Schätzung von relativen Anteilen bei Nutzung der multinomialen Dirichlet-Verteilung

Joachim Spilke
Martin-Luther-Universität
Halle-Wittenberg,
Institut für Agrar- und
Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
06120 Halle (Saale)
joachim.spilke@landw.uni-halle.de

Norbert Mielenz
Martin-Luther-Universität
Halle-Wittenberg,
Institut für Agrar- und
Ernährungswissenschaften
Karl-Freiherr-von-Fritsch-Str. 4
06120 Halle (Saale)
norbert.mielenz@landw.uni-halle.de

Zusammenfassung

Für die Schätzung von Anteilen für Kategorien innerhalb eines Objekts wird die Nutzung der multinomialen Dirichlet-Verteilung dargestellt. Die Überprüfung dieser Vorgehensweise mit Hilfe der Monte-Carlo-Simulation zeigt für die Schätzung der Parameter und deren Differenzen die Einhaltung der Treffgenauigkeit. Demgegenüber wird für die Intervallschätzung der Parameter eine Unterschreitung des nominalen Konfidenzniveaus und Überschreitung des nominalen Fehlers 1. Art beobachtet. Als wesentlicher Grund dafür ist die Unterschätzung der geschätzten Standardfehler zu sehen.

Schlüsselwörter: multinomiale Dirichlet-Verteilung, Monte-Carlo-Simulation, PROC NLMIXED

1 Einleitung und Problemstellung

Die Schätzung von mehr als zwei relativen Anteilen für Kategorien, beobachtet an einem Objekt, spielt bei vielen praktischen Anwendungen eine Rolle. Eine solche Problemstellung tritt beispielsweise auf, wenn zur Prüfung der Embryotoxizität eines Wirkstoffes der Anteil lebender, geschädigter und toter Individuen an der Gesamtzahl der Embryonen zu schätzen ist (Chen et al., 1991). Eine ähnliche Aufgabenstellung liegt vor, wenn zur Bewertung von Wirkstoffen zur Bekämpfung von Schaderregern bei Pflanzen der Anteil lebender, geschädigter sowie toter Individuen bei gegebener Gesamtanzahl von Individuen pro Objekt (Pflanze) erfasst wird und entsprechende Schätzungen anzugeben sind.
Abgeleitet von einem derartigen Anwendungsbeispiel aus dem Pflanzenschutz wird im vorliegenden Beitrag die Anwendung der multinomialen Dirichlet-Verteilung zur Schätzung der relativen Anteile mit Hilfe der Maximum-Likelihood-Methode beschrieben. Eine weiterführende Untersuchung der Wirksamkeit der Schätzmethodik und der Hypothesenprüfung erfolgt durch Monte-Carlo-Simulation. Von besonderem Interesse ist dabei der Einfluss der Datenstruktur auf Treff- und Wiederholungsgenauigkeit der Schätzungen und die Einhaltung vorgegebener statistischer Risiken bei

Inferenzaussagen. Dazu werden Bias und Standardfehler der Schätzungen sowie die Einhaltung eines nominalen Konfidenzniveaus bzw. Fehlers 1. Art untersucht.

2 Anwendungsbeispiel

Die vorliegende Untersuchung leitet sich von einem Anwendungsbeispiel aus dem Pflanzenschutz bei landwirtschaftlichen Nutzpflanzen ab (Kaiser et al., 2010). In einem Versuch soll die Wirksamkeit verschiedener Pflanzenschutzmittel gegen den Rapsglanzkäfer (*Meligethes aeneus*) verglichen werden. Dabei liegt das besondere Interesse auf der Wirkungsdauer der eingesetzten Mittel. Die Wirkungsdauer beschreibt den Zeitabstand in Tagen zwischen der Applikation des Mittels auf die Pflanze und der Ansetzung der Käfer an die Pflanze. Die untersuchten Wirkungsdauern werden von unterschiedlichen Pflanzen repräsentiert. Es liegen somit keine wiederholten Beobachtungen je Pflanze vor.

Als Untersuchungsmerkmal wird der Anteil

- überlebender
- geschädigter und
- toter Käfer

an einer Gesamtzahl je Behandlung und Wirkungsdauer erfasst. Diese Anteile sind für die Kombination Behandlung x Wirkungsdauer zu schätzen. Da die Beobachtungen stets innerhalb eines Objekts (Pflanze) erfasst werden, wird damit systematisch eine Korrelation zwischen den Anteilen erzeugt.

3 Die multinomiale Dirichletverteilung

Für die Bearbeitung der in Abschnitt 2 skizzierten Aufgabenstellung benutzen wir als Verteilungsansatz die multinomiale Dirichlet-Verteilung (Chen et al., 1991; Johnson et al., 1997).
Entsprechend der vorliegenden Aufgabenstellung betrachten wir einen Versuch mit a x b-Kombinationen (Behandlung x Wirkungsdauer) (i=1,…,a; j=1,…,b). In jeder Kombination ij werden m_{ij} Objekte (hier: Pflanzen) verwendet. Weiter sei n_{ijk} die Anzahl der Individuen (hier: Rapsglanzkäfer), angesetzt auf das k-te Objekt in Kombination ij. Wir beschränken uns nachfolgend entsprechend der praktischen Aufgabenstellung auf den Fall von drei unterschiedlichen Ausprägungen innerhalb n_{ijk}. So sollen im vorliegenden Fall x_{ijk} und y_{ijk} die Anzahl lebender bzw. geschädigter Objekte von der Gesamtzahl n_{ijk} sein. Dann repräsentiert $z_{ijk} = n_{ijk} - (x_{ijk} + y_{ijk})$ die Anzahl toter Individuen. Weiter repräsentieren p_{ijk}, q_{ijk} und r_{ijk} ($p_{ijk} + q_{ijk} + r_{ijk}=1$) Wahrscheinlichkeiten zur Angabe des Eintretens von x_{ijk}, y_{ijk} und z_{ijk} mit der gemeinsamen trinomialen Verteilung (Chen et al., 1991):

$$P(x_{ijk}, y_{ijk}, z_{ijk}) = \frac{n_{ijk}!}{x_{ijk}!\, y_{ijk}!\, z_{ijk}!} (p_{ijk})^{x_{ijk}} (q_{ijk})^{y_{ijk}} (r_{ijk})^{z_{ijk}}. \quad (1)$$

In (1) beschreibt beispielsweise p_{ijk} die Wahrscheinlichkeit, für Kombination ij auf einer Pflanze k einen Käfer mit dem Zustand „lebend“ zu finden.
Der entscheidende Gedankengang von Chen et al. (1991) besteht nun darin, die Wahrscheinlichkeiten p_{ijk}, q_{ijk} und r_{ijk} als Zufallsgrößen einer Dirichlet-Verteilung aufzufassen. Dann besitzt die gemeinsame Dichtefunktion mit den vom Index k unabhängigen Verteilungsparametern α_{ij}, β_{ij} und γ_{ij} die folgende Gestalt:

$$f(p_{ijk}, q_{ijk}, r_{ijk}) = \frac{\Gamma(\alpha_{ij}+\beta_{ij}+\gamma_{ij})}{\Gamma(\alpha_{ij})\Gamma(\beta_{ij})\Gamma(\gamma_{ij})}(p_{ijk})^{\alpha_{ij}-1}(q_{ijk})^{\beta_{ij}-1}(r_{ijk})^{\gamma_{ij}-1}, \quad \alpha_{ij}>0, \beta_{ij}>0, \gamma_{ij}>0. \tag{2}$$

Unter der Annahme der gemeinsamen Dichtefunktion in (2) leitet sich die nachfolgend dargestellte Wahrscheinlichkeitsfunktion ab. Die Wahrscheinlichkeitsfunktion für x_{ijk}, y_{ijk} und z_{ijk}, gegeben n_{ijk}, für die multinomiale (hier: trinomiale) Dirichlet-Verteilung ist nach Johnson et al. (1997, p. 80ff):

$$P(x_{ijk}, y_{ijk}, z_{ijk}) = \frac{n_{ijk}!\Gamma(\alpha_{ij}+\beta_{ij}+\gamma_{ij})\Gamma(x_{ijk}+\alpha_{ij})\Gamma(y_{ijk}+\beta_{ij})\Gamma(z_{ijk}+\gamma_{ij})}{x_{ijk}!y_{ijk}!z_{ijk}!\Gamma(n_{ijk}+\alpha_{ij}+\beta_{ij}+\gamma_{ij})\Gamma(\alpha_{ij})\Gamma(\beta_{ij})\Gamma(\gamma_{ij})}. \tag{3}$$

Für die Erwartungswerte und Korrelationen der Zufallsgrößen X_{ijk}, Y_{ijk} und Z_{ijk} gilt (Chen et al., 1991):

$$E(X_{ijk}) = n_{ijk}\cdot\mu_{ij},\ E(Y_{ijk}) = n_{ijk}\cdot\nu_{ij},\ E(Z_{ijk}) = n_{ijk}\cdot\eta_{ij},$$

$$\rho(X_{ijk}, Y_{ijk}) = -\mu_{ij}\nu_{ij}\left[\mu_{ij}(1-\mu_{ij})\nu_{ij}(1-\nu_{ij})\right]^{-\frac{1}{2}},$$

$$\rho(X_{ijk}, Z_{ijk}) = -\mu_{ik}\eta_{ik}\left[\mu_{ik}(1-\mu_{ik})\eta_{ik}(1-\eta_{ik})\right]^{-\frac{1}{2}},$$

$$\rho(Y_{ijk}, Z_{ijk}) = -\nu_{ij}\eta_{ij}\left[\nu_{ij}(1-\nu_{ij})\eta_{ij}(1-\eta_{ij})\right]^{-\frac{1}{2}}.$$

Die gesuchten Anteile μ_{ij}, ν_{ij} und η_{ij} ergeben sich aus:

$$\mu_{ij} = \alpha_{ij}/(\alpha_{ij}+\beta_{ij}+\gamma_{ij}),\ \nu_{ij} = \beta_{ij}/(\alpha_{ij}+\beta_{ij}+\gamma_{ij}),\ \eta_{ij} = \gamma_{ij}/(\alpha_{ij}+\beta_{ij}+\gamma_{ij}). \tag{4}$$

4 Effektschätzung und rechentechnische Umsetzung

Die Bereitstellung von Schätzwerten der Parameter in (4) gelingt bei Nutzung der Maximum-Likelihood-Methode, wobei die nachfolgend angegebene Likelihood, ausgedrückt als Beitrag des k-ten Objekts von Behandlung ij, zu maximieren ist:

$$\begin{aligned}\log L = \text{const} &+ \log\Gamma(\alpha_{ij}+\beta_{ij}+\gamma_{ij}) + \log\Gamma(x_{ijk}+\alpha_{ij}) + \log\Gamma(y_{ijk}+\beta_{ij}) + \log\Gamma(z_{ijk}+\gamma_{ij}) \\ &- \log\Gamma(n_{ijk}+\alpha_{ij}+\beta_{ij}+\gamma_{ij}) - \log\Gamma(\alpha_{ij}) - \log\Gamma(\beta_{ij}) - \log\Gamma(\gamma_{ij})\end{aligned} \tag{5}$$

(const = nur von den Beobachtungen abhängige Konstante).
Die numerische Umsetzung erfolgt bei Nutzung von Proc NLMIXED. Nachfolgend ist ein Auszug aus der zugehörigen Programmdatei aufgeführt (Abbildung 1).

```
PROC NLMIXED
…
 IF (behandlung=1 AND zeit=1) THEN
    ll =   LGAMMA(alpha_1_1 + beta_1_1 + gamma_1_1)
              + LGAMMA(x + alpha_1_1)
           + LGAMMA(y + beta_1_1)
           + LGAMMA(z + gamma_1_1)
           - LGAMMA(n + alpha_1_1 + beta_1_1 + gamma_1_1)
           - LGAMMA(alpha_1_1)
           - LGAMMA(beta_1_1)
              - LGAMMA(gamma_1_1);
…
 model  n ~ general(ll);
…
```

Abbildung 1: Auszug aus den Statements innerhalb Proc NLMIXED zur Gewinnung von Parameterschätzungen der trinomialen Dirichlet-Verteilung gemäß (5)

5 Stochastische Simulation

5.1 Datenstruktur und genutzte SAS-Funktionen

Zur Überprüfung der Aussagefähigkeit des von uns gewählten Auswertungsansatzes führen wir eine stochastische Simulation durch.
Die Simulation basiert auf den in Tabelle 1 zusammengestellten Parametern. Bei der Wahl der Parameter erfolgte eine Anlehnung an praktische Verhältnisse sowohl bezüglich der unterschiedlichen Wahrscheinlichkeiten je Behandlung als auch die Veränderung der Wahrscheinlichkeiten in Abhängigkeit der Wirkungsdauer. In der vorliegenden Simulation wurden die Wirkungsdauern 2 und 5 Tage beachtet.

Tabelle 1: Simulationsparameter

Behandlung	Wirkungsdauer	Kategorie	Parameter	Wirkungsdauer	Kategorie	Parameter
1	1 (2 Tage)	lebend	0.10	2 (5 Tage)	lebend	0.15
		geschädigt	0.30		geschädigt	0.45
		tot	0.60		tot	0.40
2	1 (2 Tage)	lebend	0.50	2 (5 Tage)	lebend	0.15
		geschädigt	0.25		geschädigt	0.45
		tot	0.25		tot	0.40
3	1 (2 Tage)	lebend	0.25	2 (5 Tage)	lebend	0.25
		geschädigt	0.50		geschädigt	0.60
		tot	0.25		tot	0.15

Bei der Wahl der Datenstrukturen lehnen wir uns an die oben kurz skizzierte praktische Problemstellung an. Wir simulieren den Fall von 10, 15, 20 Pflanzen mit jeweils 5, 10, 15, 20 Käfern. Somit ergeben sich 12 Strukturvarianten (Tabelle 2).

Tabelle 2: Simulationsvarianten für Anzahl Pflanzen und Käfer je Pflanze je Kombination Behandlung x Wirkungsdauer sowie die verwendeten Abkürzungen (Anzahl Pflanzen_Anzahl Käfer)

		Anzahl Käfer je Pflanze			
		5	10	15	20
Anzahl Pflanzen	10	10_5	10_10	10_15	10_20
	15	15_5	15_10	15_15	15_20
	20	20_5	20_10	20_15	20_20

Die Simulation erfolgte bei Nutzung von Proc IML der Software SAS. Da keine Funktion zur Erzeugung der multinomialen Dirichlet-Verteilung genügender Zufallszahlen vorlag, erfolgte die Simulation gemäß eines Vorschlags von Neerchal and Morel (2005) zweistufig:

1. Stufe: Simulation der Realisationen (Anteile der Kategorien) gemäß einer Dirichlet-Verteilung bei Vorgabe der in Tabelle 1 zusammen gestellten Parameter (Nutzung der Funktion RANDDIRICHLET in Proc IML)
2. Stufe: Simulation der Realisationen (Anzahlen je Kategorie) bei Vorgabe der Anzahl Versuche n je Objekt m (hier Käfer je Pflanze) bei Verwendung der in Stufe 1 erzeugten Anteile (Nutzung der Funktion RANDMULTINOMIAL in Proc IML).

Für jede Variante wurden 10000 Simulationen durchgeführt.

5.2 Simulationsergebnisse

5.2.1 Schätzung der Anteile

Die Ergebnisdarstellung wird auf Behandlung 1, Wirkungsdauer 1 (2 Tage) begrenzt. Die hier gefundenen Ergebnisse gelten entsprechend auch für die übrigen Behandlungen und Wirkungsdauern.
In Abbildung 2 sind zunächst die mittleren Schätzungen für die simulierten Strukturvarianten zusammengestellt. Dabei kann unabhängig von der Struktur eine Einhaltung der Treffgenauigkeit beobachtet werden.

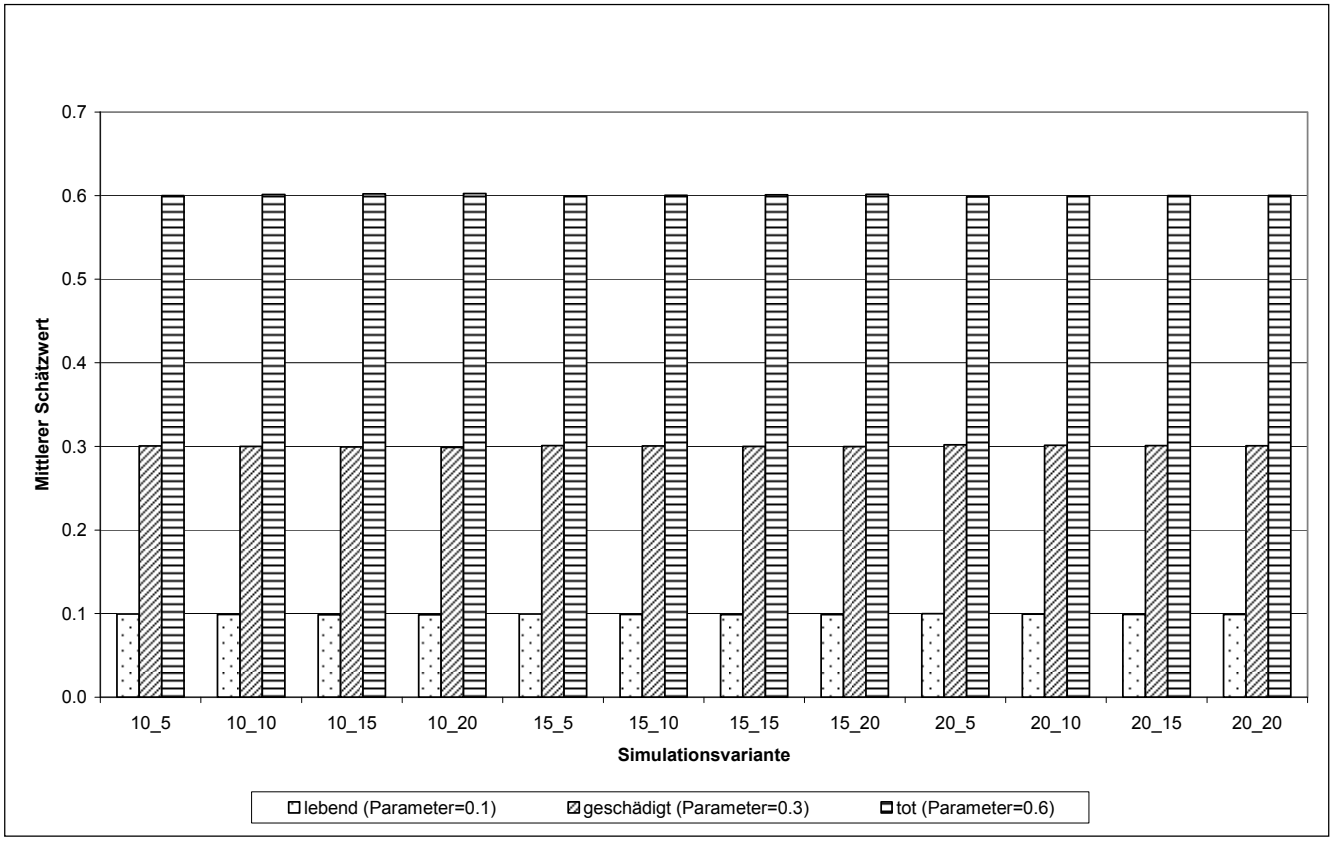

Abbildung 2: Schätzwerte der Anteile (lebend, geschädigt und tot) für Behandlung 1 und Wirkungsdauer 1 in Abhängigkeit der untersuchten Simulationsvarianten

Die in Abbildung 2 dargestellten Ergebnisse zeigen eine sehr gute Erwartungstreue, die durch Angabe des prozentualen Bias untermauert werden kann. Der Bias beträgt für alle untersuchten Varianten maximal 1.3% vom zugehörigen Parameter.
Für die Schätzungen der Erwartungswerte wurden zweiseitige Konfidenzintervalle berechnet. Es ist nun weiter bedeutsam, inwiefern das vorgegebene Konfidenzniveau $P = 1 - \alpha$ eingehalten wird. In Abbildung 3 sind die Werte für das realisierte Konfidenzniveau $\hat{P}$ zusammengestellt.
Die Ergebnisse zeigen eine durchgehende Überschreitung des nominalen $\alpha = 0.05$ und damit eine Unterschreitung des nominalen Konfidenzniveaus P=0.95. Das bedeutet, der jeweilige Parameter wird seltener überdeckt, als mit dem vorgegebenen Konfidenzniveau gefordert.

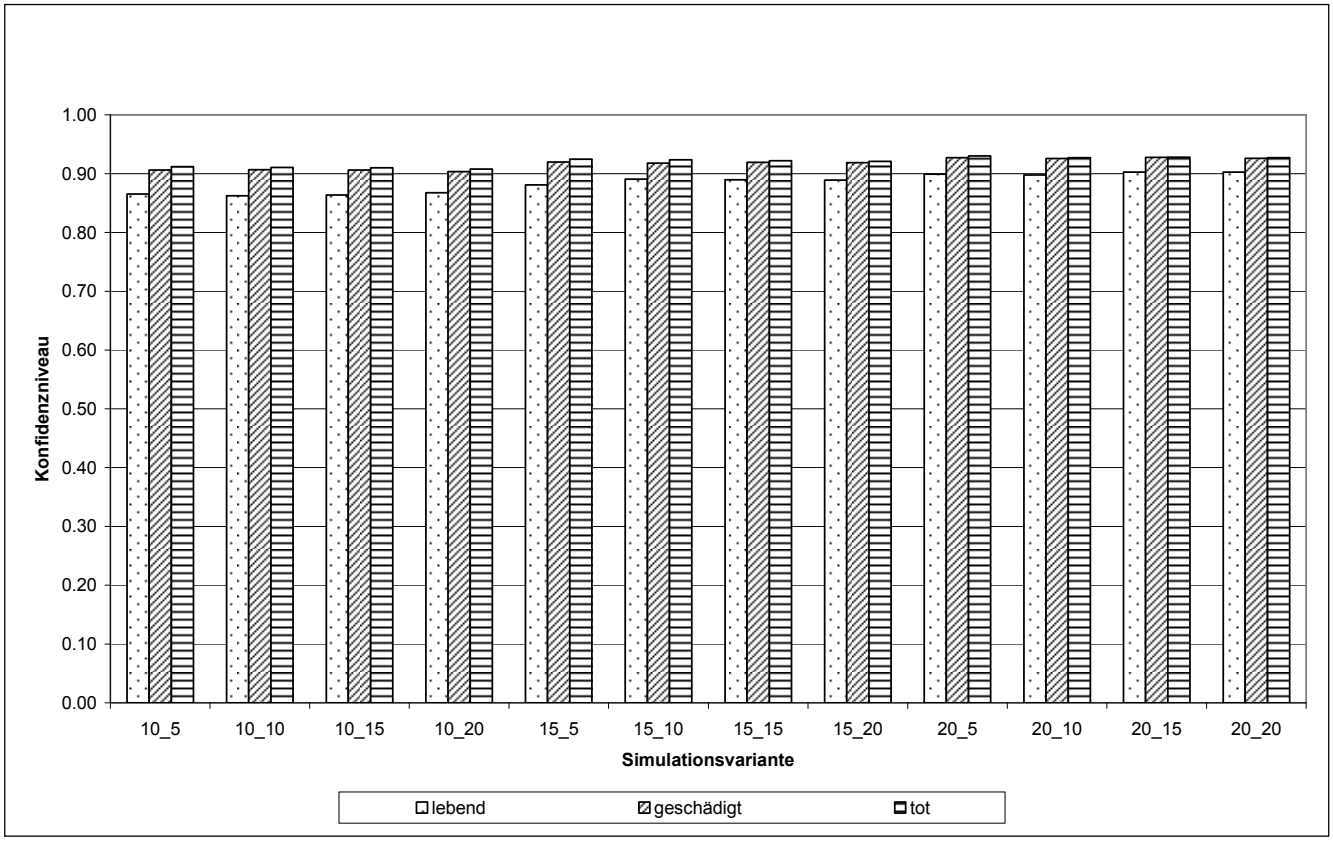

Abbildung 3: Realisierte Werte $\hat{P}$ für das Konfidenzniveau $P = 0.95$ der Schätzwerte der Anteile (lebend, geschädigt und tot) für Behandlung 1 und Wirkungsdauer 1 in Abhängigkeit der untersuchten Simulationsvarianten

Bedeutsam ist, dass die Verschätzung zwar mit Zunahme der Anzahl Objekte und Versuche je Objekt (Käfer je Pflanze) tendenziell abnimmt, aber auch bei einem vergleichsweise großem Datenumfang (20_20) noch zu beobachten ist.
Die vorliegende Simulationsstudie erlaubt weiterhin den Vergleich der geschätzten und beobachteten Standardfehler der Schätzungen. Der sog. „beobachtete Standardfehler" (SE_obs) resultiert aus der Berechnung des Standardfehlers der je Simulationslauf gewonnenen Schätzungen. D.h. im vorliegenden Fall aus 10000 Schätzwerten. Der „geschätzte Standardfehler" (SE_est) ist das Mittel der je Simulationslauf gewonnenen Standardfehler der Schätzungen. Zur besseren Vergleichbarkeit wird die Differenz bezogen auf den „beobachteten" Standardfehler dargestellt,

$$\text{Diff\_SE}(\%) = \frac{\text{SE\_est} - \text{SE\_obs}}{\text{SE\_obs}} * 100 . \qquad (6)$$

Die Bezugnahme auf den „beobachteten" Standardfehler beruht auf der Annahme, dass damit eine gute Widerspiegelung des tatsächlichen Standardfehlers der Schätzung gelingt.

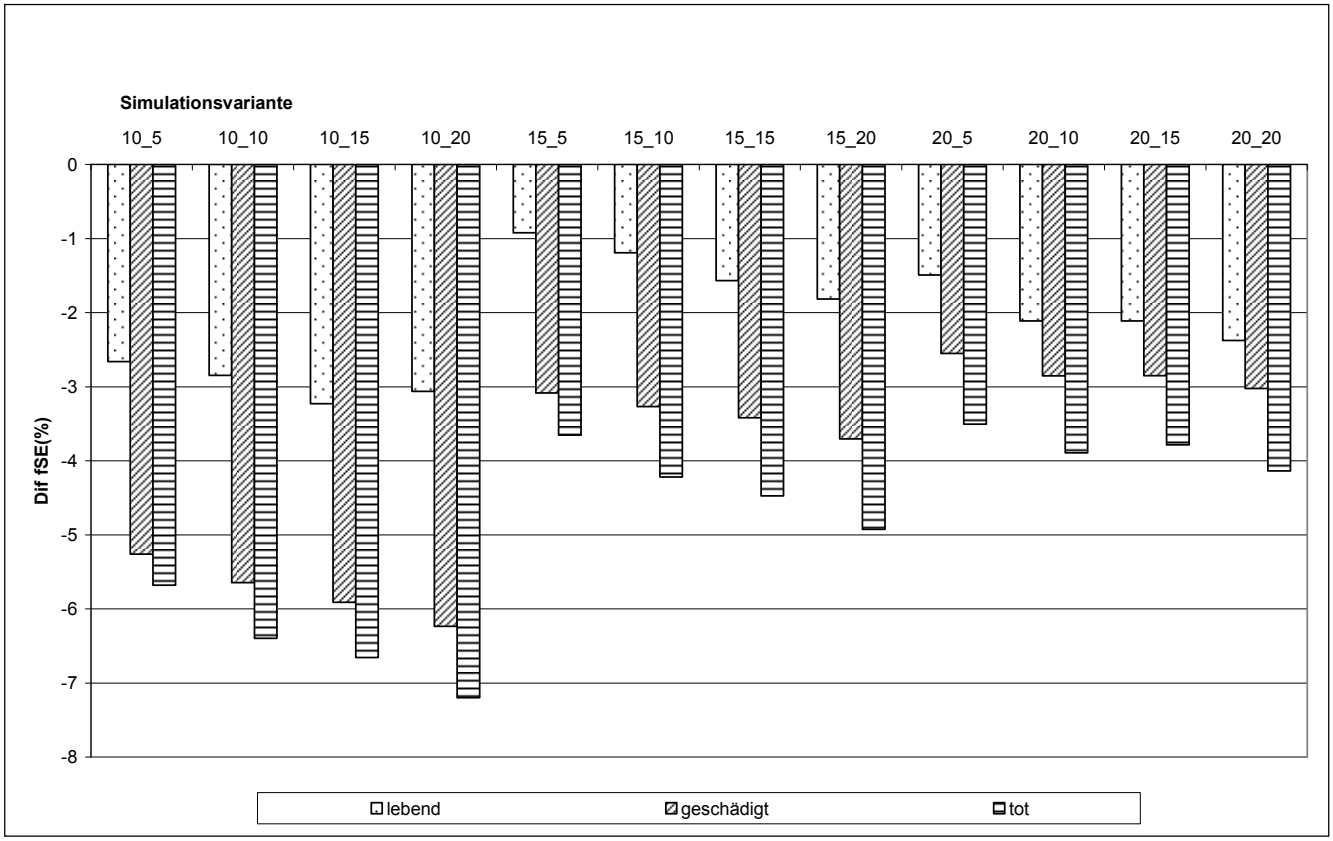

Abbildung 4: Differenz zwischen beobachtetem und geschätztem Standardfehler der Anteile (in Prozent des beobachteten Standardfehlers) für Behandlung 1 und Wirkungsdauer 1 in Abhängigkeit der untersuchten Simulationsvarianten

Die Ergebnisse in Abbildung 4 zeigen eine durchgehende Unterschätzung des „beobachteten“ Standardfehlers durch den „geschätzten“ Standardfehler. Dabei beträgt die maximale Unterschätzung etwa 7.2% und muss als wesentliche Erklärung für die beobachtete Abweichung vom nominalen Konfidenzniveau angesehen werden. Die Unterschätzung ist abhängig vom betrachteten Anteil und nimmt mit zunehmendem Stichprobenumfang ab. Die Unterschätzung für Parameter nahe 0.5 ist größer, obwohl die Verschätzung des Konfidenzniveaus für diese Fälle geringer ausfällt (vgl. Abbildung 3).

5.2.2 Schätzung von Kontrasten der Anteile

Die Ergebnisdarstellung für die Kontraste bezieht sich auf ausgewählte Hypothesen bei Gültigkeit der Nullhypothese (Tabelle 3).

Tabelle 3: Untersuchte Kontraste

Behandlung	Wirkungsdauer	Kategorie	Parameter
1-2	2 (5 Tage)	lebend	0.15-0.15= 0
1-2	2 (5 Tage)	geschädigt	0.45-0.45= 0

Für die Kontraste kann ebenfalls die Einhaltung der Treffgenauigkeit beobachtet werden (Detailergebnisse sind hier nicht gezeigt). Weiterhin ist von Interesse, wie im Fall der Gültigkeit der Nullhypothese der nominale Fehler 1. Art eingehalten wird. Die Ergebnisse in Abbildung 5 zeigen eine durchgehende Überschreitung des nominalen Risikos in Abhängigkeit des betrachteten Kontrasts und der Datenstruktur.

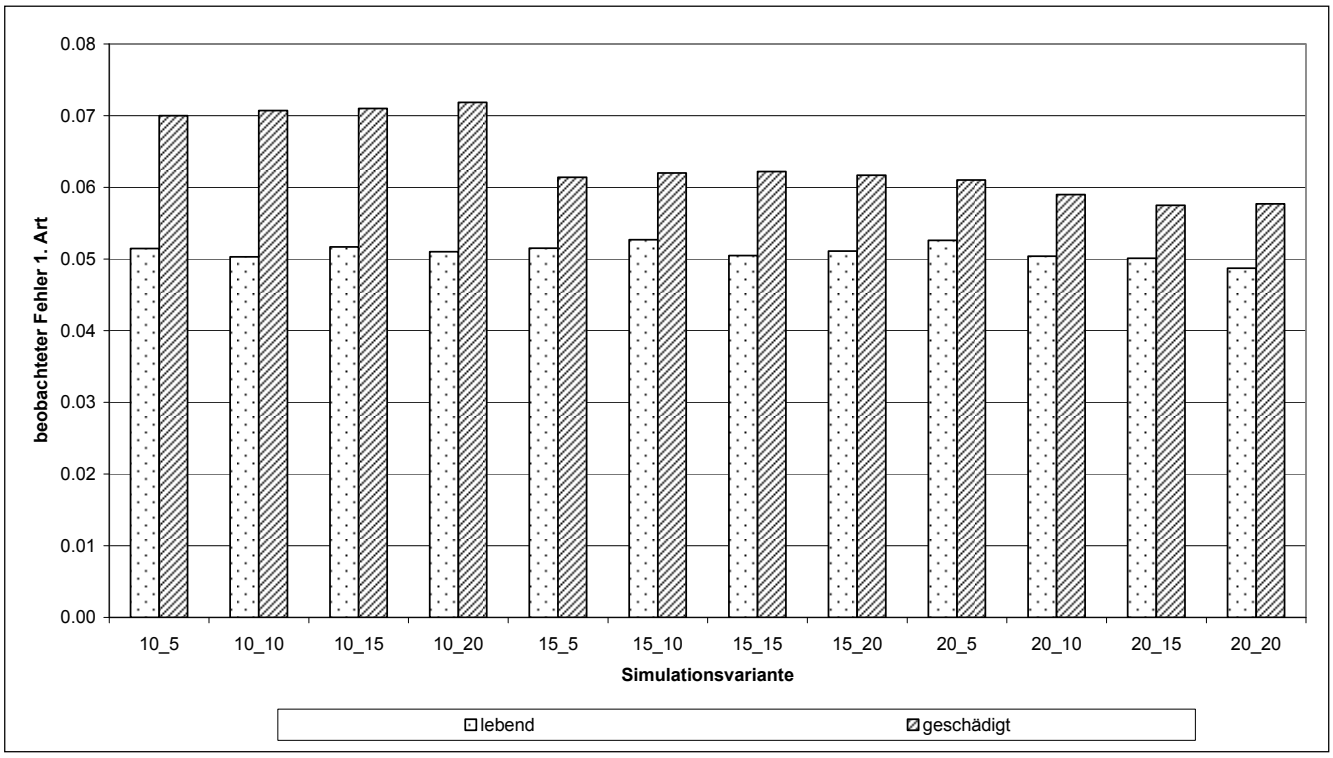

Abbildung 5: Beobachteter Fehler 1. Art (nominal $\alpha = 0.05$) für die Kontraste Behandlung 1-2 zu Termin 2 (lebend bzw. geschädigt) in Abhängigkeit der untersuchten Simulationsvarianten

Der zugehörige Vergleich zwischen beobachtetem und mittlerem geschätzten Standardfehler zeigt wiederum eine Unterschätzung in Abhängigkeit des beobachteten Kontrasts, wobei bei zunehmendem Stichprobenumfang die Unterschätzung abnimmt (Abbildung 6).

6 Diskussion und Schlussfolgerungen

Die multinomiale Dirichlet-Verteilung stellt eine gut geeignete Verteilung zur Schätzung von Anteilen für an einem Objekt erfasste Kategorien dar. Die besondere Eignung folgt auch daraus, dass die Korrelation zwischen den Beobachtungen innerhalb eines Objekts durch diesen Verteilungsansatz berücksichtigt wird (Chen et al., 1991). Die zugehörige Likelihood ist in SAS standardmäßig nicht verfügbar, kann aber bei Nutzung der Proc NLMIXED formuliert werden.
Die Überprüfung dieser Vorgehensweise durch Simulation zeigt erwartungstreue Schätzungen der Parameter und deren Differenzen. Demgegenüber werden für die realisierten Konfidenzniveaus geringere und die realisierten Fehler 1. Art höhere Werte

gegenüber den nominalen Niveaus gefunden. Als wesentliche Ursache für diesen Befund ist die teilweise bedeutsame Unterschätzung der Standardfehler anzusehen.

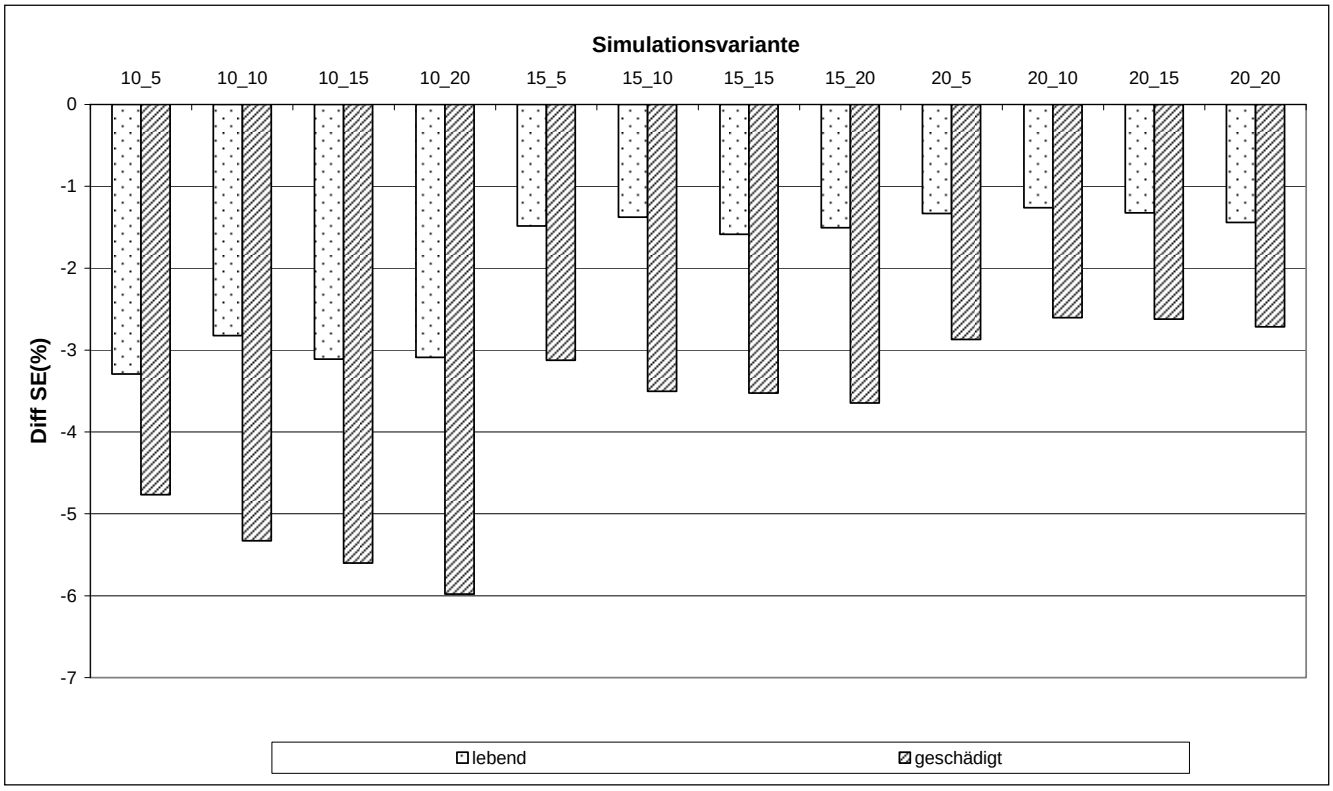

Abbildung 6: Differenz zwischen geschätztem und wahrem Standardfehler, prozentual vom wahren Standardfehler für die Kontraste Behandlung 1-2 zu Termin 2 (lebend bzw. geschädigt) in Abhängigkeit der untersuchten Simulationsvarianten

Literatur

[1] Chen J.J.; Kodell, R.L.; Howe, R.B.; Gaylor, D.W.: Analysis of trinomial responses from reproductive and developmental toxicity experiments. Biometrics 47 (1991) 1049-1058.

[2] Johnson, N.L.; Kotz, S.; Balakrishnan, N.: Discrete multivariate distributions. John Wiley & Sons. INC. New York 1997.

[3] Kaiser, C.; Grunau, S.; Müller, B.; Spilke, J.; Volkmar, C.: Wirkung von Insektiziden gegenüber dem Rapsschädling Meligethes aeneus. Julius-Kühn-Archiv 428 (2010) 503.

[4] Neerchal, K.N., Morel, J.G.: An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. Computational Statistics & Data Analysis 49 (2005) 33-43.