MASTERARBEIT

Maximal selektierte Chi2-Statistiken zur Auswertung ordinaler Zielgrößen in randomisierten klinischen Studien

Institut für Statistik
Institut für Medizinische Informationsverarbeitung,
Biometrie und Epidemiologie
LMU München

Autor: Christian Reinhold Bihl
Betreuer: Prof. Dr. Anne-Laure Boulesteix
Prof. Dr. Eva Hoster
Datum: 4. Januar 2021

Abstract

Zur Beurteilung medizinischer Studien können verschiedene statistische Tests angewandt werden. Viele dieser Tests verwenden allerdings, insbesondere bei ordinalen Strukturen, lediglich approximative Verteilungen oder berücksichtigen die ordinale Struktur überhaupt nicht.
Das Ziel dieser Arbeit ist die Bewertung der Methode der maximal selektierten Chi2-Statistiken nach Boulesteix. Dieser Test wurde im Jahr 2006 veröffentlicht und beinhaltet die Berechnung der exakten Verteilung von maximal selektierten Chi2-Statistiken bei der Auswertung ordinaler Zielgrößen.

Zur Beantwortung dieser Fragestellung, ist eine Simulationsstudie mit Hilfe eines eigenen R-Paketes durchgeführt worden. Die Methode der maximal selektierten Chi2-Statistiken nach Boulesteix ist dabei, zusammen mit drei weiteren gängigen statistischen Tests, im Bezug auf die erreichte Power und die Höhe des Fehlers 1. Art analysiert worden. Dabei sind verschiedene Szenarios mit ordinalen Strukturen verwendet worden, welche möglichst unterschiedliche Abbilder der Realität erfassen sollten.

Die Ergebnisse der Simulationsstudie zeigen, dass die Methode der maximal selektierten Chi2-Statistiken nach Boulesteix große Probleme bei der Beschränkung des Fehlers 1. Art aufweist. Als positiv kann der zusätzlich ermittelte Schwellenwert betrachtet werden sowie die erreichte Power des Tests.

Aufgrund des deutlich erhöhten Fehlers 1. Art stellt die Methode der maximal selektierten Chi2-Statistiken nach Boulesteix aktuell noch keine gleichwertige oder bessere Alternative zu herkömmlichen Tests dar. Dies könnte sich allerdings ändern, wenn beispielsweise durch eine Stetigkeitskorrektur der Fehler 1. Art erfolgreich begrenzt wird.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Statistische Methoden sind längst ein etablierter Teil unserer Wirtschaft, der Medizin und der Psychologie. Um beispielsweise einen therapeutischen Nutzen bewerten zu können, stützt sich die medizinische Forschung auf klinische Studien (Wang et al., 2007). Diese können anschließend mit Hilfe statistischer Tests analysiert werden. Im Rahmen dieser Masterarbeit wird die Methode der maximal selektierten Chi2 (Chi-Quadrat/χ^2)-Statistiken untersucht. Dabei wird speziell die Fähigkeit zur Auswertung ordinaler Zielgrößen in randomisierten klinischen Studien betrachtet und bewertet.

Randomisierte klinische Studien, welche eine spezielle Art der Kohortenstudie darstellen, zeichnen sich dadurch aus, dass die Studienteilnehmer zufällig in eine Versuchsgruppe und eine Kontrollgruppe aufgeteilt werden. Dabei werden in der Versuchsgruppe neue Behandlungsmethoden getestet, während in der Kontrollgruppe die bisherige Behandlungsmethode angewandt oder beispielsweise ein Placebo verabreicht wird. Diese zufällige Einteilung in Kontroll- und Versuchsgruppe vermeidet damit im Idealfall einen *Selection Bias*, eine Verzerrung, die durch die Zuweisung der Teilnehmer durch einen Arzt entstehen könnte. Zusätzlich werden in RCTs auch oft Verblindungen eingesetzt. Das heißt, der Studienteilnehmer weiß nicht, ob er in der Versuchs- oder in der Kontrollgruppe eingeordnet wurde. Weiß auch der behandelnde Arzt nicht, in welcher Gruppe sich der Studienteilnehmer befindet, so wird von einer Doppelblindstudie gesprochen (Stel et al., 2007).

Zur Auswertung von randomisierten klinischen Studien (RCTs) werden häufig sogenannte Unabhängigkeitstests verwendet. Dabei wird geprüft, ob zwei Merkmale stochastisch unabhängig voneinander sind. Im Zusammenhang mit RCTs wird also geprüft, ob sich die Verteilung der Ergebnisse in der Kontrollgruppe signifikant von der Verteilung der Ergebnisse in der Versuchsgruppe unterscheidet. Liegen die Daten in ordinaler Form vor, so sollten idealerweise statistische Tests verwendet werden, die diese Struktur berücksichtigen. Die Methode der maximal selektieren χ^2-Statistiken nach Boulesteix (2006) bezieht eben jene ordinale Struktur in die Berechnung mit ein und ermittelt zusätzlichen einen bestmöglichen Schwellenwert, der die ordinale Zielgröße in zwei Gruppen unterteilt. Die Methode kann als Erweiterung der Ansätze von Miller und Siegmund (1982) und Koziol (1991) betrachtet werden, die sich ebenfalls mit der Verteilung der maximal selektierten χ^2-Statistiken beschäftigten.

Den Schwerpunkt dieser Masterarbeit bildet die Untersuchung und Bewertung der Methode der maximal selektieren χ^2-Statistiken nach Boulesteix für ordinale Daten. Hierfür wurde ein veraltetes R-Paket verwendet, welches im Rahmen dieser Arbeit durch zusätzliche Funktionen erweitert und modernisiert wurde. Die Untersuchung der Methode erfolgte durch eine Simulationsstudie, welche die Einhaltung des Fehlers 1. Art und die erreichte Power des Tests ermittelte. Zur Bewertung wurden zusätzlich die ebenfalls simulierten Daten dreier weiterer gängiger Tests verwendet und mit der Methode der maximal selektieren χ^2-Statistiken verglichen.

Im Folgenden (Kapitel 2) werden mit dem *Pearson* χ^2-Test, dem *exakten Fisher*-Test und dem *Wilcoxon*-Rangsummentest drei weitere statistische Tests vorgestellt, welche zur Auswertung von ordinalen Daten verwendet werden können. Anschließend folgt in

Kapitel 3 die Methodik der maximal selektierten χ^2-Statistiken nach Boulesteix. Zur anwenderfreundlichen und einfachen Benutzung der Methode wurde zusätzlich ein R-Paket entwickelt. Dieses enthält die wichtigsten Funktionen zur Berechnung und Darstellung der maximal selektierten χ^2-Statistiken nach Boulesteix. Die verwendeten Funktionen des Paketes werden in Kapitel 4 genauer betrachtet und anschaulich erklärt. Kapitel 5 behandelt anschließend die durchgeführte Simulationsstudie. Dabei wurde die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix gemeinsam mit drei weiteren statistischen Tests verglichen. Als Kriterien für die Bewertung wurde die erreichte Power und die Begrenzung des Fehlers 1. Art angewendet. In Kapitel 6 wird die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix mit Hilfe des entwickelten R-Paketes in einem Anwendungsbeispiel verwendet. Die Masterarbeit wird mit Kapitel 7 abgeschlossen, welches das Fazit und einen Ausblick über die Thematik der maximal selektierten χ^2-Statistiken nach Boulesteix liefert. In Kapitel 8 befindet sich der Anhang dieser Arbeit und eine Auflistung aller im digitalen Anhang befindlichen Dateien.

2 Methoden zur Auswertung ordinaler Daten

In klinischen Studien geht es oft um die Frage, ob ein neues Medikament oder eine neue Behandlung ein Fortschritt zum bisherigen Goldstandard ist. Zur Beantwortung dieser Frage können diverse Studiendesigns angewandt werden. Dazu zählen beispielsweise Beobachtungsstudien wie Kohorten- oder Fall-Kontroll-Studien oder aber randomisierte kontrollierte Studien. Dabei werden besonders RCTs als Goldstandard zur Erkennung von kausalen Effekten in Behandlungen betrachtet (Noordzij et al., 2009).

Im Folgenden werden nun verschiedene Tests vorgestellt, die zur Auswertung von ordinalen Daten verwendet werden können und in der Wissenschaft relativ weit verbreitet sind. Diese Tests werden anschließend in Kapitel 5 gemeinsam mit der Methode der maximal selektierten χ^2-Statistiken nach Boulesteix bezüglich der Power und der Höhe des Fehlers 1. Art miteinander verglichen.

Notation

Die im Rahmen dieser Masterarbeit untersuchte Datenstruktur entspricht einer zweiarmigen Studie mit einer ordinalen Zielgröße. Innerhalb dieser Struktur könnten beispielsweise Beobachtungen mit $Y = 1$ der Kontrollgruppe und $Y = 2$ der Versuchsgruppe angehören. Die ordinale Zufallsvariable X entspricht dann der ordinalen Zielgröße mit den Ausprägungen $X \in \{1, \ldots, k\}$. Die entsprechende $2 \times k$-Kontingenztabelle würde sich damit ergeben durch

		X				
		1	2	...	k	
Y	1	h_{11}	h_{12}	...	h_{1k}	$h_{1.}$
	2	h_{21}	h_{22}	...	h_{2k}	$h_{2.}$
		$h_{.1}$	$h_{.2}$	...	$h_{.k}$	n

mit h_{ij} als Anzahl der Beobachtungen mit $(Y = i, X = j)$ für $i \in \{1, 2\}$ und $j \in \{1, \ldots, k\}$. Die nachfolgenden Tests sind meist in einer generalisierten Form mit Hilfe von $m \times k$-Kontingenztabellen mit $m \geq 2$ erklärt.

2.1 Pearson χ^2-Test

Der χ^2-Test wurde erstmals im Jahr 1900 von Karl Pearson veröffentlicht(Pearson, 1900). Der χ^2-Unabhängigkeitstest gehört zu den Hypothesentests, mit denen man zwei Merkmale auf ihre stochastische Unabhängigkeit prüfen kann. Handelt es sich bei den Stichprobenvariablen X und Y um kategoriale oder kategorisierte Merkmale, so lassen sie sich in einer Kontingenztafel betrachten.
Es gilt für $X \in \{1, ..., k\}$ und $Y \in \{1, ..., m\}$ mit h_{ij} als Anzahl der Beobachtungen welche die Ausprägungen $(X = i, Y = j)$ haben.

		Y			
		1	...	m	
X	1	h_{11}	...	h_{1m}	$h_{1.}$
	2	h_{21}	...	h_{2m}	$h_{2.}$
	⋮	⋮		⋮	⋮
	k	h_{k1}	...	h_{km}	$h_{k.}$
		$h_{.1}$	...	$h_{.m}$	n

(Fahrmeir et al., 2016)

Für die Hypothesen ergibt sich damit:

$$\mathbf{H_0}: \quad P(X = i, Y = j) = P(X = i) \cdot P(Y = j) \quad \text{für alle } i, j \quad \textbf{vs.}$$
$$\mathbf{H_1}: \quad P(X = i, Y = j) \neq P(X = i) \cdot P(Y = j) \quad \text{für mindestens ein } i, j.$$

Unter der Nullhypothese H_0, dass die Stichprobenvariablen X und Y unabhängig sind, lassen sich dementsprechend die erwarteten Häufigkeiten für die einzelnen Ereignisse durch die Produkte der entsprechenden Randsummen berechnen.

		Y			
		1	...	m	
X	1	$\frac{h_{1.}h_{.1}}{n}$	...	$\frac{h_{1.}h_{.m}}{n}$	$h_{1.}$
	2	$\frac{h_{2.}h_{.1}}{n}$	...	$\frac{h_{2.}h_{.m}}{n}$	$h_{2.}$
	⋮	⋮		⋮	⋮
	k	$\frac{h_{k.}h_{.1}}{n}$	...	$\frac{h_{k.}h_{.m}}{n}$	$h_{k.}$
		$h_{.1}$	...	$h_{.m}$	n

(Fahrmeir et al., 2016)

Die χ^2-Teststatistik basiert dann auf den Unterschieden zwischen den tatsächlichen Beobachtungen und den zu erwarteten Beobachtungszahlen und ergibt sich aus

$$\text{Teststatistik:} \quad \chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}ij)^2}{\tilde{h}_{ij}} \tag{1}$$

mit $\tilde{h}_{ij} = \frac{h_{i.}h_{.j}}{n}$.
Unter H_0 ist χ^2 approximativ $\chi^2((k-1)(m-1))$-verteilt (Fahrmeir et al., 2016).

2.2 Fishers exakter Test

Der *Exakte Test nach Fisher* ist ein hauptsächlich für 2×2 Kontingenztafeln entwickelter Test auf Unabhängigkeit, der allerdings auch auf $r \times c$ Kontingenztabellen erweitert werden kann (siehe unten). Im Gegensatz zum χ^2-Unabhängigkeitstest (siehe 2.1) lässt sich dieser Test allerdings auch für geringe Stichprobengrößen anwenden. Die Verteilungen der zwei betrachteten Stichproben eines binären Merkmals sind dann binomialverteilt und es gilt:

$$X_1, X_2, \ldots, X_{n_1} \text{identisch verteilt wie X}, X \sim B(1; p_1)$$
$$Y_1, Y_2, \ldots, Y_{n_2} \text{identisch verteilt wie Y}, Y \sim B(1; p_2).$$

Auch die Summen der Zufallsvariablen sind damit jeweils binomialverteilt mit:

$$X = \sum_{i=i}^{n_1} X_i \sim B(n_1; p_1)$$
$$Y = \sum_{i=i}^{n_2} Y_i \sim B(n_2; p_2)$$

(Fahrmeir et al., 2016).

Die Hypothesen, welche zum Test auf Unabhängigkeit zwischen X und Y verwendet werden, sind:

$$\mathbf{H_0}: \quad p_1 = p_2 \quad \textbf{vs.} \quad \mathbf{H_1}: \quad p_1 \neq p_2 \tag{2}$$

Mit $Z = X + Y$ folgt X, unter der Nullhypothese, einer hypergeometrischen Verteilung mit

$$X \sim H(z, n_1, n) \tag{3}$$

mit $n = n_1 + n_2$ (Fahrmeir et al., 2016).

Eine Erweiterung für $r \times c$ Kontingenztabellen ist ebenfalls möglich. Für h_{ij} als Anzahl der Beobachtungen, welche die Ausprägungen $(X = i, Y = j)$ besitzen, ist $R_i = \sum_{j=1}^{c} h_{ij}$ die Summe aller Reiheneinträge und $C_j = \sum_{i=1}^{r} h_{ij}$ die Summe aller Spalteneinträge einer $r \times c$ Kontingenztabelle H. Dann ist τ die Referenzmenge aller möglichen $r \times c$ Kontingenztabellen mit den selben marginalen Zahlen wie H.

$$\tau = \left\{ Y : Y \text{ is } r \times c, \sum_{j=1}^{c} y_{ij} = R_i, \sum_{i=1}^{r} y_{ij} = C_j, \right\}. \tag{4}$$

Unter der Nullhypothese, dass X und Y unabhängig voneinander sind, kann dann die Wahrscheinlichkeit von Y durch ein Produkt von Multinomialkoeffizienten ausgedrückt werden

$$P(Y) = \left(\prod_{j=1}^{c} \frac{C_j!}{y_{1j}! y_{2j}! \cdots y_{rj}!} \right) \Big/ \frac{T!}{R_1! R_2! \cdots R_r!} \tag{5}$$

mit $T = \sum_{i=1}^{r} R_i$ (Mehta und Patel, 1983).

Der p-Wert der Tabelle X ist dann die Summe der Wahrscheinlichkeiten aller Tabellen in τ, die mit einer höheren Wahrscheinlichkeit als X auftreten. Es gilt:

$$p = \sum_{Y \in \Lambda} P(Y), \tag{6}$$

mit $\Lambda = \{Y : Y \in \tau \text{ and } P(Y) \leq P(X)\}$ (Mehta und Patel, 1983).

2.3 Wilcoxon Rangsummentest

Der Wilcoxon Rangsummentest ist ein nichtparametrischer statistischer Test zwischen zwei unabhängigen Stichproben X und Y. Durch einen Vergleich der Mediane beider Stichproben wird getestet, ob X und Y dieselbe Verteilungsfunktion besitzen. Voraussetzung dafür ist, dass die Verteilungsfunktionen von X und Y dieselbe Form besitzen, welche lediglich um einen konstanten Betrag verschoben sein kann. Darüber hinaus muss es sich um unabhängige, ungepaarte Stichproben $X = X_1, \ldots, X_n$ und $Y = Y_1, \ldots, Y_m$ handeln (Fahrmeir et al., 2016).
Zur Berechnung der Teststatistik werden alle Beobachtungen der gepoolten Stichprobe der Größe nach sortiert. Entsprechend dieser Reihenfolge wird anschließend jeder Beobachtung ein Rang zugewiesen. Bei gleichen Werten (Bindungen) wird ein gemittelter Rangwert verwendet (Wilcoxon, 1992).
Die Überprüfung der Hypothesen

$$\mathbf{H_0}: \quad x_{med} = y_{med} \quad \textbf{vs.} \quad \mathbf{H_1}: \quad x_{med} \neq y_{med}$$

kann letztlich beantwortet werden durch die Teststatistik

$$T_W = \sum_{i=1}^{n} \text{rg}(X_i) = \sum_{i=1}^{n+m} iV_i \tag{7}$$

mit

$$V_i = \begin{cases} 1 & i\text{-te Beobachtung der geordneten gepoolten Stichprobe ist X-Variable} \\ 0 & \text{sonst} \end{cases}$$

(Fahrmeir et al., 2016).

Die Grenzwerte nach Wilcoxon (1992) für die Teststatistik T_W können in tabellierten Verteilungen abgelesen werden. Hierfür wird neben dem Signifikanzniveau α auch die Stichprobengröße von X sowie Y benötigt.

Für große Stichprobengrößen mit m oder $n > 25$ gilt, unter der Nullhypothese, approximativ

$$T_W \sim N(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{12})$$

(Fahrmeir et al., 2016).

Aufgrund der Berechnung der Teststatistik durch die Ränge der Beobachtungen kann der Wilcoxon-Test, im Gegensatz z.B. zum χ^2-Test (siehe Kapitel 2.1), eine ordinale oder stetige Struktur berücksichtigen.

3 Maximal selektierte χ^2-Statistiken

Der Ansatz für maximal selektierte χ^2-Statistiken basiert auf der Theorie des χ^2-Tests. Auch hier sollen zwei Stichprobenvariablen X und Y auf ihre Unabhängigkeit hin untersucht werden. Dafür wird ein optimaler Schwellenwert gesucht, welcher die Beobachtungen über und unter dem Schwellenwert für jede Stichprobe einteilt und den anschließend daraus zu berechnenden χ^2-Test maximiert. Für einen beliebigen Schwellenwert x^* erhält man folgende 2×2-Kontingenztabelle:

Group	$X \leq x^*$	$X > x^*$	Total
0	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+d$	N

Mit Hilfe dieser Häufigkeiten lässt sich die χ^2-Statistik berechnen durch

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}. \tag{8}$$

Der maximale Wert aller möglichen χ^2-Tests wird als maximal selektierte χ^2-Statistik bezeichnet (Koziol, 1991).

Zur Bestimmung der Verteilung der maximal selektierten χ^2-Statistik sind die Standardquantile der χ^2-Verteilung allerdings ungeeignet (Miller und Siegmund, 1982).
Stattdessen gibt es besonders für stetige Zufallsvariable einige Ansätze zur Berechnung der Verteilung. So kann für große Stichprobengrößen eine asymptotische, korrigierte χ^2-Verteilung nach Miller und Siegmund verwendet werden. Jerry Halpern dagegen simulierte und analysierte die Verteilung unter der Nullhypothese bei kleinen Stichproben (Halpern, 1982). Der Ansatz von Koziol (1991) beinhaltet die Herleitung der exakten Verteilung von maximal selektierten χ^2-Statistiken mit dem kombinatorischem Ansatz von Durbin (1971).

Bei mindestens ordinal verteilten Variablen kann die Variable X, ohne Beschränkung der Allgemeinheit, in K distinkte Level $a_1 < \cdots < a_K \in \mathbb{R}$ eingeteilt werden. Es gilt dabei $2 \leq K \leq N$ und $a_1 < \cdots < a_K$.
Für mindestens ordinal verteilte Variablen wurde von Boulesteix (2006) eine endliche Stichprobenverteilung der χ^2-Statistik erschlossen, welche eng an die Methode Koziols angelehnt ist. Dieser Ansatz ist auch zur Messung von Assoziationen zwischen einer binären Variable Y und einer nicht-stetigen Variable X mit gleichen Realisationen in der Stichprobe $(x_i, y_i)_{i=1,\ldots,N} (K < N)$ geeignet. Die folgende Aufführung der Methode nach Boulesteix basiert auf der Veröffentlichung im *Biometrical Journal: Journal of Mathematical Methods in Biosciences* aus dem Jahre 2006.

Dabei sei Y eine binäre Variable mit den Ausprägungen $Y \in \{0, 1\}$ und X eine ordinale Variable mit den Ausprägungen $a_1 < \cdots < a_K$. Für einen Schwellenwert a_k kann dann die folgende 2×2 Kontingenztabelle aufgestellt werden:

	$X \leq a_k$	$X > a_k$	$\sum$
$Y = 0$	$n_{1,\leq a_k}$	$n_{1,>a_k}$	N_1
$Y = 1$	$n_{2,\leq a_k}$	$n_{2,>a_k}$	N_2
$\sum$	$n_{.,\leq a_k} = \sum_{j=1}^{k} m_j$	$n_{.,>a_k} = \sum_{j=k+1}^{K} m_j$	N

mit
$m_k = \sum_{i=1}^{n} I(x_i = a_k)$, for $k = 1, \ldots, K$
$N_1 = \sum_{i=1}^{N} I(y_i = 0)$ bzw. $N_2 = \sum_{i=1}^{N} I(y_i = 1)$.

Die dazugehörige χ^2-Teststatistik ist dann

$$\chi^2_k = \frac{N(n_{1,\leq a_k} n_{2,>a_k} - n_{1,>a_k} n_{2,\leq a_k})^2}{N_1 N_2 n_{.,\leq a_k} n_{.,>a_k}}, \tag{9}$$

woraus sich dementsprechend die maximal selektierte χ^2-Statistik ergibt mit

$$\chi^2_{max} = \max_{k=1,\ldots,K-1} \chi^2_k. \tag{10}$$

Um die Nullhypothese (keine Assoziation zwischen X und Y) überprüfen zu können, muss die Verteilungsfunktion

$$F(d) = P_{H_0}(\chi^2_{max} \leq d)$$

berechnet werden.
Nach Miller und Siegmund (1982) kann die χ^2-Statistik der binären Variable $X^{(k)}$ formuliert werden als $\chi^2_k = A^2_k$ und

$$A_k = \frac{\frac{N}{N_1}\left(\frac{n_{2,\leq a_k}}{N_2} - \frac{n_{.,\leq a_k}}{N}\right)}{\sqrt{\frac{n_{.,\leq a_k}}{N}\left(1 - \frac{n_{.,\leq a_k}}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \tag{11}$$

für alle $k = 1, \ldots, K-1$.

Sei d beliebig mit $d \in \mathbb{R}^+$. Aus Gleichung 11 folgt, dass $\chi^2_{max} \leq d$ nur dann erfüllt ist, wenn alle Punkte mit den Koordinaten $(n_{.,\leq a_k}, n_{2,\leq a_k})$ für $k = 1, \ldots, K-1$ auf oder über der Funktion

$$\text{lower}_d(x) = \frac{N_2 x}{N} - \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{12}$$

bzw. auf oder unter der Funktion

$$\text{upper}_d(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 \sqrt{d}}{N}\sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \tag{13}$$

liegen.

Sei $N_2(i)$ die Anzahl an Realisationen mit $Y = 1$ und $X \leq x_{(i)}$, wobei $x_{(1)} \leq \cdots \leq x_{(N)}$ die geordneten Realisationen von X sind. Die Funktionen $\text{lower}_d(x)$ und $\text{upper}_d(x)$ können dann auf einem Graph $(i, N_2(i))$ dargestellt werden. Eine Voraussetzung für die Einhaltung der Gleichung $\chi^2_{max} \leq d$ ist, dass der Graph $(i, N_2(i))$ nicht durch einen der Punkte der Koordinaten (i, j) geht, mit

$$i = n_{\cdot,\leq a_k}$$

und

$$\text{upper}_d(i) < j \leq \min(N_2, i) \ \textit{oder} \ \max(0, i - N_1) \leq j < \text{lower}_d(i),$$

wobei $k = 1, \ldots, K - 1$.
Die Koordinaten $(i_1, j_1), \ldots, (i_q, j_q)$, im weiteren als Punkte $B_1, \ldots, B_q$ bezeichnet, sind geordnet nach aufsteigendem i und innerhalb darin nach aufsteigendem j. Nach dem Ansatz von Koziol (1991) ist $\mathcal{P}_s$ die Menge an Pfaden von $(0, 0)$ zu einem Punkt B_s, ohne durch einen Punkt $B_1, \ldots, B_{s-1}$ zu gehen. Die Anzahl an Pfaden in $\mathcal{P}_s$ wird mit b_s bezeichnet. Diese kann rekursiv berechnet werden mit

$$\begin{aligned} b_1 &= \binom{i_1}{j_1}, \\ b_s &= \binom{i_s}{j_s} - \sum_{r=1}^{s-1} \binom{i_s - i_r}{j_s - j_r} b_r, \qquad s = 2, \ldots, q. \end{aligned}$$

Die Anzahl an Pfaden von $(0, 0)$ zu (N, N_2), welche durch B_s, $s = 1, \ldots, q$, aber nicht durch $B_1, \ldots, B_{s-1}$ gehen, ist dann gegeben durch

$$\binom{N - i_s}{N_2 - j_s} b_s. \tag{14}$$

Dadurch folgt

$$P_{H_0}(\chi^2_{max} > d) = \binom{N}{N_2}^{-1} \sum_{s=1}^{q} \binom{N - i_s}{N_2 - j_s} b_s, \tag{15}$$

was der Wahrscheinlichkeit entspricht, dass der Pfad $(i, N_2(i))$ durch mindestens einen Punkt $B_1, \ldots, B_q$ geht. Daraus folgt die Verteilungsfunktion F unter der Nullhypothese mit

$$F(d) = 1 - \binom{N}{N_2}^{-1} \sum_{s=1}^{q} \binom{N - i_s}{N_2 - j_s} b_s. \tag{16}$$

Diese Verteilungsfunktion ist auch für den Fall $K < N$ geeignet. Ist die Anzahl der verschiedenen Ausprägungen K gleich der Stichprobengröße N, so ist die Verteilungsfunktion identisch mit der Verteilungsfunktion nach Koziol.

4 R-Paket *exactmaxsel2*

Um eine leicht zugängliche und anwendungsfreundliche Umsetzung der maximal selektierten χ^2-Statistiken (siehe Kapitel 3) zu ermöglichen, wurde im Zuge dessen das R-Paket *exactmaxsel2* entwickelt. Die in diesem Paket verwendeten Funktionen basieren zu einem großen Teil auf einem bereits archivierten R-Paket mit dem Namen *exactmaxsel*. Dieses Paket wurde im Rahmen dieser Masterarbeit durch die Anwendung von *S4*-Klassen und Funktionen, das Hinzufügen von Grafiken und durch die Möglichkeit mit höheren Zahlen zu rechnen, modifiziert. Bei *S4* handelt es sich, ebenso wie bei *S3*, um Systeme in R, welche für die objektorientierte Programmierung verwendet werden. *S4* ist dabei im Gegensatz zu *S3* deutlich strikter bezüglich der Einhaltung gewisser Strukturen. So werden beispielsweise in *S4* die Eigenschaften und Vererbungen jeglicher Klassen formal definiert (Wickham, 2019). Zur Erstellung des R-Paketes wurde die statistische Software R - Version 4.0.3 verwendet.

Klassen und Funktionen

Durch die Entwicklung von Klassen und Funktionen können Teststatistiken berechnet und die Ergebnisse grafisch dargestellt werden. Im Folgenden werden die verschiedenen Klassen und Funktionen des Paketes genauer erklärt. Funktionen, welche lediglich innerhalb des Paketes zum Tragen kommen und nicht von einem potenziellen Anwender direkt verwendet werden können/sollen, sind hier nicht aufgelistet.
Im weiteren Verlauf sei o.B.d.A. X eine ordinal skalierte Zufallsvariable mit K verschiedenen Ausprägungen. Y sei eine binär skalierte Zufallsvariable mit $Y \in \{0, 1\}$.

4.1 Maxsel-Klasse

Mit Hilfe der `Maxsel`-Klasse können Objekte beschrieben werden, welche bestimmte Eigenschaften vorweisen. Objekte der `Maxsel`-Klasse können dabei mit Hilfe der Funktion `maxsel.test` (siehe Abschnitt 4.4) erstellt werden.
Ein `Maxsel`-Objekt besitzt fünf Merkmale:

- `matrix`
- `statistic`
- `maxsel_value`
- `maxsel_p_value`
- `all_splits`

Das Merkmal `matrix` stellt dabei eine $2 \times K$-Kontingenztabelle dar. Dabei entspricht die erste Zeile der Matrix $Y = 0$ und die zweite Zeile $Y = 1$. Die Spalten stehen für K verschiedene Ausprägungen der Zufallsvariable X.
Mit `statistic` ist die verwendete Methode gemeint, die zur Messung der Assoziation zwischen X und Y verwendet wurde. Diese kann entweder die χ^2-Statistik (`statistic = chi2`) oder der *Gini gain* (`statistic = gini`) sein.
`maxsel_value` ist der maximale Wert des gewählten Assoziationsmaßes (`statistic`), welcher den optimalen Schwellenwert in der Zufallsvariable X widerspiegelt.

Da zur Auswahl des maximalen Wertes mehrere Tests durchgeführt werden, wäre ein P-Wert, der mit Hilfe der herkömmlichen χ^2-Verteilung berechnet wird, verzerrt. Aufgrund dessen wurde für das Merkmal `maxsel_p_value` die exakte Verteilung der maximal selektierten χ^2-Statistik zur Berechnung des korrigierten P-Wertes verwendet, welcher dem maximalen Wert des gewählten Assoziationsmaßes (`maxsel_value`) entspricht.
Bei dem Merkmal `all_splits` handelt es sich um eine Tabelle (*data.frame*). Diese Tabelle enthält alle möglichen Einteilungen der Kategorien in zwei Gruppen. Gilt beispielsweise $K = 4$, so enthielte die Tabelle `all_splits` insgesamt drei Zeilen. Die unterschiedlichen Einteilungen wären dann {1} vs. {2,3,4}, {1,2} vs. {3,4} und {1,2,3} vs. {4}. Zu jeder möglichen Einteilung wird zudem der dazugehörige Wert des ausgewählten Assoziationsmaßes (z.B. χ^2) und der daraus folgende unkorrigierte P-Wert angegeben. Des Weiteren ist es möglich, dass die Tabelle eine weitere Spalten mit den dazugehörigen korrigierten P-Werten enthält. Diese können mit Hilfe eines optionalen Parameters in der Funktion `maxsel.test` (siehe Abschnitt 4.4) ebenfalls berechnet werden. Dafür wird die Prozedur der *single step maxT* korrigierten P-Werte verwendet. Es ergeben sich die P-Werte aus:

$$\tilde{p}_j = Pr\left(\max_{1 \leq l \leq m} |T_l| \geq |t_j| \middle| H_0^C \right) \tag{17}$$

mit T_j als der j-ten Teststatistik und H_0^C als gesamte Nullhypothese. Mit dieser Korrektur der P-Werte wird, unter der Annahme, dass alle Nullhypothesen zutreffen, eine schwache Kontrolle der *Family-wise Error Rate* erreicht (Dudoit et al., 2003). Im Paket wird, falls alle korrigierten P-Werte erwünscht sind, der Funktion `Ford` ein Vektor mit allen gemessenen Werten der Teststatistiken überliefert (siehe Abschnitt 4.2).
Die Ausgabe eines `Maxsel`-Objektes liefert eine Zusammenfassung der wichtigsten Eigenschaften. Im folgenden `R`-Code wurde anhand eines Beispieles ein `Maxsel`-Objekt berechnet und der Output angezeigt. Dieser zeigt die Dimension der eingegebenen Matrix und die verwendete Statistik an. Zusätzlich werden der maximale Wert dieser Statistik über alle möglichen Einteilungen und der dazugehörige korrigierte P-Wert angezeigt. Zuletzt wird angegeben, ob auch für die anderen Einteilungen ein korrigierter P-Wert vorliegt.

```
library(exactmaxsel2)
x <- matrix(c(8, 10, 40, 13, 15, 4), 2, 3, byrow = TRUE)
maxsel_object <- maxsel.test(x = x, y = NULL, statistic = "chi2")
maxsel_object

## Maxsel-Object with the following properties:
##
## Matrix: 2x3 matrix
## Statistic: chi2
## Value of the maximally selected criterion:  26.31338
## Corrected p-Value of the maximally selected criterion:  1.288769e-07
## Informations about other splits are given:  FALSE
```

4.2 Ford

Die Funktion `Ford` ist möglicherweise als das Herzstück des Paketes zu beschreiben. In dieser Funktion wird die exakte Verteilung des maximalen Wertes vom gewählten Assoziationskriterium (`statistic`) berechnet. Hiermit ist es möglich, die Wahrscheinlichkeit zu berechnen, dass der maximale, beobachtete Wert des Assoziationskriterium mindestens so groß wie der beobachtete Wert ist.

Die Funktion `Ford` kann bis zu sechs verschiedene Parameterangaben verarbeiten. Diese sind:

- `c`
- n_0
- n_1
- `A`
- `statistic`
- `progress_bar`

Der Parameter `c` ist der Wert, an welchem die Verteilungsfunktion berechnet werden soll. Dieser Wert `c` entspricht somit dem maximalen beobachteten Wert des gewählten Assoziationskriteriums. Besitzt X beispielsweise vier verschiedene Kategorien, so gibt es genau drei unterschiedliche Einteilungen in eine 2×2-Kontingenztabelle. Für alle drei Einteilungen werden dann die Werte des gewählten Assoziationskriterium (χ^2 oder *Gini gain*) berechnet. Der maximale Wert dieser drei Berechnungen ist somit der gesuchte Wert `c`. Es ist zudem möglich, dass `c` aus mehreren numerischen Werten besteht. Dies kann beispielsweise angewendet werden, falls für alle möglichen Einteilungen auch der korrigierte P-Wert berechnet werden soll. Dabei ist es allerdings wichtig, dass die Werte von den selben Daten stammen, da die restlichen Parameter konstant bleiben. Dabei wird die *single step maxT* Prozedur angewandt (siehe Gleichung 17), um eine Korrektur der P-Werte aufgrund des multiplen Testens zu erwirken.
Mit n_0 wird die Anzahl an Beobachtungen mit $Y = 0$ bezeichnet. Hierbei spielt es keine Rolle, in welcher Kategorie die Beobachtung zu finden ist. Analog dazu bezeichnet n_1 die Anzahl an Beobachtungen mit $Y = 1$. Auch hier ist die Kategorie der Beobachtung irrelevant.
Der Parameter `A` besteht aus einem Vektor der Länge K. Dabei steht jedes Element des Vektors für die Anzahl an Beobachtungen, die mit $X = 1, \ldots, X = K$ in die entsprechende Kategorie eingeordnet werden kann. Handelt es sich um eine stetige Variable X, so ist `A` ein Vektor der Länge $N = n_0 + n_1$ mit allen Einträgen gleich 1.
Mit `statistic` kann die gewünschte Methode angegeben werden, welche zur Messung der Assoziation zwischen X und Y verwendet werden soll. Diese kann entweder die χ^2-Statistik (`statistic = chi2`) oder der *Gini gain* (`statistic = gini`) sein.
Bei dem letzten Parameter `progress_bar` handelt es sich um einen optionalen booleschen Parameter, welcher standardmäßig auf `progress_bar = FALSE` gesetzt ist. Besonders für zeitintensive Berechnungen mit einer großen Anzahl an Beobachtungen und/oder Kategorien sowie bei mehr als einem Wert für `c` kann es von Nutzen sein `progress_bar =`

`TRUE` zu verwenden. Dadurch wird dem Anwender während der Berechnung ein grober Überblick über die restliche Berechnungszeit gewährt.

Da die Schätzung der exakten Verteilung auf der Berechnung von Binomialkoeffizienten beruht, muss, besonders für große Stichprobengrößen, ein Umgang mit großen Zahlen gewährleistet sein. Dies wird erreicht, indem logarithmierte Binomialkoeffizienten verwendet und dadurch der Berechnungsprozess mit kleineren Zahlen durchgeführt werden kann. Die Theorie hinter der Funktion ist in Kapitel 3 ausführlich beschrieben.

Beispiel
Sei folgende Kontingenztabelle gegeben:

	Kategorie 1	Kategorie 2	Kategorie 3	Kategorie 4	$\sum$
$Y = 0$	6	12	9	15	42
$Y = 1$	2	7	11	9	29
$\sum$	8	19	20	24	71

Tabelle 1: Beispiel für eine Kontingenztabelle, welche mit Hilfe der maximal selektierten χ^2-Statistik analysiert werden soll.

Um die Verteilung an einem beliebigen Punkt `c` zu berechnen, benötigt die Funktion `Ford` die oben genannten Parameter. Diese lauten, gemäß Tabelle 1,

```
n0 <- 42
n1 <- 29
A <- c(8,19,20,24)
```

Für das Assoziationskriterium `statistic = "chi2"` und $c = 2$ ergibt sich somit:

```
statistic <- "chi2"
c <- 2
Ford(c=c,n0=n0,n1=n1,A=A,statistic=statistic)

## [1] 0.591215
```

Der Output der Funktion spiegelt somit die Verteilung der maximal selektierten χ^2-Statistik an der Stelle $c = 2$ wider. In Abbildung 1 ist eben jene Verteilung dargestellt. Zusätzlich zu $c = 2$ (rote gestrichelte Linie) ist hier ebenfalls $c = 6$ (blaue gestrichelte Linie) hervorgehoben, was einem Verteilungswert von $F(6) \approx 0.95$ entspricht.
Die Wahrscheinlichkeit, einen größeren oder gleich großen Maximalwert (wie $c = 2$) des Assoziationskriterium (*hier*: statistic = "chi2") innerhalb dieser Kontingenztabelle zu erlangen, ist:

```
p_value <- 1-Ford(c=c,n0=n0,n1=n1,A=A,statistic=statistic)
p_value

## [1] 0.408785
```

Somit entspricht $1 - F(c = 6) \approx 0.05$ der Grenze, ab der, bei einem Signifikanzniveau von $\alpha = 0.05$, ein möglicher Hypothesentest bezüglich der Unabhängigkeit von X und Y abgelehnt wird.

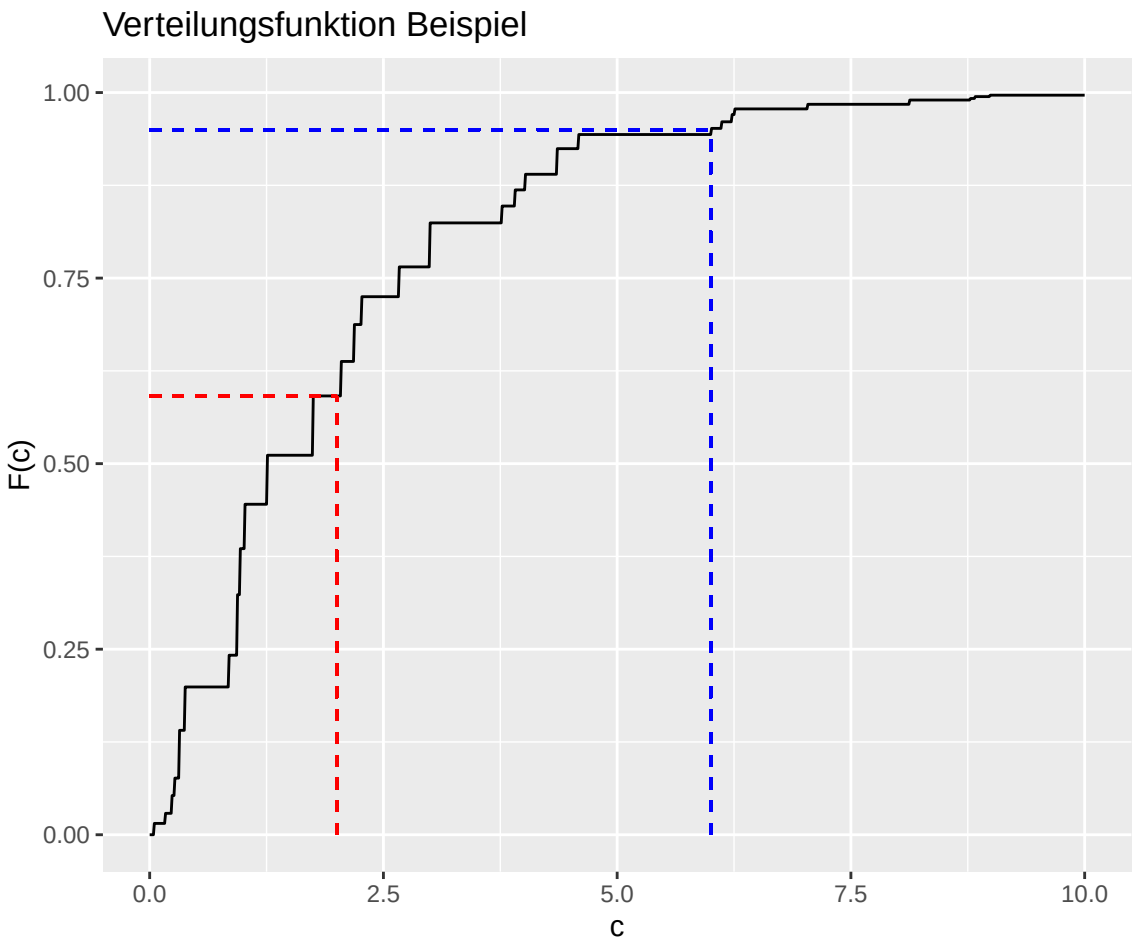


Abbildung 1: Beispiel der Verteilungsfunktion nach der Funktion Ford. Die rote gestrichelte Linie entspricht dem zufällig gewählten Wert $c = 2$, an der die Verteilungsfunktion berechnet wurde (siehe oben). Die blaue gestrichelte Linie entspricht einer Verteilung von $F(6) = 0.95$, welche oft als Grenze bezüglich der Ablehnung eines Hypothesentests verwendet wird.

4.3 maxsel.table

Mit Hilfe der Funktion `maxsel.table` ist es dem Anwender möglich, den maximalen Wert des gewählten Assoziationskriteriums zu berechnen. Die Funktion betrachtet dabei alle möglichen Aufteilungen, welche durch die gegebenen Daten in Betracht kommen.
Insgesamt können bis zu drei verschiedene Parameter angegeben werden. Diese sind:

- `x`
- `y`
- `statistic`

Der Parameter `x` gibt die Werte der Zufallsvariable X wieder. X wird dabei als numerischer Vektor der Länge n und kodiert mit $1, \ldots, K$ angegeben. Falls erwünscht, kann für `x` auch eine Matrix übermittelt werden. Diese Matrix muss dabei einer $2 \times K$ Kontingenztabelle entsprechen. Die zwei Reihen stehen dabei für $Y = 0$ bzw. $Y = 1$, während die K Spalten den Werten der Zufallsvariable X mit $X = 1, \ldots, K$ entsprechen. Wird `x` in Form einer Matrix angegeben, so darf für `x` keine weitere Angabe gemacht werden (`y = NULL`).
Der Parameter `y` gibt dagegen die Werte der Zufallsvariable Y wieder. Y wird dabei als binärer Vektor der Länge n dargestellt. Die Ausprägungen des Vektors repräsentieren die Klassen $Y \in \{0, 1\}$. Ist für den Parameter `x` eine Matrix bzw. Kontingenztabelle angegeben, so darf für `y` keine zusätzliche Angabe getätigt werden.
Mit `statistic` kann die gewünschte Methode angegeben werden, die zur Messung der Assoziation zwischen X und Y verwendet werden soll. Diese kann entweder die χ^2-Statistik (`statistic = chi2`) oder der *Gini gain* (`statistic = gini`) sein.

Beispiel
Wie bereits im Beispiel zur Funktion `Ford` sollen auch hier die Daten aus Tabelle 1 untersucht werden. Da diese bereits in Form einer Kontingenztabelle vorliegen, kann `x` ohne große Transformation als Matrix angegeben werden. Mit folgendem `R`-Code lässt sich die Kontingenztabelle in Form einer Matrix darstellen.

```
data <- matrix(c(6,12,9,15,2,7,11,9), byrow = TRUE, nrow = 2, ncol = 4)
data

##      [,1] [,2] [,3] [,4]
## [1,]    6   12    9   15
## [2,]    2    7   11    9
```

Sei unser gewähltes Assoziationskriterium zwischen den Zufallsvariablen X und Y die χ^2-Statistik. Um den maximalen Wert der χ^2-Statistik zu berechnen, können wir nun die Funktion `maxsel.table` verwenden. Diese berechnet für alle möglichen Einteilungen den dazugehörigen Wert der χ^2-Statistik und den entsprechenden P-Wert. Da es sich dabei allerdings um multiples Testen handelt, sind die vorliegenden P-Werte verzerrt. Im Output der Funktion werden die P-Werte dadurch als unkorrigierte P-Werte (*p_value_uncorrected* bezeichnet.
Aufgrund von $K = 4$ gilt für unsere Daten, dass insgesamt drei verschiedene Einteilungen in zwei Gruppen möglich sind. Wichtig dabei ist, dass die ordinale Struktur der Kategorien

berücksichtigt wird und es somit nur einen Schnittpunkt gibt. Die möglichen Einteilungen setzen sich dann zusammen aus {1} vs. {2,3,4}, {1,2} vs. {3,4} und {1,2,3} vs. {4}. Die Anwendung der Funktion ergibt damit folgenden Output:

```
statistic <- "chi2"
maxsel.table(x = data, statistic = statistic)

##      Split1    Split2 statistic p_value_uncorrected
## 1         1 2 + 3 + 4 0.9368403           0.3330918
## 2     1 + 2     3 + 4 1.0174653           0.3131210
## 3 1 + 2 + 3         4 0.1679004           0.6819846
```

Der Output wird dabei in Form eines *data.frame* mit vier Spalten dargestellt. Mit *Split1* bzw. *Split2* wird die untersuchte Einteilung beschrieben. Dabei entsprechen die Zahlen der jeweiligen Spalte in der Kontingenztabelle. Zum Beispiel wird für die zweite Reihe mit *Split1 = 1+2* und *Split2 = 3+4* eine "neue" 2×2-Kontingenztabelle bestimmt, die sich aus den Summen der beiden *Splits* ergibt. In unserem Beispiel ergibt sich:

```
data <- matrix(c(6,12,9,15,2,7,11,9), byrow = TRUE, nrow = 2, ncol = 4)
data

##      [,1] [,2] [,3] [,4]
## [1,]    6   12    9   15
## [2,]    2    7   11    9

data2 <- matrix(c(6+12, 9+15, 2+7, 11+9), byrow = TRUE, nrow = 2, ncol = 2)
data2

##      [,1] [,2]
## [1,]   18   24
## [2,]    9   20
```

Der für diese 2×2-Kontingenztabelle entsprechende Wert der χ^2-Statistik wird in der dritten Spalte (*statistic*) angezeigt. Der dazugehörige P-Wert (unkorrigiert, siehe oben) befindet sich in Spalte vier (*p_value_uncorrected*).

4.4 maxsel.test

Bei `maxsel.test` handelt es sich um die Funktion, welche vermutlich am häufigsten vom Anwender verwendet werden wird. Diese Funktion kombiniert dabei die beiden bereits vorgestellten Funktionen `Ford` und `maxsel.table`. Im Zusammenspiel zwischen diesen Funktionen ist es dadurch möglich, den maximalen Wert des gewünschten Assoziationskriteriums zu finden und gleichzeitig auch den dazugehörigen korrigierten P-Wert zu berechnen.

Die Parameter der `maxsel.test`-Funktion sind:

- `x`
- `y`
- `statistic`
- `all_splits`
- `progress_bar`

Die meisten Parameter der `maxsel.test`-Funktion wurden bereits in gleicher bzw. ähnlicher Funktion zuvor verwendet und erklärt. Der Vollständigkeit halber werden diese Parameter allerdings an dieser Stelle nochmals erwähnt und erklärt.
Der Parameter `x` gibt die Werte der Zufallsvariable X wieder. X wird dabei als numerischer Vektor der Länge n und kodiert mit $1, \ldots, K$ angegeben. Falls erwünscht, kann für `x` auch eine Matrix übermittelt werden. Diese Matrix muss dabei einer $2 \times K$ Kontingenztabelle entsprechen. Die zwei Reihen stehen dabei für $Y = 0$ bzw. $Y = 1$, während die K Spalten den Werten der Zufallsvariable X mit $X = 1, \ldots, K$ entsprechen. Wird `x` in Form einer Matrix angegeben, so darf für `y` keine weitere Angabe gemacht werden (`y = NULL`).
Der Parameter `y` gibt die Werte der Zufallsvariable Y wieder. Y wird dabei als binärer Vektor der Länge n dargestellt. Die Ausprägungen des Vektors repräsentieren die Klassen $Y \in c(0, 1)$. Ist für den Parameter `x` eine Matrix bzw. Kontingenztabelle angegeben, so darf für `y` keine zusätzliche Angabe getätigt werden.
Mit `statistic` kann die gewünschte Methode angegeben werden, die zur Messung der Assoziation zwischen X und Y verwendet werden soll. Diese kann entweder die χ^2-Statistik (`statistic = chi2`) oder der *Gini gain* (`statistic = gini`) sein.
Der optionale Parameter `all_splits` kann, falls gewünscht (`all_splits = TRUE`), dafür sorgen, dass der korrigierte P-Wert für alle möglichen Einteilungen berechnet wird. Die Standardeinstellung von `all_splits = FALSE` berechnet dagegen lediglich den korrigierten P-Wert des maximalen Wertes des Assoziationskriteriums.
Bei dem letzten Parameter `progress_bar` handelt es sich um einen weiteren optionalen booleschen Parameter, welcher standardmäßig auf `progress_bar = FALSE` gesetzt ist. Besonders für zeitintensive Berechnungen mit einer großen Anzahl an Beobachtungen und/oder Kategorien sowie bei der Parametereinstellung `all_splits = TRUE` kann es von Nutzen sein `progress_bar = TRUE` zu verwenden. Dadurch wird dem Anwender während der Berechnung ein grober Überblick über die restliche Berechnungszeit gewährt.

Wie bereits in der Übersicht zur `Maxsel`-Klasse (siehe Abschnitt 4.1), wird auch in folgendem Beispiel die Funktion `maxsel.test` verwendet. Der Output der Funktion entspricht dabei der `Maxsel`-Klasse. Bei Verwendung der Standardeinstellung von `all_splits = FALSE` wird lediglich der korrigierte P-Wert für den maximalen Wert des Assoziationskriterium berechnet.

```
data <- matrix(c(6,12,9,15,2,7,11,9), byrow = TRUE, nrow = 2, ncol = 4)
maxsel_object <- maxsel.test(x = data, y = NULL, statistic = "chi2")
maxsel_object

## Maxsel-Object with the following properties:
##
## Matrix: 2x4 matrix
## Statistic: chi2
## Value of the maximally selected criterion:  1.017465
## Corrected p-Value of the maximally selected criterion:  0.5547055
## Informations about other splits are given:  FALSE
```

Mit Hilfe des Parameters `all_splits` lassen sich allerdings auch für die nicht maximalen Werte des Assoziationskriteriums einen korrigierten P-Wert berechnen.

```
data <- matrix(c(6,12,9,15,2,7,11,9), byrow = TRUE, nrow = 2, ncol = 4)
maxsel_object <- maxsel.test(x = data, y = NULL, statistic = "chi2",
                             all_splits = TRUE)
maxsel_object

## Maxsel-Object with the following properties:
##
## Matrix: 2x4 matrix
## Statistic: chi2
## Value of the maximally selected criterion:  1.017465
## Corrected p-Value of the maximally selected criterion:  0.5547055
## Informations about other splits are given:  TRUE
```

Die berechneten, korrigierten P-Werte können dann im `Maxsel`-Objekt enthaltenen *data.frame* "`all_splits`" aufgerufen werden.

```
maxsel_object@all_splits

##      Split1    Split2 statistic p_value_uncorrected p_value_corrected
## 2     1 + 2     3 + 4 1.0174653           0.3131210         0.5547055
## 1         1 2 + 3 + 4 0.9368403           0.3330918         0.6765613
## 3 1 + 2 + 3         4 0.1679004           0.6819846         0.9711485
```

Die verschiedenen Einteilungen in *Split1* und*Split2* sind dabei nach absteigender Größe der χ^2-Statistik sortiert.

Für `all_splits = FALSE` würde in diesem Fall die letzte Spalte (`p_value_corrected`) fehlen. Der Zugriff auf den maximalen Wert des Assoziationskriteriums und den dazugehörigen P-Wert kann allerdings auch direkt mit Hilfe des @-Symbols erfolgen. Dies gilt auch für die verwendete Matrix und das gewählte Assoziationskriterium.

```
maxsel_object@matrix

##      [,1] [,2] [,3] [,4]
## [1,]    6   12    9   15
## [2,]    2    7   11    9

maxsel_object@statistic

## [1] "chi2"

maxsel_object@maxsel_value

## [1] 1.017465

maxsel_object@maxsel_p_value

## [1] 0.5547055
```

4.5 maxsel.plot

Die Funktion `maxsel.plot` ist zur grafischen Darstellung eines `Maxsel`-Objektes gedacht. Mit verschiedenen Parametereinstellungen sind dadurch mehrere deskriptive Analysen des Objektes möglich. Die Funktion produziert dabei gruppierte Säulendiagramme, welche die Daten des `Maxsel`-Objektes wiedergeben. Zusätzlich werden die entsprechenden P-Werte angegeben. Diese liegen, je nach Informationslage, entweder in korrigierter oder in unkorrigierter Weise vor und sind dementsprechend auch als solche gekennzeichnet. Inklusive des `Maxsel`-Objektes können folgende Parameter angegeben werden:

- `x`
- `split`
- `col`
- `cat_names`
- `row_names`
- `main`
- `xlab`
- `ylab`
- `title_size`
- `axis_title_size`
- `axis_text_size`
- `text_size`
- `legend`

Der Parameter `x` muss ein `Maxsel`-Objekt enthalten und dementsprechend die Anforderungen einer `Maxsel`-Klasse erfüllen. Es können sowohl `Maxsel`-Objekte ohne als auch mit korrigierten P-Werten verwendet werden. Hierbei werden allerdings unterschiedliche Grafiken produziert. Weiteres wird im Parameter `split` erklärt.
Mit Hilfe des Parameters `split` lässt sich die gewünschte Darstellungsform angeben. Dabei kann zwischen drei verschiedenen Möglichkeiten gewählt werden. Wird keine Angabe bezüglich des Parameters `split` gemacht, oder aber der Parameter wird auf `split = „all"` gesetzt, so wird die Standardeinstellung verwendet. Diese zeigt jede Kategorie als gruppiertes Säulendiagramm ($Y = 0$ vs. $Y = 1$) an. Die gestrichelten vertikalen Linien zwischen den Kategorien zeigen einen möglichen Split mit dem dazugehörigen P-Wert an. Falls es sich um den optimalen Split handelt, so ist die Linie grün, ansonsten rot eingefärbt. Mit `split = „optimal"` wird der optimale Split erhalten. Hierfür werden Summen innerhalb der Einteilung gebildet und diese werden entsprechend im Säulendiagramm dargestellt. Die grüne vertikale Linie ist mit dem P-Wert beschriftet. Als dritte Option kann der Anwender selbst entscheiden, an welcher Stelle die Kategorien eingeteilt werden sollen. Hierfür kann ein numerischer Vektor angegeben werden mit $split = c(1, \ldots, k)$ und $k \leq K$. Die Grafik summiert dann jeweils die Zahlen von $1, \ldots, k$ und $k+1, \ldots, K$ auf und zeigt diese mit dem dazugehörigen P-Wert für diesen Split an. Die Farbe der vertikalen Trennlinie entspricht auch hier grün, falls es sich um einen optimalen Split handelt, und andernfalls um die Farbe rot.
Zur optischen Gestaltung der Grafik kann der Parameter `col` verwendet werden. Dieser benötigt zwei Einträge von Farben, welche für die Darstellung des Säulendiagramms verwendet werden. Die Standardeinstellung der Farben sind `"gray30"` und `"gray60"`.
Der Parameter `cat_names` ist dazu geeignet, den Kategorien individuelle Namen zuzuweisen. Der Vektor muss dabei die Länge K besitzen. Als Standardeinstellung werden die Bezeichnungen *Category 1*, ..., *Category K* verwendet.

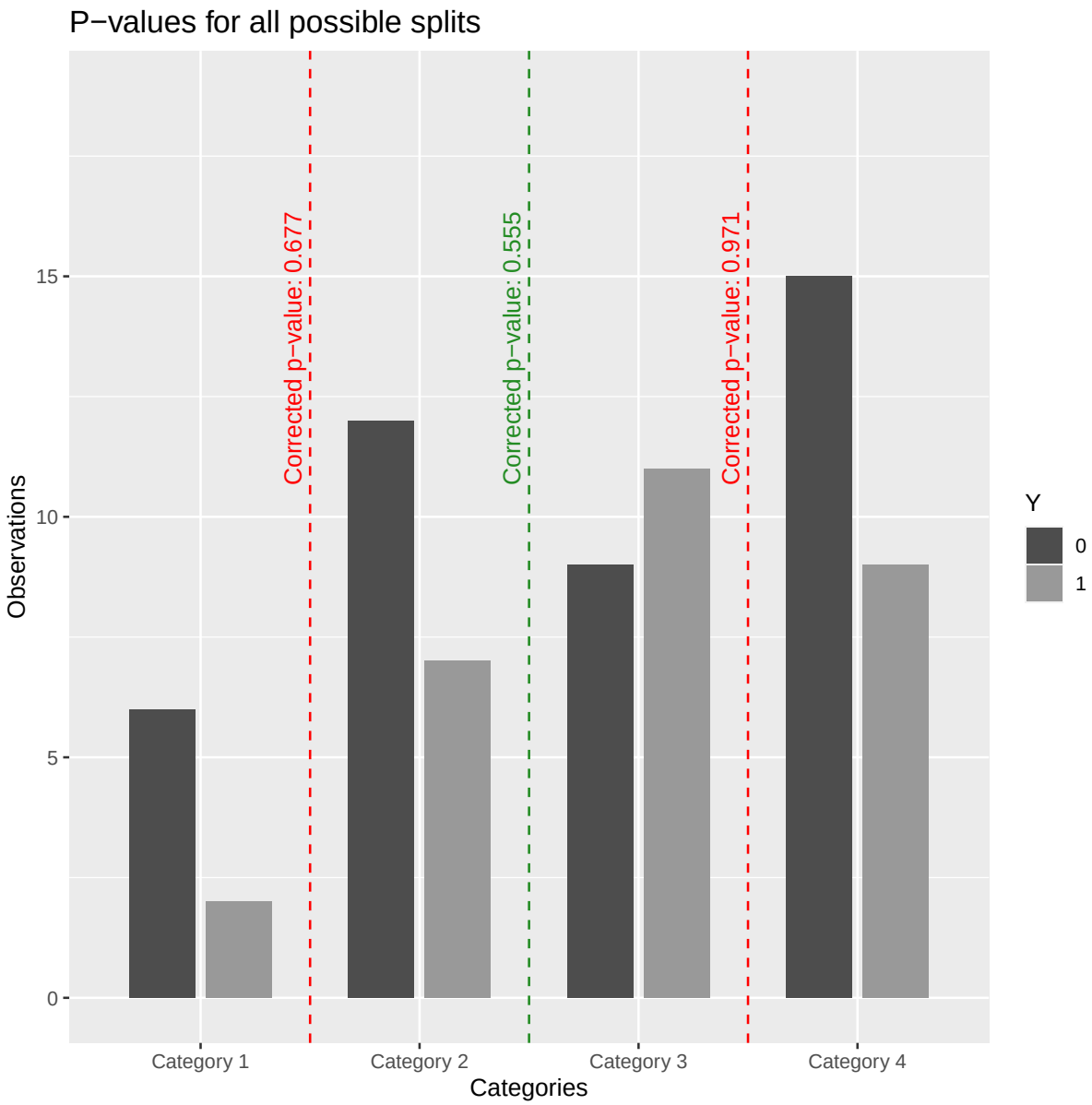
Mit dem Parameter `row_names` dagegen kann der Anwender individuelle Namen für die Variable Y angeben. Der Vektor muss dabei die Länge zwei besitzen. Während der erste Eintrag $Y = 0$ entspricht, so gilt der zweite Eintrag für $Y = 1$.
Um Änderungen an der Beschriftung vorzunehmen, können die Parameter `main, xlab` und `ylab` verwendet werden. Während mit dem ersten Parameter eine eigene Überschrift gebildet werden kann, sind die Parameter `xlab` und `ylab` dafür geeignet, die X- bzw. Y-Achsenbeschriftung zu verändern. Die Größe der Schrift kann letztlich noch mit den Parametern `title_size`, `axis_title_size`, `axis_text_size` und `text_size` modifiziert werden. Während mit `title_size` die Schriftgröße der Überschrift festlegt wird, kann mit `axis_title_size` auch die Schriftgröße der Achsenbeschriftung verändert werden. `axis_text_size` kann dazu verwendet werden, die Schriftgröße für die Skala auf der Y-Achse und die Namen der Kategorien auf der X-Achse anzupassen. Zuletzt kann mit `text_size` die Schriftgröße der Anmerkungen neben den vertikalen Linien individuell angepasst werden.
Aus unterschiedlichen Gründe kann es von Vorteil sein, dass die Legende der Grafik nicht angezeigt wird. Dies kann mit der Parametereinstellung `legend = FALSE` verwirklicht werden.

Um die bereits zuvor genutzten Daten auch grafisch zu analysieren, kann die Funktion `maxsel.plot` verwendet werden. Essenziell für die Funktion ist ein `Maxsel`-Objekt, welches bereits zuvor mit der Funktion `maxsel.test` erstellt wurde.

```
# Anzeigen der Eigenschaften des Maxsel-Objektes
maxsel_object

## Maxsel-Object with the following properties:
##
## Matrix: 2x4 matrix
## Statistic: chi2
## Value of the maximally selected criterion:  1.017465
## Corrected p-Value of the maximally selected criterion:  0.5547055
## Informations about other splits are given:  TRUE

maxsel_object@all_splits

##      Split1    Split2 statistic p_value_uncorrected p_value_corrected
## 2     1 + 2     3 + 4 1.0174653           0.3131210         0.5547055
## 1         1 2 + 3 + 4 0.9368403           0.3330918         0.6765613
## 3 1 + 2 + 3         4 0.1679004           0.6819846         0.9711485
```

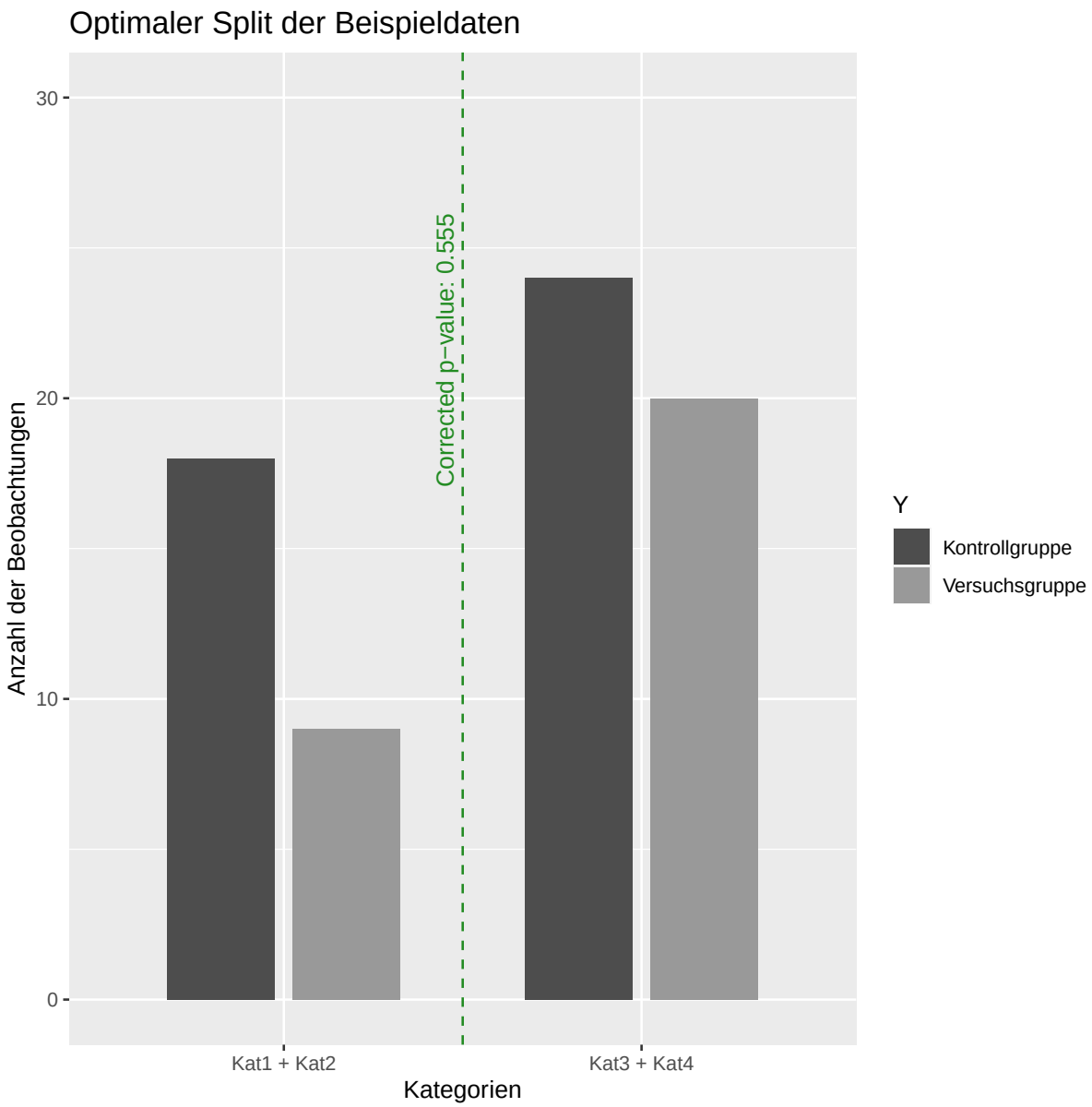
Wie der vorangegangene R-Output darstellt, besitzt das `Maxsel`-Objekt insgesamt vier verschiedene Kategorien. Der geringste P-Wert entspricht einer Einteilung der Kategorien in {Kategorie 1, Kategorie 2} vs. {Kategorie 3, Kategorie 4}. Eine grafische Unterstützung der Ergebnisse lässt sich mit der Funktion `maxsel.plot` erreichen.

```
# Grafische Darstellung des Maxsel-Objektes
maxsel.plot(maxsel_object)
```



Wie bereits erwähnt, gilt die Standardeinstellung `split = „all"`. Ist der Anwender allerdings nur an der optimalen Aufteilung der Kategorien interessiert, so kann diese mit `split = „optimal"` erreicht werden. Zusätzlich sind im folgenden auch Veränderungen für die Beschriftungen innerhalb der Grafik hinzugefügt.

```
maxsel.plot(maxsel_object, split = "optimal",
                ylab = "Anzahl der Beobachtungen", xlab = "Kategorien",
                main = "Optimaler Split der Beispieldaten",
                cat_names = c("Kat1", "Kat2", "Kat3", "Kat4"),
                row_names = c("Kontrollgruppe", "Versuchsgruppe"))
```



Falls ein `Maxsel`-Objekt keine korrigierten P-Werte für alle möglichen Aufteilungen besitzt, werden stattdessen die unkorrigierten P-Werte angezeigt. Diese werden, bei einer individuellen Angabe des *Split*-Parameters, folgendermaßen angezeigt.

```
# Erstellen der gleichen Matrix wie bei den Beispielen zuvor
data <- matrix(c(6,12,9,15,2,7,11,9), byrow = TRUE, nrow = 2, ncol = 4)
# Es werden nicht für alle Aufteilungen der korrigierte P-Wert berechnet
maxsel_object2 <- maxsel.test(x = data, y = NULL, statistic = "chi2",
                              all_splits = FALSE)
# Data.frame enthält nur die unkorrigierten P-Werte
maxsel_object2@all_splits

##      Split1    Split2 statistic p_value_uncorrected
## 2     1 + 2     3 + 4 1.0174653           0.3131210
## 1         1 2 + 3 + 4 0.9368403           0.3330918
## 3 1 + 2 + 3         4 0.1679004           0.6819846
```

Für die gewünschte Aufteilung der Kategorien von {Kat1, Kat2, Kat3} vs. {Kat4} ergibt sich die Grafik mit:

```
maxsel.plot(maxsel_object2, split = c(1,2,3),
                ylab = "Anzahl der Beobachtungen", xlab = "Kategorien",
                main = "Optimaler Split der Beispieldaten",
                cat_names = c("Kat1", "Kat2", "Kat3", "Kat4"),
                row_names = c("Kontrollgruppe", "Versuchsgruppe"),
                col = c("orange", "blue"),
                title_size = 16)
```

Die vertikale Linie ist nun mit dem Schriftzug *Uncorrected p-value:...* versehen. Zusätzlich wurde mit dem Parameter `col` eine unterschiedliche Farbkombination gewählt, welche in den Säulen und in der Legende der Grafik zu sehen ist. Mit dem Parameter `title_size` wurde zudem die Schriftgröße für die Überschrift leicht erhöht.

5 Simulationsstudien

Die vorherigen Kapitel erklärten sowohl die bereits weit verbreiteten Teststatistiken (χ^2-Test, Wilcoxon-Test, exakter Fisher Test), als auch die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix. Um die Perfomance der maximal selektierten χ^2-Teststatistik mit denen der gängigen Tests zu vergleichen, wurden verschiedene Simulationen durchgeführt. Zur Durchführung der Simulation ist die statistische Software R - Version 4.0.3 verwendet worden. Im Mittelpunkt stehen dabei die Überprüfung der Höhe des **Fehlers 1. Art** und der **Power** (Trennschärfe) der Tests.

Mit dem *Fehler 1. Art* werden Fälle bezeichnet, bei denen die Nullhypothese (H_0) fälschlicherweise abgelehnt wird. Statistische Tests werden so entworfen, dass eine obere Schranke den Fehler 1. Art begrenzt. Diese Schranke wird als Signifikanzniveau bezeichnet. Ob die verschiedenen Tests den Fehler 1. Art entsprechend begrenzen und damit die Schranke einhalten, wird in den Simulationen untersucht.
Die *Power* bzw. die Trennschärfe der Tests ergibt sich aus $1 - \beta$, wobei β wiederum der Fehler 2. Art ist. Mit dem Fehler 2. Art wird die Wahrscheinlichkeit bezeichnet, dass eine Nullhypothese fälschlicherweise nicht abgelehnt wird. Die Power eines Tests ergibt sich somit aus der Wahrscheinlichkeit, dass der Test einen Unterschied zwischen den Stichprobenpopulationen erkennt, wenn dieser auch tatsächlich vorhanden ist. Die Power eines Tests hängt sowohl von dem Signifikanzniveau α wie auch vom Stichprobenumfang n ab (Fahrmeir et al., 2016). Unterschreitet ein Test das vorgegebene Signifikanzniveau (*hier: 0.05*), geht dies zu Lasten der erreichten Power. Dementsprechend wird in der folgenden Simulation und in den Bewertungen bezüglich der Höhe des Fehlers 1. Art ein Test positiver bewertet, wenn die Höhe des Fehlers 1. Art relativ nah an der 5%-Grenze liegt. Im Gesamtkonzept ist ein Test aber trotzdem besser, falls er, trotz deutlicher Unterschreitung der 5%-Grenze, eine höhere Power aufweist als ein anderer Test. Ein Nachteil liegt dann höchstens darin, dass er nicht seine maximal mögliche Power ausschöpft.
Die Simulationen basieren auf dem *Monte-Carlo-Simulations*-Verfahren. Zu Beginn werden die Wahrscheinlichkeiten für die Kategorien in Kontroll- und Versuchsgruppe gewählt. Hierfür sollten die Wahrscheinlichkeiten möglichst verschiedene Szenarios der Realität gut abdecken. Sei $\frac{n}{2} = n_K = n_V$ die Anzahl an Beobachtungen in der Kontroll- wie auch in der Versuchsgruppe. Für jeden Simulationsdurchgang werden dann pro Gruppe $\frac{n}{2}$ Beobachtungen gemäß der vorgegebenen Wahrscheinlichkeiten in die Kategorien in Kontroll- und Versuchsgruppe aufgeteilt. Dieser Vorgang wiederholt sich anschließend m-mal, wobei die endgültigen Ergebnisse letztlich die gemittelten Werte über alle m Simulationsdurchgänge darstellen.

5.1 Simulationsaufbau

Für die Simulationen wurden Parameter verwendet, die möglichst viele unterschiedliche Szenarios abdecken sollen. Das Ziel ist die Stärken und Schwächen der unterschiedlichen Testverfahren aufzuzeigen.

Zur Überprüfung der Einhaltung des Signifikanzniveaus werden sowohl in der Kontroll- als auch in der Versuchsgruppe gleiche Wahrscheinlichkeiten für die Zugehörigkeit zu einer Kategorie angenommen.

Die Hypothesen

$\mathbf{H_0}$: Verteilung der abhängigen Variable in beiden Gruppen ist identisch

$\mathbf{H_1}$: Verteilung der abhängigen Variable in beiden Gruppen ist **nicht** identisch

werden anschließend mit den verschiedenen Teststatistiken überprüft. Aufgrund des Simulationsaufbaus mit gleichen Wahrscheinlichkeiten für die Gruppenzugehörigkeit, sollte die relative Anzahl der abgelehnten Hypothesen dem gewählten Signifikanzniveau entsprechen. Zu große Abweichungen nach oben (erhöhte Wahrscheinlichkeit für Fehler 1. Art) oder nach unten (konservatives Verhalten zu Lasten der Power) sind nicht erwünscht.
Nach einem ähnlichen Muster wird die Power (Trennschärfe) der Tests überprüft. Der Unterschied besteht darin, dass für die Versuchsgruppe unterschiedliche Wahrscheinlichkeiten für die einzelnen Kategorien angenommen werden. Dadurch wäre die theoretisch richtige Testentscheidung, dass H_0 abgelehnt wird. Die Power der Tests ergibt sich dann als die relative Anzahl aller (richtig) abgelehnten Nullhypothesen H_0.

Parameterwahl

Da sowohl die Einhaltung des Signifikanzniveaus als auch die Power von mehreren Parametern abhängen, werden diese in der Simulation ebenfalls berücksichtigt.

Das **Signifikanzniveau** α bestimmt ab welcher Grenze eine bestimmte Nullhypothese ablehnt werden soll. Typische Werte für das Signifikanzniveau sind beispielsweise 0.1, 0.05 und 0.01 (Fahrmeir et al., 2016). Für die Durchführung der Simulationen wurde deshalb ein Signifikanzniveau von

$$\alpha = 0.05$$

festgelegt.

Die **Anzahl an Simulationsläufen** soll die Varianz der Ergebnisse möglichst gering halten. Für eine relativ hohe Zahl an Simulationen tendiert der daraus resultierende Schätzwert gegen den wahren Wert. Um ausreichend gesicherte Daten zu erhalten wurden in dieser Simulation

$$m = 100000$$

Iterationen durchgeführt.

Die **Stichprobengröße** hängt sowohl mit dem Fehler 1. Art als auch mit dem Fehler 2. Art zusammen und hat dementsprechend Auswirkungen auf beide Untersuchungsergebnisse. Um die Performance der Teststatistiken für kleinere und für größere Stichprobengrößen zu testen, wurden fünf verschiedene Gruppengrößen verwendet. Die Anzahl an Beobachtungen in Kontroll- und Versuchsgruppe ist dabei immer identisch. Somit gilt

$$\frac{n}{2} = n_K = n_V = 30, 60, 100, 150, 300$$

Ebenfalls untersucht wurde die **Anzahl an Kategorien**. Da es sich um eine ordinal skalierte Einflussvariable handelt, spielt auch die Anzahl an verschiedenen Ausprägungen (Kategorien) eine Rolle. Um verschiedene Szenarios der Realität ausreichend abzudecken,

wurden die Simulationen mit unterschiedlichen Anzahlen an Kategorien durchgeführt. Es wurden letztlich mit

$$k = 3, 5, 8$$

drei verschiedene Parameterausprägungen verwendet.

Auch die Wahl der Szenarios und der damit einhergehenden **Wahl der Wahrscheinlichkeiten** für die einzelnen Kategorien ist von großer Bedeutung. Im Rahmen dieser Masterarbeit wurden drei bzw. vier Szenarios betrachtet. Sie wurden so ausgewählt, dass möglichst viele verschiedene Facetten der Realität abgebildet werden konnten.

Simulation bezüglich des Fehlers 1. Art

Zur Simulation bezüglich des Fehlers 1. Art wurden insgesamt drei verschiedene Szenarios betrachtet. Da in der Kontroll- und Versuchsgruppe dieselben Wahrscheinlichkeiten für die Kategorien angenommen werden (siehe Kapitel 5), lassen sich die Szenarios folgendermaßen zusammenfassen:

1. Szenario: Gleiche Wahrscheinlichkeit für alle Kategorien
2. Szenario: Wahrscheinlichkeit für Kategorie steigt monoton an
3. Szenario: erhöhte Wahrscheinlichkeit für eine der mittleren Kategorien.

Tabelle 2 zeigt die verwendeten Wahrscheinlichkeiten je nach Anzahl an Kategorien. Es werden somit neun verschiedene Unter-Szenarios bezüglich der Einhaltung des Signifikanzniveaus berechnet. Da jedes dieser neun Unter-Szenarios nochmals mit fünf verschiedenen Stichprobengrößen berechnet wird, werden insgesamt 45 Simulationen mit jeweils 100 000 Iterationen durchgeführt.

		Kategorie							
		1	2	3	4	5	6	7	8
Szenario	1a	1/3	1/3	1/3					
	1b	1/5	1/5	1/5	1/5	1/5			
	1c	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
	2a	0.1	0.2	0.7					
	2b	0.1	0.1	0.2	0.2	0.4			
	2c	0.05	0.05	0.075	0.1	0.1	0.1	0.225	0.3
	3a	0.2	0.5	0.3					
	3b	0.05	0.2	0.4	0.25	0.1			
	3c	0.05	0.1	0.15	0.3	0.2	0.1	0.05	0.05

Tabelle 2: Wahrscheinlichkeit der Zugehörigkeit zu einer Kategorie für die untersuchten Szenarios. Diese Wahrscheinlichkeiten werden zur Untersuchung des Fehlers 1. Art verwendet.

Simulation bezüglich der Power

Zur Simulation bezüglich der Power wurden insgesamt vier verschiedene Szenarios betrachtet. Hierbei handelt es sich nun um verschiedene Wahrscheinlichkeiten für die Kategorien von Kontroll- und Versuchsgruppe. Die Wahrscheinlichkeiten zur Simulation des Fehlers 1. Art dienen hierbei als die Wahrscheinlichkeiten für die Kontrollgruppe. Diese werden mit abweichenden Wahrscheinlichkeiten der Versuchsgruppe verglichen. Die vier Szenarios behandeln dabei folgende Sachverhalte:

1. Szenario: gleiche Wahrscheinlichkeiten *vs.* linearer Anstieg der Wahrscheinlichkeiten
2. Szenario: monotoner Anstieg mit Schwellenwert *vs.* gleiche Wahrscheinlichkeiten
3. Szenario: erhöhte Wahrscheinlichkeit für eine der mittleren Kategorien *vs.* monotoner Anstieg der Wahrscheinlichkeit
4. Szenario: Vergleich relativ ähnlicher Kontroll- und Versuchsgruppen mit jeweils erhöhter Wahrscheinlichkeit für eine der mittleren Kategorien

Diese vier Szenarios werden dann mit jeweils drei verschiedenen Anzahlen an Kategorien und fünf verschiedenen Stichprobengrößen berechnet. Somit werden insgesamt 60 Simulationen mit jeweils 100 000 Iterationen durchgeführt. Tabelle 3 zeigt die verwendeten Wahrscheinlichkeiten je nach Anzahl an Kategorien. Hierbei sind lediglich die Wahrscheinlichkeiten für die Versuchsgruppen dargestellt. Die dazugehörigen Wahrscheinlichkeiten für die Kontrollgruppe sind in Tabelle 2 aufgelistet.

		Kategorie							
		1	2	3	4	5	6	7	8
Szenario	1a	1/9	3/9	5/9					
	1b	1/15	2/15	3/15	4/15	5/15			
	1c	1/36	2/36	3/36	4/36	5/36	6/36	7/36	8/36
	2a	1/3	1/3	1/3					
	2b	1/5	1/5	1/5	1/5	1/5			
	2c	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
	3a	0.2	0.3	0.5					
	3b	0.1	0.1	0.2	0.25	0.35			
	3c	0.05	0.05	0.075	0.1	0.125	0.15	0.2	0.25
	4a	0.3	0.4	0.3					
	4b	0.1	0.2	0.3	0.35	0.05			
	4c	0.05	0.05	0.2	0.2	0.25	0.1	0.1	0.05

Tabelle 3: Wahrscheinlichkeit der Zugehörigkeit zu einer Kategorie für die untersuchten Szenarios. Diese Wahrscheinlichkeiten werden zur Untersuchung der Power verwendet.

Dabei gehören alle Szenarios mit gleicher Kennzeichnung zusammen. Für Szenario 4 wurden die gleichen Wahrscheinlichkeiten für die Kontrollgruppe verwendet wie in Szenario 3. Aufgrund dessen findet sich in Tabelle 2 keine Auflistung für Szenario 4. Da es sich in Szenario 3 und Szenario 4 um die gleichen Wahrscheinlichkeiten für die Kategorien der Kontrollgruppe handelt, wurde in Szenario 4 auf die erneute Untersuchung des Fehlers 1. Art verzichtet. Diese wäre identisch mit jener aus Szenario 3.

In Abbildung 2 sind die dazugehörigen Wahrscheinlichkeitsfunktionen grafisch dargestellt. Dabei entspricht die Abbildung genau genommen der Simulation für die Power der einzelnen Tests. Die Zahlen der Kontroll- und Versuchsgruppe aus Abbildung 2 entsprechen denen aus Tabelle 2 und Tabelle 3. Für die Simulation des Fehlers 1. Art entspricht die dargestellte Kontrollgruppe den verwendeten Wahrscheinlichkeiten, welche in der Simulation sowohl für die Kontroll- als auch für die Versuchsgruppe verwendet wurden.

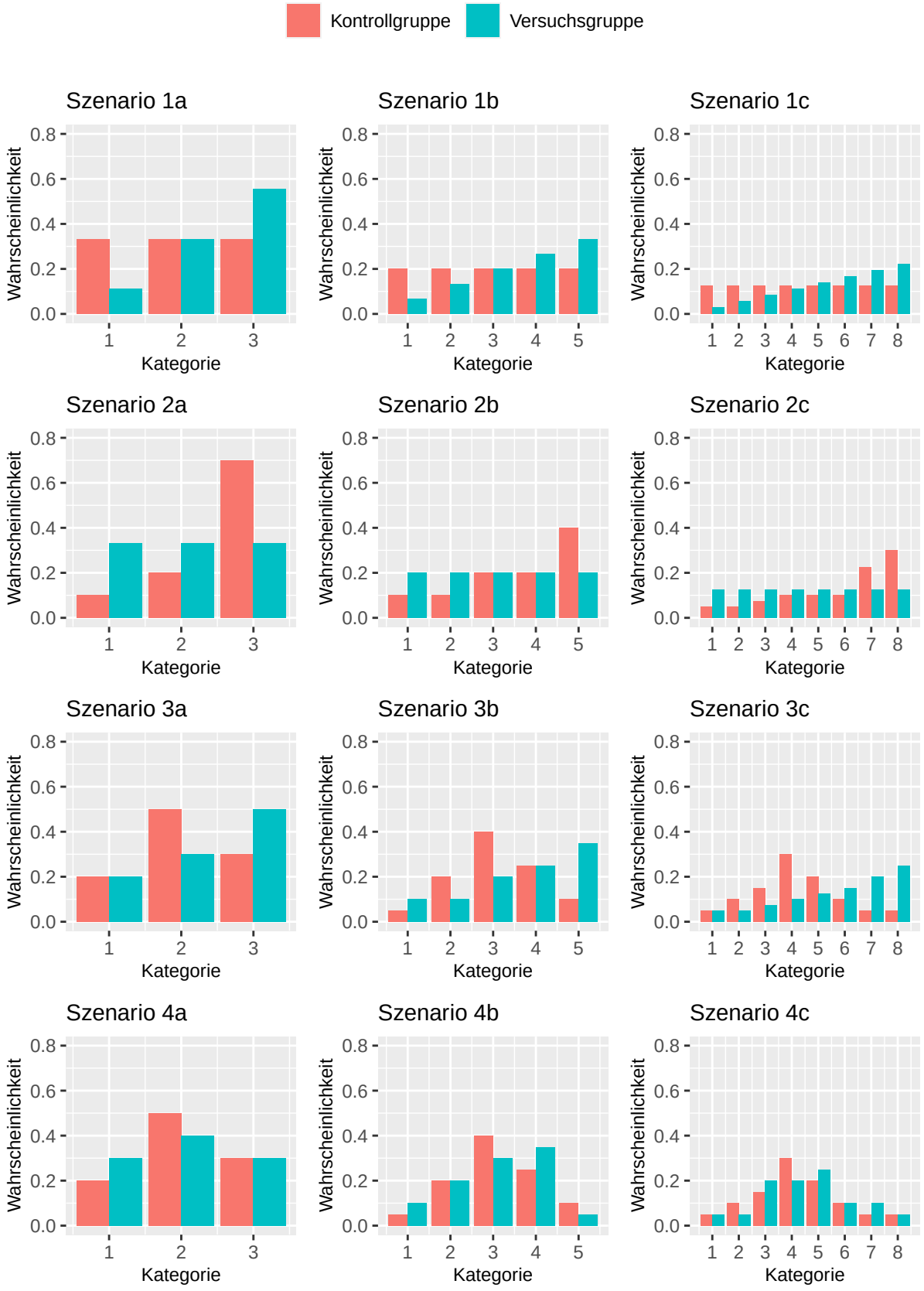


Abbildung 2: Gegenüberstellung der Dichte in Kontroll- und Versuchsgruppe innerhalb der getesteten Szenarios.

5.2 Ergebnisse

Um die Ergebnisse der Simulationen genauer zu betrachten, werden sie im folgenden zuerst getrennt nach Szenario untersucht. Anschließend werden auch allgemeingültige Aussagen und eine Zusammenfassung der vorherigen Ergebnisse vorgestellt. Alle durchgeführten Simulationen sind für spätere Untersuchungen reproduzierbar.

5.2.1 Szenario 1

In Szenario 1 sollte untersucht werden, welcher Test die besten Ergebnisse erzielt, bei einer gleichverteilten Wahrscheinlichkeit der Kategorien in der Kontrollgruppe und einer ansteigenden Wahrscheinlichkeit der Kategorien in der Versuchsgruppe (siehe Abbildung 2). Dieses Szenario beleuchtet damit den Fall, dass in der Kontrollgruppe die Kategorie keinen erkennbaren Zusammenhang mit der Wahrscheinlichkeit der Kategorie hat. In der Versuchsgruppe lässt sich dagegen erkennen, dass mit steigender Kategorie auch die Wahrscheinlichkeit der Kategorie zunimmt. Dabei wurde eine gleichbleibende Steigung der Wahrscheinlichkeit ohne konkreten Schwellenwert angenommen.

Ergebnisse bezüglich des Fehlers 1. Art

Um zu ermitteln, ob die untersuchten Tests den Fehler 1. Art erfolgreich begrenzen, wurden für die Kontroll- und Versuchsgruppe gleiche Wahrscheinlichkeiten für die Kategorien angenommen. Die relative Anzahl aller fälschlicherweise abgelehnten Nullhypothesen sollte dann unter der 5%- Linie liegen. Um die verschiedenen Einflüsse von der *Anzahl der Kategorien* und der *Gruppengröße* zu analysieren, wurden diese einzeln betrachtet.

In Abbildung 3 sind vier verschiedene Grafiken zu sehen, welche den Verlauf des Fehlers 1. Art mit ansteigender Gruppengröße skizzieren. Während in der ersten Grafik alle Daten enthalten sind, sind die folgenden drei Grafiken untergliedert in die jeweilige Anzahl an Kategorien. Somit ist Grafik 1 der Mittelwert der drei restlichen Grafiken. Für jede Kombination von Gruppengröße und Anzahl an Kategorien wurden insgesamt 100 000 Simulationen durchgeführt und die hier dargestellten Verläufe stellen die Mittelwerte dieser Simulationen dar.
Besonders auffällig ist, dass der *Maxsel*-Test, ungeachtet von der Anzahl an Kategorien, das vorgegebene Signifikanzniveau von 5% nicht einhält. Für eine höhere Anzahl an Kategorien lässt sich allerdings ein im Mittel geringerer Fehler 1. Art feststellen. Dieser Fehler sinkt meist auch mit steigender Gruppengröße. Für eine Gruppengröße von $n = 300$ und acht verschiedenen Kategorien ist der Fehler 1. Art im Mittel nur noch leicht über der Grenze bei 0.05.
Der χ^2-Test ist besonders bei wenigen Beobachtungen pro Kategorie (z.B. acht Kategorien und 30 Beobachtungen pro Gruppe) relativ konservativ. Das unter diesen Umständen (zu) strenge Einhalten des Signifikanzniveau kann sich dadurch mit einem Verlust der Power bemerkbar machen.
Während bei dem *Wilcoxon*-Test und dem *exakte Fisher*-Test relativ kleine Abweichungen bei geringer Gruppengröße zu erkennen sind, so schneiden diese, zusammen mit dem χ^2-Test, doch deutlich besser ab als der *Maxsel*-Test.

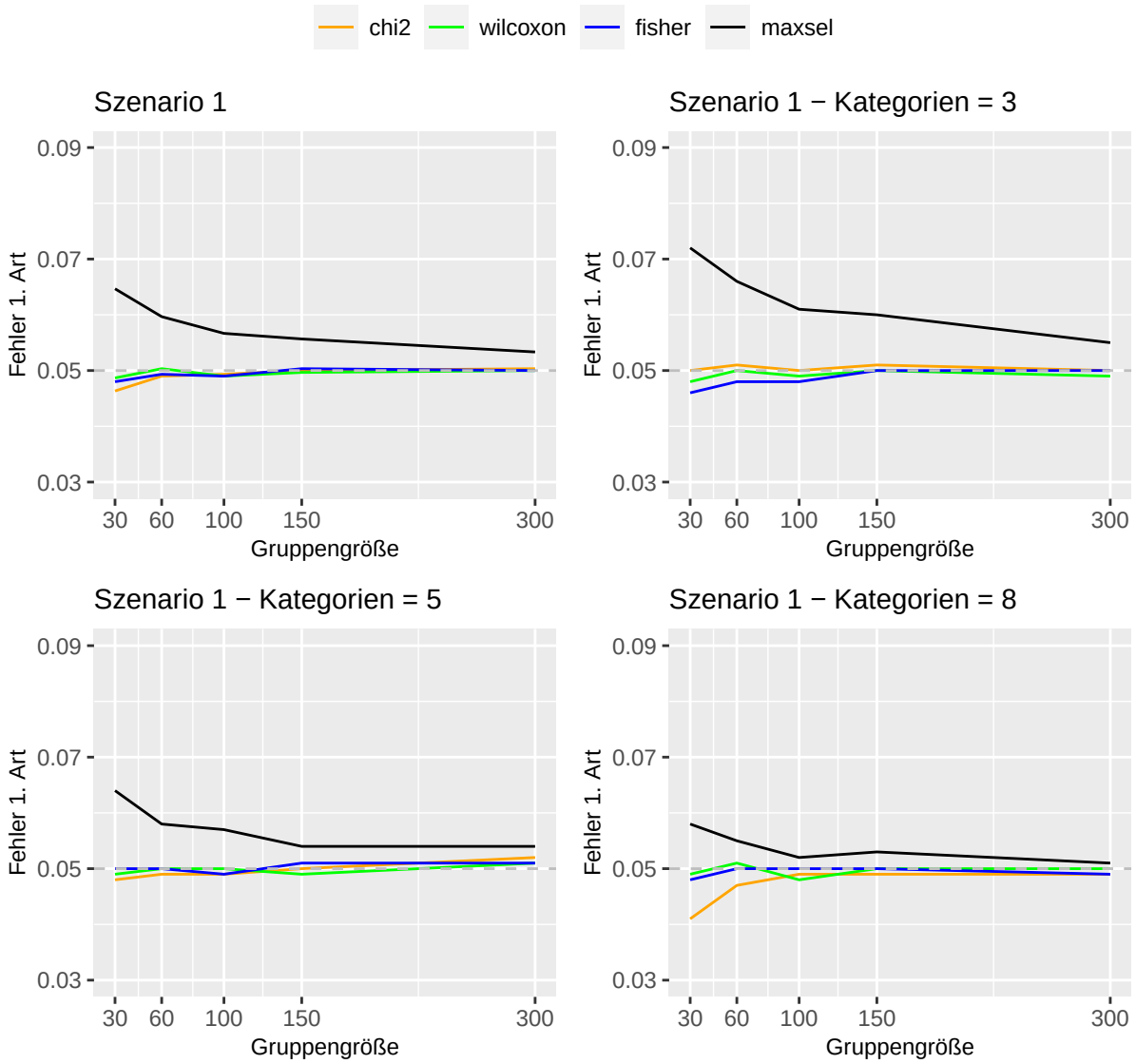


Abbildung 3: Verlauf der Simulation des Fehlers 1. Art in Szenario 1. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

In Abbildung 4 verstärkt sich dieser Eindruck noch weiter. Die erste Grafik zeigt auch hier eine Zusammenfassung über alle Gruppengrößen im Verlauf mit steigender Anzahl an Kategorien. Die folgenden Grafiken sind dagegen aufgeteilt und zeigen den Verlauf für jeweils eine einzelne Gruppengröße an.
Hierbei zeigt der *Maxsel*-Test erneut, insbesondere bei kleinen Gruppengrößen, seine Schwierigkeiten bei der Einhaltung des Signifikanzniveaus. Allerdings gilt dies auch mit Abstrichen für den χ^2-Test. Der *Wilcoxon*-Test und der *exakte Fisher*-Test dagegen zeigen relativ konstant gute Ergebnisse bezogen auf den Fehler 1. Art.
Wie bereits in Abbildung 3 ist auch hier zu sehen, dass für große Gruppengrößen und einer großen Anzahl an Kategorien die besten Ergebnisse für den *Maxsel*-Test erzielt werden können.
Bezüglich des Fehlers 1. Art in Szenario 1 lässt sich letztlich festhalten, dass insbesondere für kleine Gruppengrößen und wenige Kategorien der *Maxsel*-Test das vorgegebene Signifikanzniveau nicht einhält.

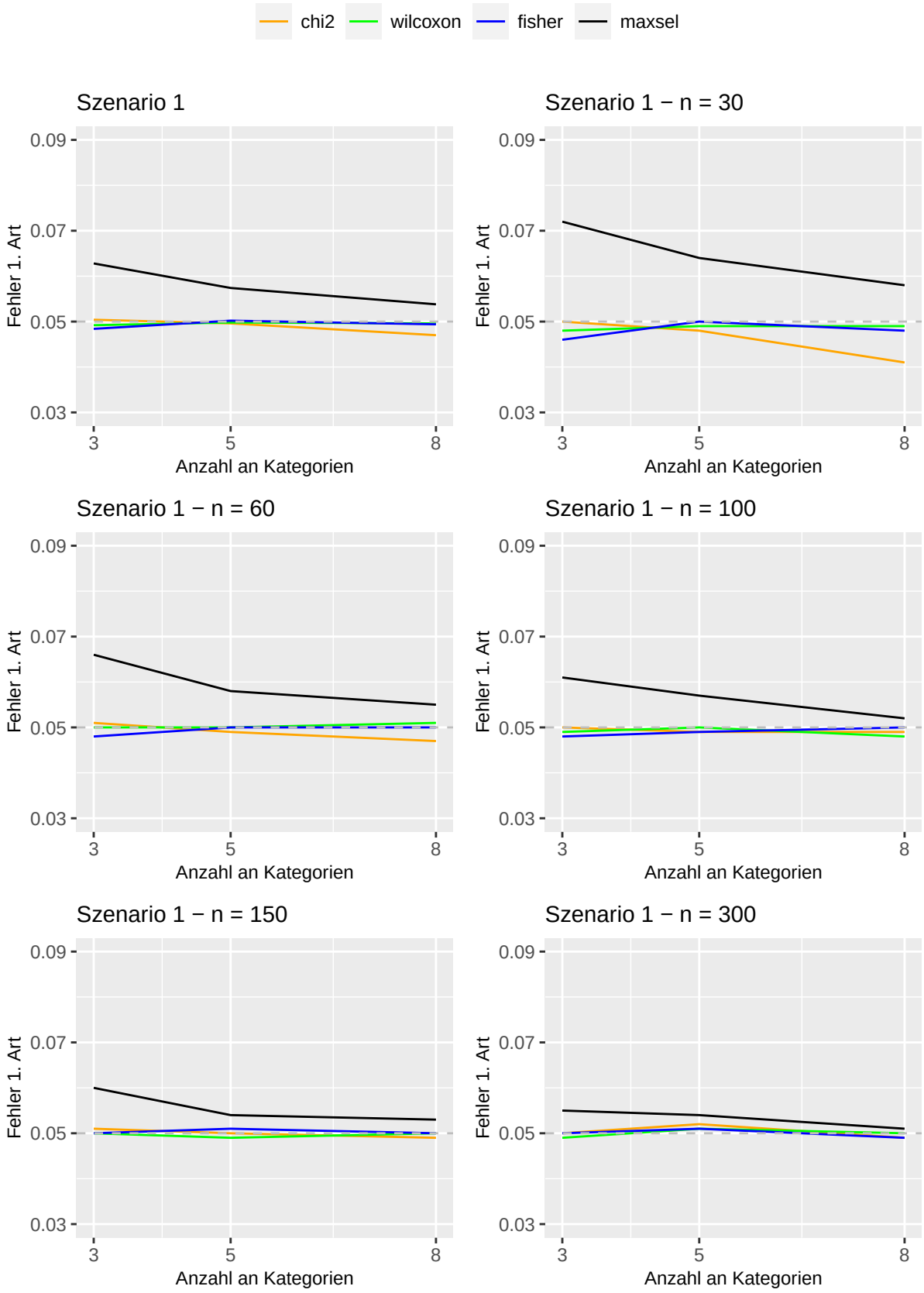


Abbildung 4: Verlauf der Simulation des Fehlers 1. Art in Szenario 1. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

Ergebnisse bezüglich der Power

Neben dem Fehler 1. Art wurde auch die Power der Tests in Szenario 1 getestet. Dabei wurden insgesamt fünf verschiedene Gruppengrößen und drei verschiedene Anzahlen an Kategorien verglichen. Zum Vergleich der Power wurden für die Kontrollgruppe und die Versuchsgruppe unterschiedliche Wahrscheinlichkeiten für die Kategorien angenommen (siehe Tabelle 2 und Tabelle 3). Je mehr Nullhypothesen, welche eine gleiche Verteilung in Kontroll- und Versuchsgruppe annehmen, abgelehnt werden, desto größer die Power und desto besser der jeweilige Test.

Für eine größere Gruppengröße in Kontroll- und Versuchsgruppe lässt sich, wie in Abbildung 5 zu sehen, eine deutliche Zunahme der Power über alle Tests erkennen. Spätestens ab einer Gruppengröße von 300 erreichen alle Tests eine Power von 1, was bedeutet, dass alle Nullhypothesen korrekterweise abgelehnt wurden.

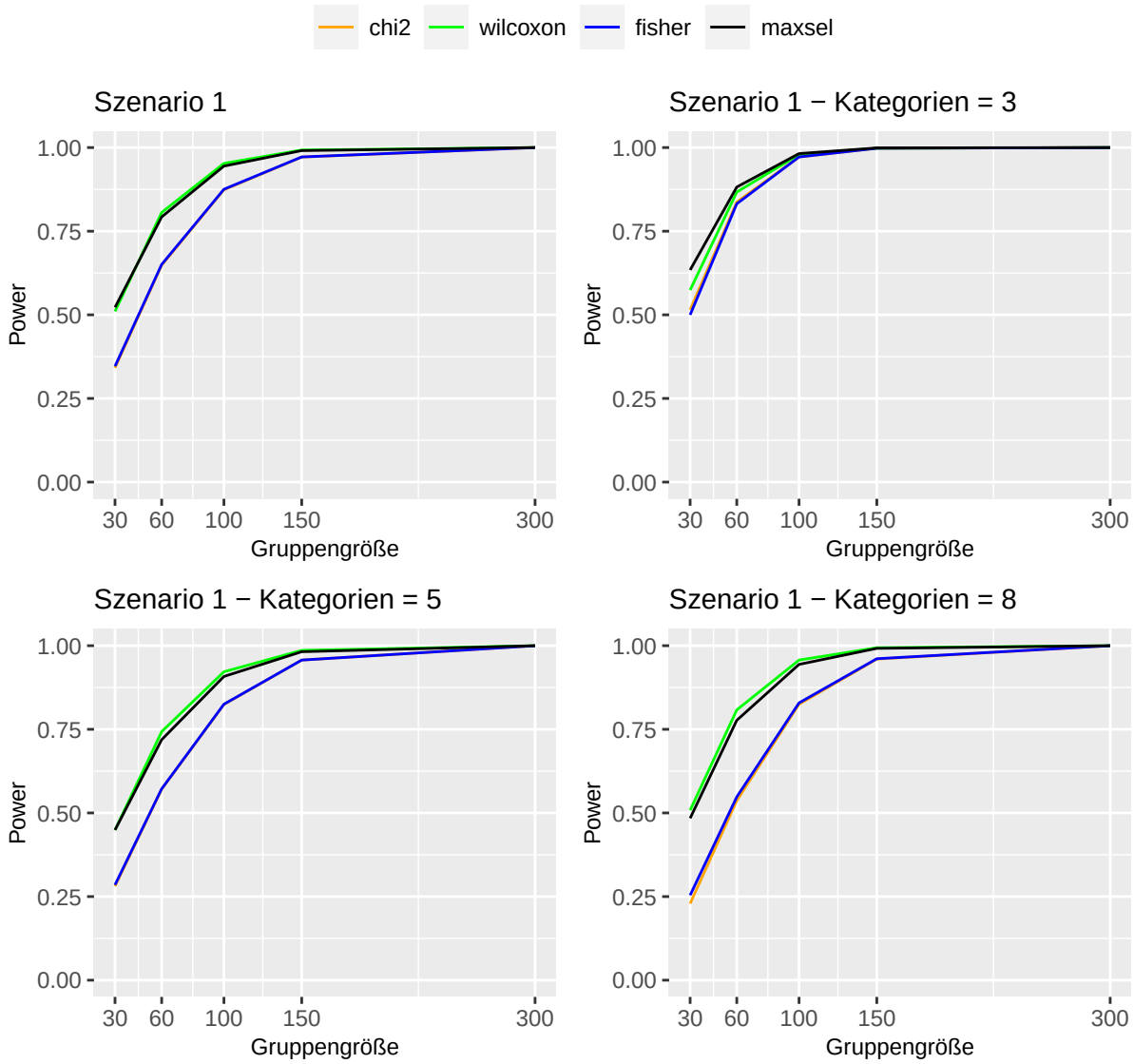


Abbildung 5: Verlauf der Simulation der Power in Szenario 1. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

Bei Daten, die aus drei ordinalen Kategorien stammen, lassen sich auch für geringere Gruppengrößen nur geringe Unterschiede zwischen der Power der Tests erkennen. Für eine höhere Anzahl an Kategorien lassen sich allerdings zwei Lager ausmachen. So weisen hier der *Wilcoxon*-Test und der *Maxsel*-Test in kleineren Gruppengrößen eine erkennbar bessere Power auf, als der *Fisher*-Test und der χ^2-Test.
In Abbildung 6 sind die Grafiken nach Gruppengröße aufgeteilt. Wie bereits in Abbildung 5 zu sehen war, steigt die Power mit steigender Gruppengröße an. Zudem ist auch hier gut zu erkennen, dass die Power in Szenario 1 besonders hoch für den *Wilcoxon*-Test und den *Maxsel*-Test ist. Fast alle Tests schneiden zudem am Besten ab, wenn die Anzahl der Kategorien bei drei liegt, welches der geringste getestete Wert war. Für den *Wilcoxon*-Test und den *Maxsel*-Test gilt, dass acht Kategorien meist besser oder zumindest gleich gut abschneiden wie fünf Kategorien. Für den *Fisher*-Test und den χ^2-Test lässt sich dagegen nur ein geringer Abwärtstrend zwischen fünf und acht Kategorien erkennen.

Bezogen auf die Power der Tests in Szenario 1 schneiden der *Wilcoxon*-Test und der *Maxsel*-Test am Besten ab. Sie kommen sowohl mit kleinen als auch mit größeren Gruppengrößen besser zurecht. Bei einer entsprechend großen Stichprobe lassen sich allerdings auch mit dem *Fisher*-Test und dem χ^2-Test eine Power von 1 erreichen.

Zwischenfazit: Szenario 1

Insgesamt lässt sich durch die Simulationen in Szenario 1 beobachten, dass besonders der *Wilcoxon*-Test gut mit den Anforderungen zurecht kommt. So schneidet er sowohl bei der Einhaltung des Signifikanzniveaus als auch bei der erreichten Power als einer der Besten ab. Der *Fisher*-Test und der χ^2-Test erreichen zwar die Zielsetzung für den Fehler 1. Art meist gut, allerdings ist bei diesen Tests eine deutlich geringere Power zu erkennen. Der *Maxsel*-Test erreicht zwar bezüglich der Power sehr gute Werte auf einem ähnlichen Niveau wie der *Wilcoxon*-Test, allerdings hat er besonders bei geringen Gruppengrößen und wenigen Kategorien große Probleme den Fehler 1. Art zu begrenzen.

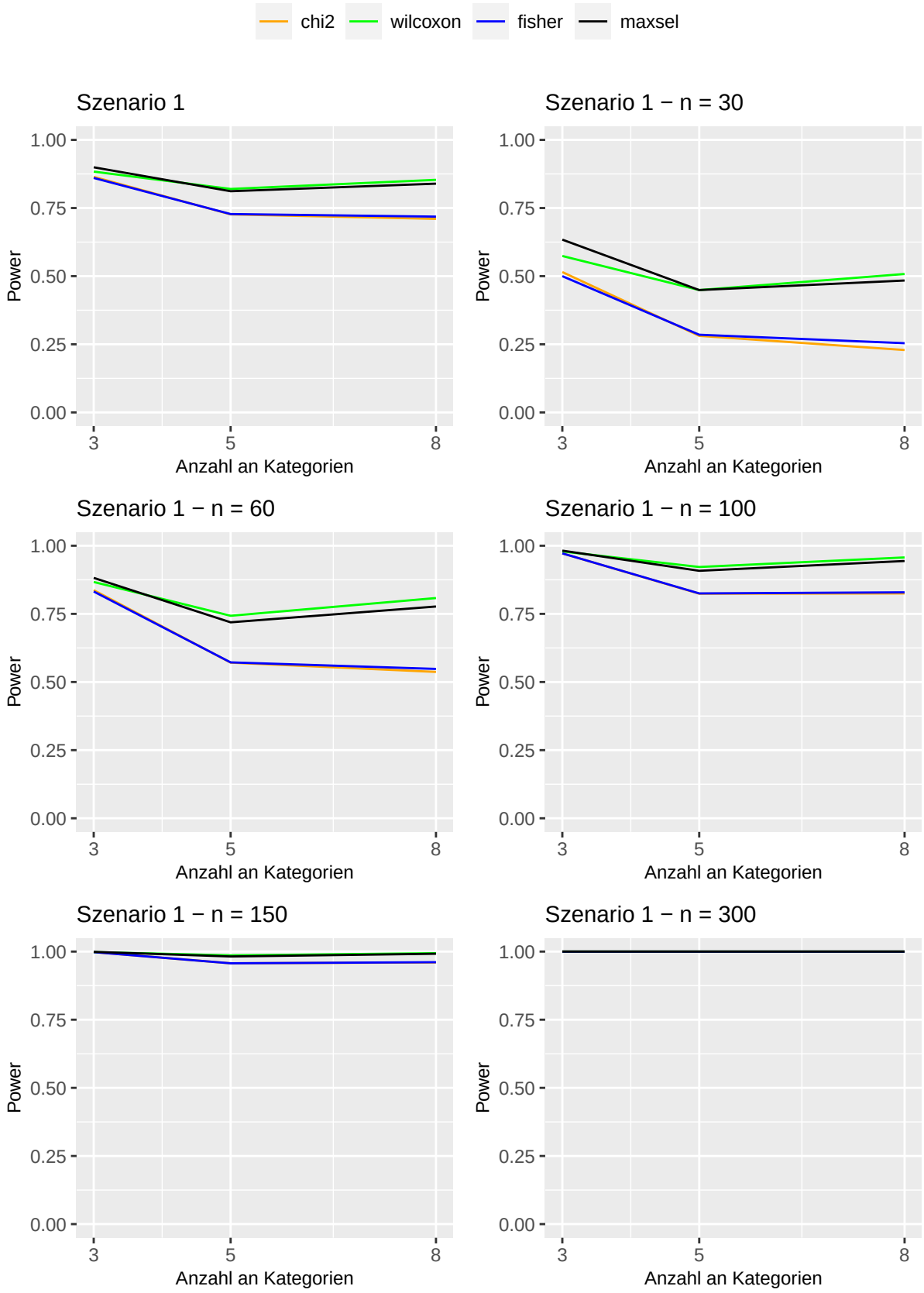


Abbildung 6: Verlauf der Simulation der Power in Szenario 1. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

5.2.2 Szenario 2

Szenario 2 ist eine nur leicht veränderte Variation von Szenario 1. Hier weist die Versuchsgruppe eine konstante Wahrscheinlichkeit für die Kategorien auf, während die Kontrollgruppe einem monoton ansteigenden Verlauf über die Kategorien folgt. Der Unterschied zwischen den Szenarios liegt darin, dass es in Szenario 2 einen konkreten Schwellenwert in der Kontrollgruppe gibt. Dabei steigt die Wahrscheinlichkeit zwischen zwei aufeinanderfolgenden Kategorien relativ stark an. Für Szenario 2a liegt dieser Schwellenwert zwischen Kategorie 2 und 3. Für Szenario 2b liegt er zwischen Kategorie 4 und 5 und für Szenario 2c liegt er zwischen Kategorie 6 und 7 (siehe Tabelle 2). Da zur Analyse des Fehlers 1. Art lediglich die Wahrscheinlichkeiten der Kategorien in der Kontrollgruppe verwendet werden, wurden in Szenario 2 die Wahrscheinlichkeiten zwischen Kontrollgruppe und Versuchsgruppe im Vergleich zu Szenario 1 vertauscht.

Ergebnisse bezüglich des Fehlers 1. Art

Zur Simulation des Fehlers 1. Art wurden insgesamt 100 000 Simulationsdurchgänge für jede Kombination der Gruppengröße und der Anzahl an Kategorien durchgeführt. Die Ergebnisse hierfür sind in den Abbildungen 7 und 8 dargestellt. Zur Feststellung des Fehlers 1. Art wurden gleiche Wahrscheinlichkeiten der Kategorien in Kontroll- und Versuchsgruppe angenommen (siehe Tabelle 2). Der Fehler 1. Art ergibt sich dementsprechend aus der relativen Anzahl aller fälschlicherweise abgelehnten Nullhypothesen.

In Abbildung 7 ist, neben dem mittleren Verlauf des Fehlers 1. Art über verschiedene Gruppengrößen, auch eine Untergliederung in die verschiedenen Anzahlen an Kategorien zu sehen. Auch hier lässt sich feststellen, dass der *Maxsel*-Test große Schwierigkeiten aufweist, sich an die Beschränkung des Fehlers 1. Art zu halten. Wie in Szenario 1 lässt sich dieser Sachverhalt besonders für kleine Gruppengrößen und wenige Kategorien beobachten. Selbst für relativ große Gruppengrößen von $n_{KG} = n_{VG} = 300$ und bei einer Anzahl von acht ordinalen Kategorien lässt sich noch ein im Mittel leicht erhöhter Fehler 1. Art beobachten.
Auch der χ^2-Test hat unter bestimmten Umständen Probleme, das erwünschte 5%-Ziel zu erreichen. Anders als der *Maxsel*-Test trifft der χ^2-Test, wie bereits in Szenario 1 zu sehen war, in einigen Fällen eine zu konservative Testentscheidungen. Besonders bei wenigen Beobachtungen für viele Kategorien trifft dies zu.
Das Problem eines zu konservativen Tests lässt sich auch in Ansätzen für den *exakten Fisher*-Test beobachten, welches besonders bei kleiner Gruppengröße und wenigen Kategorien auftritt.
Der *Wilcoxon*-Test hält als einziger Test, von einzelnen geringen Schwankungen abgesehen, die Marke von 0.05 sehr gut ein. Dabei kommt er meistens auch mit einer geringen Gruppengröße gut aus.

Abbildung 8 zeigt für die verschiedenen Gruppengrößen den Verlauf bezüglich der Anzahl an Kategorien. Dabei lassen sich unterschiedliche Auswirkungen auf die Tests beobachten. Der *Maxsel*-Test zeigt für jegliche Gruppengrößen einen positiven Verlauf in Richtung 5%-Fehlerquote bei steigender Anzahl an Kategorien. Besonders mit nur drei Kategorien ist der Fehler deutlich erhöht. Dieser wird zwar bereits mit fünf Kategorien merklich abgeschwächt, allerdings zeigen die Simulationen mit acht Kategorien die besten Ergebnisse.

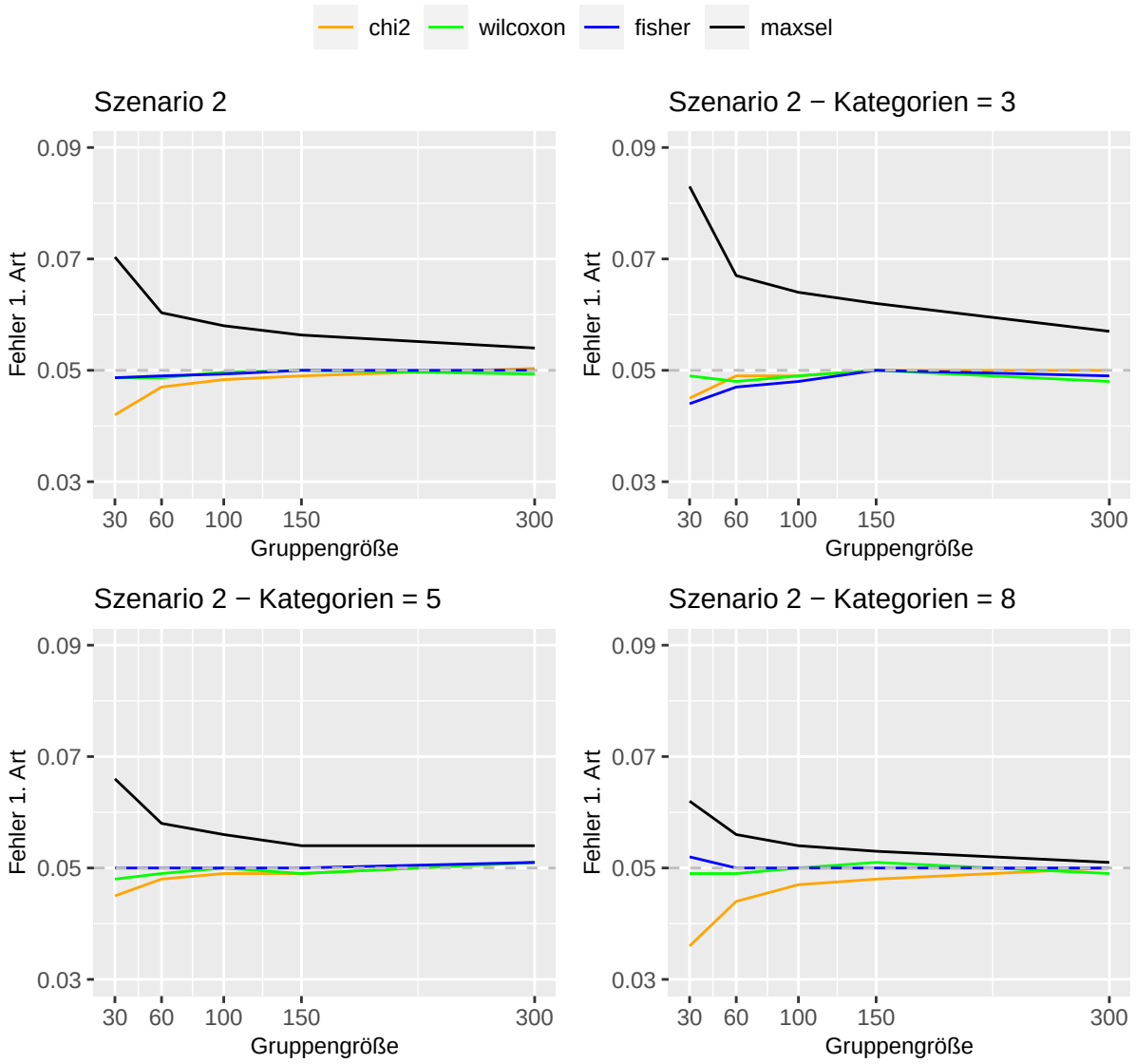


Abbildung 7: Verlauf der Simulation des Fehlers 1. Art in Szenario 2. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

Der *exakte Fisher*-Test verhält sich hier ähnlich wie der *Maxsel*-Test. Der Unterschied ist, dass er die 5%-Grenze nicht überschreitet, sondern zum Teil darunter liegt. Das Problem tritt ebenfalls hauptsächlich bei einer niedrigen Anzahl an Kategorien auf.
Der χ^2-Test trifft, wie bereits zum Teil zuvor gesehen, besonders für eine hohe Anzahl an Kategorien eine zu konservative Testentscheidung. Dies liegt vermutlich daran, dass für kleine Stichproben mit vielen Kategorien nur wenige Beoachtungen pro Kategorie zu verzeichnen sind.
Wie bereits in Abbildung 7 gesehen, erzielt der *Wilcoxon*-Test die deutlich besten Ergebnisse bezüglich des Fehlers 1. Art. Dabei sind keine Muster im Blick auf die Gruppengröße und die Anzahl der Kategorien ersichtlich.

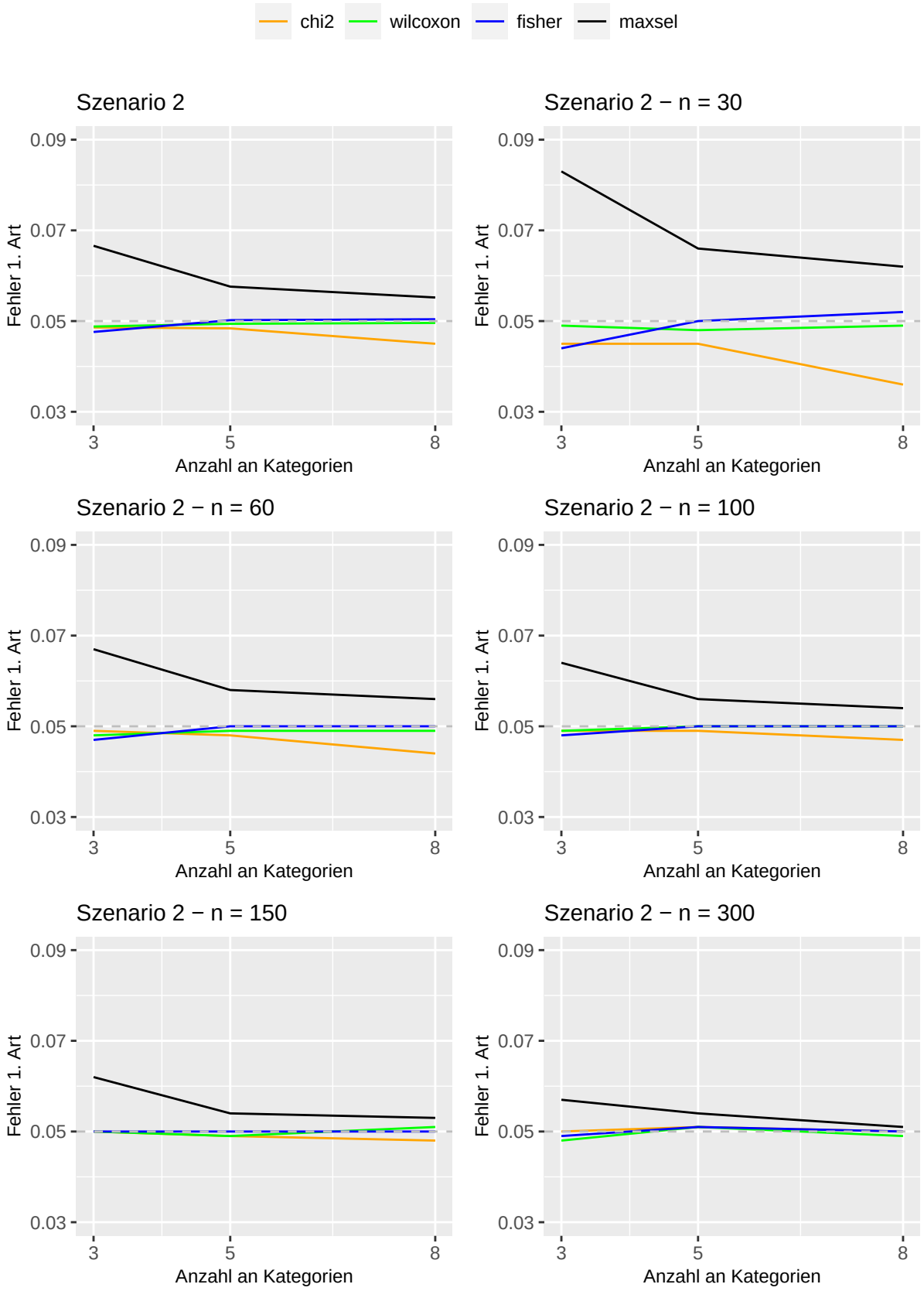


Abbildung 8: Verlauf der Simulation des Fehlers 1. Art in Szenario 2. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

Ergebnisse bezüglich der Power

Zur Testung der erreichten Power in Szenario 2 wurden wiederum verschiedene Wahrscheinlichkeiten der Kategorien in Kontroll- und Versuchsgruppe verwendet. Dabei wurde einer Versuchsgruppe mit gleichen Wahrscheinlichkeiten für alle Kategorien eine Kontrollgruppe mit ansteigenden Wahrscheinlichkeiten und einem Schwellwert gegenübergestellt (siehe Abbildung 2). Für die Power eines Tests ist die relative Anzahl an korrekterweise abgelehnten Nullhypothesen, welche keinen Unterschied zwischen den Wahrscheinlichkeiten der Kategorien in Kontroll- und Versuchsgruppe annehmen, entscheidend.

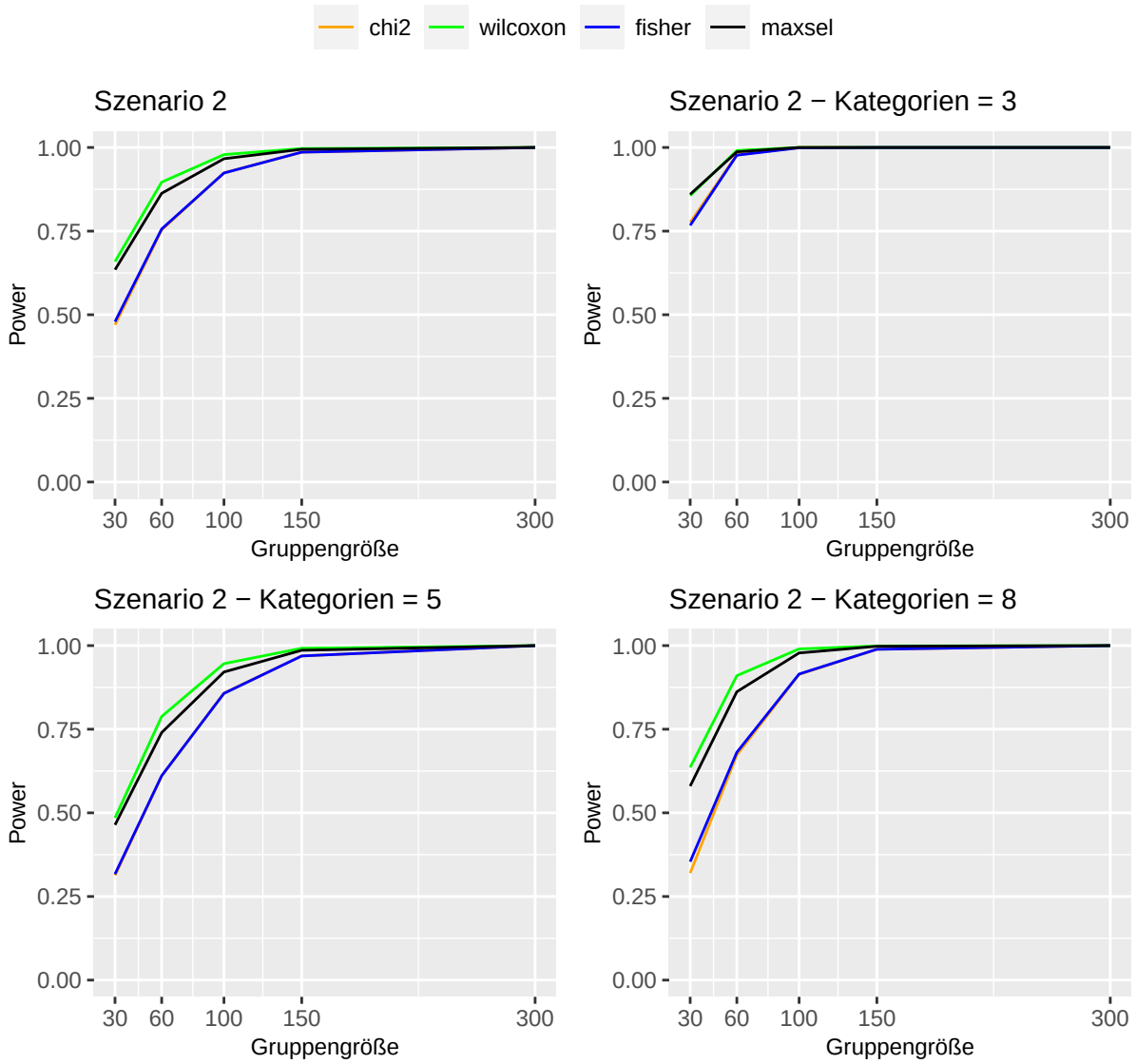


Abbildung 9: Verlauf der Simulation der Power in Szenario 2. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

Zur Analyse der Power wurden verschiedene Gruppengrößen und Anzahlen an Kategorien verwendet. In Abbildung 9 ist zu Beginn (oben links) der mittlere Verlauf der Power über verschiedene Gruppengrößen zu sehen. Hierfür wurden insgesamt 300 000 Simulationen pro Gruppengröße durchgeführt. Jeweils 100 000 Simulationen pro Gruppengröße stam-

men dabei aus den Teilsimulationen, welche in den restlichen drei Grafiken abgebildet sind. Dabei wurden die Ergebnisse nach der Anzahl der Kategorien separiert.

Die Ergebnisse erreichen dabei ähnliche Ausmaße wie bei Szenario 1. Auch hier ist zu beobachten, dass besonders der *Wilcoxon*-Test und der *Maxsel*-Test eine hohe Power vorweisen können. Für Simulationen mit lediglich drei Kategorien und mindestens 60 Beobachtungen pro Gruppe wurde für alle Tests im Mittel eine Power von annähernd 1 erreicht. Bei einer höheren Anzahl an Kategorien von fünf oder gar acht Kategorien haben alle Tests einen gewissen Verlust bei kleineren Gruppengrößen zu verzeichnen. Allerdings ist dieser Verlust an Power für den *Wilcoxon*-Test und den *Maxsel*-Test deutlich geringer als für den *exakten Fisher*-Test und den χ^2-Test.

In Abbildung 10 ist ebenfalls die erreichte Power der Tests dargestellt. Hierbei zeigen allerdings die Grafiken den Verlauf für eine unterschiedliche Anzahl an Kategorien. Während die erste Grafik eine Zusammenfassung darstellt, ist in den restlichen Grafiken eine Unterteilung in die jeweiligen Gruppengrößen erfolgt. Die Abbildung zeigt deutlich, dass sich die Power im Mittel für alle Tests erhöht, wenn eine höhere Anzahl an Beobachtungen pro Gruppe vorliegt. Interessanterweise fällt zudem auf, dass bei allen Tests ein schlechteres Abschneiden bei fünf Kategorien erzielt wird als bei drei oder acht. Der Grund für diesen Unterschied könnte allerdings auch an der Wahl der Wahrscheinlichkeiten für die Kategorien liegen. Trotz einer sehr großen Ähnlichkeit zwischen Szenario 2a, 2b und 2c (siehe Abbildung 2) können hier auch leichte Unterschiede eine Auswirkung auf die geschätzte Power der Tests haben. Ebenso wie zuvor für die Anzahl der Kategorien lässt sich hier erkennen, dass der *Wilcoxon*-Test und *Maxsel*-Test besonders für kleine Gruppengrößen besser abschneiden als der *exakte Fisher*-Test und der χ^2-Test.

Zwischenfazit: Szenario 2

Die Simulationen lassen erkennen, dass besonders der *Wilcoxon*-Test für das Szenario 2 geeignet ist. Sowohl bei der Beschränkung des Fehlers 1. Art als auch bei der erreichten Power erzielt der *Wilcoxon*-Test die besten Ergebnisse. Bezüglich der Beschränkung des Fehlers 1. Art ist, im Vergleich zu Szenario 1, bei fast allen Tests eine leicht größere Schwankung erkennbar. Während der *exakte Fisher*-Test und der χ^2-Test unter gewissen Umständen für ein zu strenges Beschränken des Fehlers 1. Art sorgen, so übersteigt der *Maxsel*-Test im Mittel die Beschränkung deutlich. Bei der Power wurden dagegen im Mittel bessere Ergebnisse erzielt als in Szenario 1, welches ohne deutlichen Schwellenwert gebildet wurde. Hier schneiden der *Wilcoxon*-Test und der *Maxsel*-Test am besten ab. Besonders für kleine Gruppengrößen und einer hohen Anzahl an Kategorien zeigen der *exakte Fisher*-Test und der χ^2-Test schlechtere Ergebnisse bezüglich der Power.

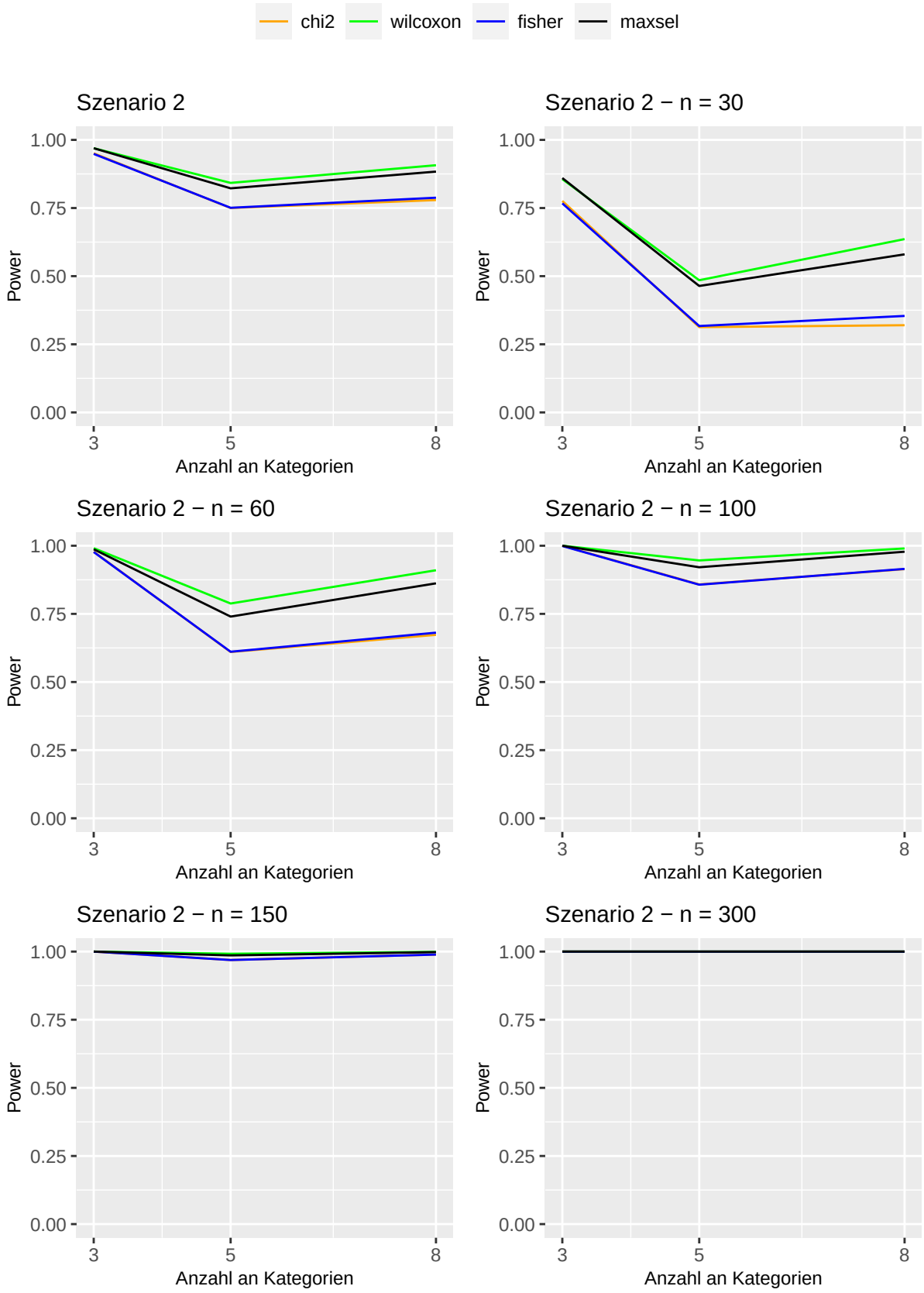


Abbildung 10: Verlauf der Simulation der Power in Szenario 2. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

5.2.3 Szenario 3

Szenario 3 behandelt den Fall, dass die mittleren Kategorien der Kontrollgruppe eine erhöhte Wahrscheinlichkeit aufweisen. Der Simulationsaufbau für die Überprüfung zur Beschränkung des Fehlers 1. Art basiert dabei wieder darauf, dass die gleiche Verteilung in Kontroll- und Versuchsgruppe vorliegen. In Abbildung 2 entspricht diese für beide Gruppen der dargestellten Kontrollgruppe in Szenario 3a, 3b und 3c. Für die Simulation der Power wird zusätzlich eine Versuchsgruppe mit ansteigender Wahrscheinlichkeit für höhere Kategorien angenommen.

Ergebnisse bezüglich des Fehlers 1. Art

Um die Höhe des Fehlers 1. Art in Szenario 3 zu testen, wurden sowohl für die Kontroll-, als auch für die Versuchsgruppe eine gleiche Verteilung der Beobachtungen angenommen (siehe Tabelle 2). In Abbildung 11 sind die gemittelten Ergebnisse von insgesamt 300 000 Simulationsdurchläufen pro Gruppengröße abgebildet. Während in der ersten Grafik die gemittelten Werte aller Simulationsdurchläufe für Szenario 3 dargestellt sind, zeigen die restlichen drei Grafiken eine Untergliederung in die drei verschiedenen Ausprägungen der Variable *Anzahl an Kategorien*. In jene drei Teilgrafiken fließen somit die Informationen von jeweils 100 000 Simulationsdurchgängen pro Gruppengröße mit ein.
Die Ergebnisse aus Abbildung 11 decken sich größtenteils mit den Erkenntnissen aus den Szenarios 1 und 2. Besonders für kleine Gruppengrößen und einer geringen Anzahl an Kategorien überschreitet der *Maxsel*-Test die 0.05-Grenze deutlich. Mit steigender Gruppengröße und einer höheren Anzahl an Kategorien tendiert der Fehler 1. Art zwar in Richtung 0.05, erreicht diesen aber auch bei einer Gruppengröße von $n_{KG} = n_{VG} = 300$ und acht Kategorien im Mittel nicht.
Die anderen drei Tests, von geringen Schwankungen abgesehen, zeigen dagegen im Mittel keine Überschreitung der vorgegebenen Höhe des Fehlers 1. Art. Insbesondere der χ^2-Test zeigt aber für kleine Stichprobengrößen und einer hohen Anzahl an Kategorien eine konservative Testentscheidung. So liegt der mittlere Fehler 1. Art bei acht Kategorien und 30 Beobachtungen pro Gruppe bei etwa 0.033.
Auch der *exakte Fisher*-Test zeigt bei einer kleinen Anzahl an Kategorien geringe Schwankungen unterhalb der 0.05 Marke. Diese sind aber, besonders im Vergleich zum χ^2-Test, deutlich unbedeutsamer.
Für den *Wilcoxon*-Test sind dagegen kaum Schwankungen und Einflüsse bezüglich der unterschiedlichen Anzahl der Kategorien zu erkennen. So hält der Test im Mittel die Beschränkung von 5% der Testentscheidungen, welche die Nullhypothese fälschlicherweise ablehnen, sehr gut ein.

Abbildung 12 zeigt ebenfalls den Verlauf des Fehlers 1. Art in Szenario 3. Hierbei sind die Verläufe über die verschiedenen Anzahlen an Kategorien dargestellt und in die getesteten Gruppengrößen untergliedert. Wie in Abbildung 11 ist auch hier deutlich ersichtlich, dass der *Maxsel*-Test größere Schwierigkeiten mit der Einhaltung der vorgegebenen Grenze von 5% für den Fehler 1. Art hat. Mit einer höheren Anzahl an Beobachtungen pro Gruppe tendiert der Fehler 1. Art allerdings in Richtung 0.05-Marke. Bei einer Gruppengröße von $n_{KG} = n_{VG} = 300$ ist der Fehler 1. Art damit nur noch leicht über der gegebenen Grenze.

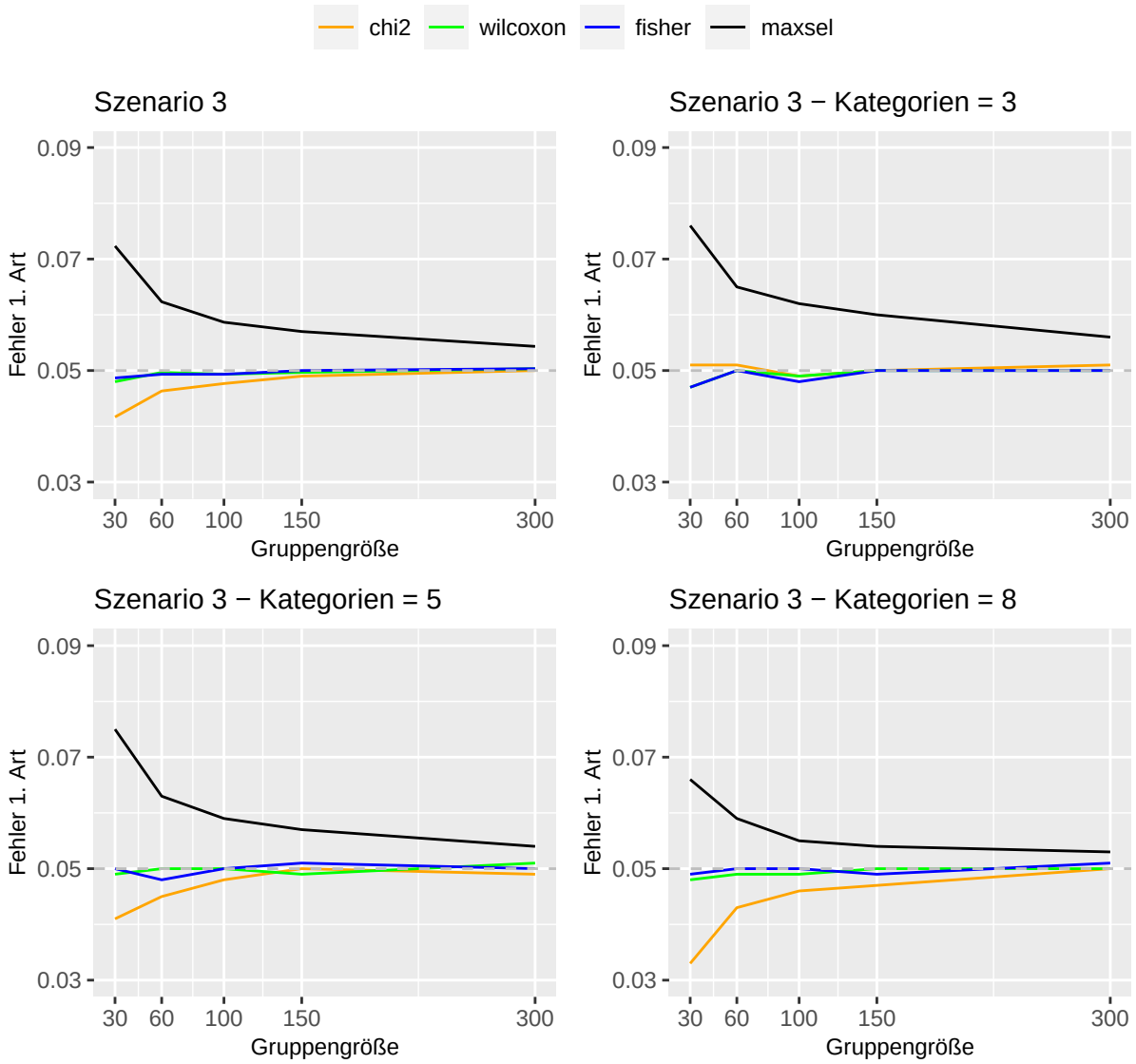


Abbildung 11: Verlauf der Simulation des Fehlers 1. Art in Szenario 3. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

Auch der χ^2-Test erzielt die besten Ergebnisse bei einer Gruppengröße von $n_{KG} = n_{VG} = 300$. Für weniger Beobachtungen unterschreitet der Test die 0.05-Grenze mitunter eindeutig.
Für den *exakten Fisher*-Test und den *Wilcoxon*-Test lassen sich in Abbildung 12 keine Auffälligkeiten im Bezug auf die Gruppengröße erkennen.

Insgesamt ist wieder zu erkennen, dass der *Maxsel*-Test, besonders für wenige Beobachtungen und einer kleinen Anzahl an Kategorien, einen erhöhten Fehler 1. Art vorweist. Dieser verringert sich mit einer erhöhten Gruppengröße, aber auch mit mehreren Kategorien. Während der χ^2-Test die gewünschte 5%-Marke auch nicht immer erreicht, zeigen der *exakte Fisher*-Test und der *Wilcoxon-Test* relativ konstant gute Ergebnisse bezüglich des Fehlers 1. Art.

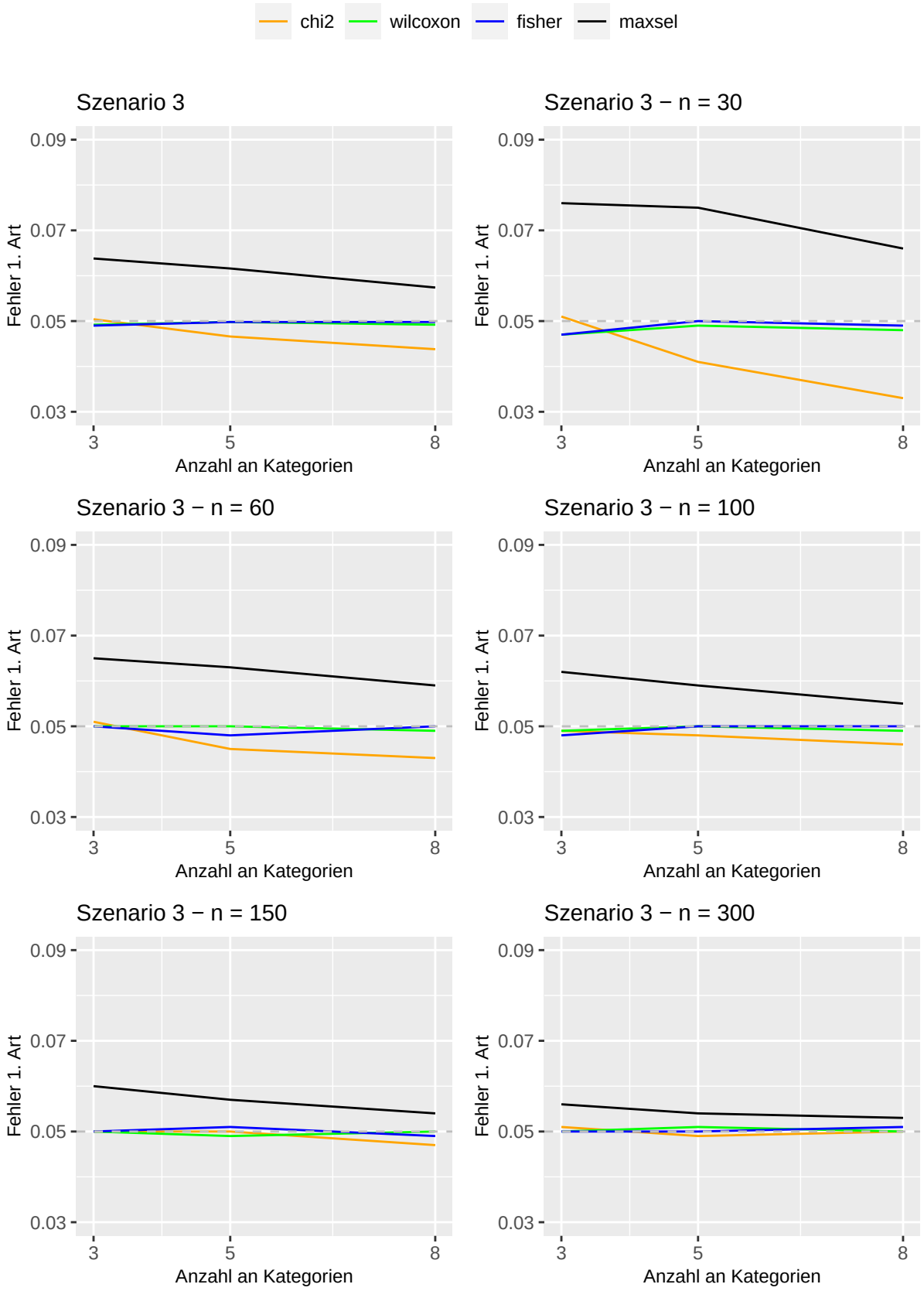


Abbildung 12: Verlauf der Simulation des Fehlers 1. Art in Szenario 3. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

Ergebnisse bezüglich der Power

Zur Simulation der Power in Szenario 3 wurden wieder unterschiedliche Wahrscheinlichkeiten für die Kategorien in Kontroll- und Versuchsgruppe angenommen. Die relative Anzahl aller richtigerweise abgelehnter Nullhypothesen ergibt somit die gesuchte Trennschärfe. Während in der Kontrollgruppe der Modus der Wahrscheinlichkeiten in einer mittleren Kategorie zu finden ist, steigt in der Versuchsgruppe die Wahrscheinlichkeit für eine Kategorie monoton mit ansteigender Kategorie (siehe Abbildung 2).

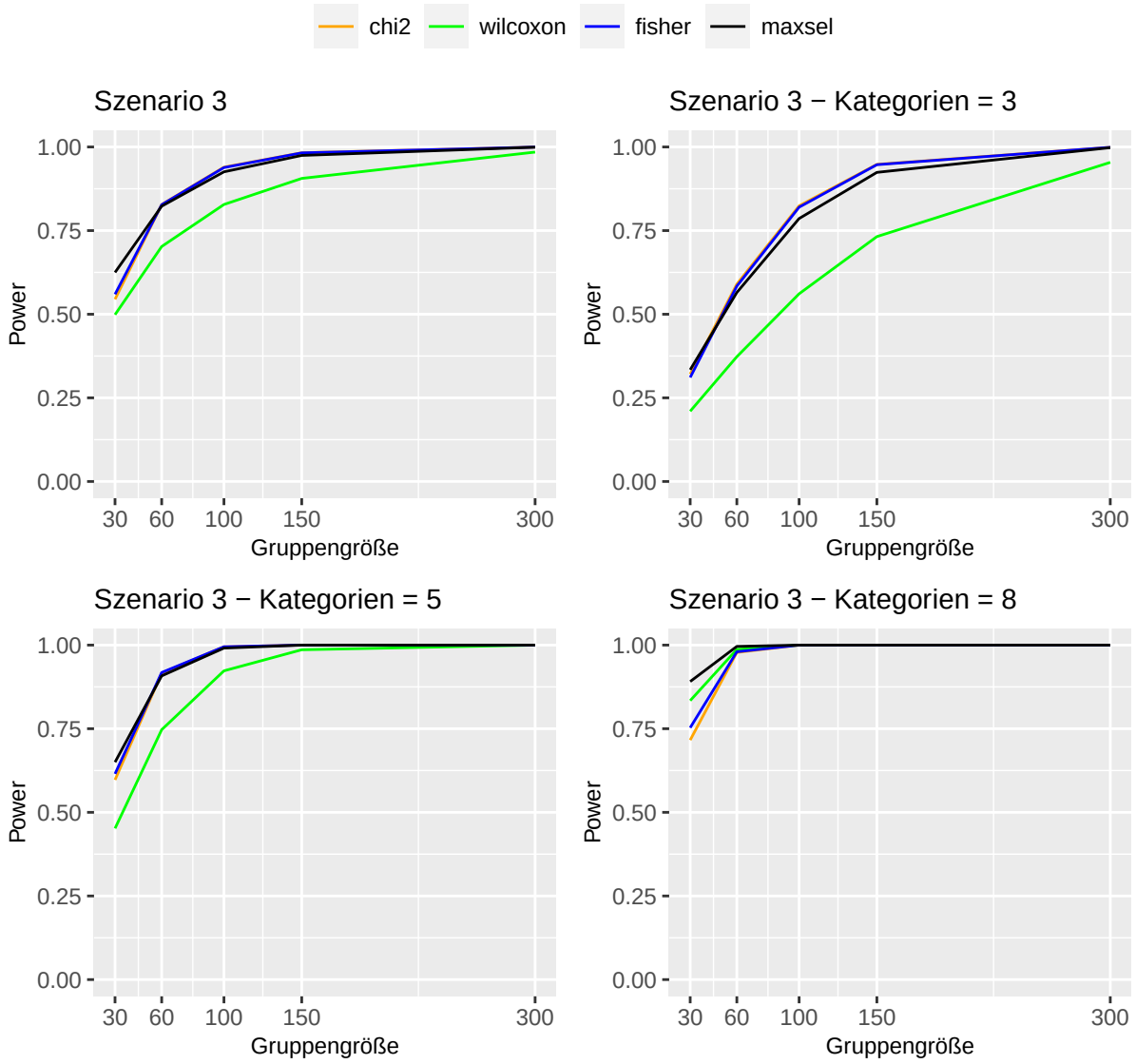


Abbildung 13: Verlauf der Simulation der Power in Szenario 3. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

In Abbildung 13 ist die ermittelte Power im Verlauf über die Gruppengröße dargestellt. Zusammen mit den drei Teilauswertungen, abhängig von der *Anzahl an Kategorien*, sind eben jene Teilauswertungen in der ersten Grafik zusammengefasst. Dabei wurde pro Kombination von Gruppengröße und Anzahl an Kategorien eine Simulation mit 100 000 Iterationen durchgeführt. Die Ergebnisse entsprechen den ermittelten Mittelwerten. Die Ergebnisse

zeigen, dass mit einer höheren Anzahl an Kategorien auch die Power der verwendeten Tests steigt. Fällt die Anzahl der Kategorien von acht auf fünf beziehungsweise auf drei, so fällt besonders die Power des *Wilcoxon*-Tests ab. Auch beim χ^2-Test, dem *exakten Fisher*-Test und dem *Maxsel*-Test lassen sich Einbußen bei der Power erkennen. Dabei bleiben diese drei Tests allerdings immer auf dem etwa gleichen Niveau.

Abbildung 14 zeigt den Verlauf über die Anzahl der Kategorien. Dabei sind die Grafiken jeweils unterteilt in die verschiedenen Gruppengrößen und die dazugehörige Zusammenfassung. Gut erkennbar ist, dass für eine größere Stichprobe auch eine höhere Power erzielt werden kann. Während bei einer Gruppengröße von $n_{KG} = n_{VG} = 300$ bei fast allen Tests eine Power von 1 erzielt wurde, ist das Ergebnis für eine Gruppengröße von $n_{KG} = n_{VG} = 30$ deutlich schlechter. So erzielt hier, bei drei verschiedenen Kategorien, kein Test eine höhere Power als 0.334. Allerdings ist bei einer Gruppengröße von $n_{KG} = n_{VG} = 30$ zu sehen, dass der *Maxsel*-Test unter diesen Umständen am Besten abschneidet.

Zwischenfazit: Szenario 3

Auch in Szenario 3 hat der *Maxsel*-Test große Schwierigkeiten mit der Beschränkung des Fehlers 1. Art. Dieser wird im Mittel in keiner Kombination von Gruppengröße und Anzahl an Kategorien eingehalten. Während der χ^2-Test bei wenigen Beobachtungen pro Kategorie die 5%-Grenze deutlich unterschreitet, sind für den *exakten Fisher*-Test und den *Wilcoxon*-Test kaum Schwankungen zu erkennen.
Bezüglich der erreichten Power erreicht der *Maxsel*-Test im Mittel die besten Ergebnisse. Auch der *exakte Fisher*-Test und der χ^2-Test schneiden relativ ähnlich gut ab. Besonders bei einer geringen Anzahl an Kategorien fällt der *Wilcoxon*-Test dagegen deutlich ab.
Die konstantesten Ergebnisse in Szenario 3 liefert somit der *exakte Fisher*-Test. Dieser schneidet sowohl beim Vergleich des Fehlers 1. Art als auch bei der erreichten Power als einer der Besten ab. Obwohl der *Maxsel*-Test die im Schnitt höchste Power aller Tests erreicht, ist auch in diesem Szenario der deutlich zu hohe Fehler 1. Art zu beobachten. Während der χ^2-Test Probleme mit dem Fehler 1. Art hatte, ist die erreichte Power des *Wilcoxon*-Test deutlich hinter den anderen Tests einzustufen.

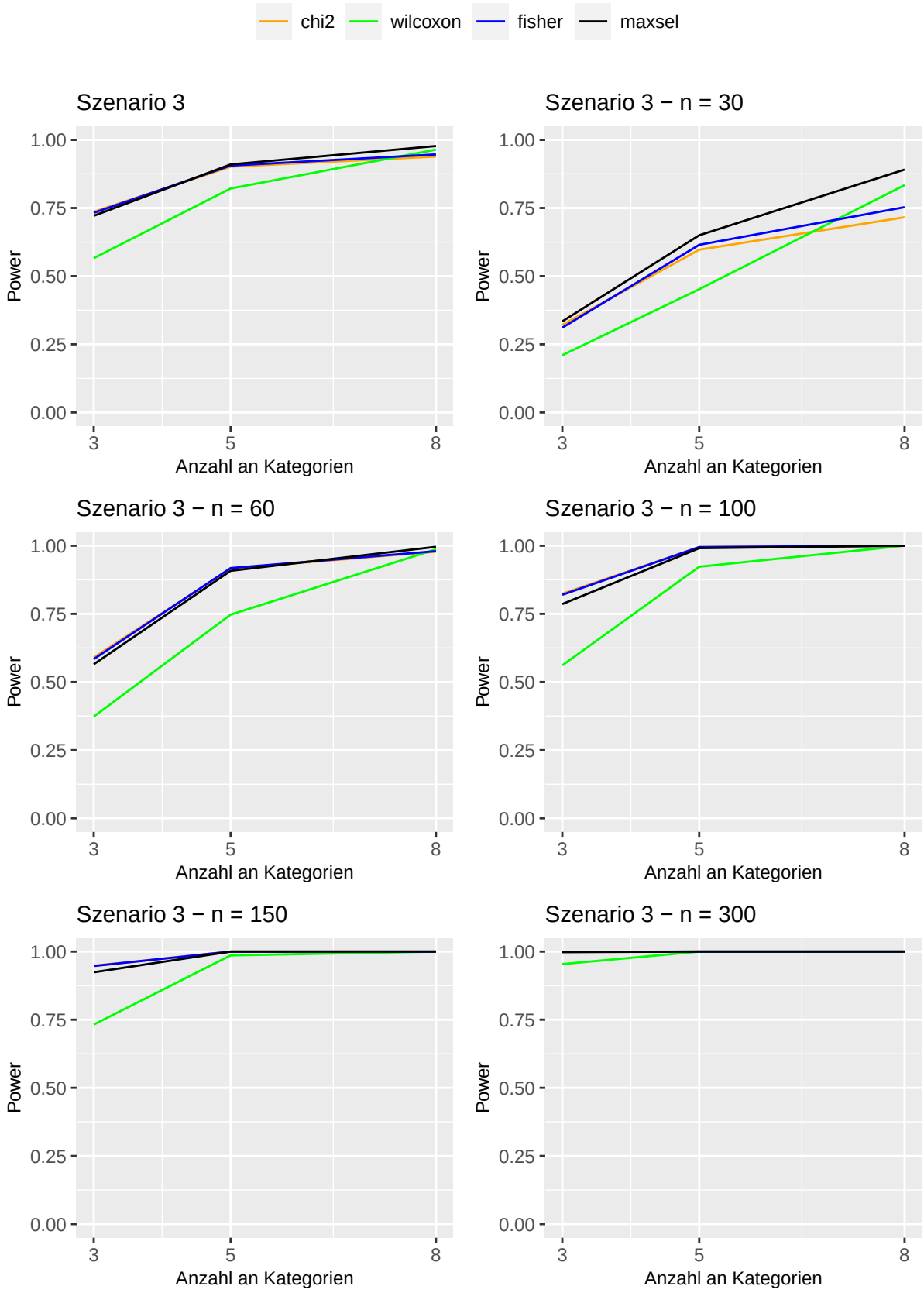


Abbildung 14: Verlauf der Simulation der Power in Szenario 3. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

5.2.4 Szenario 4

In Szenario 4 werden die gleichen Wahrscheinlichkeiten der Kategorien für die Kontrollgruppe verwendet wie in Szenario 3. Aufgrund dessen wird an dieser Stelle lediglich die erreichte Power der Tests untersucht. Der Fehler 1. Art für die Kontrollgruppe wäre hier identisch mit dem aus Szenario 3 (siehe Abbildung 2). Ziel dieser Untersuchung in Szenario 4 ist es, die Power bei relativ ähnlichen, aber dennoch unterschiedlichen Verteilungen in Kontroll- und Versuchsgruppe zu untersuchen.

Ergebnisse bezüglich der Power

Um die Tests auf die erreichte Power in Szenario 4 zu untersuchen, wurden pro Kombination von *Gruppengröße* und *Anzahl an Kategorien* 100 000 Simulationsdurchläufe getätigt.

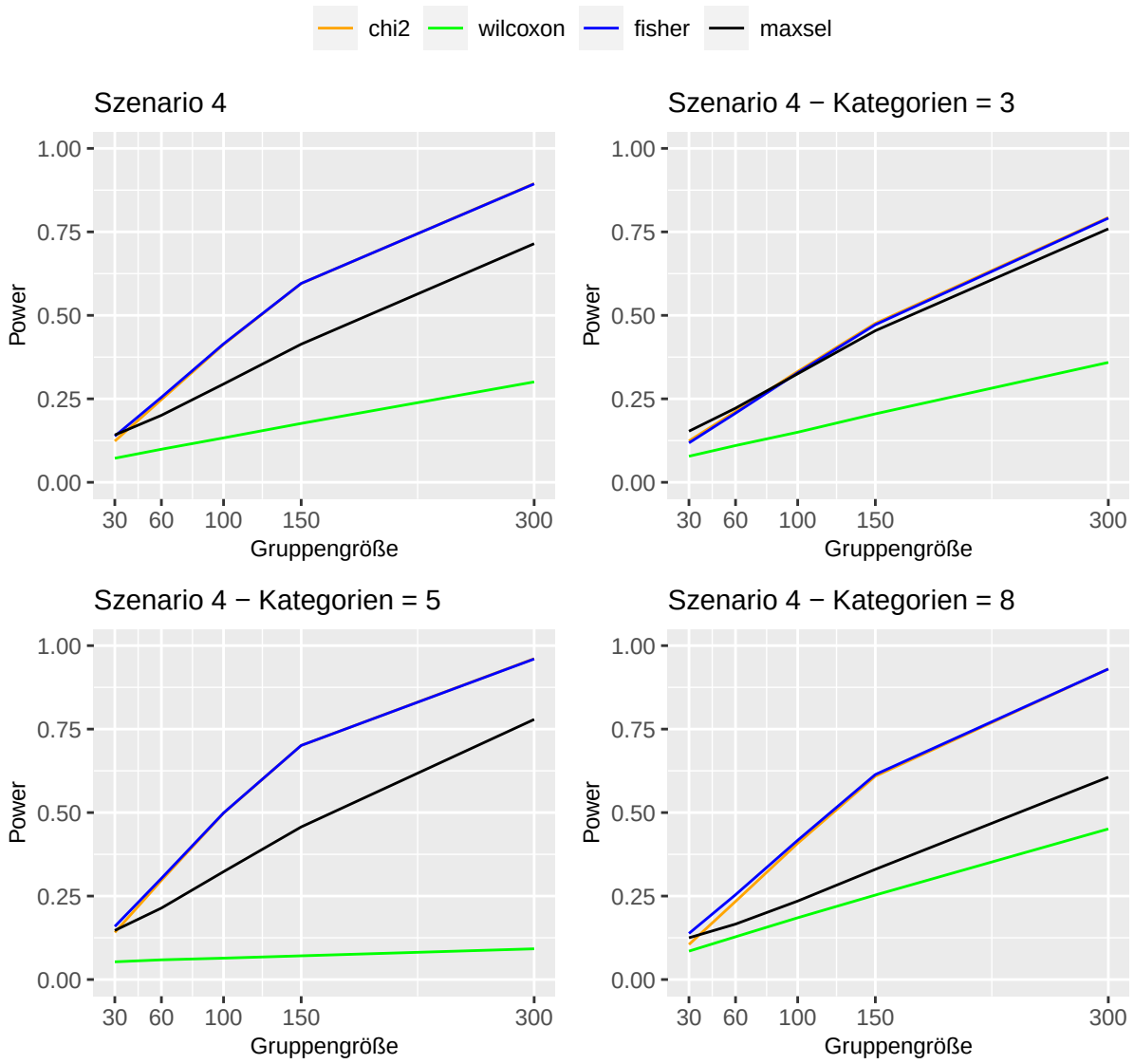


Abbildung 15: Verlauf der Simulation der Power in Szenario 4. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken gemäß der Anzahl an Kategorien unterteilt sind.

In Abbildung 15 ist die erreichte Power in Form von Mittelwerten dargestellt. Die erste Grafik zeigt die Zusammenfassung der folgenden drei Grafiken, in denen jeweils nach der *Anzahl an Kategorien* separiert wurde. Dabei ist ersichtlich, dass die Power bei allen geprüften Tests meist deutlich niedriger ist, als bei den Szenarios 1, 2 und 3. Dieses erwartbare Ergebnis ist vermutlich größtenteils darin begründet, dass die Verteilungen von Kontroll- und Versuchsgruppe sehr ähnlich sind. Dadurch ist es für die Tests deutlich schwieriger einen signifikanten Unterschied zu erkennen.
Der *exakte Fisher*-Test und der χ^2-Test erreichen über alle möglichen Anzahlen an Kategorien die besten Ergebnisse. Hierbei ist kein eindeutiger Trend erkennbar, welcher mit der Anzahl der Kategorien zusammenhängen könnte.
Der *Maxsel*-Test erreicht meist nicht ganz die Power dieser zwei Tests. Besonders im Szenario 4 mit acht Kategorien ist ein deutlicher Abfall der Power zu erkennen.
Am schlechtesten schneidet der *Wilcoxon*-Test ab, welcher im Mittel immer an letzter Stelle liegt. Insbesondere Szenario 4 mit fünf Kategorien zeigt das unterschiedliche Verhalten der Tests in diesem Szenario. Während hier sowohl der *exakte Fisher*-Test und der χ^2-Test ihre im Mittel besten Ergebnisse erzielen, erreicht der *Wilcoxon*-Test seine schlechtesten und damit eine nur sehr geringe Power.

Abbildung 16 zeigt ebenfalls die erreichte Power in Szenario 4. Hierbei sind die Ergebnisse, neben einer Zusammenfassung, in die verschiedenen Gruppengrößen aufgeteilt. Hierbei ist, wie auch bereits in Abbildung 15 zu sehen, ein klarer Trend zu erkennen. So ist bei allen Tests eine zum Teil deutlich höhere Power bei größeren Gruppengrößen zu beobachten. Besonders der *exakte Fisher*-Test und der χ^2-Test können bei höheren Gruppengrößen die erreichte Power steigern.

Zwischenfazit: Szenario 4

Die Power in Szenario 4 ist, wie zu erwarten, deutlich niedriger wie die Ergebnisse aus Szenario 1, 2 und 3. Allerdings sind eindeutige Unterschiede in der Performance der Tests in Szenario 4 zu erkennen. Besonders der *exakte Fisher*-Test und der χ^2-Test kommen mit den sehr ähnlichen Verteilungen von Kontroll- und Versuchsgruppe vergleichsweise gut zurecht. Während der *Maxsel*-Test bei lediglich drei Kategorien relativ ähnliche Ergebnisse erzielt, fällt dieser bei einer größeren Anzahl an Kategorien weiter ab. Am wenigsten kommt der *Wilcoxon*-Test mit Szenario 4 zurecht, welcher in allen Situationen die niedrigste mittlere Power aufweist.

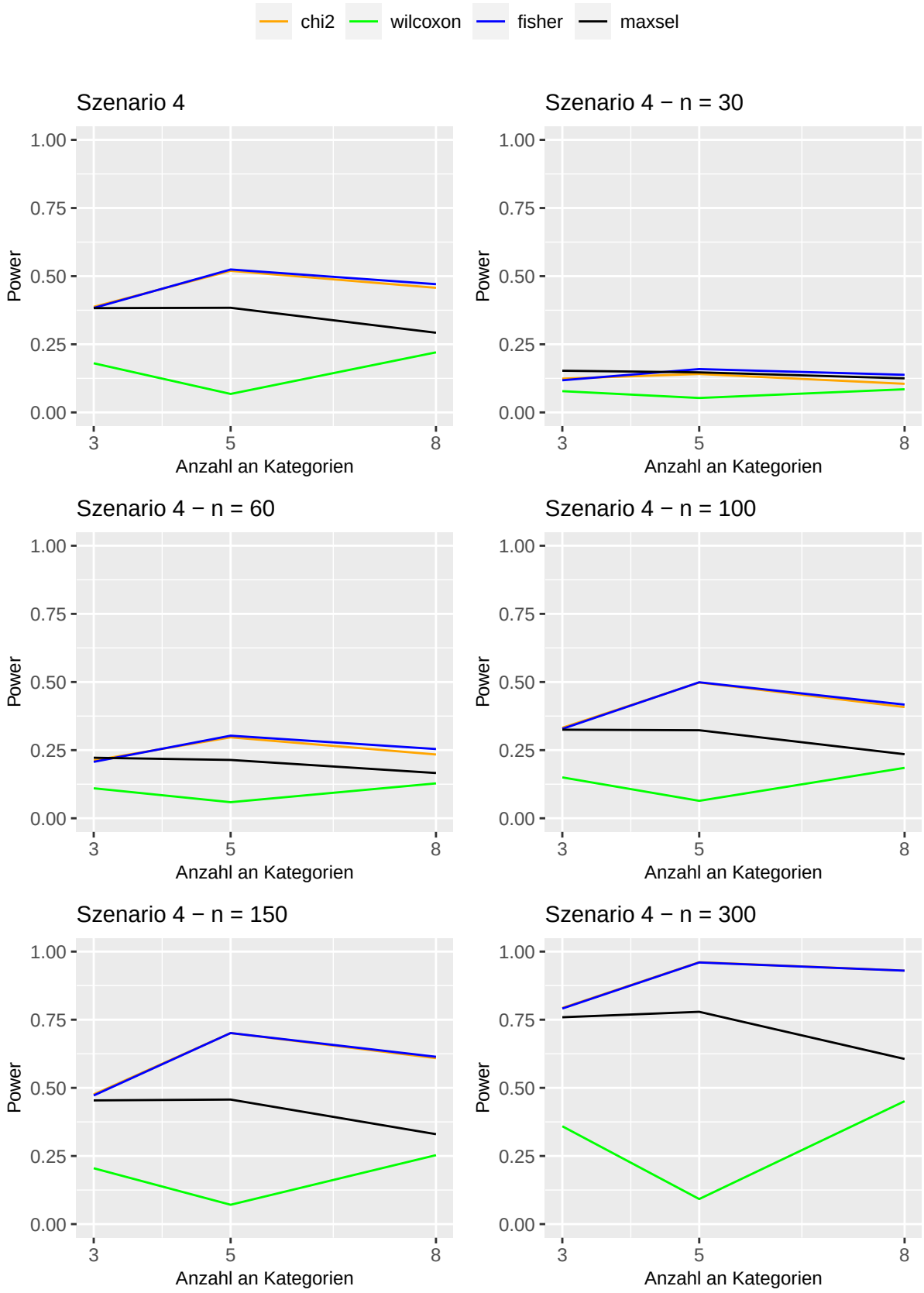


Abbildung 16: Verlauf der Simulation der Power in Szenario 4. Grafik 1 zeigt den mittleren Verlauf der Ergebnisse, während die restlichen Grafiken bezüglich der Gruppengröße unterteilt sind.

5.3 Zusammenfassung

In vier verschiedenen Szenarios wurden der χ^2-Test, der *exakte Fisher*-Test, der *Wilcoxon*-Test und der *Maxsel*-Test auf ihre Power und die Höhe des Fehlers 1. Art getestet. Dabei schnitten die Tests auch innerhalb dieser vier Szenarios sehr unterschiedlich ab. Dabei ist kein Test zu beobachten, welcher in allen Szenarios ausschließlich gut abgeschnitten hat.

Der χ^2-Test erreichte in Szenario 1 einen Fehler 1. Art, der relativ konstant an der 5%-Marke lag. Für Szenario 2 und Szenario 3 zeigte er allerdings ein eher konservatives Verhalten. Vor allem bei einer hohen Anzahl an Kategorien und einer kleinen Stichprobengröße lag der Fehler 1. Art zum Teil deutlich unter der gewünschten Grenze von 5%. Im Blick auf die erreichte Power erzielte der χ^2-Test besonders in Szenario 3, trotz konservativem Verhalten bezüglich des Fehlers 1. Art, und den Umständen entsprechend auch in Szenario 4 ein gutes Ergebnis. So ist die erreichte Power in Szenario 4 zwar insgesamt vergleichsweise relativ niedrig, allerdings erzielte hier der χ^2-Test gemeinsam mit dem *exakten Fisher*-Test die besten Ergebnisse. In Szenario 1 und Szenario 2 erzielte der χ^2-Test zwar eine relativ hohe Power, allerdings erreichten insbesondere der *Wilcoxon*-Test und der *Maxsel*-Test hier ein besseres Ergebnis.

Der *Wilcoxon*-Test erreichte die besten Ergebnisse bezüglich der Höhe des Fehlers 1. Art. Abgesehen von minimalen Schwankungen liegen die Testergebnisse aller Szenarios relativ eng bei der 5%-Grenze. Bei den Simulationen bezüglich der Power des *Wilcoxon*-Test lassen sich allerdings große Unterschiede ausmachen. In Szenario 1 und insbesondere in Szenario 2 erreichte der *Wilcoxon*-Test eine gute bis sehr gute Power und schnitt im Vergleich zu den anderen Tests am besten ab. Im Gegensatz dazu erreichte der *Wilcoxon*-Test in Szenario 3 und Szenario 4 im Mittel die niedrigste Power. Besonders bei relativ ähnlichen Verteilungen in Kontroll- und Versuchsgruppe, wie sie in Szenario 4 vorliegen, erreichte der Test eine deutlich niedrigere Power als alle anderen Tests.

Der *exakte Fisher*-Test erzielte relativ ähnliche Resultate wie der χ^2-Test. Aufgrund der relativ ähnlichen Berechnungsweise der beiden Tests sind in allen Szenarios somit relativ identische Ergebnisse zu beobachten. Lediglich bei der Höhe des Fehlers 1. Art in Szenario 2 und Szenario 3 schöpfte der *exakte Fisher*-Test die vorgegebene Grenze von 5% besser aus. Auch in Szenario 1, abgesehen von kleinen Schwankungen bei kleinen Stichprobengrößen und einer geringen Anzahl an Kategorien, erreichte der *exakte Fisher*-Test einen Fehler 1. Art, welcher relativ nah an der erwünschten Grenze war. Betrachtet man die Ergebnisse der erreichten Power, so sind die Ergebnisse im Mittel mehr oder weniger identischen mit denen des χ^2-Tests. So erzielt der *exakte Fisher*-Test in Szenario 1 und Szenario 2 zwar eine niedrigere Power als der *Wilcoxon*-Test und der *Maxsel*-Test, welche aber besonders für hohe Stichprobengrößen und einer kleinen Anzahl an Kategorien vergleichbar erscheint. In Szenario 3 und Szenario 4 dagegen erzielt der *exakte Fisher*-Test, gemeinsam mit dem χ^2-Test, die höchste Power aller Tests. Auch hier ist aber in Szenario 4 eine insgesamt niedrigere mittlere Power zu beobachten.

Im besonderen Fokus dieser Arbeit steht das Abschneiden des *Maxsel*-Tests. Dabei kann festgehalten werden, dass der *Maxsel*-Test enorme Probleme bei der Einhaltung des Fehlers 1. Art vorweist. So zeigt sich in allen Szenarios ein relativ ähnliches Bild. Besonders bei relativ kleinen Stichprobengrößen und einer geringen Anzahl an Kategorien ist der

Fehler 1. Art deutlich erhöht. Lediglich in Szenario 1 und Szenario 2 befinden sich die mittleren Simulationsergebnisse, bei einer Gruppengröße von $n_{KG} = n_{VG} = 300$ mit acht verschiedenen Kategorien, relativ knapp über der 5%-Grenze. Dies spiegelt sich auch in den entsprechenden Verteilungen der P-Werte, welche im Anhang zu finden sind (siehe Abbildungen 21, 22 und 23). Anstatt eine Gleichverteilung abzubilden handelt es sich hier meist um eine rechtsschiefe Verteilung mit einem Trend zu eher niedrigeren P-Werten.
Betrachtet man das Abschneiden des *Maxsel*-Tests im Bezug auf die erreichte Power, lassen sich hier deutlich bessere Ergebnisse beobachten. So erreichte er, zum Teil gemeinsam mit anderen Tests, sowohl in Szenario 1, Szenario 2 und auch Szenario 3 die im Mittel besten Ergebnisse. Allenfalls in Szenario 4 fällt die mittlere erreichte Power des *Maxsel*-Tests etwas hinter der des *exakten Fisher*-Tests und des χ^2-Tests zurück. Die Verteilungen der P-Werte in der Simulation bezüglich der erreichten Power sind sowohl im Anhang (Abbildungen 24, 25 und 26), als auch im digitalen Anhang zu finden.

Zusammenfassend lässt sich festhalten, dass der *Maxsel*-Test im Schnitt die vermutlich besten mittleren Ergebnisse aller Tests im Bezug auf die erreichte Power erzielte. Dies geht allerdings einher mit einem im Mittel deutlich erhöhten Fehler 1. Art. Hierbei ist es für den Anwender wichtig, in welchen Situationen dieser Fehler besonders erhöht ist. Ob der *Maxsel*-Test auch bei einer Einhaltung der Höhe des Fehlers 1. Art eine gute Power erreichen würde, ist im Rahmen dieser Simulationsstudien nicht zu beantworten.
Bei der Betrachtung der anderen drei Tests ist in Szenario 1 und Szenario 2 die Anwendung des *Wilcoxon*-Tests und in Szenario 3 und Szenario 4 die des *exakte Fisher*-Test und des χ^2-Tests zu empfehlen.

6 Anwendungsbeispiel

In diesem Kapitel soll die Methode der maximal selektieren χ^2-Statistiken auf Daten aus veröffentlichten Studien angewandt werden. Es handelt sich dabei um insgesamt vier verschiedene Studien, welche sich um Patienten mit Lymphomen drehen. Dabei handelt es sich um Tumore, die im Immunsystem zu finden sind (Shankland et al., 2012). Während sich die Studien von Hermine et al. (2016) und Kluin-Nelemans et al. (2012) hauptsächlich mit dem Mantelzell-Lymphom (MCL) auseinandersetzen, stammen die Daten aus den Studien von Nickenig et al. (2006) und Hiddemann et al. (2005) von Patienten, welche ein follikuläres Lymphom vorweisen. Beide Lymphome gehören dabei zu den sogenannte *Non-Hodgkin-Lymphomen* (Shankland et al., 2012; Witzig, 2005). Die im Folgenden verwendeten Daten entsprechen dabei nicht immer den publizierten Daten der Studienergebnisse. Dies ist darin begründet, dass die hier verwendeten Daten mit Hilfe eines anderen Stichtages gewählt wurden.

Während es sich bei den Studien von Hermine et al., Kluin-Nelemans et al. und Hiddemann et al. um sogenannte *Intention-to-treat*(ITT) Analysen handelt, ist in der Studie von Nickenig et al. eine *Per-Protokoll* (PP)-Analyse durchgeführt worden.
Bei einer ITT-Analyse werden die Daten aller Personen in die Analyse miteinbezogen. Dabei werden sie der Gruppe zugeordnet, in die sie zu Beginn per Zufall eingeteilt wurden. Eine PP-Analyse wertet dagegen nur Patienten aus, die ihre zufällig zugewiesene Behandlung vollständig erhalten haben. Patienten, die im Laufe der Studienzeit ausscheiden oder die Behandlungsgruppe wechseln, werden somit nicht berücksichtigt (Ranganathan et al., 2016).

Bei allen vier Studien handelt es sich um randomisierte klinische Studien, in denen eine neue Behandlungsmethode mit der herkömmlichen Behandlung verglichen werden soll. Zur Messung des Behandlungseinflusses werden insgesamt sechs Kategorien verwendet, welche einer ordinalen Struktur folgen. Bei den Kategorien handelt es sich in entsprechender Reihenfolge um:

- CR = complete remission (komplette Remission)
- CRu = complete remission, unconfirmed (unbestätigte, komplette Remission)
- PR = partial remission (partielle Remission)
- SD = stable disease (stabile Erkrankung)
- PD = progressive disease (Progess)
- ED = early death (Tod während der Induktionstherapie)

In Abbildung 17 sind die Ergebnisse aus den Studien von Hermine et al. (2016) und Kluin-Nelemans et al. (2012) dargestellt. Beide Studien untersuchten den Einfluss einer neuen Behandlungsmethode bei Patienten mit dem Mantelzell-Lymphom. Die Daten der Studie von Hermine et al. (2016) umfassen insgesamt 442 Patienten, wovon 223 auf die Kontrollgruppe und 219 auf die Versuchsgruppe entfallen. Dabei wurden lediglich Patienten eingebunden, welche höchstens 65 Jahre alt waren (*MCL-Younger*).

Die Daten der Studie von Kluin-Nelemans et al. (2012) umfassen hier insgesamt 505 Patienten. 247 Patienten wurden der Kontrollgruppe zugeteilt und 258 der Versuchsgruppe. Im Gegensatz zur *MCL-Younger*-Studie wurden hier nur Patienten berufen, welche ein Mindestalter von 60 Jahren aufweisen (*MCL-Elderly*).

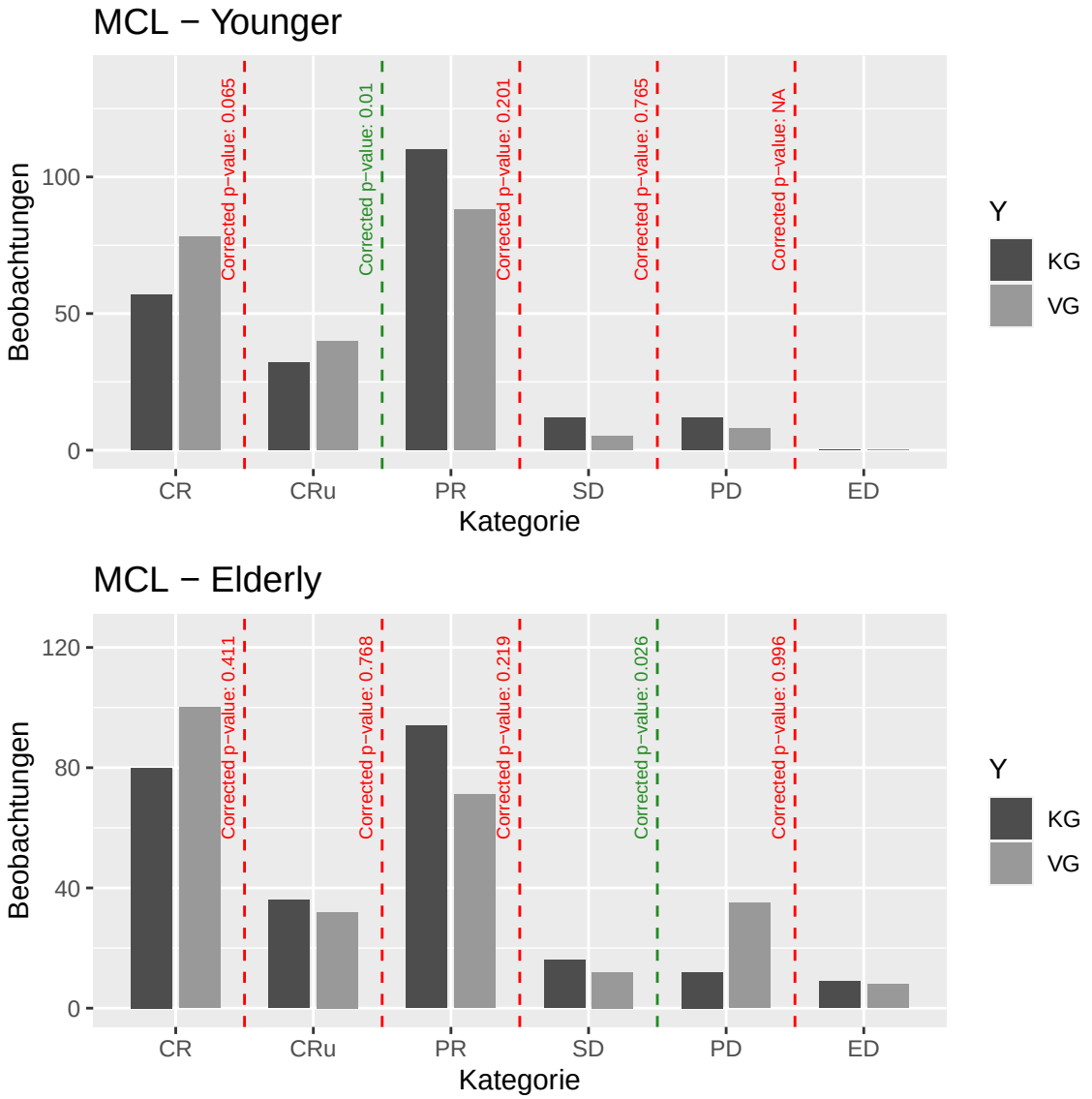


Abbildung 17: Darstellung der *MCL-Younger* und der *MCL-Elderly*-Studie mit Hilfe der maxsel.plot-Funktion. Die verwendeten Daten können leichte Abweichungen zu den publizierten Ergebnissen aufweisen, was in der Wahl des Stichtages begründet ist. Beide Studien untersuchten den Einfluss neuer Behandlungsmethoden bei Patienten mit dem Mantelzell-Lymphom.

Beide Studien zeigen signifikante Unterschiede zwischen Kontrollgruppe und Versuchsgruppe bei einem verwendeten Signifikanzniveau von $\alpha = 0.05$. Während in der *MCL-Younger*-Studie eine optimale Trennung zwischen den Kategorien *CRu* und *PR* kalkuliert wurde, liegt sie bei der *MCL-Elderly*-Studie zwischen den Kategorien *SD* und *PD*.
In Abbildung 18 sind die summierten Balken dargestellt, welche sich durch den jeweiligen optimalen Split ergeben. Dabei lassen sich unterschiedliche Eigenschaften innerhalb der zwei Studien erkennen. In der *MCL-Younger*-Studie (links) hat die neue Behandlungs-

methode (VG) deutlich mehr Beobachtungen, bei denen eine (unbestätigte) komplette Remission auftritt, als die herkömmliche Behandlungsmethode (KG). Entsprechend hat die Kontrollgruppe eine höhere Anzahl an Beobachtungen, welche in die Kategorien *PR, SD* und *PD* eingeordnet werden können, als die Versuchsgruppe.
Die *MCL-Elderly*-Studie weist dagegen ein anderes Muster auf. Hier ist die Kontrollgruppe mit einer höheren Anzahl an Beobachtungen in den Kategorien *CR, CRu, PR* und *SD* vertreten. Die Versuchsgruppe ist dagegen im Verhältnis zur Kontrollgruppe häufiger in den Kategorien *PD* und *ED* zu finden.

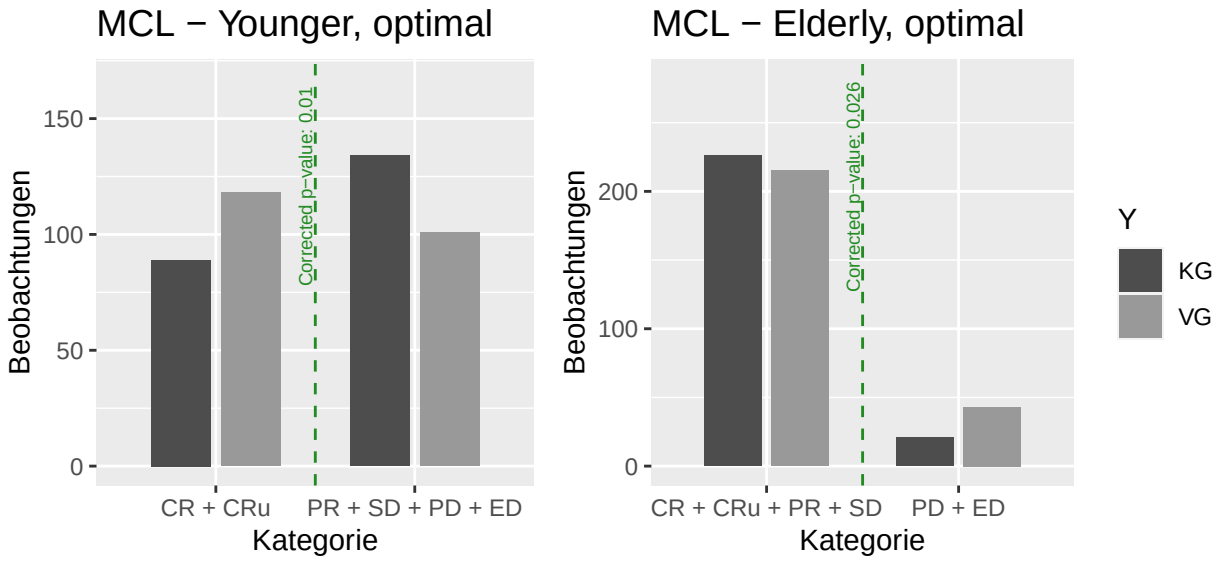


Abbildung 18: Darstellung der *MCL-Younger* und der *MCL-Elderly*-Studie mit Hilfe der maxsel.plot-Funktion. Die verwendeten Daten können leichte Abweichungen zu den publizierten Ergebnissen aufweisen, was in der Wahl des Stichtages begründet ist. Beide Studien untersuchten den Einfluss neuer Behandlungsmethoden bei Patienten mit dem Mantelzell-Lymphom. Die Darstellung entspricht der optimalen Aufteilung der Kategorien in zwei Gruppen.

Die Studienergebnisse von Nickenig et al. (2006) und Hiddemann et al. (2005) sind mit Hilfe der `maxsel.plot`-Funktion in Abbildung 19 grafisch dargestellt. Die Studien stammen beide aus der *Deutschen Studiengruppe Niedrigmaligne Lymphome* (GLSG). Dabei wurden jeweils eine neue Behandlungsmethode für Patienten mit föllikulärem Lymphom einer herkömmlichen Behandlungsmethode gegenübergestellt und getestet. Die Studie von Nickenig et al. (*FL-1996*) enthält Daten von 415 Patienten mit einem föllikulärem Lymphom, die im Zeitraum von Mai 1996 bis Dezember 1998 rekrutiert wurden. Davon entfallen 198 Patienten auf die Kontrollgruppe und 217 Patienten auf die Versuchsgruppe. Die Studie von Hiddemann et al. (*FL-2000*) beinhaltet insgesamt 602 Patienten, welche ein föllikuläres Lymphom aufweisen. Die Studie wurde als prospektive randomisierte Studie im Jahr 2000 gestartet und besteht aus 296 Patienten aus der Kontrollgruppe und 306 Patienten aus der Versuchsgruppe.

Abbildung 19 zeigt, dass in beiden Studien bei einem sehr großer Anteil der Patienten eine (unbestätigte) komplette Remission (*CR* & *CRu*) aufgetreten ist. Wie in den Studien *MCL-Younger* und *MCL-Elderly* zeigen sich auch hier signifikante Unterschiede zwischen Kontroll- und Versuchsgruppe bei einen Signifikanzniveau von $\alpha = 0.05$. Die optimale

Aufteilung der Gruppen in der *FL-1996*-Studie liegt hier zwischen der vierten Kategorie *stabile Erkrankung* (SD) und der fünften Kategorie *Progress* (PD). Die Studie *FL-2000* zeigt dagegen eine optimale Aufteilung zwischen der zweiten Kategorie *unbestätigte komplette Remission* (CRu) und der dritten Kategorie *partielle Remission* (PR).

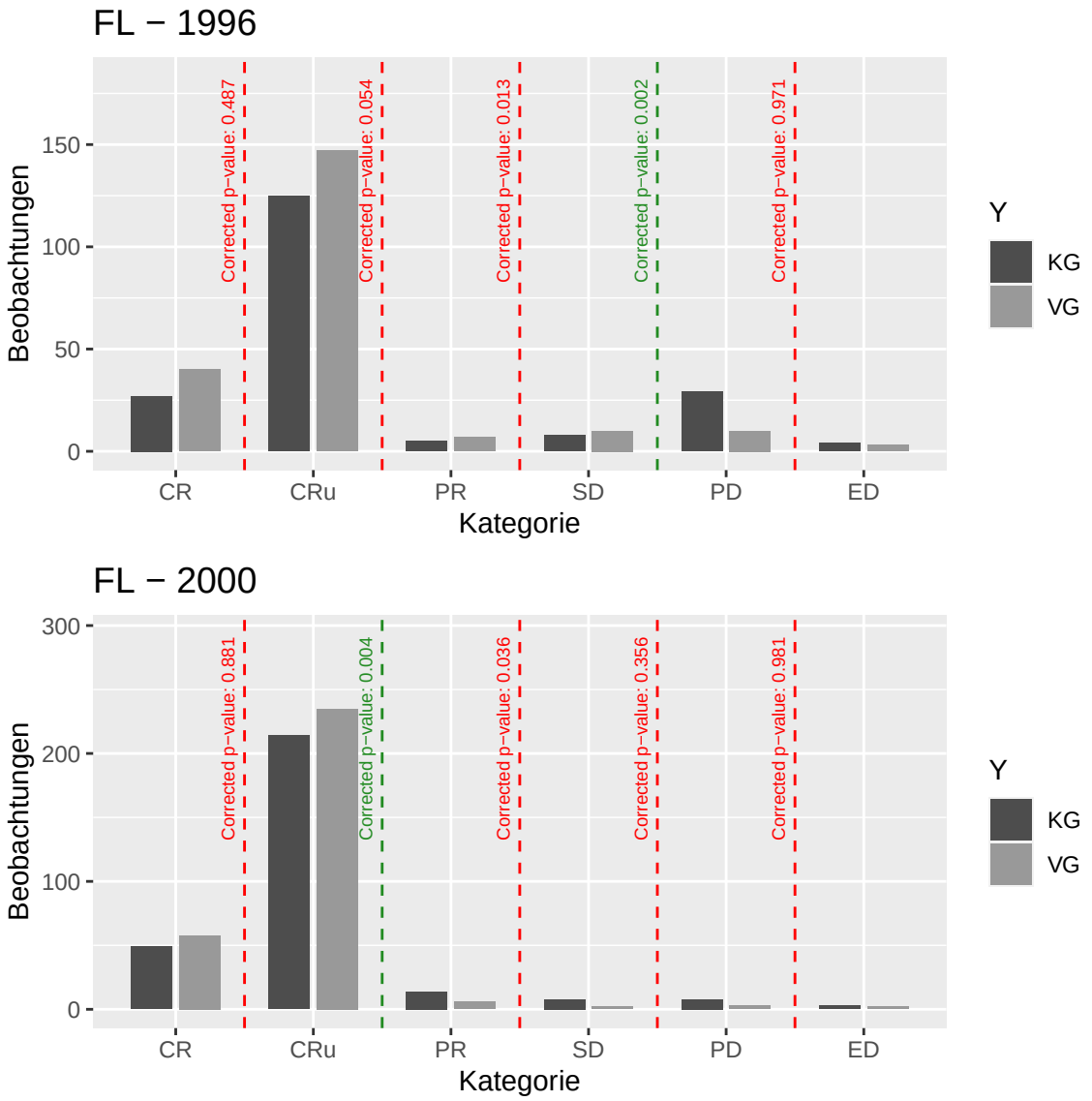


Abbildung 19: Darstellung der *FL-1996* und der *FL-2000*-Studie mit Hilfe der maxsel.plot-Funktion. Die verwendeten Daten können leichte Abweichungen zu den publizierten Ergebnissen aufweisen, was in der Wahl des Stichtages begründet ist. Beide Studien untersuchten den Einfluss neuer Behandlungsmethoden bei Patienten mit follikulärem Lymphom.

In Abbildung 20 findet sich die optimale Aufteilung der insgesamt sechs Kategorien in zwei disjunkte Gruppen. Hierbei ist, wie bereits in Abbildung 19 zu sehen war, der dazugehörige (signifikante) P-Wert mit Hilfe der grünen gestrichelten Linie dargestellt. In beiden Grafiken ist ein ähnliches Verhalten in den Gruppen zu erkennen. Für die Studie *FL-1996* gilt, dass der Anteil an Beobachtungen der Versuchsgruppe in den Kategorien *CR, CRu, PR* und *SD* höher ist als in der Kontrollgruppe. Dafür ist der Anteil an Beobachtungen in der Kontrollgruppe, welche den Kategorien *PD* und *ED* zugeordnet werden können, höher als in der Versuchsgruppe.

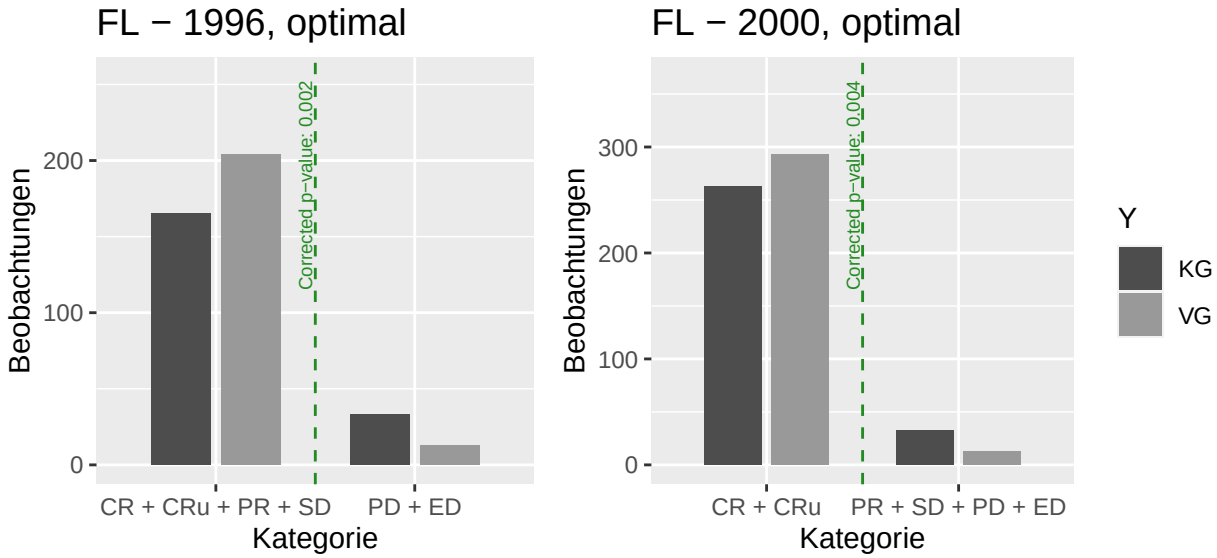


Abbildung 20: Darstellung der *FL-1996* und der *FL-2000*-Studie mit Hilfe der maxsel.plot-Funktion. Die verwendeten Daten können leichte Abweichungen zu den publizierten Ergebnissen aufweisen, was in der Wahl des Stichtages begründet ist. Beide Studien untersuchten den Einfluss neuer Behandlungsmethoden bei Patienten mit follikulärem Lymphom. Die Darstellung entspricht der optimalen Aufteilung der Kategorien in zwei Gruppen.

Das gleiche Muster lässt sich auch für die *FL-2000*-Studie erkennen. Allerdings führt hier ein anderer Schwellenwert zu einer leicht veränderten Aufteilung der Gruppen. Diese betreffen eine veränderte Zuteilung der Kategorien *PR* und *SD*, welche beide, besonders im Vergleich zu den Kategorien *CR* und *CRu*, nur eine relativ geringe Beobachtungsanzahl aufweisen. Somit werden hier anteilsmäßig mehr Personen der Versuchsgruppe in die Kategorien *CR* und *CRu* eingeteilt werden. Für die Summe der Beobachtungen in den Kategorien *PR, SD, PD* und *ED* wiederum ist ein größerer Anteil aus der Kontrollgruppe zu erkennen.

Zusammenfassung

Insgesamt bildet hier die Auswertung mit Hilfe der maximal selektierten χ^2-Statistiken eine gute Alternative zu herkömmlichen Auswertungsmethoden. Besonders vorteilhaft erscheint, dass neben einem herkömmlichen Ergebnis eines Unabhängigkeitstests (P-Wert), auch ein optimaler Schwellenwert zur Aufteilung der Kategorien in zwei disjunkte Gruppen berechnet wird. Alle Auswertungen der Studien durch die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix zeigten hier einen signifikanten Unterschied zwischen Kontroll- und Versuchsgruppe.

7 Fazit und Ausblick

Im Rahmen dieser Masterarbeit sollte untersucht werden, ob die Methode der maximal selektierten χ^2-Statistiken zur Auswertung ordinaler Zielgrößen in randomisierten klinischen Studien geeignet erscheint. Im Gegensatz zu bisherigen Ansätzen verwendet die Methode nach Boulesteix die exakte Verteilung der maximal selektierten χ^2-Statistiken.
Mit Hilfe einer Simulationsstudie wurde sowohl die erreichte Power als auch die Beschränkung des Fehlers 1. Art untersucht und bewertet. Als Vergleich dienten mit dem χ^2-Test, dem *Wilcoxon*-Test und dem *exakten Fisher*-Test drei weit verbreitete Tests, welche ebenfalls bei ordinalen Strukturen angewandt werden können.

Die Durchführung der Simulation wurde mit Hilfe der statistischen Software `R` durchgeführt. Hierfür wurde eigens ein veraltetes R-Paket modernisiert und mit neuen Funktionen ausgestattet. Das Paket dient zur Auswertung von maximal selektierten χ^2-Statistiken mit ordinalen Zielgrößen.
Durch die Verwendung einer *S4*-Klasse und leicht verständlichen Funktionen kann die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix somit einer breiten Öffentlichkeit zugänglich gemacht werden. Dabei sind auch grafische Darstellungen der Ergebnisse möglich und individuell anpassbar. Mit Hilfe inkludierter, externer Pakete ist es außerdem möglich, dass auch Daten mit einer großen Anzahl an Beobachtungen und Kategorien verwendet und analysiert werden können. Neben einem P-Wert liefert das Paket auch den bestmöglichen Schwellenwert, um die ordinale Variable in zwei Gruppen zu unterteilen. Somit kann die ordinale Struktur verwendet und gleichzeitig intuitiv vom Anwender interpretiert werden.

Die Ergebnisse der Simulationsstudie lassen allerdings erkennen, dass Auswertungen mit Hilfe der maximal selektierten χ^2-Statistiken nach Boulesteix aktuell noch keine vollwertige Alternative zu den herkömmlichen Tests darstellen. Dies ist hauptsächlich darin begründet, dass die Beschränkung des Fehlers 1. Art in fast jedem getesteten Szenario zum Teil deutlich überschritten wird. Ein erhöhter Fehler 1. Art, was einer erhöhten Wahrscheinlichkeit für die Ablehnung einer wahren Nullhypothese entspricht, könnte besonders in medizinischen Studien ein Problem darstellen. Durch solche falsch positiven Ergebnisse kann beispielsweise ein Patient als krank diagnostiziert werden, obwohl er eigentlich gesund ist.
Bezüglich der erreichten Power schnitt der *Maxsel*-Test dagegen relativ gut ab. So erreichte er, auch im Vergleich zu den anderen geprüften statistischen Tests, durchschnittlich meist sehr gute Ergebnisse. Besonders bei einer relativ hohen Anzahl an Kategorien der ordinalen Variable und einer geringen Anzahl an Beobachtungen waren die Ergebnisse bezüglich der erreichten Power vergleichsweise gut. Im Vergleich mit den anderen Tests zeigte der *Maxsel*-Test in allen vier getesteten Szenarios die konstantesten Ergebnisse hinsichtlich der erreichten Power.
Allerdings kann ein erhöhter Fehler 1. Art eine Verringerung des Fehlers 2. Art bewirken, was damit einer Erhöhung der erreichten Power gleichkommt. Aufgrund dessen müssen die Ergebnisse dieser Simulationen mit Vorsicht betrachtet werden. Ob der *Maxsel*-Test auch bei einer erfolgreichen Beschränkung des Fehlers 1. Art gute Ergebnisse bezüglich der erreichten Power erzielen würde, kann anhand dieser Ergebnisse nicht beurteilt werden.

Das Anwendungsbeispiel zeigte, dass Auswertungen mit der Methode der maximal selektierten χ^2-Statistiken nach Boulesteix durch das R-Paket *exactmaxsel2* einfach und intuitiv möglich sind. Mit Hilfe der grafischen Darstellung kann das methodische Vorgehen und Verständnis gegenüber der Methode weiter vereinfacht werden. Die Darstellung der P-Werte, welche durch unterschiedliche Aufteilungen der Kategorien in zwei disjunkte Gruppen zustande kommt, kann bei der Wahl eines Grenzwertes innerhalb der Kategorien von großer Hilfe sein. Allerdings lassen sich die Ergebnisse der Auswertung mit Hilfe des *Maxsel*-Tests nur bedingt mit den veröffentlichten Ergebnissen der Studien vergleichen. Durch eine Abweichung in der Wahl des Stichtages ist die Datengrundlage leicht verändert, wie sie in den veröffentlichten Studien dargestellt wurden. Dadurch ist ein korrekter Vergleich der Ergebnisse nicht möglich. Zudem ist aus den Simulationsstudien die Problematik bezüglich des Fehler 1. Art bekannt, welche sich auch in den Ergebnissen des Anwendungsbeispiels zeigen könnten.

Insgesamt zeigen die Erkenntnisse dieser Arbeit, dass durch die Methode der maximal selektierten χ^2-Statistiken nach Boulesteix eine intuitive und leicht interpretierbare Möglichkeit zur Auswertung ordinaler Zielgrößen geschaffen wurde. Um die Problematik des erhöhten Fehlers 1. Art zu lösen, müsste der Tests allerdings noch weiter modifiziert werden. Hierbei wäre eventuell eine Stetigkeitskorrektur von Nutzen, welche die exakte und damit diskrete Verteilungsfunktion zu einer stetigen Verteilungsfunktion approximiert. Ein Lösen der Problematik bezüglich des Fehlers 1. Art würde anschließend die Möglichkeit schaffen, die Methode nochmals mit diversen Tests zu vergleichen und zu bewerten. Auch die Wahl neuer Szenarios könnte dazu führen, dass weitere Erkenntnisse bezüglich des Verhaltens des *Maxsel*-Tests in bestimmten Situationen gesammelt werden können.

8 Anhang

Im Anhang befindet sich weiteres Material zu dieser Masterarbeit. Dazu gehören sowohl weitere Abbildungen und Tabellen, als auch die verwendeten R-Codes.

8.1 Abbildungen und Tabellen

Hier finden sich alle im vorhergehenden Text erwähnten Abbildungen und Tabellen.

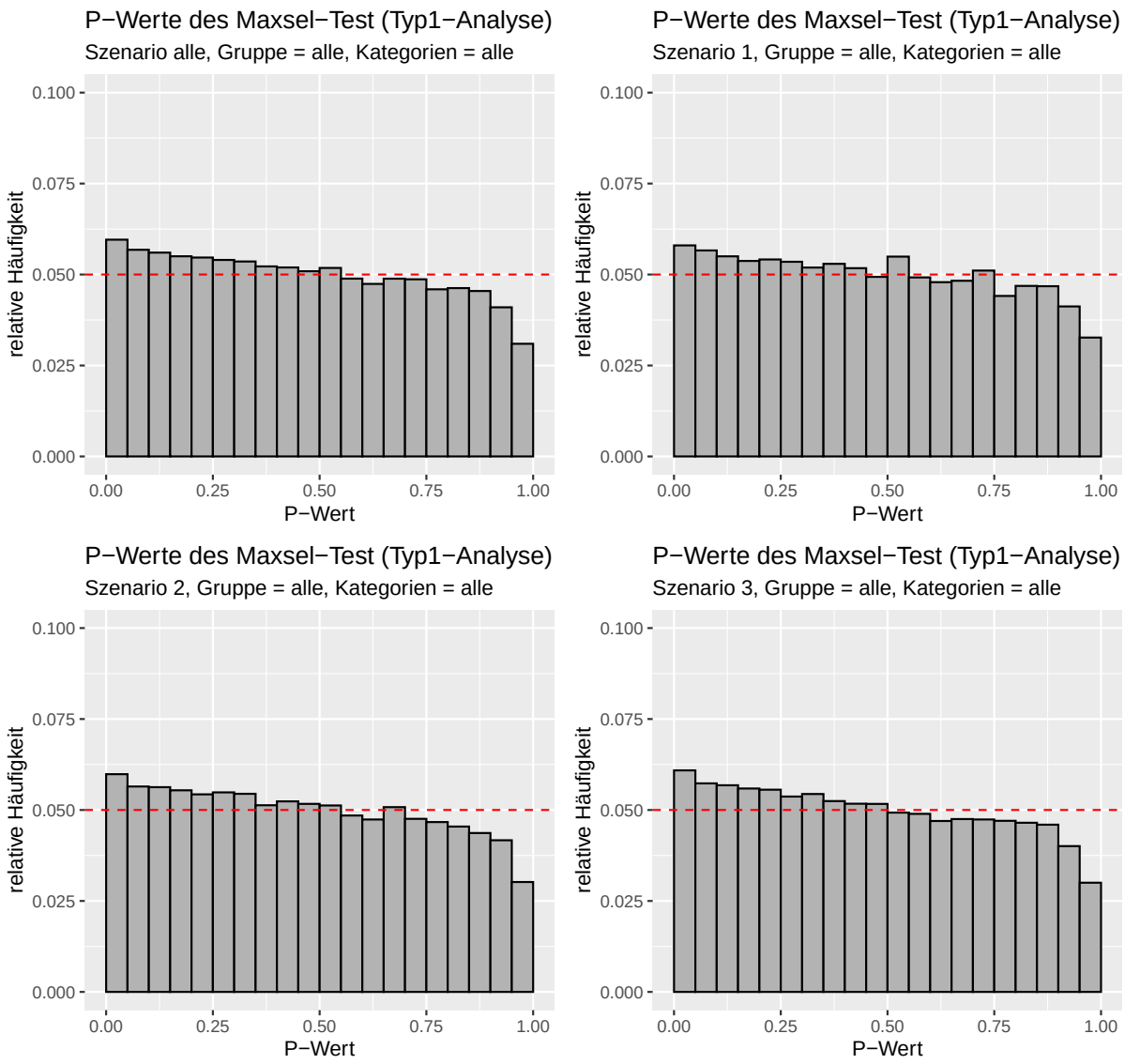


Abbildung 21: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich des Fehlers 1. Art, separiert gemäß der verschiedenen Szenarios. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte. Der erste Balken entspricht dem aufgetretenen Fehler 1. Art.

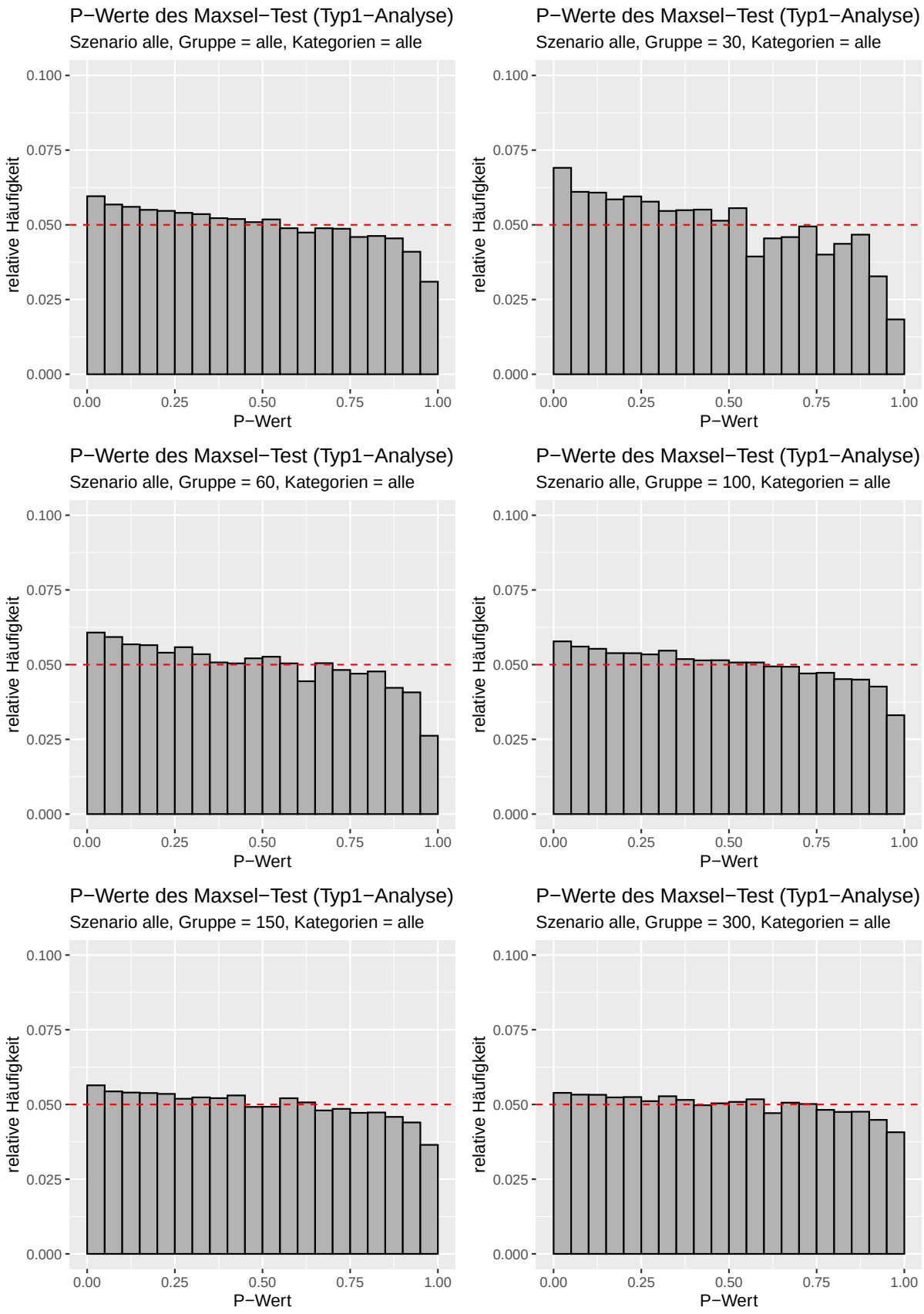


Abbildung 22: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich des Fehlers 1. Art, separiert gemäß der Anzahl an Beobachtungen in den Gruppen. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte. Der erste Balken entspricht dem aufgetretenen Fehler 1. Art.

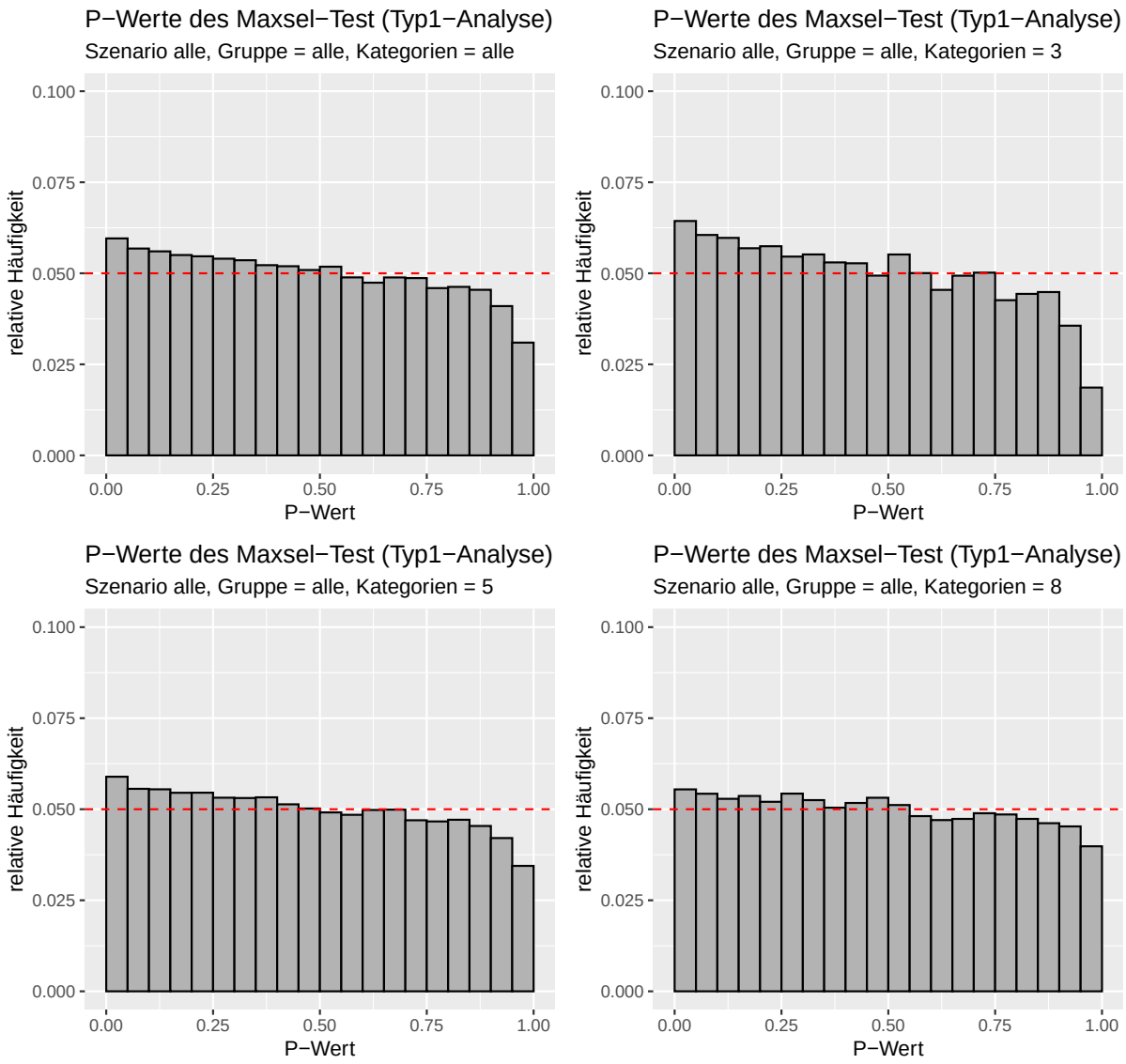


Abbildung 23: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich des Fehlers 1. Art, separiert gemäß der Anzahl an Kategorien in den Gruppen. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte. Der erste Balken entspricht dem aufgetretenen Fehler 1. Art.

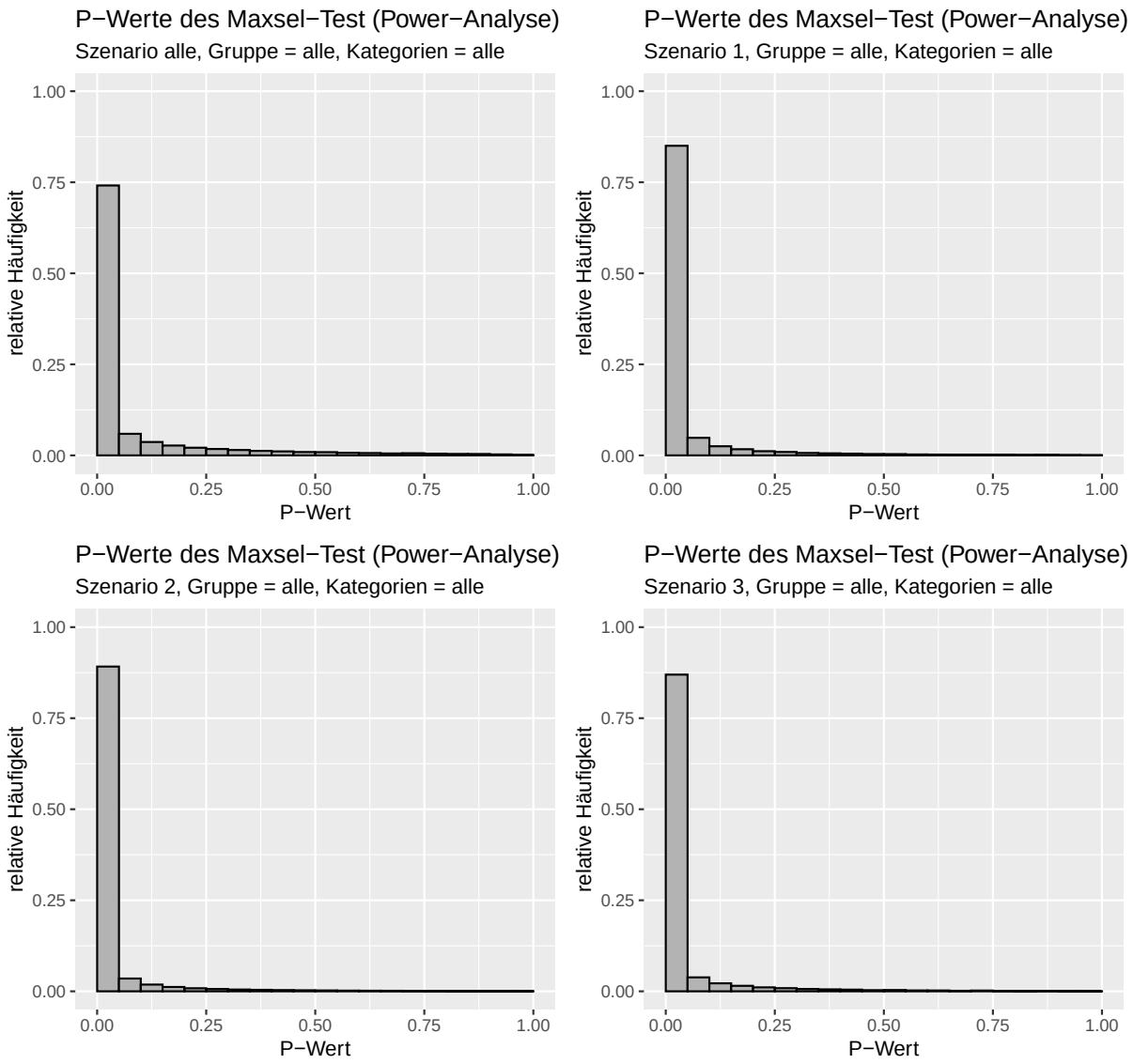


Abbildung 24: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich der erreichten Power, separiert gemäß der verschiedenen Szenarios. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte.

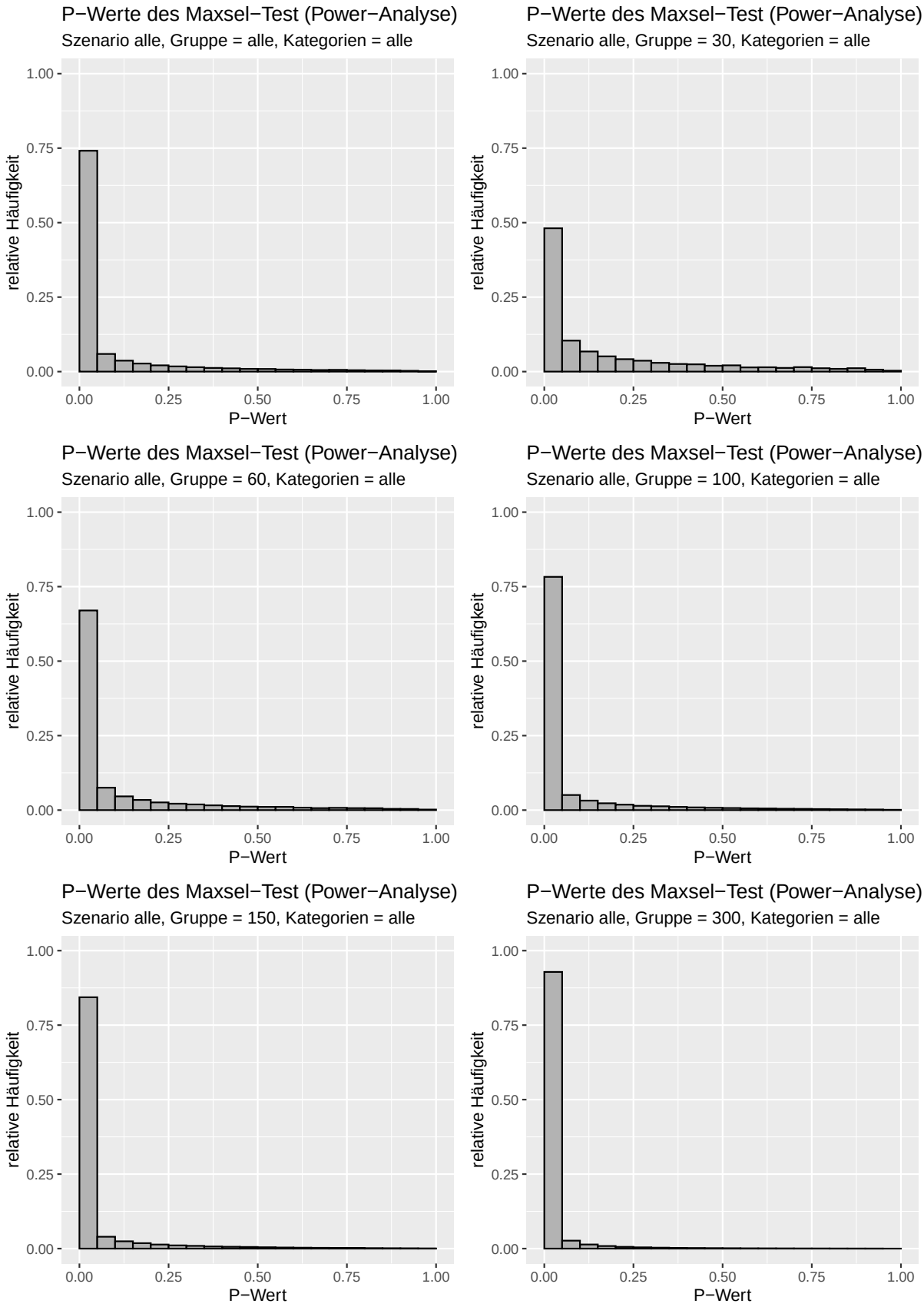


Abbildung 25: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich der erreichten Power, separiert gemäß der Anzahl an Beobachtungen in den Gruppen. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte. Der erste Balken entspricht dem aufgetretenen Fehler 1. Art.

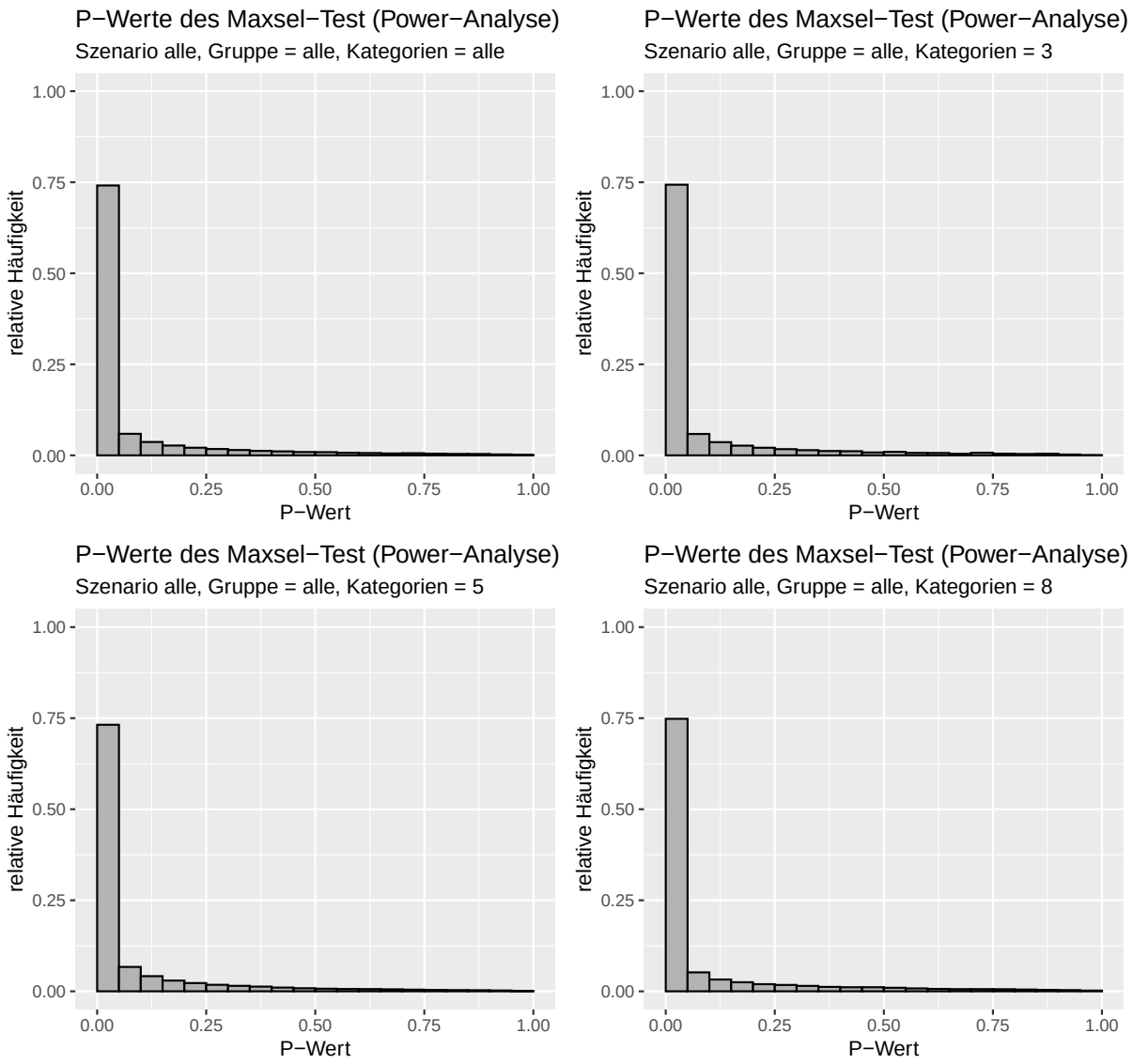


Abbildung 26: Verteilung der P-Werte des Maxsel-Tests in den Simulationen bezüglich der erreichten Power, separiert gemäß der Anzahl an Kategorien in den Gruppen. Die rote Linie entspricht einer Gleichverteilung, welche im Idealfall erreicht werden sollte. Der erste Balken entspricht dem aufgetretenen Fehler 1. Art.

8.2 Digitaler Anhang

Dieser Stick enthält den digitalen Anhang dieser Masterarbeit. Darin befinden sich alle geschriebenen R-Codes, das erstellte R-Paket, die verwendeten Abbildungen und die Masterarbeit im PDF-Format. Zusätzlich sind weitere Abbildungen vorhanden, die aus Zeitgründen nicht in der Masterarbeit verwendet wurden.

Folgende Dateien enthält der digitale Anhang:

- *Masterarbeit_Christian_Bihl.pdf*: Die Masterarbeit im PDF-Format
- R-Paket
  - *Start.R*: R-Skript um das exactmaxsel2-Paket zu installieren
  - exactmaxsel2: Ordner mit dem erstellten R-Paket inklusive den verschiedenen R-Skripten der Funktionen, die entsprechenden Hilfeseiten, eine Beispielanwendung (Vignette) und weiteres
  - *exactmaxsel2_1.0.1.tar.gz*: GZ-Datei, welche zur Installation des Paketes benötigt wird
- Simulation
  - Daten/*Results.Rdata*: Rdata-Objekt, das die Simulationsergebnisse enthält
  - Grafiken: Alle grafischen Darstellungen der Simulationsergebnisse (inklusive der Verteilungen der P-Werte)
  - *Auswertung_Allgemein.R*: R-Skript zur allgemeinen Auswertung der Simulationsergebnisse
  - *Auswertung_p_Werte_Power.R*: R-Skript zur Auswertung der Verteilungen der P-Werte bei der Simulation der Power
  - *Auswertung_p_Werte_Typ1.R*: R-Skript zur Auswertung der Verteilungen der P-Werte bei der Simulation des Fehlers 1. Art
  - *Auswertung_Szenario1(2/3/4).R*: R-Skripte zur spezifische Auswertung für jedes Szenario
  - *load_data.R*: R-Skript zum Einlesen und Aufbereiten der Simulationsergebnisse
  - *Simulation.R*: R-Skript, dass die Simulationsdurchführung enthält
- Anwendungsbeispiel
  - Daten/*Daten_Anwendungsbeispiel.R*: R-Skript, welches die verwendeten Daten der vier Studien enthält
  - Grafiken: Ordner mit allen grafischen Darstellungen der Ergebnisse
  - *Auswertung_Anwendungsbeispiel.R*: R-Skript zur Auswertung der Anwendungsbeispiele
  - *Klinische_Daten_Response_GLSG.pdf*: PDF-Dokument mit weiteren Erklärungen zu den verwendeten Studien

Literatur

Boulesteix, A.-L. (2006). Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(3):451–462.

Dudoit, S., Shaffer, J. P., und Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Durbin, J. (1971). Boundary-crossing probabilities for the brownian motion and poisson processes and techniques for computing the power of the kolmogorov-smirnov test. *Journal of Applied Probability*, 8(3):431–453.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., und Tutz, G. (2016). *Statistik: Der Weg zur datenanalyse*. Springer-Verlag.

Halpern, J. (1982). Maximally selected chi square statistics for small samples. *Biometrics*, 38(4):1017–1023.

Hermine, O., Hoster, E., Walewski, J., Bosly, A., Stilgenbauer, S., Thieblemont, C., Szymczyk, M., Bouabdallah, R., Kneba, M., Hallek, M., et al. (2016). Addition of high-dose cytarabine to immunochemotherapy before autologous stem-cell transplantation in patients aged 65 years or younger with mantle cell lymphoma (mcl younger): a randomised, open-label, phase 3 trial of the european mantle cell lymphoma network. *The Lancet*, 388(10044):565–575.

Hiddemann, W., Kneba, M., Dreyling, M., Schmitz, N., Lengfelder, E., Schmits, R., Reiser, M., Metzner, B., Harder, H., Hegewisch-Becker, S., et al. (2005). Frontline therapy with rituximab added to the combination of cyclophosphamide, doxorubicin, vincristine, and prednisone (chop) significantly improves the outcome for patients with advanced-stage follicular lymphoma compared with therapy with chop alone: results of a prospective randomized study of the german low-grade lymphoma study group. *Blood*, 106(12):3725–3732.

Kluin-Nelemans, H., Hoster, E., Hermine, O., Walewski, J., Trneny, M., Geisler, C., Stilgenbauer, S., Thieblemont, C., Vehling-Kaiser, U., Doorduijn, J., et al. (2012). Treatment of older patients with mantle-cell lymphoma. *New England Journal of Medicine*, 367(6):520–531.

Koziol, J. A. (1991). On maximally selected chi-square statistics. *Biometrics*, 47(4):1557–1561.

Mehta, C. R. und Patel, N. R. (1983). A network algorithm for performing fisher's exact test in r× c contingency tables. *Journal of the American Statistical Association*, 78(382):427–434.

Miller, R. und Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics*, 38(4):1011–1016.

Nickenig, C., Dreyling, M., Hoster, E., Pfreundschuh, M., Trumper, L., Reiser, M., Wandt, H., Lengfelder, E., Unterhalt, M., und Hiddemann, W. (2006). Combined cyclophosphamide, vincristine, doxorubicin, and prednisone (chop) improves response rates but

not survival and has lower hematologic toxicity compared with combined mitoxantrone, chlorambucil, and prednisone (mcp) in follicular and mantle cell lymphomas: results of a prospective randomized trial of the german low-grade lymphoma study group. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 107(5):1014–1022.

Noordzij, M., Dekker, F. W., Zoccali, C., und Jager, K. J. (2009). Study designs in clinical research. *Nephron Clinical Practice*, 113(3):c218–c221.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Ranganathan, P., Pramesh, C., und Aggarwal, R. (2016). Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspectives in clinical research*, 7(3):144.

Shankland, K. R., Armitage, J. O., und Hancock, B. W. (2012). Non-hodgkin lymphoma. *The Lancet*, 380(9844):848 – 857.

Stel, V., Jager, K., Zoccali, C., Wanner, C., und Dekker, F. (2007). The randomized clinical trial: an unbeatable standard in clinical research? *Kidney international*, 72(5):539–542.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., und Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194.

Wickham, H. (2019). *Advanced r*. CRC press.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, New York, NY.

Witzig, T. E. (2005). Current treatment approaches for mantle-cell lymphoma. *Journal of Clinical Oncology*, 23(26):6409–6414.

Eigenständigkeitserklärung

Hiermit erkläre ich, Christian Reinhold Bihl, dass ich die vorliegende Masterarbeit eigenständig ohne fremde Hilfe verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 04.01.2021	
Ort,Datum	Unterschrift