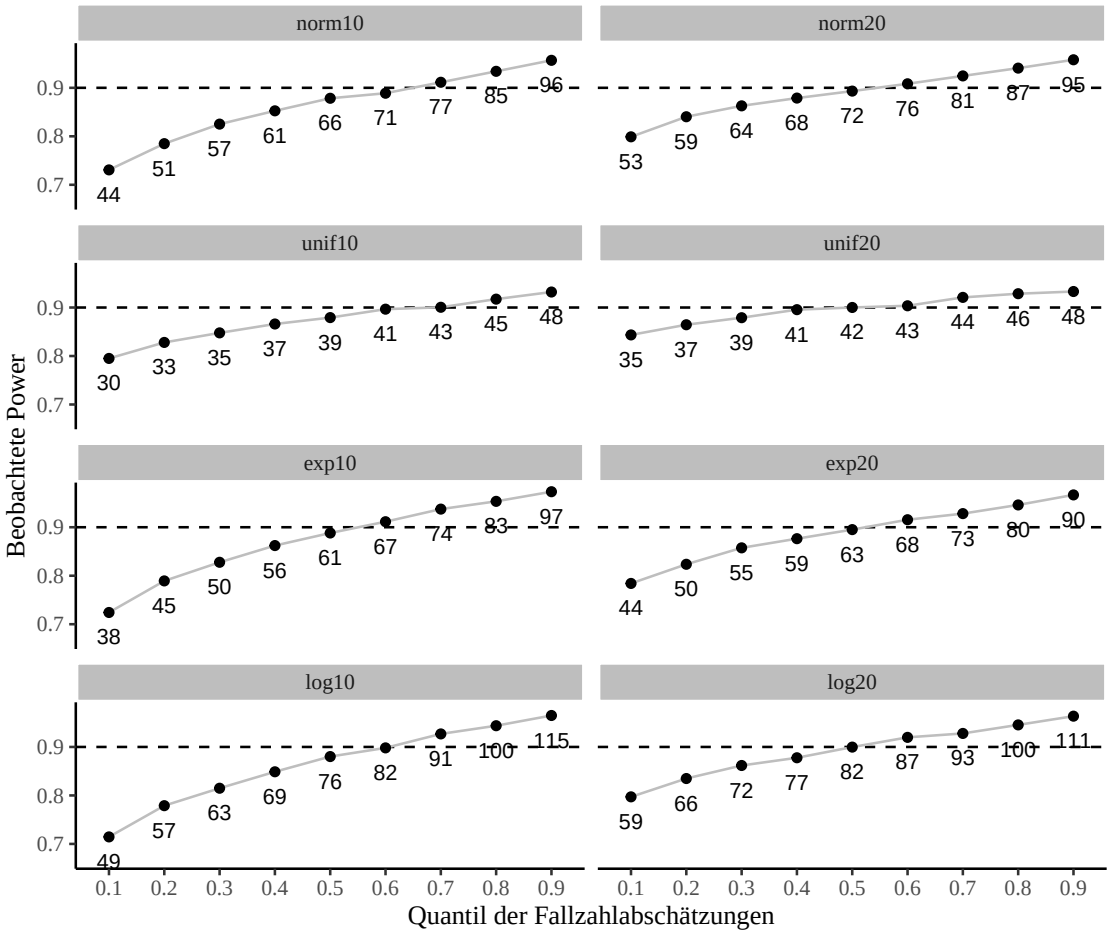

Abbildung B.4 Beobachtete Power für die Quantile der NECDF-Fallzahlabschätzungen. Die beobachtete Power basiert jeweils auf 10 000 Simulationsdurchläufen. Die Zahlen neben den Punkten zeigen die verwendete Fallzahl in *einer* Gruppe.

C. Reproduzierbarkeit und Paket

Unter dem Link https://osf.io/z5vtf/ sind der Quellcode und die Daten unserer Simulationsstudie verfügbar, so dass unsere Arbeit vollständig reproduziert und das Paket genutzt werden kann.

Zur Reproduzierung unseres Codes ist R Version 4.1 erforderlich, das Paket skewsamp kann aber auch mit älteren R-Versionen verwendet werden.

Der Quellcode zum Paket ist auf GitHub verfügbar: https://github.com/jobrachem/skewsamp