Strategien der Power-Vergabe in flexiblen Designs

Dissertation

zur Erlangung des akademischen Grades

eines Doktors der Naturwissenschaften

der Technischen Universität Dortmund

Der Fakultät Statistik

der Technischen Universität Dortmund

vorgelegt von

Susanne Menzler

aus Essen

München 2008

Erstgutachter: Prof. Dr. Joachim Hartung

Zweitgutachter: Prof. Dr. Roland Fried

Tag der mündlichen Prüfung: 17. November 2008

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die strikte Vorgehensweise bei der Planung und Umsetzung klinischer Studien wird dem Wunsch von Forschern und Sponsoren nach Reaktionsmöglichkeiten auf veränderte äußere Umstände sowie unerwartete Entwicklungen während der Studiendurchführung nicht gerecht. Die Forderung nach flexiblen Studiendesigns ist dabei zum einen durch den ethischen Nutzen motiviert. Deutlich überlegene Therapien sollen schneller allen Patienten zur Verfügung gestellt werden, wenn dies bereits in einer frühen Phase der Studie deutlich wird. Zum anderen soll wirtschaftlicher Schaden durch eine zu optimistisch geplante Studie, die bei höherem Stichprobenumfang zur Zulassung eines wirksamen Medikamentes geführt hätte, vermieden werden. Insbesondere bei Studien zu Psychopharmaka ist sowohl der erwartete Therapieeffekt als auch die Varianz innerhalb des beobachteten Patientenkollektivs schwer vorhersagbar und eine Kontrolle mit Option zur Anpassung des Studiendesigns im Studienverlauf wünschenswert. Daher stellt seit einigen Jahren die Entwicklung und Diskussion flexibler Studiendesigns einen wichtigen Bereich in der biometrischen Forschung dar.

Den Vorteil, eine Studie frühzeitig mit Erfolg, also dem Verwerfen der Nullhypothese zu beenden, bieten bereits gruppensequentielle Designs, die jeweils nach einer festen Anzahl an beobachteten Patienten eine Zwischenauswertung ermöglichen. Zu diesen festgelegten Zeitpunkten kann eine vorzeitige Ablehnung der Nullhypothese erfolgen, wobei insgesamt die vorgegebene Irrtumswahrscheinlichkeit erster Art eingehalten wird. Größere Veränderungen am Studiendesign während des Studienverlaufs, wie z.B. die Anpassung des Stichprobenumfangs oder die Wahl einer anderen Teststatistik, ermöglichen jedoch erst die sogenannten adaptiven Designs. Die Veränderungen am Design können dabei entweder auf einer verblindeten oder einer unverblindeten Interimsanalyse beruhen. Diese Arbeit beschränkt sich auf Designs, die einen Test der Nullhypothese zu den Interimszeitpunkten erlauben und damit eine Entblindung verlangen. Auch wenn theoretisch viele Adaptionen

des Studiendesigns betrachtet werden, liegt die wichtigste mögliche Veränderung in der Anpassung des Stichprobenumfangs.

Viele Ansätze zur Konstruktion adaptiver Designs wurden entwickelt, die jeweils unterschiedliche Eigenschaften und Einschränkungen aufweisen. Bei einigen Verfahren müssen feste Vorgaben für den Stichprobenumfang oder das Gewicht pro Studienabschnitt eingehalten werden. Ein großer Teil der Designs setzt die maximale Anzahl an durchführbaren Stufen zu Beginn fest. Andere erlauben trotz Entblindung nur eine Anpassung des Stichprobenumfangs und keinen Test auf die Nullhypothese oder sind mit hohem Rechenaufwand bei der Bestimmung der kritischen Werte auf den einzelnen Stufen verbunden.

Das adaptive Design gemäß der Inverse-χ^2-Methode von Hartung und Knapp (2003) erlaubt nach jeder Stufe bei einfacher Berechnung eines universalen kritischen Werts und offener Anzahl an Interimsanalysen eine Anpassung des Stichprobenumfangs und des Gewichts für jeden folgenden Studienabschnitt. Dabei wird zu jeder Zwischenauswertung ein Test der Nullhypothese durchgeführt.

Der Schwerpunkt vieler Arbeiten beruht auf der Konstruktion eines adaptiven Verfahrens, das das globale Niveau einhält. Konkrete Vorschläge für die Umsetzung, wie die Wahl der bedingten Power auf den einzelnen Stufen, werden nur selten formuliert. Die grundlegende Eigenschaft der Einhaltung des Niveaus ist durch die Inverse-χ^2-Methode gewährleistet, so dass in dieser Arbeit neben der Beschreibung der weiteren Eigenschaften des Verfahrens Strategien zur praktischen Umsetzung des Designs entwickelt werden.

Dazu erfolgt zunächst eine Beschreibung ausgewählter Verfahren in Kapitel 2, die verschiedene Konstruktionsweisen gruppensequentieller und adaptiver Designs gegenüberstellt. Die zugrundeliegenden Annahmen und Beschränkungen der Designs werden aufgezeigt und verglichen.

In Kapitel 3 wird das Prinzip der Inverse-χ^2-Methode in adaptiven Designs erläutert. Die im Originalartikel beschriebene Umsetzung der proportionalen Wahl von Stichprobenumfang und Gewicht wird dargestellt und kommentiert.

Basierend auf den vorangegangenen Kapiteln erfolgt in Kapitel 4 eine Einordnung der Inverse-χ^2-Methode in die bestehenden Designs. Durch die Beziehung zu rekursiven Kombinationstests und die Deutung im Rahmen des äußerst flexiblen Prinzips von Müller und Schäfer (2001) ist die Grundlage für einige theoretische Erweiterungen gelegt. Sowohl ein globaler P-Wert als auch optionale Schranken zum Abbruch wegen Aussichtslosigkeit zu Beginn jeder Stufe lassen sich dadurch konstruieren. Der Zusammenhang zwischen Designparametern und den Fehlerraten erster und zweiter Art wird anhand der Konstanthaltung jeweils einzelner Parameter beschrieben. Weiter wird in Abschnitt 4.3 für eine

vorgegebene bedingte Power durch Lösung eines Fixpunktproblems der Stichprobenumfang bei proportionaler Gewichtung exakt bestimmt. In Abschnitt 4.5 werden vier Schätzer vorgestellt, deren Eignung zur Anpassung des Stichprobenumfangs in den Strategien im folgenden Kapitel 5 mit intensiven Simulationen untersucht wird.

Die Auswahl der betrachteten Strategien zur Anpassung des Stichprobenumfangs orientiert sich hinsichtlich der Powervergabe unter anderem an in der Literatur häufig genannten Verfahren. So sollen zum einen Powerverläufe bekannter gruppensequentieller Designs nachgeahmt werden, zum anderen werden Strategien mit konstanter bedingter Power sowie gleichmäßigem Anwachsen der erzielten Power betrachtet. Die Entwicklung der Designs erfolgt schrittweise, ausgehend vom ursprünglichen Design bei vorläufiger Beschränkung auf fünf Stufen. Dabei werden zunächst bei konstantem Parameterwert zur Anpassung des Stichprobenumfangs verschiedene Strategien hinsichtlich der bedingten Power für den verbleibenden Teil der Studie untersucht. Anschließend erfolgen Simulationen zu Strategien, die den Stichprobenumfang an einen der vier vorgeschlagenen Schätzer anpassen. Basierend auf den Parameterschätzungen wird in einem weiteren Strategievorschlag rückwirkend die vergebene Power und daraus die verbleibende Power bestimmt. Schließlich werden Strategien wie die konstante bedingte Powervergabe oder der geplante konstante Powerzuwachs betrachtet, für die eine offene Stufenzahl ohne Wechsel der Strategie nach der fünften Stufe möglich ist. Die unterschiedlichen Strategien werden hinsichtlich des benötigten mittleren und medianen Stichprobenumfangs und der mittleren Anzahl durchgeführter Stufen sowie der erzielten Power verglichen. Für die betrachteten Schätzer werden die Auswirkungen auf die entsprechenden Kennwerte der Strategien untersucht und die Varianz sowie die mittlere Verzerrung am Studienende erfasst.

Schließlich erfolgen in Kapitel 6 eine Zusammenfassung der Ergebnisse und ein Ausblick auf weitere mögliche Anwendungen und Erweiterungen.

Kapitel 2

Grundlegende flexible Designs

Der Wunsch von Forschern und Sponsoren, den strengen Verlauf klinischer Studien von der Stichprobenplanung auf Basis von Vorwissen über die Rekrutierung von Studienteilnehmern bis zur abschließenden Auswertung zu durchbrechen, führte zu der Entwicklung verschiedener flexibler Studiendesigns. Ihr wesentliches Kennzeichen ist die Möglichkeit, bei Einhaltung eines vorgegebenen globalen Niveaus eine oder mehrere Zwischenauswertungen vorzunehmen. Je nach Design sind eine vorzeitige Testentscheidung oder eine Anpassung des Studiendesigns möglich. In diesem Kapitel werden verschiedene Typen flexibler Designs vorgestellt und ihre Eigenschaften erläutert.

Die innerhalb einer Studie interessierende Responsevariable wird im Folgenden mit X bezeichnet und sei stetig verteilt. Die Indizierung X_{kij} weist innerhalb einer mehrstufigen und mehrarmigen Studie auf den Response des j-ten Patienten im i-ten Behandlungsarm im k-ten Studienteil hin. Die Studienhypothese wird bezüglich des Parameters θ formuliert und vergleicht im Allgemeinen die Erwartungswerte der Responsevariablen verschiedener Behandlungsarme bzw. Gruppen. Mit α und β, jeweils $\in (0, 1)$, werden die globalen Wahrscheinlichkeiten für den Fehler erster und zweiter Art bezeichnet, wobei zusätzliche, im Folgenden erläuterte Indizes lokale und bedingte Fehlerraten innerhalb von Studienabschnitten kennzeichnen.

2.1 Gruppensequentielle Pläne

Sequentielle Verfahren wurden zum einen entwickelt, um neue wirksamere Therapien schneller allen Patienten zugänglich zu machen, zum anderen um dem Bedürfnis der Forscher nach Information über den laufenden Stand einer Studie nachzukommen. Die im Rahmen eines gruppensequentiellen Tests durchgeführten Interimsanalysen zielen vorwie-

gend auf einen vorgezogenen Studienabbruch bei besonders überzeugenden Therapieergebnissen ab und liefern einen Einblick in den Verlauf der Studie. Sie dienen jedoch nicht der Anpassung an nicht erfüllte Annahmen oder Voraussetzungen bzgl. der Verteilung der Responsevariablen, die in der Planung des Studiendesigns berücksichtigt wurden. Die Stichprobenumfänge auf den einzelnen Stufen werden so gewählt, dass für einen vorgegebenen Wert aus der Alternativhypothese eine vorher festgelegte Power von $1-\beta$ erreicht wird. Im Folgenden werden einige der bekanntesten gruppensequentiellen Verfahren erläutert.

Ausgehend von sequentiellen Verfahren nach Armitage et al. (1969), schlägt Pocock (1977) vor, statt der organisatorisch schwierigen kontinuierlichen Erhebung von Patienten jeweils Signifikanztests nach längeren gleichdauernden Intervallen durchzuführen, zum Beispiel nach Vorliegen des Response von jeweils n Personen. Die maximale Anzahl K von Stufen wird im Voraus festgelegt und somit auch der maximale Stichprobenumfang $N = nK$. Die ersten gruppensequentiellen Pläne legen die Normalverteilung der Response-Variablen mit bekannter Varianz zugrunde.

Das Verfahren nach Pocock (1977) lässt sich sowohl auf den Test zweiseitiger als auch einseitiger Hypothesen anwenden. Im Folgenden wird der einseitige Fall des Vergleichs zweier Gruppen dargestellt. Unter Annahme der Normalverteilung der beobachteten Variablen X_{kij}, mit den Indizes $k = 1, \ldots, K$ für den Studienabschnitt, $i = 1, 2$ für die Gruppe bzw. Behandlung, $j = 1, \ldots, n/2$ für den Patienten, mit den unbekannten Erwartungswerten μ_1 und μ_2 und bekannter Varianz σ^2 soll die folgende Hypothese zum Niveau α getestet werden:

$$\mathrm{H}_0 : \ \theta = \mu_1 - \mu_2 = 0 \ \text{ vs. } \ \mathrm{H}_1 : \ \theta > 0. \tag{2.1}$$

Aus den unter H_0 standardnormalverteilten, unabhängigen Teststatistiken der einzelnen Stufen

$$z_k = \frac{\bar{X}_{k1} - \bar{X}_{k2}}{\sigma} \cdot \sqrt{\frac{n}{4}}, \quad k = 1, \ldots, K \tag{2.2}$$

$$\text{mit } \ \bar{X}_{ki} = \frac{2}{n} \cdot \sum_{j=1}^{n/2} X_{kij}, \ i = 1, 2,$$

ergibt sich die Teststatistik für Stufe k zu

$$Z_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} z_i. \tag{2.3}$$

Zur Erhaltung eines globalen Niveaus α muss der kritische Wert $u_{1-\alpha'}$, der dem $(1-\alpha')$-Quantil der Standardnormalverteilung entspricht, aus folgender Gleichung numerisch

bestimmt werden.

$$1 - \mathrm{P}_{\mathrm{H}_0}\left(\bigcap_{k=1}^{K}\{Z_k < u_{1-\alpha'}\}\right) = \alpha. \tag{2.4}$$

Das nominale Niveau α' ist kleiner als das globale Niveau α, um den Effekt des multiplen Testens auszugleichen. Die Studie endet mit der Ablehnung der Nullhypothese auf der k-ten Stufe, wenn $Z_k \geq u_{1-\alpha'}$, andernfalls wird sie mit der Beobachtung weiterer n Objekte fortgesetzt. Erfolgt auch in der letzten Stufe K keine Ablehnung, endet der Versuch mit der Annahme der Nullhypothese.

Unter Vorgabe der Fehlerraten erster und zweiter Art, α und β, sowie der maximal geplanten Anzahl Stufen K lässt sich der benötigte Stichprobenumfang n pro Stufe ebenfalls numerisch ermitteln.

Alternativ zum Ansatz nach Pocock, dessen kritische Schranken auf jeder der K Stufen konstant sind, ist jede beliebige Wahl kritischer Schranken $(u_1, \ldots, u_K)$ denkbar, solange das globale Niveau α eingehalten wird. O'Brien und Fleming (1979) wählen für die Teststatistik aus (2.3) eine monoton fallende Folge kritischer Schranken, die sich als Produkt einer Konstanten und dem Kehrwert der Wurzel aus der jeweiligen Stufenzahl k darstellen lässt. Somit ergeben sich konstante kritische Schranken für die transformierte Teststatistik

$$Z_k^* = \sqrt{k} Z_k. \tag{2.5}$$

Dadurch wird die Ablehnung auf einer frühen Stufe erschwert, während der Test auf der letzten Stufe K nahe des globalen Niveaus α durchgeführt wird.

Der mittlere benötigte Stichprobenumfang unter der Alternative wird häufig als Vergleichskriterium für unterschiedliche gruppensequentielle Verfahren herangezogen. Pocock (1982) bestimmt die kritischen Schranken bei Verwendung der Teststatistik (2.3) so, dass dieser minimal wird. Für $K = 5$ sind diese kritischen Schranken für jede der Stufen nahezu konstant und ähneln somit den ursprünglich von Pocock (1977) vorgeschlagenen Schranken. Der Stichprobenumfang auf den einzelnen Stufen und somit der maximale Stichprobenumfang ist jedoch bei gleicher Power bei O'Brien und Fleming (1979) kleiner als bei Pocock (1982). Für beide Designs gilt, dass bei größerer maximaler Stufenzahl K zwar der Umfang n auf den einzelnen Stufen sinkt, aber der maximale Gesamtumfang nK steigt.

Anhand der Verdeutlichung der Unterschiede der beiden Pläne werden im Folgenden einige Begriffe und Notationen festgelegt. Beispielhaft werden die Eigenschaften für eine Studie zum Testproblem (2.1) mit $K = 5$ Stufen erläutert. Zum globalen Niveau von $\alpha = 0.025$

Tabelle 2.1: *Kritische Schranken u_k für das einseitige Testproblem bei 5 Stufen nach Pocock (1977) und nach O'Brien und Fleming (1979) bei globalem Niveau von $\alpha = 0.025$, vergebenes Niveau bis zur k-ten Stufe, $\alpha(k)$, und in der k-ten Stufe α_k mit $\sum_{i=1}^{k} \alpha_i = \alpha(k)$*

	Pocock			O'Brien und Fleming		
Stufe	u_k	$\alpha(k)$	α_k	u_k	$\alpha(k)$	α_k
1	2.4132	0.0079	0.0079	4.5617	< 0.0001	< 0.0001
2	2.4132	0.0138	0.0059	3.2256	0.0006	0.0006
3	2.4132	0.0183	0.0045	2.6337	0.0045	0.0038
4	2.4132	0.0219	0.0037	2.2809	0.0128	0.0083
5	2.4132	0.0250	0.0031	2.0401	0.0250	0.0122

lassen sich mittels der Module SEQSCALE und SEQ der Software SAS/IML, Version 9.1 die kritischen Schranken u_k, der bis zur k-ten Stufe vergebene Fehler erster Art $\alpha(k)$ und der auf der Stufe k vergebene Fehler $\alpha_k = \alpha(k) - \alpha(k-1)$ mit $\alpha(0) = 0$ für $k = 1, \ldots, K$ berechnen.

Aus Tabelle 2.1 ist zu entnehmen, dass die konstanten Schranken von Pocock (1977) zu einer schnelleren Vergabe des Fehlers erster Art im Vergleich zum Design nach O'Brien und Fleming (1979) führen. Dies wird insbesondere am Niveau, das auf der fünften Stufe noch zu vergeben ist, deutlich. Während bei Pocock dies lediglich 0.0031 beträgt, wird bei O'Brien und Fleming mit 0.0122 noch etwa die Hälfte des Gesamtniveaus auf der letzten Stufe vergeben.

Entsprechend zur Vergabe des Niveaus enthält Tabelle 2.2 den Verlauf der Power, wobei die kumulierte Power bis einschließlich der k-ten Stufe mit $\pi(k)$ und der Powerzuwachs in der Stufe k mit π_k bezeichnet wird. Wenn ein standardisierter Effekt $\theta = 0.5$ in einem Zweigruppen-Vergleich mit einer globalen Power von $\pi = 1-\beta = 0.9$ nachgewiesen werden soll, ergibt sich für das Design nach Pocock (1977) ein Stichprobenumfang $n = 40$ pro Stufe, bei O'Brien und Fleming (1979) ein Umfang pro Stufe von $n = 34$. Die Aufteilung der globalen Power $1 - \beta$ ist unabhängig von der Höhe des nachzuweisenden Effekts θ. Während im fünfstufigen Design mit konstanten kritischen Schranken bereits 67% Power auf den ersten drei Stufen erzielt werden, liegen beim Design mit abfallenden kritischen Schranken 78% Power auf den letzten drei Stufen.

Wang und Tsiatis (1987) führen eine Klasse von kritischen Schranken für die Teststatistik Z_k aus (2.3) ein, die neben der Stufenzahl K und dem Niveau α von einem Parameter Δ abhängig sind. Die kritischen Schranken beim einseitigen Test des Problems (2.1) sind

Tabelle 2.2: *Erreichte Power $\pi(k)$ und pro Stufe erzielte Power π_k für das einseitige Testproblem im 5-stufigen Plan nach Pocock (1977) und O'Brien und Fleming (1979) bei globalem Niveau $\alpha = 0.025$ und globaler Power $\pi = 0.9$ für standardisierten Effekt $\theta = 0.5$*

	Pocock		O'Brien und Fleming	
Stufe	$\pi(k)$	π_k	$\pi(k)$	π_k
1	0.2027	0.2027	0.0010	0.0010
2	0.4600	0.2573	0.1222	0.1213
3	0.6679	0.2079	0.4600	0.3378
4	0.8091	0.1411	0.7446	0.2846
5	0.8957	0.0866	0.8957	0.1511

gegeben durch:

$$u_k = c(K, \alpha, \Delta) k^{\Delta - 0.5}, \quad k = 1, \ldots, K, \tag{2.6}$$

wobei $c(K, \alpha, \Delta)$ für eine von K, α und Δ abhängige Konstante steht. Die Wahl von Δ erfolgt in Abhängigkeit einer vorgegebenen Power $1 - \beta$ für ein festes $\theta_0 > 0$ so, dass der mittlere Stichprobenumfang (ASN = Average Sample Number) minimiert wird. Da der ASN dieser Schranken, wie auch die kritischen Werte selbst, mit dem der optimalen Schranken aus den Berechnungen von Pocock (1982) nahezu übereinstimmt, werden die Grenzen der Δ-Klasse approximativ optimal genannt. Bei einer Power um 0.5 bei einem vorgegebenen Wert aus der Alternative ergibt sich mit $\Delta = 0$ das Design von O'Brien und Fleming (1979) als approximativ optimal, während die Schranken von Pocock (1977) mit $\Delta = 0.5$ bei einer Power von ungefähr 0.95 bestimmt werden.

Die bisher beschriebenen Verfahren gehen von gleichen Stichprobenumfängen n auf jeder der K Stufen aus, was im tatsächlichen Studienverlauf zum Teil nicht zu verwirklichen ist, da eher eine Zwischenauswertung nach einem bestimmten Zeitintervall als nach einer bestimmten Menge an beobachteten Objekten realisierbar ist. In diesem Fall ist die Anzahl der bis zum jeweiligen Auswertungszeitpunkt erfassten Objekte unbekannt und kann jeweils unterschiedlich groß sein. Lan und DeMets (1983) sowie Kim und DeMets (1987) entwickelten daher den sogenannten *use function*- oder auch α *spending function*-Ansatz, bei dem weder der Zeitpunkt noch die Anzahl der Zwischenauswertungen festgesetzt ist, sondern nur der maximale Stichprobenumfang N.

Vor Beginn der Studie wird eine stetige monoton wachsende Funktion $\alpha(t)$ auf dem Intervall $[0, 1]$ definiert, die an der Stelle 1 den Wert α annimmt. Dabei gibt der Wert $\alpha(t) \leq \alpha$ zum Zeitpunkt t die bis dahin aufgebrauchte Wahrscheinlichkeit für den Fehler erster Art

an. Dabei entspricht t dem Anteil der bisher erhobenen Daten am maximalen Umfang N.

Die stetigen kritischen Grenzen u_k, $k = 1, 2, \ldots, K$, ergeben sich in Abhängigkeit von den einzelnen Auswertungszeitpunkten $t_1, t_2, \ldots, t_K$ rekursiv aus folgenden Gleichungen:

$$\begin{aligned} \mathrm{P}_{\mathrm{H}_0}\left(Z(t_1) \geq u_1\right) &= \alpha(t_1) \\ \mathrm{P}_{\mathrm{H}_0}\left(\bigcap_{j=1}^{i-1}\left(Z(t_j) < u_j\right), Z(t_i) \geq u_i\right) &= \alpha(t_i) - \alpha(t_{i-1}). \end{aligned} \tag{2.7}$$

Lan und DeMets (1983) schlagen unter anderem zwei α vergebende Funktionen vor, die bei gleichmäßigen Abständen zwischen den Auswertungen $t_i = i/K$ zu Schranken führen, die mit Pocock (1977) bzw. O'Brien und Fleming (1979) vergleichbar sind. Kim und DeMets (1987) empfehlen die Wahl von $\alpha(t)$ in Abhängigkeit von den Wünschen der Studienplaner. So kann entweder ein frühes Stoppen bei sehr großen Effekten oder ein konservativeres Vorgehen mit einer höheren Power bei niedrigen Effekten favorisiert werden. Die gruppensequentiellen Pläne nach Pocock (1977) und O'Brien und Fleming (1979) stellen dabei die Extrempositionen hinsichtlich dieser beiden Absichten dar, was Tabelle 2.2 anhand der erzielten Power pro Stufe in den beiden Designs verdeutlicht. Kritische Schranken bei Kenntnis bzw. Vorgabe der ungleichen Stichprobenumfänge pro Stufe werden u.a. bei Wassmer (2001) beschrieben.

Die Konstruktion der kritischen Schranken in den bisher beschriebenen Testverfahren beruht auf Anforderungen an die kritischen Schranken selbst, wie Konstanz, oder auf der Vorgabe des pro Stufe vergebenen Fehlers erster Art. Bauer (1992) berechnet kritische Schranken für normalverteilte Zufallsvariablen mit bekannter Varianz, die das globale Niveau α einhalten und bei gleich großen Stichprobenumfängen pro Stufe einen vorgegebenen Anteil der globalen Power bei einem Wert θ aus der Alternative vergeben. Für Studien mit zwei und drei Testzeitpunkten werden kritische Schranken bestimmt, wobei die Power entweder konstant oder proportional zur Inversen der Standardabweichung oder proportional zur Inversen der Varianz eines klassischen einstufigen Effektschätzers vergeben wird. In der angegebenen Reihenfolge führen die drei Strategien der Powervergabe bei konstantem maximalen Stichprobenumfang zu steigendem erwarteten Stichprobenumfang bei richtiger Verwerfung der Nullhypothese und steigender Power.

Umsetzungen des bereits von Pocock (1977) erwähnten Stopps wegen Aussichtslosigkeit hinsichtlich der Verwerfung der Nullhypothese liefern u.a. DeMets und Ware (1982) und Pampallona und Tsiatis (1994). Die Einführungen von Grenzen in Plänen zum einseitigen Testproblem, bei deren Unterschreitung die Studie zu Gunsten der Nullhypothese abgebrochen wird, führt zu einer Absenkung der kritischen Schranken.

Pocock (1982) empfiehlt, die maximale Anzahl an Stufen K nicht größer als fünf zu wählen, da die Erhöhung der Anzahl zu keinem bemerkenswerten Powergewinn bzw. nicht zu einer deutlichen Reduktion des mittleren Stichprobenumfangs (siehe auch McPherson, 1982) unter der Alternative führt. Der organisatorische Aufwand der Durchführung einer weiteren Zwischenauswertung ist im Verhältnis dazu sehr groß.

Die vorgestellten gruppensequentiellen Verfahren setzen die Normalverteilung der Zielvariablen mit bekannter Varianz voraus, werden jedoch auf weitere Fälle wie mit unbekannter Varianz, binärem Response oder nichtparametrische Tests übertragen bzw. approximativ verwendet (siehe z.B. O'Brien und Fleming, 1979; Pocock, 1977). Jennison und Turnbull (1991) bestimmen mittels Rekursion und numerischer Integration exakte kritische Schranken für gruppensequentielle Versuche mit t-, χ^2- oder F-Tests. Sie zeigen, dass sich die Fehlerrate erster Art bei Verwendung der kritischen Schranken für fünf Stufen von Pocock (1977) bei unbekannter Varianz bei einem geplanten Signifikanzniveau von 0.05 auf 0.056 erhöht. Die exakten kritischen Werte für den t-Test sind daher nur leichte Modifikationen der kritischen Schranken eines gruppensequentiellen Designs für normalverteilte Zufallsvariablen mit bekannter Varianz.

2.2 Adaptive Verfahren

In den bisher beschriebenen gruppensequentiellen Plänen besteht bereits die Möglichkeit, aufgrund eindeutiger Effekte eine Studie frühzeitig zu beenden. Die Anpassung des Designs, z.B. durch Fallzahlkorrektur, für die folgenden Stufen aufgrund von Erkenntnissen aus den vorangegangen Zwischenauswertungen ist jedoch nicht möglich. Adaptive Designs erlauben Änderungen am Studiendesign im Verlauf einer Studie unter Einbeziehung der bereits gewonnen Erkenntnisse bei Einhaltung des vorgegebenen Signifikanzniveaus. In der Literatur finden sich Vorschläge zur Änderung der Teststatistik, Auswahl von Behandlungsarmen und vor allem zur Anpassung des Stichprobenumfangs.

Die theoretische Grundlage für entsprechende Pläne geht auf Bauer (1989a) zurück, der auch ähnliche Überlegungen zur Kombination der Ergebnisse aufeinander folgender Studien, ähnlich wie bei einer Meta-Analyse, jedoch bei Abhängigkeit der Studiendesigns, anstellt (Bauer, 1989b). Bauer (1989a) schlägt vor, für eine Folge von maximal K Testproblemen $TP_{(1)}, \ldots, TP_{(K)}$, einen Kombinationstest auf Grundlage der P-Werte p_k, $k = 1, \ldots, K$, aus den einzelnen Testproblemen zum globalen Niveau α durchzuführen. Dabei werden die Testprobleme voneinander abhängig aus einer endlichen Menge von Testproblemen $\{TP_i, i \in I\}$, $I = \{1, \ldots, m\}$, mit zugehörigen Nullhypothesen H_{0i} gewählt. Der Kombinationstest richtet sich auf die globale Nullhypothese $\mathrm{H}_0 = \bigcap_{i=1}^{K} \mathrm{H}_{0i}$,

den Schnitt der individuellen Nullhypothesen zu den Testproblemen $TP_{(i)}$. Bei der Wahl des auf $TP_{(i)}$ folgenden Testproblems können alle Informationen verwendet werden, die bis dahin gesammelt wurden. Wesentlich ist dabei, dass für jedes der Testprobleme eine eigene Stichprobe gezogen wird, so dass die P-Werte unter H_0 unabhängig und identisch gleichverteilt auf dem Intervall $[0, 1]$ sind. Dies gilt für Teststatistiken mit stetigen Verteilungen. Bauer schlägt als geeigneten Kombinationstest den häufig in Meta-Analysen angewandten Test von Fisher vor, der das Produkt der P-Werte verwendet. Die globale Nullhypothese wird abgelehnt, wenn

$$\prod_{i=1}^{K} p_i \leq \exp\left(-\frac{1}{2}\chi^2(2K)_{1-\alpha}\right), \tag{2.8}$$

wobei $\chi^2(\nu)_\gamma$ das γ-Quantil der χ^2-Verteilung mit ν Freiheitsgraden darstellt.

Unter der Nullhypothese beträgt bei Gleichverteilung der P-Werte auf $[0, 1]$ die Wahrscheinlichkeit, mindestens eine der wahren Nullhypothesen H_{0i} fälschlicherweise zu verwerfen, höchstens α. Gilt die Annahme der identischen Gleichverteilung nicht, so hält der Kombinationstest nach Fisher das globale Niveau ein, wenn die Verteilungen der P-Werte unter H_0 stochastisch größer oder gleich der Gleichverteilung sind. Bei Brannath et al. (2002) wird die Forderung bzgl. der Verteilung der P-Werte erweitert. Die Verteilung des P-Werts p_j auf der j-ten Stufe, $j \geq 2$, bedingt auf die P-Werte der vorangegangen Stufen $p_1, \ldots, p_{j-1}$, muss unter der Nullhypothese stochastisch größer oder gleich der Gleichverteilung sein, damit das globale Niveau eingehalten wird, d.h.

$$P_{H_0}(p_j \leq \alpha | p_{j-1}, \ldots, p_1) \leq \alpha \quad \text{für alle} \quad 0 \leq \alpha \leq 1. \tag{2.9}$$

Diese Eigenschaft wird bei Brannath et al. (2002) mit „p clud" (clud = *conditionally larger than uniform distribution*) bezeichnet.

Dem Vorschlag von Bauer (1989a) schließt sich eine vielschichtige Diskussion an u.a. hinsichtlich der Realisierbarkeit der Auswahl aller Studienhypothesen bzw. Testprobleme TP_i, $i = 1, \ldots, m$, vor Beginn der Studie sowie der Interpretierbarkeit der globalen Hypothese. Der sehr allgemein formulierte Ansatz wird in der Literatur nur selten für den Test unterschiedlicher Hypothesen auf den einzelnen Stufen konkretisiert, zum Beispiel beim Wechsel zwischen Überlegenheits- und Nicht-Unterlegenheitshypothese (siehe Shih et al., 2004; Wang et al., 2001). Vorwiegend wird der Ansatz zur Anpassung des Stichprobenumfangs beim wiederholten Testen derselben Nullhypothese verwendet. Im Folgenden werden neben der Erweiterung des adaptiven Designs unter Verwendung des Kombinationstests nach Fisher noch die Inverse-Normalmethode sowie alternative Konstruktionsmethoden für adaptive Designs vorgestellt.

2.2.1 Kombinationstests

Kombinationstest nach Fisher, Inverse-χ^2-Methode

Bauer und Köhne (1994, 1996) führen die Idee von Bauer (1989a) für zwei Stufen weiter aus, um Ergebnisse und Beobachtungen einer sogenannten internen Pilotstudie in die Planung einer zweiten Stufe zu integrieren. Dabei sind Design- und Hypothesenänderungen denkbar, wie z.B. der Übergang von einer Dosis-Wirkungs-Studie auf der ersten Stufe zu einem Vergleich einer Dosis mit einem Placebo auf der zweiten Stufe oder der Wechsel des primären Endpunkts. Werden ausschließlich die Stichprobenumfänge angepasst, sind die (einseitigen) Nullhypothesen H_{0i} aus der allgemeinen Hypothese $H_0 = \bigcap_{i=1}^{K} H_{0i} = H_{01}$ identisch. Die Wahl einseitiger Testprobleme empfiehlt sich, um in der Richtung widersprüchliche Ergebnisse auf einzelnen Stufen zu vermeiden.

Voraussetzung für die Anwendbarkeit des Kombinationstests zum Test der globalen Nullhypothese ist die Unabhängigkeit und Gleichverteilung der P-Werte auf dem Intervall $[0, 1]$ unter der Nullhypothese. Diese Voraussetzung wird durch die strenge Trennung der Daten vor und nach der Interimsanalyse durch Ziehung unabhängiger Stichproben für jede Stufe und die Verwendung einer stetigen Teststatistik gewährleistet. Dabei ist zu bemerken, dass bei einer Adaption der Stichprobengrößen diese nicht als Gewichtung für die P-Werte in den Kombinationstest eingehen dürfen.

Beim Kombinationstest nach Fisher (2.8) ist zu beachten, dass bei Unterschreitung der kritischen Schranke durch das Produkt der P-Werte zu einer Stufe $k < K$ die globale Nullhypothese bereits vor der Durchführung aller K Stufen abgelehnt werden kann, da das Produkt der P-Werte auf allen weiteren Stufen wegen $p_i \leq 1$, für alle $i = 1, \ldots, K$, stets kleiner oder gleich $\prod_{i=1}^{k} p_i$ ist. Dieses Ereignis wird mit nicht-stochastischer Verkürzung (*non-stochastic curtailment*) bezeichnet.

Zunächst für den zweistufigen Fall mit einem globalen Signifikanzniveau von α führen Bauer und Köhne (1994) neben einer kritischen Schranke α_1 mit $0 \leq \alpha_1 < \alpha$ für frühzeitiges Verwerfen von H_0 auf der ersten Stufe eine obere Schranke α_0 mit $\alpha \leq \alpha_0 \leq 1$ für das Stoppen bei ungenügenden Effekten ein. Die Studie endet so bereits nach der ersten Stufe mit der Annahme von H_0, wenn $p_1 \geq \alpha_0$ ist, bzw. mit der Ablehnung von H_0, sofern $p_1 \leq \alpha_1$. Bei Vorgabe der globalen Fehlerwahrscheinlichkeit erster Art α berechnet sich der kritische Wert auf der zweiten Stufe für das Produkt der P-Werte als $c_\alpha = \exp(-0.5 \cdot \chi^2(4)_{1-\alpha})$ gemäß (2.8). Bei einem gewählten α_0 ergibt sich der kritische Wert α_1 für die erste Stufe im Intervall $[c_\alpha; \alpha]$ aus der Gleichung für die gesamte

Fehlerwahrscheinlichkeit erster Art,

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^{c_\alpha/p_1} \partial p_2 \partial p_1 = \alpha_1 + c_\alpha \cdot (\ln \alpha_0 - \ln \alpha_1)\,. \tag{2.10}$$

Nach Konstruktion hält dieses Testverfahren das globale Niveau α ein, unabhängig von der Wahl des Stichprobenumfangs n_2 für die zweite Stufe.

Bauer und Köhne (1994) konstruieren auf ähnliche Weise einen Test mit drei Stufen. Ein Abbruch der Studie nach der ersten oder zweiten Stufe mit Annahme der Nullhypothese erfolgt, wenn p_1 oder p_2 größer als der Wert α_0 ausfallen. Die Studie endet mit der Ablehnung der Nullhypothese auf der ersten Stufe, wenn $p_1 \leq \alpha_1$ gilt. Der kritische Wert auf der dritten Stufe für das Produkt der P-Werte der drei Stufen ist $c_\alpha = \exp(-0.5 \cdot \chi^2(6)_{1-\alpha})$. Zur Vermeidung von widersprüchlichen Ergebnissen, wird der kritische Wert $c_{\alpha_2} = \exp(-0.5 \cdot \chi^2(4)1 - \alpha_2)$ auf der zweiten Stufe gleich c_α/α_0 gesetzt, so dass bei Vorgabe von α_0 und α nur das Niveau der ersten Stufe α_1 bestimmt werden muss. Wassmer (1999) betrachtet Designs mit mehr als zwei Zwischenauswertungen und unterschiedlicher Vergabe des globalen Niveaus bei der Möglichkeit eines Abbruchs wegen Aussichtslosigkeit, indem die kritischen Schranken des Kombinationstests nach Fisher $c_{\alpha_k} = \exp(-0.5 \cdot \chi^2(2k)_{1-\alpha_k})$ durch Wahl der nominalen Niveaus α_k, $k = 1, \ldots, K$ auf den einzelnen Stufen variiert werden.

Die Kombinationsregel nach Fisher (2.8) lässt sich unter Anwendung der Inversen der Verteilungsfunktion der χ^2-Verteilung mit zwei Freiheitsgraden $F^{-1}_{\chi^2(2)}(x)$ auf die P-Werte der einzelnen Stufen umschreiben. Die globale Nullhypothese wird zum Niveau α abgelehnt, wenn eine der folgenden Ungleichungen erfüllt ist:

$$\begin{aligned} \prod_{i=1}^{K} p_i &\leq \exp\left(-\frac{1}{2}\chi^2(2K)_{1-\alpha}\right) \\ -2\sum_{i=1}^{K} \ln(1-(1-p_i)) &\geq \chi^2(2K)_{1-\alpha} \\ \sum_{i=1}^{K} F^{-1}_{\chi^2(2)}(1-p_i) &\geq \chi^2(2K)_{1-\alpha}. \end{aligned} \tag{2.11}$$

Mittels dieser auch Inverse-χ^2-Methode genannten Transformation der P-Werte konstruiert Hartung (2006) kritische Schranken eines adaptiven Designs für die kumulierten Teststatistiken $S_k = \sum_{i=1}^{K} F^{-1}_{\chi^2(\nu_i)}(1-p_i)$, $k = 1, \ldots, K$, vergleichbar mit (2.5) in den gruppensequentiellen Designs. Die Freiheitsgrade ν_k, $k = 1, \ldots, K$, werden konstant entweder gleich 1 oder 2 gewählt. Die kritischen Schranken werden durch numerische Integration so bestimmt, dass sie ähnlich den gruppensequentiellen Designs von Pocock (1977) und O'Brien und Fleming (1979) bei steigender Stufenzahl fallend bzw. konstant sind. Im De-

sign vom sogenannten O'Brien und Fleming-Typ ergibt sich ein globaler kritischer Wert $\chi^2(\nu_i \cdot K)_{1-\alpha}$ auf jeder der K Stufen.

Inverse-Normalmethode

Eine weitere Möglichkeit, die P-Werte der einzelnen Stufen zu kombinieren, ist der Ansatz von Lehmacher und Wassmer (1999), der ebenfalls aus der Meta-Analyse stammt (Hedges und Olkin, 1985). Für die P-Werte der K maximal durchzuführenden Stufen werden wieder die Unabhängigkeit und die Gleichverteilung unter H_0 auf dem Intervall $[0, 1]$ vorausgesetzt. Zur Bewahrung der Richtung der Beobachtungen werden sowohl bei der einseitigen als auch bei der zweiseitigen Fragestellung die einseitigen P-Werte verwendet. Der P-Wert jeder Stufe wird mit der inversen Verteilungsfunktion der Standardnormalverteilung Φ^{-1} zu $z_k = \Phi^{-1}(1 - p_k)$, $k = 1, \ldots, K$, transformiert. Da die Zufallsvariablen z_k unter der Nullhypothese standardnormalverteilt sind, lässt sich analog zur Teststatistik (2.3) in den gruppensequentiellen Designs eine Teststatistik auf der k-ten Stufe konstruieren:

$$Z_k = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \Phi^{-1}(1 - p_i). \tag{2.12}$$

Die Statistik Z_k weist unter der Nullhypothese die gleichen Verteilungseigenschaften wie die Teststatistik (2.3) auf. Die Verwendung von kritischen Schranken aus bekannten gruppensequentiellen K-stufigen Tests wie zum Beispiel nach O'Brien und Fleming (1979) oder Pocock (1977) ermöglicht die Durchführung eines exakten Tests zum vorgegebenen Niveau α. Die Verteilungseigenschaft der Teststatistiken z_k auf den einzelnen Stufen und damit das Niveau des Kombinationstests bleiben bei adaptiver Wahl der Stichprobenumfänge auf den einzelnen Stufen gültig.

Die Kombinationsstatistik (2.12) führt wie auch der Kombinationstest nach Fisher zu gleicher Gewichtung der P-Werte aller betrachteten Stufen, unabhängig vom eingehenden Stichprobenumfang. Eine Gewichtung der Stufen kann durch im Voraus festgelegte Gewichte w_k, $k = 1, \ldots, K$, erfolgen. Für die Inverse-Normalmethode ergibt sich eine gewichtete Teststatistik auf der k-ten Stufe

$$Z_k^w = \sum_{i=1}^{k} \frac{w_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \cdot \Phi^{-1}(1 - p_i), \tag{2.13}$$

wobei für die nichtnegativen Gewichte $\sum_{i=1}^{K} w_i^2 = 1$ gilt (Wassmer et al., 2001). Werden entsprechend der Wahl der Gewichte Stichprobenumfänge gemäß einem gruppensequentiellen Design geplant, gelten die Eigenschaften des gruppensequentiellen Verfahrens, sofern

ein normalverteilter Response mit bekannter Varianz erhoben wird und nach einer Interimsanalyse keine Änderung des Designs erfolgt. Bei Vorgabe der anfangs geplanten Stichprobenumfänge $n_1, \ldots, n_K$ auf den K Stufen ergibt sich für den transformierten P-Wert der k-ten Stufe das Gewicht $w_k = \sqrt{n_k / \sum_{i=1}^{K} n_i}$. Mit Hilfe dieser konstanten Gewichte konstruieren auch Cui et al. (1999) für normalverteilte Zufallsvariablen mit bekannter Varianz eine Erweiterung gruppensequentieller Pläne. Die Stichprobenumfänge können beliebig und in Abhängigkeit der bereits beobachteten Daten gewählt werden, während die Gewichte im Verlauf der Studie nicht mehr verändert werden dürfen. Dies ist sowohl bei der Planung der Studie zu Beginn als auch bei der Anpassung des Stichprobenumfangs zu berücksichtigen.

Basierend auf den kombinierten Teststatistiken (2.12) und (2.13) lassen sich adaptive Designs mittels kritischer Schranken aus gruppensequentiellen Plänen u.a. mit Abbruch wegen Aussichtslosigkeit konstruieren. Bei Beibehaltung der im Voraus geplanten Stichprobenumfänge erzielen diese Designs mindestens eine Power $1 - \beta$ für einen festgelegten Wert aus der Alternative.

Eine allgemeine Beschreibung der Konstruktion adaptiver Designs mittels Kombinationstests geben Schäfer et al. (2006). Zum Erhalt des globalen Fehlers erster Art beim Test einer einzelnen Punktnullhypothese auf allen Stufen werden folgende grundlegende Voraussetzungen genannt:

1. die stochastische Unabhängigkeit der kombinierten Teststatistiken der K Stufen,
2. der Erhalt der anfänglich geplanten Randverteilungen der Teststatistiken und
3. die stochastische Unabhängigkeit der Teststatistik der i-ten Stufe von der gesamten Information, die für Designänderungen vor dem Zeitpunkt i verwendet wurde.

2.2.2 Bedingte Fehlerfunktion

Proschan und Hunsberger (1995) stellen eine weitere Konstruktionsmethode vor, eine Studie um eine zweite Stufe zu erweitern und das globale Niveau α einzuhalten, indem sie die sogenannte bedingte Fehlerfunktion (*conditional error function*) einführen. Dabei handelt es sich um eine je nach Design zu bestimmende Funktion $A(z_1)$ der normalverteilten Teststatistik z_1 der ersten Stufe. Die Funktion $A(z_1)$ muss stetig sein und den Wertebereich $[0, 1]$ besitzen. Weiter muss der Erwartungswert der bedingten Fehlerfunktion bezüglich z_1 unter der Nullhypothese dem globalen Signifikanzniveau α entsprechen:

$$\alpha = \int_{-\infty}^{\infty} A(z_1)\phi(z_1)\partial z_1. \tag{2.14}$$

Dabei steht $\phi(x)$ für die Dichte der Standardnormalverteilung. Der Wert der bedingten Fehlerfunktion gibt die bedingte Irrtumswahrscheinlichkeit erster Art für die zweite Stufe in Abhängigkeit vom Ergebnis z_1 der ersten Stufe an. Diese Funktion der bedingten Irrtumswahrscheinlichkeit ist jeweils vor Beginn einer Studie festzulegen.

Die Konstruktion führen Proschan und Hunsberger (1995) für den einseitigen Vergleich der Mittelwerte zweier Normalverteilungen (vgl. (2.1)) mit gleicher Varianz σ^2 und für zwei Stufen aus. Nach $n_1/2$ Beobachtungen pro Gruppe auf der ersten Stufe erhält man die Teststatistik $z_1 = \sqrt{n_1/(4\hat{\sigma}_1^2)}(\bar{X}_{11} - \bar{X}_{12})$, wobei $\bar{X}_{ij}$ für den Mittelwert der j-ten Behandlungsgruppe auf der i-ten Stufe steht. Unter der Annahme, dass die geschätzte Varianz $\hat{\sigma}_1^2$ der ersten Stufe gleich der unbekannten Varianz σ^2 ist, lässt sich die bedingte Fehlerwahrscheinlichkeit erster Art auf der zweiten Stufe bei Verwendung einer gepoolten Teststatistik berechnen und in Abhängigkeit vom Stichprobenumfang der zweiten Stufe maximieren. Die maximale Erhöhung des Fehlers erster Art auf $\alpha + \exp(-u_{1-\alpha}^2/2)/4$ beim zweimaligen Test zum Niveau α wird durch Anpassung des kritischen Werts c_A und durch die Einführung einer unteren Schranke $u_{1-\alpha_0}$ korrigiert. Die Unterschreitung der Schranke $u_{1-\alpha_0}$ durch z_1 führt zu einem Abbruch der Studie wegen Aussichtslosigkeit mit Annahme der Nullhypothese nach der ersten Stufe.

Aus der Bestimmung des maximalen Fehlers erster Art leitet sich die Klasse der zirkulären bedingten Fehlerfunktionen A_{cir} ab:

$$A_{\text{cir}}(z_1) = \begin{cases} 0, & \text{wenn } z_1 < u_{1-\alpha_0}, \\ 1 - \Phi(\sqrt{c_A^2 - z_1^2}), & \text{wenn } u_{1-\alpha_0} \leq z_1 < c_A, \\ 1, & \text{wenn } z_1 \geq c_A. \end{cases} \tag{2.15}$$

Nach Vorgabe von α_0 wird der kritische Wert c_A bestimmt, so dass $A_{\text{cir}}(z_1)$ die Gleichung (2.14) erfüllt. Unabhängig von der Wahl des Stichprobenumfangs n_2 wird das vorgegebene Niveau α nicht überschritten, jedoch, sofern n_2 nicht gemäß den Maximierungsüberlegungen gewählt wird, häufig nicht voll ausgeschöpft.

Um das Niveau voll auszuschöpfen, wird eine kritische Schranke c für die zweite Stufe in Abhängigkeit des Stichprobenumfangs n_2 so bestimmt, dass die folgende Gleichung erfüllt ist:

$$\text{CP}_0(n_2, c|z_1) = A(z_1). \tag{2.16}$$

CP_θ steht dabei für die bedingte Wahrscheinlichkeit, auf der zweiten Stufe die Nullhypothese abzulehnen, wenn θ der wahre Parameter ist und z_1 bereits vorliegt. Für die zirkuläre bedingte Fehlerfunktion ergibt sich der kritische Wert in Abhängigkeit vom Wert der be-

dingten Fehlerfunktion an der Stelle z_1 und n_2 zu

$$c = \frac{\sqrt{n_1}z_1 + \sqrt{n_2} \cdot u_{1-A(z_1)}}{\sqrt{n_1 + n_2}}, \tag{2.17}$$

wobei $u_{1-A(z_1)}$ das $(1 - A(z_1))$-Quantil der Standardnormalverteilung ist. Wird der Stichprobenumfang der zweiten Stufe als $n_2 = n_1 \cdot ((c_A/z_1)^2 - 1)$ gewählt, ergibt sich der maximale kritische Wert $c = c_A$ aus (2.15).

Als weitere Klasse von Fehlerfunktionen stellen Proschan und Hunsberger (1995) die linearen bedingten Fehlerfunktionen A_l vor. Sie besitzen die Form

$$A_l(z_1) = \begin{cases} 0, & \text{wenn } z_1 < u_{1-\alpha_0}, \\ \Phi(a + bz_1), & \text{wenn } u_{1-\alpha_0} \leq z_1 < c_A, \\ 1, & \text{wenn } z_1 \geq c_A, \end{cases} \tag{2.18}$$

wobei a, b, α_0 und c_A Bedingung (2.14) erfüllen müssen. Es ist zu berücksichtigen, dass bei unbedachter Wahl der Parameter, ein Ergebnis auf der zweiten Stufe, das im Wertebereich der Nullhypothese liegt, zu einer Ablehnung der Nullhypothese führen kann. Proschan und Hunsberger (1995) bestimmen keine optimale bedingte Fehlerfunktion, sondern empfehlen die Wahl der Funktion vom Grad der Unsicherheit bezüglich des Zielparameters und der Störgrößen abhängig zu machen.

Posch und Bauer (1999) greifen die Idee der bedingten Fehlerfunktion auf und betten die auf Kombinationstests basierenden Verfahren von Bauer und Köhne (1994) und Lehmacher und Wassmer (1999) (siehe Abschnitt 2.2.1) in diese Theorie ein, indem sie für diese adaptiven Designs die bedingten Fehlerfunktionen für zwei Stufen bestimmen.

Für die Inverse-Normalmethode lautet die zugehörige bedingte Fehlerfunktion mit kritischen Schranken $u_{1-\alpha_i} = \Phi^{-1}(1 - \alpha_i)$, $i = 0, 1, 2$:

$$A_{\text{LW}}(z_1) = \begin{cases} 0, & \text{wenn } z_1 < u_{1-\alpha_0}, \\ 1 - \Phi(\sqrt{2}u_{1-\alpha_2} - z_1), & \text{wenn } u_{1-\alpha_0} \leq z_1 < u_{1-\alpha_1}, \\ 1, & \text{wenn } z_1 \geq u_{1-\alpha_1}. \end{cases} \tag{2.19}$$

Für das Verfahren nach Bauer und Köhne (1994) ergibt sich eine bedingte Fehlerfunktion in Abhängigkeit von z_1

$$A_{\text{BK}}(z_1) = \begin{cases} 0, & \text{wenn } z_1 < u_{1-\alpha_0}, \\ \frac{c_{\alpha_2}}{1-\Phi(z_1)}, & \text{wenn } u_{1-\alpha_0} \leq z_1 < u_{1-\alpha_1}, \\ 1, & \text{wenn } z_1 \geq u_{1-\alpha_1}. \end{cases} \tag{2.20}$$

Wassmer (2001) und Wassmer et al. (2001) stellen eine verallgemeinerte Form der bedingten Fehlerfunktion vor, in der sie als nicht steigende Funktion im P-Wert p_1 der ersten Stufe formuliert wird.

Ebenso ist die Darstellung von Designs, die gemäß einer bedingten Fehlerfunktion konstruiert werden, als Kombinationstest möglich. Für die zirkuläre Fehlerfunktion aus (2.15) ergibt sich die kombinierte Teststatistik

$$Z_{\mathrm{PH}} = \sqrt{(\Phi^{-1}(1-p_1))^2 + (\Phi^{-1}(1-p_2))^2} \tag{2.21}$$

(siehe Wassmer, 2001). Proschan (2003) interpretiert die bedingten Fehlerfunktionen im positiven Quadranten des Koordinatensystems für die normalverteilten Zufallsvariablen z_1 und z_2 eines zweistufigen Designs geometrisch. Es zeigt sich, dass die zirkuläre Fehlerfunktion und die nach Bauer und Köhne (1994) ähnliche Eigenschaften besitzen. Während die lineare Fehlerfunktion (2.18) mit $b = 1$ bei einem Verhältnis von z_1/z_2 nahe 1 häufiger zur Ablehnung führt, wird bei größeren Unterschieden zwischen z_1 und z_2 bei den beiden anderen Fehlerfunktionen häufiger die Nullhypothese abgelehnt. Daher empfiehlt Proschan (2003), die lineare Fehlerfunktion dann anzuwenden, wenn nur moderate Abweichungen von a priori gleich groß geplanten Stichprobenumfängen $n_1 = n_2$ zu erwarten sind, da dann die Erwartungswerte von z_1 und z_2 gleich groß sind.

Die bisher vorgestellten Verfahren sind jeweils durch eine vorgegebene Anzahl an maximal durchzuführenden Stufen charakterisiert. Sie erlauben das Testen der Nullhypothese zu lokalen Signifikanzniveaus auf den einzelnen Stufen sowie die Einführung von Schranken zum Abbruch zugunsten der Nullhypothese wegen Aussichtslosigkeit. Die im folgenden Abschnitt beschriebenen Designs ermöglichen hingegen eine adaptive Wahl der Anzahl Studienabschnitte in Abhängigkeit der bereits realisierten Stufen.

2.2.3 Varianzvergabe

Der sogenannte *Self-Designing*-Ansatz nach Fisher (1998) und Shen und Fisher (1999) dient zunächst der Regulierung der Studiendauer und des Stichprobenumfangs während des Studienverlaufs und ist durch eine offene Anzahl an Studienabschnitten gekennzeichnet. Das Ende der Studie wird durch adaptive Vergabe von Gewichten für die einzelnen Studienabschnitte, die die Varianz einer finalen Teststatistik aufteilen (*variance spending*), und entsprechende Wahl der Stichprobenumfänge gesteuert. Nach vollständiger Aufteilung der Varianz erfolgt ein Test auf die Nullhypothese, wobei der kritische Wert zu Beginn der Studie bezüglich der finalen Teststatistik gewählt wird. Das primäre Ziel des Self-Designing-Ansatzes liegt in der Anpassung des Studiendesigns an die gewonnenen Informationen aus der Interimsanalyse. Ein vorzeitiger Abbruch mit Ablehnung der Nullhypothese kann nur indirekt durch Verminderung des Stichprobenumfangs nach einer Interimsanalyse erreicht werden. Schranken für einen Abbruch zugunsten der Nullhypothese wegen Aussichtslosigkeit lassen sich integrieren. Ein weiteres Kennzeichen ist die

meist sehr einfache Bestimmung des kritischen Werts für die abschließende Testentscheidung nach Vergabe der Gesamtvarianz.

Shen und Fisher (1999) betrachten den einseitigen Test beim Vergleich der Mittelwerte zweier Normalverteilungen mit bekannter Varianz σ^2 gemäß (2.1). Ausgehend von einer vorgegebenen Folge von Stichprobenumfängen n_i, $i = 1, 2, 3, \ldots$, pro Stufe wird auf jeder Stufe die Teststatistik $z_i = \sqrt{n_i/(4\sigma^2)}(\bar{X}_{i1} - \bar{X}_{i2})$ berechnet. Die tatsächlich ausgeführte Anzahl Stufen K ist in diesem Design eine endliche Zufallsvariable. Die finale Teststatistik, in die in Abhängigkeit der Ergebnisse der vorangegangenen Interimsanalysen gewählte Gewichte $w_i = w_i(z_1, \ldots, z_{i-1})$, $i = 1, \ldots, K$, eingehen, lautet

$$Z_K = \sum_{i=1}^{K} w_i z_i = \sum_{i=1}^{\infty} w_i z_i. \tag{2.22}$$

Für die Gewichte muss wie bei der gewichteten Inverse-Normalmethode (vgl. (2.13)) gelten, dass sie nichtnegativ sind. Vorausgesetzt wird, dass ein positives endliches K existiert mit $\sum_{i=1}^{K} w_i^2 = 1$. Die Gewichte w_k, $k > K$, sind in (2.22) gleich Null zu setzen. Ist die Existenz eines endlichen K gesichert, ist die Teststatistik Z_K unter der Nullhypothese standardnormalverteilt. Daraus lässt sich die Entscheidungsregel ableiten, dass nach vollständiger Vergabe der Gesamtvarianz 1 durch die Gewichte w_i, $i = 1, \ldots, K$, die Nullhypothese zum Niveau α auf der K-ten Stufe abgelehnt wird, wenn $Z_K > u_{1-\alpha}$.

Für einen Abbruch zugunsten der Nullhypothese werden vor Beginn der Studie Schranken u_{Lk} vorgegeben. Werden diese durch eine Teststatistik

$$U_k = \sum_{i=1}^{k} \sqrt{n_i} z_i / n_\Sigma(k) \text{ mit } n_\Sigma(k) = \sum_{i=1}^{k} n_i \tag{2.23}$$

auf der k-ten Stufe unterschritten, führt dies zum Studienende wegen Aussichtslosigkeit hinsichtlich des Verwerfens des Nullhypothese, auch wenn die Gesamtvarianz der finalen Teststatistik noch nicht vergeben ist.

Ein Vorschlag von Shen und Fisher (1999) für die Wahl der Gewichte orientiert sich am Stichprobenumfang des einstufigen Designs bei Vorgabe der Fehlerraten erster und zweiter Art sowie dem zu entdeckenden Unterschied θ

$$N_{\text{fix}} = 4\sigma^2 \frac{(u_{1-\alpha} + u_{1-\beta})^2}{\theta^2}. \tag{2.24}$$

Das Gewicht der ersten Stufe wird gemäß dem Stichprobenumfang der ersten Stufe n_1 im Verhältnis zu N_{fix} mit $w_1 = \sqrt{n_1/N_{\text{fix}}}$ gewählt. Auf den folgenden Stufen wird, sofern die Studie nicht wegen Aussichtslosigkeit auf der $(k-1)$-ten Stufe abgebrochen wird, ein Stichprobenumfang N_k bestimmt, der bedingt auf den bereits beobachteten Ergebnissen zu einer bedingten Power von $1 - \beta$ auf der k-ten Stufe führt, sofern nach dem k-ten

Studienabschnitt die finale Testentscheidung getroffen wird. Dabei kann als nachzuweisender Wert ein Effektschätzer $\hat{\theta}_{k-1}$ verwendet werden, der auf den Ergebnissen der $(k-1)$ durchgeführten Stufen basiert. Der Stichprobenumfang N_k berechnet sich gemäß

$$N_k = \frac{4\sigma^2}{\hat{\theta}^2_{k-1}} \cdot \left(\frac{u_{1-\alpha} - \sum_{i=1}^{k-1} w_i z_i}{\sqrt{1 - \sum_{i=1}^{k-1} w_i^2}} + u_{1-\beta} \right)^2 \tag{2.25}$$

und das vorgeschlagene Gewicht bei Verwendung des vorgegebenen Stichprobenumfangs n_k gemäß

$$w_k = \sqrt{\frac{n_k \cdot \left(1 - \sum_{i=1}^{k-1} w_i^2\right)}{N_k}}. \tag{2.26}$$

Sobald die bedingte Power bei Verwendung des Stichprobenumfangs n_k größer oder gleich $1-\beta$ ist, wird das verbleibende Gewicht $w_K = w_k = 1 - \sum_{i=1}^{k-1} w_i$ vergeben und die letzte Stufe der Studie durchgeführt.

Eine alternative Gewichtung mit $w_k = \sqrt{n_k/(2N_k)}$ für $k \geq 2$ führt zu stärkerer Gewichtung der zu Beginn durchgeführten Studienabschnitte. Beide Gewichtungen führen bei positivem Behandlungseffekt zu einem schnellen Studienende, da ein großer Effekt eine Senkung des benötigten Stichprobenumfangs N_k und damit große Gewichte w_k nach sich zieht.

Shen und Fisher (1999) schlagen vor, andere Testprobleme durch Transformation in eine entsprechende Struktur zu überführen und so das Verfahren approximativ zu erweitern. Die Auswirkungen auf das Niveau bei Übertragung dieses Ansatzes auf Transformationen beim Vergleich von Wahrscheinlichkeiten oder Verwendung des log-Rank-Tests sind nicht näher beschrieben.

Hartung (2001) überträgt durch die Kombination der P-Werte mittels der Inverse-Normalmethode die Ideen von Fisher (1998) auf beliebige Zielgrößen und zeigt eine Möglichkeit, neben der adaptiven Wahl der Gewichte auch die Umfänge auf den einzelnen Stufen an die vorangegangenen Daten anzupassen. Die Anzahl an durchgeführten Stufen wird wie bei Shen und Fisher (1999) durch eine endliche Zufallsvariable K beschrieben und ein einseitiges Testproblem bzgl. des Parameters θ betrachtet. Dabei wird vorausgesetzt, dass die Stufen disjunkt und die P-Werte auf den einzelnen Stufen unter der Nullhypothese auf dem Intervall $[0, 1]$ gleichverteilt sind. Der P-Wert der k-ten Teststatistik wird mit der Inversen der Normalverteilung zur Statistik $z_k = \Phi^{-1}(1 - p_k)$ transformiert, die somit unter der Nullhypothese standardnormalverteilt ist. Jeder Stufe wird in Abhängigkeit der Ergebnisse der vorangegangen Stufen bzw. auf Grundlage von Vorinformation auf der ersten Stufe ein Gewicht $w_k = w_k(z_1, \ldots, z_{k-1})$, $k = 1, \ldots, K$, zugeordnet, mit

$\sum_{k=1}^{K} w_k^2 = \sum_{k=1}^{\infty} w_k^2 = 1$. Die kombinierte Statistik, die unter der Nullhypothese standardnormalverteilt ist, lautet wie in (2.22) $Z_K = \sum_{k=1}^{K} w_k z_k = \sum_{k=1}^{\infty} w_k z_k$. Entsprechend wird die Studie mit dem Verwerfen der Nullhypothese zum Niveau α nach der K-ten Stufe abgebrochen, falls $Z_K > \Phi^{-1}(1-\alpha)$, oder bei Vorgabe unterer Schranken U_{Lk} wegen Aussichtslosigkeit auf Stufe k beendet. Die Verteilungseigenschaften von p_k, z_k und Z_K bleiben unter der Nullhypothese bei adaptiver Wahl des Stichprobenumfangs auf den einzelnen Stufen bestehen.

Nach Wahl des Stichprobenumfangs n_1 und des Gewichts w_1 erfolgt die Bestimmung von n_k und w_k nach vorgegebenen Regeln, die das Design der Studie in Abhängigkeit der Ergebnisse der Zwischenauswertungen steuern. Neben den globalen Fehlerraten erster und zweiter Art, α und β, wird eine bedingte Fehlerrate zweiter Art $\beta_{g,k} > \beta$ konstant oder in Abhängigkeit der Stufenzahl k und ein minimales Gewicht $w_{\min} > 0$ festgelegt.

Analog zu Shen und Fisher (1999) werden die Gewichte in Abhängigkeit des Stichprobenumfangs M_k gewählt, der benötigt wird, um nach Durchführung von $(k-1)$ Stufen auf der k-ten Stufe bei Vergabe des restlichen Gewichts $w_k = \sqrt{1 - \sum_{i=1}^{k-1} w_i^2}$ eine bedingte Power von β zu erzielen. Dabei wird die bedingte Fehlerrate erster Art

$$\alpha_k^* = 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha) - Z_{k-1}}{\sqrt{1 - \sum_{i=1}^{k-1} w_i^2}}\right), \tag{2.27}$$

die einer bedingten Fehlerfunktion (siehe Abschnitt 2.2.2) entspricht, in einer Stichprobenumfangsfunktion $f_k(a, b, \delta)$ zum behandelten Testproblem verwendet. Die gewichtete Teststatistik ergibt sich mit $Z_{k-1} = \sum_{i=1}^{k-1} w_i z_i$. Neben den Fehlerraten erster und zweiter Art, a und b, und dem nachzuweisenden Effekt δ können in die Funktion $f_k(a, b, \delta)$ Störgrößen eingehen, die wie der Effekt aus den bereits vorliegenden Daten geschätzt werden. Die Stichprobenumfänge M_k für eine bedingte Power von $1-\beta$ und m_k für eine bedingte Power von $1-\beta_{g,k}$ stellen sich somit dar als

$$M_k = f_k(\alpha_k^*, \beta, \hat{\theta}_{k-1}) \quad \text{und} \quad m_k = f_k(\alpha_k^*, \beta_{g,k}, \hat{\theta}_{k-1}). \tag{2.28}$$

Für den Stichprobenumfang m_k kann implizit aus $m_k = f_k(\alpha_{\beta,k}^*, \beta, \hat{\theta}_{k-1})$ ein Niveau $\alpha_{\beta,k}^*$ bestimmt werden, dass bei Vorliegen von $\hat{\theta}_{k-1}$ zu einer bedingten Power von $1-\beta$ führt.

Hartung (2001) schlägt bezüglich der Gewichtsvergabe auf der k-ten Stufe folgende drei

Möglichkeiten von Anteilen $\epsilon_k = \epsilon_k^i$, $i = 1, 2, 3$, am verbleibenden Gewicht vor:

$$
\begin{aligned}
W(k) &= \sqrt{1 - \sum_{i=1}^{k-1} W(j)^2 \cdot \epsilon_k} \quad \text{mit} \\
\epsilon_k^1 &= \frac{\Phi^{-1}\left(1 - \frac{\alpha^*_{\beta,k}}{2}\right)}{\Phi^{-1}\left(1 - \frac{\alpha^*_k}{2}\right)}, \\
\epsilon_k^2 &= \frac{m_k}{M_k} \quad \text{oder} \\
\epsilon_k^3 &= \frac{\Phi^{-1}(1 - p_k^{E,m_k})}{\Phi^{-1}(1 - p_k^{E,M_k})},
\end{aligned}
\tag{2.29}
$$

wobei $p_k^{E,m}$ den P-Wert des Erwartungswerts der gewählten Teststatistik bei Verwendung des Umfangs m und bei Vorliegen des Effekts $\hat{\theta}_{k-1}$ darstellt. Die Gewichte und Stichprobenumfänge werden unter Berücksichtigung des minimalen Gewichts $w_{\min}$ wie folgt festgelegt:

$$
w_k = \begin{cases} W(k), & \text{wenn } W(k) > w_{\min}, \\ \sqrt{1 - \sum_{j=1}^{k-1} w_j^2}, & \text{sonst}, \end{cases}
\tag{2.30}
$$

$$
n_k = \begin{cases} m_k, & \text{wenn } W(k) > w_{\min}, \\ M_k, & \text{sonst}. \end{cases}
\tag{2.31}
$$

Die Wahl von ϵ_k^1 führt im Allgemeinen zu einer niedrigeren Stufenzahl als die beiden anderen Gewichtungsvarianten. Sofern ein minimales Gewicht vorgegeben wird und nicht gemäß der Wahl der Folge ϵ_k die Endlichkeit von K gesichert ist, liegt die maximale Stufenzahl unter $2 + (1 - w_1^2)/w_{\min}$. In der Formulierung des Algorithmus von Hartung (2001) ist es möglich, dass die letzte Stufe ein kleineres Gewicht als $w_{\min}$ erhält. Um dies zu vermeiden, kann eine entsprechende Forderung in (2.30) und (2.31) aufgenommen werden.

2.2.4 Weitere Verfahren

Neben den beschriebenen Verfahren existieren u.a. adaptive Designs für konkrete Anwendungen wie die Überlebenszeitanalyse (z.B. Li et al., 2005; Schäfer und Müller, 2001) oder mit dem Hypothesenwechsel nach einer Zwischenauswertung, wie z.B. beim Vergleich mehrerer Dosen bzw. Behandlungsarme und der Auswahl der „besten" Behandlung

(Bauer und Kieser, 1999; Bauer und Röhmel, 1995; Bischoff und Miller, 2005) oder dem Wechsel zwischen Überlegenheits- und Nicht-Unterlegenheitshypothese (Shih et al., 2004; Wang et al., 2001).

Weitere Vorschläge zur Konstruktion flexibler Designs beruhen auf der Idee, gruppensequentielle Pläne adaptiv zu erweitern. Von Müller und Schäfer (2001) stammt ein Ansatz, Studien ohne festgelegte Stufenanzahl zu planen, indem nach jeder Stufe entschieden wird, ob die Studie als gruppensequentieller Test oder als einstufiger Test mit dem verbleibenden Niveau fortgesetzt wird. Dazu wird der gruppensequentielle Test als Kombinationstest dargestellt und die Funktion der bedingten Fehlerwahrscheinlichkeit berechnet. Nach der Interimsanalyse des gruppensequentiellen Plans kann das Design der Studie geändert werden, sofern die bedingte Fehlerwahrscheinlichkeit bewahrt wird. So kann z.B. die Studie als neuer gruppensequentieller Plan mit der vorgegebenen bedingten Fehlerwahrscheinlichkeit erster Art fortgesetzt werden. Müller und Schäfer (2001) weisen darauf hin, dass mit ihrem Verfahren erstmals eine Änderung der α vergebenden Funktion (siehe Kapitel 2.2.2) im Verlauf einer adaptiven Studie ermöglicht wird. Aufgrund der Möglichkeit, nach jeder Interimsanalyse den Stichprobenumfang und das bedingte Niveau der folgenden Stufe zu verändern, bis in einem letzten Studienabschnitt der verbleibende bedingte Fehler verbraucht wird, ergibt sich ein sehr flexibles Design.

Eine Studie als Kombination aus Varianzvergabe-Prinzip und gruppensequentiellen Designs beschreiben Yin und Shen (2005). Nach der ersten Stufe wird gemäß dem gruppensequentiellen zweistufigen Design nach Pocock (1977) ein Test der Nullhypothese zugelassen bzw. geprüft, ob wegen Aussichtslosigkeit die Studie abgebrochen wird. Liegt die Teststatistik im Fortsetzungsbereich wird die Studie gemäß Shen und Fisher (1999) mit adaptiver Wahl von Gewichten fortgeführt und nach vollständiger Vergabe der Varianz eine finale Testentscheidung gefällt. Ein Abbruch wegen Aussichtslosigkeit ist auf allen Stufen möglich.

Das rekursive Anwenden von zweistufigen Kombinationstest, das von Bauer et al. (2001) und Brannath et al. (2002) beschrieben wird, ähnelt dem Vorgehen von Müller und Schäfer (2001). Nach der Durchführung der ersten Stufe eines Kombinationstests wird entschieden, ob die folgende Stufe als zweistufiges Design unter Verwendung eines Kombinationstests oder als finale Stufe durchgeführt wird. Das Verfahren wird allgemein für eine Kombinationsfunktion $C(p_1, p_2)$ der P-Werte auf den zwei Stufen beschrieben, die in beiden Argumenten steigend, in wenigstens einem streng monoton und in p_2 links-stetig ist. Um das Verfahren anzuwenden wird eine Berechnungsvorschrift für den globalen P-Wert der zweistufigen Testprozedur benötigt. Mit den Grenzen für den Abbruch wegen Aussichtslosigkeit α_0 und Verwerfen der Nullhypothese α_1 auf der ersten Stufe wird für das

zweistufige Testverfahren ein P-Wert wie folgt definiert

$$\tilde{p}(p_1, p_2) = \begin{cases} p_1, & \text{wenn } p_1 \leq \alpha_1 \text{ oder } p_1 > \alpha_0, \\ \alpha_1 - \int_{\alpha_1}^{\alpha_0} \int_0^1 1_{[C(x,y) \leq C(p_1,p_2)]} \partial y \partial x, & \text{sonst.} \end{cases} \tag{2.32}$$

Dabei ist $1_{[C(x,y)\leq c]}$ gleich 1, falls $C(x, y) \leq c$, sonst 0. Für den Kombinationstest nach Fisher wird eine geschlossene Darstellung des P-Werts angegeben, während für andere Kombinationstests ein P-Wert meist mittels numerischer Integration bestimmt werden muss. Der P-Wert p_2 in (2.32) kann durch einen P-Wert, der sich aus einem zweistufigen Design gemäß dem Kombinationstest ergibt, ersetzt werden. Wiederholt man die Konstruktion eines zweistufigen Tests als zweite Stufe, so ergibt sich nach dem Studienende nach Stufe K ein rekursiv definierter globaler P-Wert für die gesamte Studie

$$p = \tilde{p}_1(p_1, \tilde{p}_2(p_2, \tilde{p}_3(\ldots, \tilde{p}_{K-2}(p_{K-2}, \tilde{p}_{K-1}(p_{K-1}, p_K))))). \tag{2.33}$$

Der P-Wert $q_k = \tilde{p}_k(p_k, q_{k+1})$, $k = 1, \ldots, K-1$ ist der gemäß (2.32) definierte P-Wert aus dem Ergebnis der k-ten Stufe und den nachfolgenden Stufen. Bei geeigneter Wahl der Ablehn- und Annahmeregionen in den rekursiv geplanten Studienabschnitten führt die Ablehnung der Nullhypothese auf einer einzelnen Stufe zu $p \leq \alpha$ und der Stopp wegen Aussichtslosigkeit bzw. die Annahme der Nullhypothese nach einer letzten geplanten Stufe zu $p > \alpha$.

Ein sehr allgemeiner Ansatz von Müller und Schäfer (2004) beruht auf der zum Teil sehr aufwändigen Berechnung der bedingten Wahrscheinlichkeit des Verwerfens der Nullhypothese in einer der folgenden geplanten Zwischenauswertungen, gegeben den Wert der Teststatistik bei der aktuellen Zwischenauswertung, wenn die Nullhypothese wahr ist. Das Design erlaubt zu ungeplanten Interimsanalysen Designänderungen hinsichtlich Stichprobenumfang, Anzahl der geplanten Zwischenauswertungen, Art der Niveau α vergebenden Funktion, der Zielvariable oder der Teststatistik sowie die Einführung von Grenzen für den Stopp wegen Aussichtslosigkeit. Bei geplanten Zwischenauswertungen ist neben den genannten Designänderungen auch der Test der Nullhypothese zulässig. Posch et al. (2004) untersuchen die bedingte Ablehnwahrscheinlichkeit beim Einstichproben-t-Test, wenn nach n_1 von insgesamt n geplanten Beobachtungen eine ungeplante Interimsanalyse erfolgt, und geben verschiedene Möglichkeiten der Fortsetzung an. Wird die Ablehnwahrscheinlichkeit bedingt auf die Teststatistik der ersten Stufe betrachtet, kann sie unabhängig von der unbekannten Varianz bestimmt werden. Bei der Bedingung auf den Mittelwert und die empirische Varianz auf der ersten Stufe hängt die bedingte Ablehnwahrscheinlichkeit von der unbekannten Varianz ab und muss geschätzt werden.

In den Ansatz von Müller und Schäfer (2004) lassen sich die meisten der beschriebenen Testverfahren einbetten (vgl. Schäfer et al., 2006). Die bedingte Ablehnwahrscheinlichkeit

wird z.B. zur Wahl der oberen und unteren Schranken bei der Konstruktion adaptiver Designs mittels rekursiver Kombinationstests verwendet, um die Gleichheit der Aussagen aus globalem P-Wert und den Testentscheidungen auf den einzelnen Stufen zu gewährleisten (siehe Brannath et al., 2002).

2.3 Parameterschätzer

In gruppensequentiellen Verfahren ist der übliche Maximum-Likelihood-Schätzer verzerrt (Jennison und Turnbull, 2000) und es existiert kein Schätzer mit gleichmäßig kleinster Varianz (Liu und Hall, 1999). Dies gilt analog für Effektschätzungen in adaptiven Designs. Es wird generell vorausgesetzt, dass auf jeder Stufe des adaptiven Designs Daten gleicher Art und Qualität zur Schätzung des Parameters θ erhoben werden. Wird ein Effekt ausschließlich basierend auf den Daten der zuletzt durchgeführten Stufe geschätzt, so ist er bei Verwendung eines im einstufigen Design unverzerrten Schätzers ebenfalls unverzerrt. Bei Vermischung der Stufen geht diese Unverzerrtheit verloren, da Beobachtungen aus den Stufen $k \geq 2$ nur dann entstehen, wenn die Teststatistiken der vorangegangen Stufen stets im Fortsetzungsbereich der Studie gelegen haben. Das bedeutet, dass keine extremen Werte aus der Alternative aus bereits durchgeführten Stufen vorliegen, da dies zur vorzeitigen Ablehnung der Nullhypothese geführt hätte. Entsprechend treten keine extremen Werte der entgegengesetzten Richtung auf, wenn ein Stopp wegen Aussichtslosigkeit eingeplant ist.

Im adaptiven Design führt die zufällige Wahl des Stichprobenumfangs zu Schwierigkeiten bei der Bestimmung der Verzerrung. Um die Gesamtverzerrung des Verfahrens berechnen zu können, wird eine exakte Regel hinsichtlich der Wahl der Stichprobenumfänge benötigt. Für das Design von Proschan und Hunsberger (1995) berechnet Denne (2000) die Gesamtverzerrung des gepoolten Mittelwerts als Schätzer des Erwartungswerts normalverteilter Zufallsvariablen bei festen Regeln für die Wahl des Stichprobenumfangs auf der zweiten Stufe. In der einseitigen Testsituation $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ werden kleine und negative Werte θ unterschätzt und größere positive Werte überschätzt. Die Verzerrung fällt umso kleiner aus, je größer der Stichprobenumfang der ersten Stufe ist. Für die Regel, den Stichprobenumfang der zweiten Stufe so zu wählen, dass für den Wert des Schätzers $\hat{\theta}_1$ aus der ersten Stufe eine vorgegebene konstante bedingte Power $1-\beta_2$ erzielt wird, gibt Denne (2000) einen korrigierten, approximativ unverzerrten Schätzer an.

Coburger und Wassmer (2001) schlagen eine Adjustierung des Schätzers des Erwartungswerts einer Normalverteilung nach Studienende in Abhängigkeit von der Anzahl durchgeführter Stufen vor, da die Qualität des Bias auf den einzelnen Stufen unterschiedlich

ist. Die Korrektur, die im gruppensequentiellen Fall, d.h. bei Kenntnis des Stichprobenumfangs auf jeder Stufe, zu unverzerrten Schätzern führt, dient im adaptiven Fall als Approximation, da die adaptiv gewählten Stichprobenumfänge wie zu Beginn festgelegte behandelt werden. Im Gegensatz zum Verfahren nach Denne (2000) wird keine feste Regel bzgl. der Wahl der Stichprobengrößen benötigt. Die Korrektur führt im Vergleich zum Maximum-Likelihood-Schätzer der gepoolten Daten zu einer deutlichen Reduzierung der Verzerrung, jedoch zu einer größeren Varianz.

Einen Median-unverzerrten Schätzer nach dem Studienende konstruieren Brannath et al. (2002) für den Erwartungswert einer Normalverteilung mit bekannter Varianz mit Hilfe eines einseitigen monotonen 50%-Konfidenzintervalls im Rahmen rekursiver Kombinationstests.

In den Designs gemäß dem Prinzip der Varianzvergabe (vgl. Abschnitt 2.2.3) beeinflussen die Option zum Stopp wegen Aussichtslosigkeit sowie die Blockgrößen der einzelnen Stufen die Größe der Verzerrung. Für den Erwartungswertschätzer gewichten Cheng und Shen (2004) die Stufenmittelwerte mit dem Produkt aus gewähltem Gewicht und der Quadratwurzel des Stichprobenumfangs der einzelnen Stufen. Dieser Schätzer ist Median-unverzerrt (siehe Brannath et al., 2006a), jedoch führt die Möglichkeit zum Abbruch wegen Aussichtslosigkeit zu einer Überschätzung im Erwartungswert, wenn Studien nicht frühzeitig wegen Aussichtslosigkeit beendet werden. Um dies zu korrigieren, wird ein Schrumpfungsfaktor in die Bestimmung des Schätzers aufgenommen. Er ist abhängig von den Stichprobenumfängen und den Gewichten auf den einzelnen Stufen sowie vom erwünschten Behandlungseffekt, der die Schranken für den Abbruch wegen Aussichtslosigkeit bestimmt.

Einen umfangreichen Überblick über Schätzer und Konfidenzintervalle im zweistufigen adaptiven Design geben Brannath et al. (2006a). Sie betrachten die Verzerrung und den mittleren quadratischen Fehler von Schätzern, die durch unterschiedliche Gewichtungen der Mittelwerte auf den zwei Stufen gebildet werden. Die untersuchten Verfahren unterteilen sich in Designs mit zwingender Fortsetzung nach der ersten Stufe und mit der Möglichkeit zum Abbruch nach der ersten Stufe. Bei beiden Vergleichskriterien führen der Mittelwert über die gepoolten Daten und der Median-unverzerrte Schätzer zu akzeptablen Ergebnissen, während unverzerrte Schätzer, konstruiert als a priori gewählte Linearkombination der Mittelwerte der Stufen, einen deutlich größeren mittleren quadratischen Fehler aufweisen.

2.4 Kontrolle der Power

2.4.1 Vergleich mit einstufigen und gruppensequentiellen Designs

Die beschriebenen adaptiven Verfahren sind so konstruiert, dass ein vorgegebenes Signifikanzniveau eingehalten wird, während die Gesamtpower durch Änderungen am Stichprobenumfang im Gegensatz zu einstufigen Studienplänen verändert werden kann. Bauer (1989a) weist darauf hin, dass die Verwendung eines optimalen einstufigen Tests der gepoolten Daten aller Stufen einer Studie mehr Power besitzt als der Kombinationstest für die unabhängigen P-Werte. Dieser Powerverlust bei gleichem Stichprobenumfang ist der Preis für die Adaptionsmöglichkeit. Je flexibler die Verfahren werden, umso schwieriger ist es, die Gesamtpower zu berechnen. Eine feste Regel zur Bestimmung des Stichprobenumfangs, wie z.B. die Vorgabe einer bedingten Power, wird zur Berechnung der Gesamtpower eines Verfahrens benötigt.

Bauer und Köhne (1994) vergleichen den Kombinationstest nach Fisher mit dem gleichmäßig besten Test. Unter Normalverteilungsannahme der Teststatistik mit bekannter Varianz untersuchen sie unterschiedliche Aufteilungen des gesamten Stichprobenumfangs eines einstufigen Tests auf zwei Stufen. Sie beobachten unter diesen relativ strengen Bedingungen, dass der Powerverlust insgesamt gering ausfällt. Am kleinsten ist er bei gleichmäßiger Aufteilung des Stichprobenumfangs auf die zwei Stufen. Ein zusätzlicher Powerverlust entsteht, wenn ein Stopp zu Gunsten der Nullhypothese auf der ersten Stufe zugelassen wird. Wird $\alpha_0 \geq 0.5$ als Grenze für den Stopp wegen Aussichtslosigkeit gewählt, fällt der Verlust im genannten Beispiel relativ klein aus. Für den Vergleich zweier Erwartungswerte bei unbekannter Varianz zeigen Banik et al. (1996), dass der Powerverlust im Verhältnis zum einstufigen t-Test in der Regel klein und nur bei starker Unbalanciertheit in den Behandlungsgruppen groß ausfällt.

Proschan und Hunsberger (1995) untersuchen den erwarteten prozentualen Anstieg des Stichprobenumfangs bei Verwendung von Designs mit der zirkulären bedingten Fehlerfunktion (2.15) im Vergleich zum einstufigen Design. Auf der ersten Stufe wird dabei der Stichprobenumfang eines einstufigen Designs mit Irrtumswahrscheinlichkeiten von α und β bei einem angenommenen Effekt von θ_0 und der Umfang auf der zweiten Stufe in Abhängigkeit einer konstanten bedingten Power gewählt. Die Power im einstufigen Design fällt für Werte aus der Alternative $\lambda\theta_0$, $0 < \lambda \leq 1$, stets geringer aus als in den betrachteten zweistufigen Designs. Dabei ist der zusätzlich benötigte Stichprobenumfang im zweistufigen Design jeweils kleiner als der zusätzlich benötigte Stichprobenumfang im einstufigen Design, um die gleiche Power zu erzielen.

Cheng und Shen (2004) zeigen für das Design von Shen und Fisher (1999), dass nur bei einer Studie ohne Stopp wegen Aussichtslosigkeit und einem konstanten Verhältnis zwischen Stufengewicht und Quadratwurzel aus dem Stichprobenumfang über alle Stufen hinweg die gleiche Power wie im einstufigen Design erzielt werden kann. Bei Abweichung von den optimalen Gewichten wird die Erhöhung des durchschnittlich benötigten Stichprobenumfangs im Vergleich zum Stichprobenumfang des einstufigen Designs mit gleicher Power bei Verwendung kleiner Blockgrößen als gering angegeben.

Tsiatis und Mehta (2003) führen einen formalen Beweis, dass bei vorgegebenen Annahme- und Ablehnwahrscheinlichkeiten unter der Nullhypothese zu jedem adaptiven Test ein gruppensequentieller Test mit folgender Eigenschaft konstruiert werden kann: Für jeden Wert aus der Alternative verwirft der gruppensequentielle Test die Nullhypothese mit einer gleichmäßig höheren Wahrscheinlichkeit bzw. bei Werten, die nicht aus der Alternative stammen, nimmt er mit gleichmäßig höherer Wahrscheinlichkeit die Nullhypothese an. Dabei wird vorausgesetzt, dass der gruppensequentielle Test zu jedem Informationszeitpunkt, eine Interimsanalyse ermöglichen muss, an dem der adaptive Test eine Interimsanalyse zulässt. Der Informationszeitpunkt entspricht dem Anteil der beobachteten Patienten an der vorgegebenen Maximalzahl. Im Vergleich bleibt die Anzahl der durchzuführenden Interimsanalysen unberücksichtigt, die in den nach Tsiatis und Mehta (2003) optimalen gruppensequentiellen Verfahren höher als im adaptiven Design ist. Jennison und Turnbull (2003) beschreiben gruppensequentielle Designs, die eine höhere Power bei niedrigerem erwarteten Stichprobenumfang als zweistufige Designs gemäß dem Prinzip der Varianzvergabe von Fisher (1998) besitzen. Der Powerverlust entsteht dadurch, dass nach Gewichtung der Teststatistiken auf den zwei Stufen keine suffiziente Teststatistik verwendet wird. Für die zum Vergleich herangezogenen gruppensequentiellen Pläne wird jedoch ein kleinerer Wert aus der Alternative zur Planung verwendet. So erfolgt der Vergleich im Nachhinein und der Vorteil der adaptiven Designs, eine Studie mit zu geringer Power zu retten, wird vernachlässigt. Brannath et al. (2006b) legen dar, dass die Wahl der Stichprobenumfänge sowie die Wahrscheinlichkeiten für Annahme und Ablehnung der Nullhypothese durch Kosten und Gewinne gesteuert werden. Kosten fallen für falsch positive oder falsch negative Entscheidungen sowie für zusätzliche Interimsanalysen, Gewinne bei einer richtigen Entscheidung an. Die a priori Festlegung der Stichprobenumfänge und der Abbruchsregeln zur Konstruktion des effizienten gruppensequentiellen Plans entspricht daher nicht der Idee der adaptiven Designs (Brannath et al., 2006b).

Shih (2006) weist in einem Kommentar zu gruppensequentiellen und adaptiven Designs ebenfalls darauf hin, dass ein Vergleich der Effizienz beider Designs gemäß Tsiatis und Mehta (2003) aufgrund der unterschiedlichen Zielsetzungen nicht sinnvoll ist. Während

die gruppensequentiellen Designs auf ein frühes Ende der Studie abzielen, ermöglichen adaptive Studienpläne, eine Studie durch Anpassung des Designs hinsichtlich falscher Annahmen zu korrigieren und zu einem befriedigenden Abschluss zu führen. Der Verlust eines gleichmäßig besten Tests, der nur im Rahmen der gruppensequentiellen Tests besteht (Li et al., 2005), ist im Vergleich zu einer falsch geplanten und dadurch nicht verwertbaren Studie gering. Ähnliche Argumente finden sich bei Shen und Fisher (1999), die ihr Verfahren den gruppensequentiellen Verfahren von Pocock (1977) und O'Brien und Fleming (1979) gegenüberstellen, dabei aber auf die unterschiedliche Zielsetzung hinweisen.

2.4.2 Bedingte Power und Wahl des Stichprobenumfangs nach Interimsanalysen

Die bedingte Power wird als die Wahrscheinlichkeit definiert, die Nullhypothese im weiteren Verlauf der Studie zu verwerfen, gegeben die Teststatistik der vorausgegangenen Stufen, bei einem Wert aus der Alternative. Die bedingte Power hängt vom Stichprobenumfang für den folgenden Studienabschnitt bzw. dem verbleibenden Teil der Studie ab. Außerdem beeinflusst die Wahl des Parameters aus der Alternative, der entweder ein vorgegebener klinisch relevanter Wert, eine Schätzung aus den bereits durchgeführten Stufen oder eine Kombination aus beiden sein kann, die bedingte Power. Sie wird weiterhin vom gewählten adaptiven Design beeinflusst. Bauer und König (2006) untersuchen die Dichte der bedingten Power in einem zweistufigen gruppensequentiellen Design in Abhängigkeit vom Anteil des Stichprobenumfangs der ersten Stufe am Gesamtumfang und vom wahren Parameterwert. Dabei verwenden sie den Effektschätzer bzw. einen a priori definierten konstanten Parameterwert. Die Dichte bei Verwendung des Effektschätzers ist stets U-förmig, während bei Verwendung des Designparameters im Fall von kleinen Anteilen des Stichprobenumfangs der ersten Stufe die Dichte eingipflig ist mit dem Modus in der Nähe der ursprünglichen Power. Für größere Anteile nähern sich die Verteilungen an. Die Grenzverteilung der bedingten Power unter der Alternative, wenn der Anteil der ersten Stufe gegen 1 geht, entspricht einer Bernoulli-Verteilung mit Parameter $p = 1 - \beta$.

Posch und Bauer (1999) vergleichen die zweistufigen Designs von Bauer und Köhne (1994) und Lehmacher und Wassmer (1999) hinsichtlich der bedingten Power auf der zweiten Stufe in Abhängigkeit des Testergebnisses der ersten Stufe. Dabei ist bei gleichem Stichprobenumfang die bedingte Power für den Kombinationstest mittels der Inverse-Normalmethode bei kleinen Werten der Teststatistik der ersten Stufe größer als beim Kombinationstest nach Fisher, während bei Werten nahe der kritischen Schranke auf der ersten Stufe sich das Verhältnis umkehrt.

Meist wird bei der Durchführung adaptiver Designs eine feste, auf die bis zum jeweiligen Zeitpunkt erhobenen Daten bedingte Power von z.B. 80% oder 90% für die einzelnen Studienabschnitte vorgeschlagen (z.B. Brannath et al., 2002; Lehmacher und Wassmer, 1999; Li et al., 2002; Proschan und Hunsberger, 1995), wobei der geschätzte interessierende Effekt als wahr angenommen wird. Die Power der Verfahren insgesamt wird dadurch nicht mehr kontrolliert. Regeln für die Neubestimmung des Stichprobenumfangs bei Verwendung des Kombinationstests nach Fisher in einem zweistufigen Design finden sich bei Posch und Bauer (2000). Bauer und König (2006) vergleichen die erwarteten Stichprobenumfänge in einem zweistufigen adaptiven Design bei Verwendung entweder des Effektschätzers aus der ersten Stufe oder eines a priori festgelegten Werts θ_d zur Planung des Stichprobenumfangs der zweiten Stufe, um eine bedingte Power von 0.8 zu erzielen, und in einem entsprechenden gruppensequentiellen Design mit dem Umfang eines einstufigen Designs. Wird der Effektschätzer verwendet, erhöht sich die Power insbesondere für wahre Parameter unterhalb θ_d deutlich. Diese Erhöhung ergibt sich aus einem im Vergleich zum einstufigen Designs zum Teil mehr als vervierfachten erwarteten Stichprobenumfang. Bei Verwendung von θ_d unterscheidet sich das adaptive Design in Power und Stichprobenumfang nur wenig von einem für eine Power von 0.8 bei θ_d geplanten gruppensequentiellen Design. Entsprechend ergibt sich für Parameter kleiner als θ_d eine deutlich reduzierte Power bei nur geringer Erhöhung des erwarteten Umfangs im Vergleich zum einstufigen Design. Bei größeren wahren Parametern wird der Umfang, verursacht durch Abbruch nach der ersten Stufe, reduziert und die Power vergrößert.

In adaptiven Designs ist eine exakte Kontrolle der bedingten und der globalen Power nur für eine zu Beginn festgelegte Alternative bzw. eine vorgegebene Verteilung von Alternativen zu erreichen. Unter diesen Vorgaben bestimmen Brannath und Bauer (2004) optimale bedingte Fehlerfunktionen für zweistufige Designs bei Normalverteilung mit bekannter Varianz, die den erwarteten Stichprobenumfang minimieren. Es zeigt sich, dass der benötigte Stichprobenumfang auf der zweiten Stufe bei einer konstanten bedingten Power im optimierten Design eine konkave Funktion der Teststatistik der ersten Stufe ist. Dies ist auch bei der zirkulären Fehlerfunktion und beim Kombinationstest nach Fisher der Fall, während bei linearen bedingten Fehlerfunktionen der Graph konvex verläuft. Obwohl der Unterschied in den maximal benötigten Umfängen groß ist, fällt der Unterschied in den erwarteten Stichprobenumfängen zwischen den optimierten Designs und Designs mit linearer bedingter Fehlerfunktion relativ gering aus. Die optimierten Designs liefern größere Stichprobenumfänge für mittlere Effekte und kleinere bei kleinen und großen Effekten. Ähnliche Ergebnisse erhalten Posch et al. (2003), die den erwarteten Stichprobenumfang im zweistufigen adaptiven Design ausgehend von einem gruppensequentiellen Design für eine Auswahl von Parameterwerten minimieren, indem sie den Stichprobenumfang der

zweiten Stufe als Polynom vierten Grades der Teststatistik der ersten Stufe bestimmen. Weiterhin empfehlen sie die Einführung einer oberen Schranke für den Stichprobenumfang auf der zweiten Stufe.

Vorschläge für vollständig sich selbst planende Studien, sogenannte Designautomaten, die konkrete Regeln für die Wahl der Umfänge und der Gewichte vorgeben, sofern letztere adaptiv festzulegen sind, stammen von Shen und Fisher (1999), Hartung (2000, 2001, 2006) und Hartung und Knapp (2003). Dabei wird wie bei den meisten vorgeschlagenen Designs für den zukünftigen Teil der Studie eine bedingte Power gleich der global gewünschten Power festgesetzt.

Eine Erweiterung des Versuchs ist mit zusätzlichen Kosten durch die Verlängerung der Studie bzw. Verlusten durch die verzögerte Markteinführung bei Studienerfolg verbunden. Deshalb optimieren Mehta und Patel (2006) innerhalb eines ursprünglich zweistufigen gruppensequentiellen Designs nach Lan und DeMets (1983) mit Hilfe eines Bayes-Ansatzes den Stichprobenumfang der zweiten Stufe hinsichtlich des erwarteten Gewinns bei Markteinführung des Produkts bzw. Verlusts bei fehlgeschlagenem Nachweis der Wirksamkeit. Im von den Autoren gewählten Beispiel entspricht der empfohlene Stichprobenumfang in etwa der Hälfte der Anzahl Patienten, die zur Erreichung einer bedingten Power von 80% benötigt werden. Ähnliche Berechnungen führen Thach und Fisher (2002) für ein zweistufiges Design nach Shen und Fisher (1999) durch, indem sie für verschiedene a priori Verteilungen des Parameters bzgl. der erwarteten Kosten ein optimales Gewicht und einen optimalen Stichprobenumfang für die erste Stufe bestimmen. Das optimierte Design wird mit zweistufigen gruppensequentiellen Designs hinsichtlich der erwarteten Kosten und des erwarteten Stichprobenumfangs verglichen. Kostenfunktion und erwarteter Stichprobenumfang des optimierten Designs und des Designs nach Pocock (1977) sind für die betrachteten Parameterwerte und Designvorgaben nahezu identisch.

An Hand des Designs nach Hartung und Knapp (2003), das im folgenden Kapitel näher beschrieben wird, werden verschiedene Strategien und Ideen zur Wahl der bedingten Power vorgestellt.

Kapitel 3

Verallgemeinerte gewichtete Inverse-χ^2-Methode

Das im Folgenden vorgestellte Verfahren verwendet als Konstruktionsidee die Varianzvergabe wie Fisher (1998) und Shen und Fisher (1999), indem die Varianz einer finalen χ^2-verteilten Teststatistik auf die einzelnen Studienabschnitte aufgeteilt wird. Gleichzeitig lässt sich die Methode durch die Anwendung der Inversen von χ^2-Verteilungen auf die P-Werte in die Gruppe der Kombinationstests einordnen, wobei sie das Design nach Bauer und Köhne (1994) als Spezialfall enthält. Das Verfahren kennzeichnet sich durch die Möglichkeit des Testens der Nullhypothese auf jeder der durchgeführten Stufen und einer offenen Anzahl von Zwischenauswertungen.

3.1 Grundprinzip

Ausgangspunkt für die Konstruktion einer Studie gemäß dem Varianzvergabe-Prinzip nach Hartung (2000) und Hartung und Knapp (2003) ist ein einseitiges Testproblem bezüglich eines reellwertigen Parameters θ:

$$\mathrm{H}_0\colon \theta = 0 \quad \text{gegen} \quad \mathrm{H}_1\colon \theta > 0. \tag{3.1}$$

Es wird angenommen, dass sich die Studie formal in K disjunkte Studienteile zerlegen lässt, von denen aufgrund der Möglichkeit eines vorzeitigen Abbruchs mit Ablehnung der Nullhypothese nach dem k^*-ten Studienabschnitt, $k^* < K$, jedoch nicht immer alle durchgeführt werden.

Auf jeder Stufe k, $k = 1, \ldots, K$, wird eine geeignete Teststatistik T_k für das einseitige Testproblem (3.1) gewählt, deren Wahrscheinlichkeitsverteilung $F_{k,0}$ unter der Nullhypothese stetig ist. Führen große Werte von T_k zu einer Ablehnung der Nullhypothese, so

wird in jedem Studienteil k der P-Wert wie folgt definiert:

$$p_k = 1 - F_{k,0}(T_k), \quad k = 1, \ldots, K. \tag{3.2}$$

Aufgrund der stetigen Verteilung der Teststatistiken sind die P-Werte unter der Nullhypothese gleichverteilt auf dem Intervall $(0, 1)$. Transformiert man die P-Werte mit der inversen Verteilungsfunktion $F^{-1}_{\chi^2(\nu_k)}$ einer χ^2-Verteilung mit ν_k Freiheitsgraden, so erhält man

$$q_k(\nu_k) = F^{-1}_{\chi^2(\nu_k)}(1 - p_k). \tag{3.3}$$

Die Zufallsvariable $q_k(\nu_k)$ ist unter der Nullhypothese χ^2-verteilt mit ν_k Freiheitsgraden.

Die kombinierte Teststatistik bis zum k-ten Studienabschnitt ergibt sich aus der Summe der transformierten P-Werte $q_j(\nu_j)$. Bei einer vorgegebenen Folge von Freiheitsgraden ν_j, $j = 1, \ldots, K$, und Unabhängigkeit der disjunkten Studienteile gilt unter der Nullhypothese:

$$S_k = \sum_{j=1}^{k} q_j(\nu_j) \overset{\text{H}_0}{\sim} \chi^2(\nu_\Sigma(k)), \quad \nu_\Sigma(k) = \sum_{j=1}^{k} \nu_j, \quad k = 1, \ldots, K. \tag{3.4}$$

Da in jedem Studienabschnitt ein nichtnegativer Wert $q_j(\nu_j) \geq 0 \; \forall j$ zur Teststatistik S_{j-1} addiert wird, kann die Nullhypothese zum Niveau α verworfen werden, sobald gilt:

$$S_{k^*} \geq \chi^2(\nu_\Sigma(K))_{1-\alpha}, \quad \text{für ein } k^* \in \{1, \ldots, K\}. \tag{3.5}$$

Dabei bezeichnet $\chi^2(\nu_\Sigma(K))_{1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $\nu_\Sigma(K)$ Freiheitsgraden. Die nach der Überschreitung des kritischen Werts durch die kombinierte Teststatistik noch ausstehenden $K - k^*$ Stufen werden nicht mehr ausgeführt.

Wird die Anzahl der Freiheitsgrade auf den einzelnen Stufen konstant gleich zwei gesetzt, für die Anzahl der globalen Freiheitsgrade $2K$ gewählt und nur der Stichprobenumfang n_k in Abhängigkeit der Ergebnisse der Zwischenauswertungen bestimmt, ergibt sich der Kombinationstest nach Fisher bzw. das Verfahren nach Bauer und Köhne (1994) ohne Stopp wegen Aussichtslosigkeit. Dies gilt, da die χ^2-Verteilung mit zwei Freiheitsgraden der Exponentialverteilung mit Parameter 0.5 entspricht, vgl. (2.11).

3.2 Adaptives Vorgehen

Um die Verteilungseigenschaft der finalen Teststatistik und damit das Signifikanzniveau zu erhalten, werden bei adaptiver Wahl der Stichprobenumfänge n_k und der Gewichte ν_k zu Beginn der Studie die Gesamtfreiheitsgrade ν_G des gesamten sequentiellen Versuchs

festgelegt werden. Durch die Wahl der Gesamtfreiheitsgrade ν_G ergibt sich der globale kritische Wert als $cv_\alpha = \chi^2(\nu_G)_{1-\alpha}$.

Hartung und Knapp (2003) fordern wie Fisher (1998), dass mit Wahrscheinlichkeit 1 die Studie nach einer zufälligen, endlichen Anzahl Stufen K endet. Seien ν_G, ν_1 nichtnegative Konstanten und sei $\nu_i := \nu_i(p_1, \ldots, p_{i-1})$, $i = 2, 3, \ldots$, eine nicht endliche, zufällige Folge von nichtnegativen Zahlen, für die mit Wahrscheinlichkeit Eins für ein zufälliges, endliches K gilt

$$\sum_{i=1}^{K} \nu_i = \sum_{i=1}^{\infty} \nu_i = \nu_G. \tag{3.6}$$

Dann gilt für die kombinierte Statistik unter der Nullhypothese

$$S_K = \sum_{i=1}^{K} q_i(\nu_i) \overset{\mathrm{H}_0}{\sim} \chi^2(\nu_G). \tag{3.7}$$

Eine Möglichkeit, die Zahl der Stufen zu begrenzen und die Voraussetzung zu erfüllen, ist, ein minimales Gewicht $\nu_{\min}$ für zu realisierende Studienabschnitte festzulegen. Wählt man z.B. $\nu_{\min} = 1$ und $\nu_G = K^*$, ergibt sich ein relatives minimales Stufengewicht von $1/K^*$ an den Gesamtfreiheitsgraden und eine maximale Anzahl von K^* Studienabschnitten.

Zu Beginn werden die ν_G Freiheitsgrade in ν_1, $\nu_{\min} \leq \nu_1 \leq \nu_G$, als Gewicht der ersten Stufe und $\nu_2^* = \nu_G - \nu_1 \geq \nu_{\min}$ für den verbleibenden Studienteil aufgeteilt. Nach der Realisierung der ersten Stufe wird die Studie mit der Ablehnung der Nullhypothese beendet, falls $S_1 = q_1(\nu_1) \geq \chi^2(\nu_G)_{1-\alpha}$ gilt. Andernfalls wird die Studie fortgesetzt, indem die Freiheitsgrade ν_2^* auf die zweite Stufe und den verbleibenden Studienteil aufgeteilt oder vollständig für den zweiten Studienabschnitt verwendet werden. Generell ist bei der Aufteilung der Freiheitsgrade darauf zu achten, dass für den folgenden Studienteil mindestens $\nu_{\min}$ Freiheitsgrade zurückgehalten werden. Stehen schließlich für einen Studienabschnitt k weniger als $2\nu_{\min}$ Freiheitsgrade zur Verfügung, werden diese aufgebraucht und die Studie endet nach dem k-ten Studienabschnitt.

Nach der Durchführung von $(k-1)$ Studienabschnitten gilt unter der Nullhypothese bei Verbrauch aller Freiheitsgrade für die kombinierte Teststatistik:

$$q_1(\nu_1) + \left[q_2(\nu_2) + \left\{ q_3(\nu_3) + \left(\ldots + \left[q_{k-1}(\nu_{k-1}) + q_k \left(\nu_G - \sum_{j=1}^{k-1} \nu_j \right) \right] \right) \right\} \right] \overset{\mathrm{H}_0}{\sim} \chi^2(\nu_G) \tag{3.8}$$

mit $\nu_j \geq \nu_{\min}$, $i = 1, \ldots, k-1$, und $\nu_G - \nu_\Sigma(k-1) \geq \nu_{\min}$.

Somit wird weiterhin die Entscheidungsregel aus Abschnitt 3.1 verwendet, dass H_0 nach k^* Stufen verworfen und die Studie beendet wird, falls $S_{k^*} \geq \chi^2(\nu_G)_{1-\alpha}$.

Für mögliche verbleibende Freiheitsgrade $\nu_G - \nu_\Sigma(k^*) > \nu_{\min}$ wird nach einem Abbruch zu Gunsten der Alternative ein beliebiger Stichprobenumfang n_{k^*+1} gewählt, der zugehörige (k^*+1)-te Studienabschnitt jedoch nicht mehr durchgeführt.

Die Studie endet in jedem Fall nach dem Studienteil k^*, in dem alle verbleibenden Freiheitsgrade mit $\nu_\Sigma(k^*) = \nu_G - \nu_\Sigma(k^*-1)$ verbraucht werden.

3.3 Wahl der Freiheitsgrade und des Stichprobenumfangs

Während die Informationen zur Bestimmung von Umfang und Gewicht auf der ersten Stufe aus dem Vorwissen über Effekt- und Störgrößen basieren, stehen nach der Durchführung von $(k-1)$ Interimsanalysen Schätzer für den interessierenden Parameter $\hat{\theta}_{k-1}$ sowie für andere Parameter wie Varianzen zur Verfügung. Diese können zur Berechnung des Stichprobenumfangs in einer geeigneten Stichprobenfunktion $f_k(\alpha, \beta, \theta)$, entsprechend der Teststatistik T_k, verwendet werden. Nach Studienabschnitt $(k-1)$ lässt sich der Stichprobenumfang M_k berechnen:

$$M_k = f_k(\alpha_k^*, \beta, \hat{\theta}_{k-1}). \tag{3.9}$$

Dieser wird benötigt, wenn im k-ten Studienabschnitt die Studie nach Vergabe der verbleibenden Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ endet und der neu festgesetzte Wert des Parameters aus der Alternative $\hat{\theta}_{k-1}$ mit einer bedingten Power von $1-\beta$ nachgewiesen werden soll. Dabei gibt

$$\alpha_k^* = 1 - F_{\chi^2(\nu_G - \nu_\Sigma(k-1))}(cv_\alpha - S_{k-1}) \tag{3.10}$$

den bedingten Fehler erster Art für den verbleibenden Teil der Studie an. Wenn noch mindestens zwei Studienabschnitte durchgeführt werden sollen, kann mit einer geringeren Power $1-\beta_{g,k} < 1-\beta$ für den k-ten Studienabschnitt der folgende Stichprobenumfang berechnet werden:

$$m_k = f_k(\alpha_k^*, \beta_{g,k}, \hat{\theta}_{k-1}). \tag{3.11}$$

Hartung und Knapp (2003) schlagen vor, die Freiheitsgrade ν_k anteilig an den verbleibenden Freiheitsgraden gemäß dem Verhältnis

$$\epsilon_k = \frac{m_k}{M_k} \tag{3.12}$$

zu wählen. Bei Vorgabe eines minimalen Gewichts $\nu_{\min}$ und eines minimalen Stichprobenumfangs $n_{\min}$ ergibt sich zunächst folgende Gewichtsfunktion:

$$W(k) = \max\left(\nu_{\min}, (\nu_G - \nu_\Sigma(k-1)) \cdot \max\left(\epsilon_k, \frac{n_{\min}}{\max(n_{\min}, M_k)}\right)\right). \tag{3.13}$$

Um ein minimales Gewicht für die Stufe $k+1$ zu erhalten, sofern in der k-ten Stufe nicht das vollständige restliche Gewicht vergeben wird, definiert man das Gewicht

$$\nu_k = \begin{cases} W(k), & \text{wenn } \nu_G - W(k) - \nu_\Sigma(k-1) > \nu_{\min}, \\ \nu_G - \nu_\Sigma(k-1), & \text{sonst} \end{cases} \tag{3.14}$$

und den Stichprobenumfang

$$n_k = \max\left(M_k \cdot \frac{\nu_k}{\nu_G - \nu_\Sigma(k-1)}, n_{\min}\right). \tag{3.15}$$

Allgemein kann eine Folge ϵ_k, $0 < \epsilon_k \leq 1$, zur Bestimmung des Stichprobenumfangs $n_k = \epsilon_k \cdot M_k$ und des Gewichts $\nu_k = \epsilon_k \cdot (\nu_G - \nu_\Sigma(k-1))$ definiert werden. Um ein Studienende zu sichern, muss die Folge bei einem zufälligen endlichen Zeitpunkt den Wert 1 annehmen.

Hartung und Knapp (2003) beschreiben zunächst die Verwendung ganzzahliger Freiheitsgrade und somit χ^2-Verteilungen bei der Konstruktion einer globalen Teststatistik. Bei Verwendung der oben genannten Anteile ist die Ganzzahligkeit der Freiheitsgrade für den nächsten Studienabschnitt jedoch nicht immer gewährleistet. Die obigen Aussagen gelten jedoch entsprechend für die Verwendung inverser Gammaverteilungen mit den Parametern $\nu_j/2$ und $1/2$ anstelle der χ^2-Verteilungen mit ν_j Freiheitsgraden. Um die Notation beizubehalten, werden im Folgenden auch nicht ganzzahlige Freiheitsgrade für die χ^2-Verteilungen zugelassen.

3.4 Simulationsstudie aus Hartung und Knapp (2003)

Hartung und Knapp (2003) vergleichen in einer Simulationsstudie verschiedene Strategien bezüglich der Wahl der bedingten Power $(1 - \beta_{g,k})$ in (3.11). Die Teststatistik T_k auf den einzelnen Stufen ist die des einseitigen Gauß-Tests, wobei die Varianz der Zufallsvariablen gleich 1 gesetzt wird. Die Gesamtanzahl Freiheitsgrade und das minimale Gewicht werden mit $\nu_G = 10$ und $\nu_{\min} = 1$ gewählt. Für die den Anteil bestimmenden Umfänge m_k und M_k werden obere Schranken angegeben, die bei der Berechnung von (3.12) und (3.13) berücksichtigt werden. Die globalen Fehlerraten lauten $\alpha = 0.025$ und $\beta = 0.10$. In einer ersten Simulationsstudie wird eine konstante bedingte Power $1 - \beta_{g,k} = 0.2$ zur Berechnung von

Tabelle 3.1: *Empirische Power (in %), durchschnittliche Stichprobengröße (ASN) und durchschnittliche Stufenanzahl ($\bar{k}$) für zwei Strategien im adaptiven Design mit der Inverse-χ^2-Methode, Simulation aus Hartung und Knapp (2003)*

θ	Strategie I			Strategie II		
	$(1-\hat{\beta})100\%$	ASN	$\bar{k}$	$(1-\hat{\beta})100\%$	ASN	$\bar{k}$
0.30	54.3	190.0	3.5	51.8	188.9	3.7
0.35	66.9	180.3	3.6	65.3	177.0	3.7
0.40	77.5	165.4	3.6	76.5	162.5	3.7
0.45	85.0	151.0	3.5	84.3	146.3	3.6
0.50	89.9	136.4	3.4	89.9	130.7	3.5
0.55	92.6	122.3	3.4	93.2	117.3	3.4
0.60	94.5	109.3	3.2	95.0	104.8	3.2
0.65	95.9	97.8	3.1	96.1	94.1	3.1
0.70	96.6	89.2	2.9	96.8	85.7	2.9

m_k festgelegt. Auf der ersten Stufe wird das Gewicht $\nu_1 = 2$ und der Stichprobenumfang $n_1 = 34$ gewählt. Der Stichprobenumfang entspricht dem Stichprobenumfang pro Stufe in einem fünfstufigen gruppensequentiellen Design bei einseitiger Fragestellung nach O'Brien und Fleming (1979) bei einer Power von etwa 90% und einem standardisierten Effekt von $\theta = 0.5$. Die betrachteten Strategien unterscheiden sich in den oberen Grenzen $M_{\max}$ und $m_{\max}$ für M_k und m_k, $k \geq 2$. In Strategie I werden $M_{\max} = m_{\max} = 138$, in Strategie II $M_{\max} = 106$ und $m_{\max} = 90$ gewählt. Zusätzlich wird ein minimaler Stichprobenumfang $n_{\min} = 18$ pro Stufe festgelegt. Der Parameter θ wird nach der k-ten Stufe durch die mit der inversen Varianz gewichteten Schätzer auf den einzelnen Stufen $\hat{\theta}(i)$, $i = 1, \ldots, k$, geschätzt mit

$$\hat{\theta}_k = \frac{\sum_{i=1}^{k} n_i \cdot \hat{\theta}(i)}{\sum_{i=1}^{k} n_i}. \tag{3.16}$$

Um bei der Berechnung der Stichprobenumfänge einen Wert aus der Alternative zu verwenden, wird $\hat{\theta}_k$ gleich 0.001 gesetzt, sofern ein kleinerer Schätzwert als 0.001 auftritt. Die simulierte Power, der mittlere benötigte Stichprobenumfang sowie die mittlere benötigte Stufenzahl bis zum Abbruch der Studie aus 10000 Simulationsläufen sind in Tabelle 3.1 zusammengefasst, wobei der wahre Parameter θ zwischen 0.3 und 0.7 variiert wird. Für den Parameterwert $\theta = 0.5$ wird bei beiden Strategien eine Power von etwa 90% erreicht. Die Strategien führen bei den gewählten Parameterwerten nur zu sehr geringen Unterschieden in den beobachteten Kennwerten.

In einer entsprechenden Simulation bei wahren Effekten $\theta = 0.3, 0.5, 0.7$ ergeben sich die

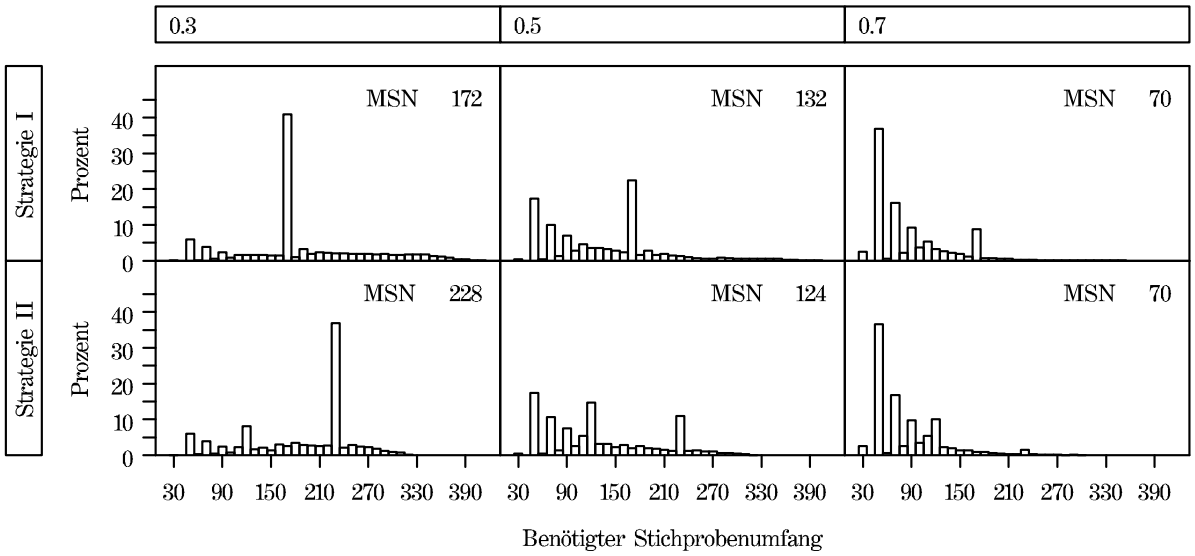

Abbildung 3.1: *Verteilung des benötigten Stichprobenumfangs und medianer Stichprobenumfang (MSN) bei Verwendung der Strategien I und II aus Hartung und Knapp (2003), $\theta = 0.3, 0.5, 0.7$*

in Abbildung 3.1 dargestellten Verteilungen der benötigten Stichprobenumfänge, die den Einfluss der gewählten Strategie zeigen. Je nach Strategie ergeben sich Häufungspunkte an den Stellen, die sich aus der Summe des ersten Stichprobenumfangs $n_1 = 34$, einem Vielfachen der Grenze $n_{\min}$ und $m_{\max}$ sowie $M_{\max}$, beim Verbrauch aller Freiheitsgrade auf der letzten Stufe, berechnen lassen. Die Häufungen bei 52, 70 und 88 ergeben sich bei beiden Strategien durch den Minimalwert $n_{\min}$. Unter Strategie I findet sich ein weiterer Häufungspunkt bei $172 = n_1 + m_{\max} = n_1 + M_{\max}$, während unter Strategie II Häufungspunkte bei $124 = n_1 + m_{\max}$ und bei $230 = n_1 + m_{\max} + M_{\max}$ auftreten. Bei einem wahren Wert $\theta = 0.7$ sind die Verteilungen bei beiden Strategien sehr ähnlich und durch den Minimalwert dominiert. Der mediane Stichprobenumfang bei beiden Strategien liegt bei 70 Patienten. Für $\theta = 0.3$ ergibt sich unter Strategie I mit dem höheren maximalen Stichprobenumfang bei gleichzeitig schnellerer Vergabe der Freiheitsgrade ein niedrigerer Median von 172 im Vergleich zu 228 unter der zweiten Strategie, die pro Stufe weniger Patienten zulässt, jedoch häufiger zu einer weiteren Stufe mit der höchstmöglichen Anzahl führt. Eine ähnliche Beobachtung lässt sich für $\theta = 0.5$ machen, wobei sich der Anteil Studien mit 230 Patienten unter Strategie II von etwa 35% bei $\theta = 0.3$ auf 10% reduziert, so dass der mediane Stichprobenumfang bei 124 liegt, während in Strategie I etwa 21% der Studien einen Umfang von 172 haben und der Median 132 beträgt. Hartung und Knapp (2003) vergleichen die Strategien I und II hinsichtlich des Einflusses der bedingten Power $1 - \beta_{g,k}$ und der oberen Grenze $M_{\max}$ auf die globale empirische Power und den durchschnittlichen Stichprobenumfang pro Stufe mit jeweils einer entsprechend gewählten

Variante. Verwendet man in Strategie I $\beta_{g,k} = 0.25$, ergibt sich für Parameterwerte θ im Bereich von 0.2 bis 0.5 eine niedrigere Power als unter $\beta_{g,k} = 0.8$, die mit einer Reduktion im mittleren Stichprobenumfang verbunden ist. Für $\theta > 0.5$ sind erzielte Power und mittlerer Stichprobenumfang nahezu gleich. In einer zweiten Variante von Strategie I werden die oberen Grenzen auf $M_{\max} = m_{\max} = 180$ erhöht, was für alle betrachteten Parameterwerte zu einer höheren Power und einem höheren ASN im Vergleich zur ursprünglichen Grenze von $M_{\max} = m_{\max} = 138$ führt. Der Unterschied in der Power beträgt bis zu 8% für $\theta = 0.3$. Für Werte nahe der Nullhypothese ist der Zuwachs im ASN erheblich. Für steigende Parameterwerte verringert sich der Unterschied in der beobachteten globalen Power und dem ASN. In einer dritten Simulation wird in Strategie II die obere Grenze verdoppelt, d.h. $M_{\max} = 212$. Die Erhöhung der oberen Schranke führt zu einem beträchtlichen Powergewinn für $\theta < 0.6$ bei einem zum Teil mehr als verdoppelten ASN für $\theta < 0.2$. Für $\theta > 0.6$ verlaufen Power und ASN ähnlich. Die Anzahl der benötigten Stufen erhöht sich bei Verwendung von $M_{\max} = 212$ deutlich. Für $\theta = 0.3$ beträgt die mittlere Stufenzahl statt 3.7 beinahe 5.

Kapitel 4

Eigenschaften und Erweiterungen

In diesem Kapitel wird die verallgemeinerte Inverse-χ^2-Methode als Konstruktionsverfahren für adaptive Designs in die bestehenden Verfahren eingeordnet bzw. Bezug zu diesen hergestellt. Dadurch wird die Grundlage geschaffen, die Einschränkung (3.6) an die Folge von Freiheitsgraden aufzuheben und das Verfahren um optionale Schranken zum Abbruch wegen Aussichtslosigkeit auf jeder Stufe zu erweitern. Weiterhin wird eine Berechnungsmethode für den Stichprobenumfang auf einer einzelnen Stufe vorgestellt, um bei proportionaler Wahl von Umfang und Gewicht eine exakte bedingte Power für einen festgelegten Parameterwert aus der Alternative zu erzielen.

4.1 Bezug zu anderen Verfahren

Die verallgemeinerte gewichtete Inverse-χ^2-Methode, die als Varianz vergebendes Design konstruiert ist, lässt sich als Kombinationstest interpretieren. Sie erweitert das Design nach Bauer und Köhne (1994), indem die einzelnen Stufen unterschiedlich stark gewichtet werden, vgl. (2.11). Betrachtet man das Design unter einem anderen Blickwinkel, ergibt sich ein Bezug zu den allgemeineren Designs nach Müller und Schäfer (2001, 2004), die auf der Berechnung und Aufteilung des bedingten verbleibenden Fehlers erster Art beruhen. Für den jeweils verbleibenden Studienteil kann in Abhängigkeit der Ergebnisse aus den bereits durchgeführten Studienteilen eine bedingte Fehlerfunktion formuliert werden (vgl. Abschnitte 2.2.2 und 2.2.4). So wie die Verwendung der Inverse-Normalmethode (vgl. Abschnitt 2.2.1) die Benutzung der Schranken gruppensequentieller Designs, die auf normalverteilten Zufallsvariablen mit bekannter Varianz beruhen, bei anderen Zufallsvariablen ermöglicht, ist mit Hilfe der Inverse-χ^2-Methode eine Veränderung der α vergebenden Funktion bei beliebigen stetigen Teststatistiken möglich. Die zum Teil sehr aufwändige Berechnung des verbleibenden bedingten Niveaus für den folgenden Teil der

Studie basierend auf der ursprünglichen Teststatistik wie zum Beispiel beim t-Test (vgl. Posch et al., 2004) wird durch die Verwendung von χ^2-Verteilungen zur Transformation der P-Werte vereinfacht. Umgekehrt lässt die Argumentation von Müller und Schäfer (2001) die Aufhebung der formalen Konstruktion einer Folge von Zufallsvariablen mit einer zufälligen endlichen Anzahl von Elementen gemäß (3.6) zu. Gleichzeitig lässt sich das Verfahren in die Gruppe der rekursiven Kombinationstests gemäß Brannath et al. (2002) einordnen, wie im Folgenden deutlich wird.

Durch a priori festgelegte Gewichte ν_1 und $\nu_2^* = \nu_G - \nu_1$ und Stichprobenumfänge n_1 und n_2^* ist ein zweistufiger gruppensequentieller Plan über die Kombinationsstatistik der P-Werte auf den einzelnen Stufen mit $S_1 = q_1(\nu_1)$ und $S_2 = q_1(\nu_1) + q_2(\nu_2^*)$, $q_i(\nu) = F^{-1}_{\chi^2(\nu)}(1 - p_i)$, $i = 1, 2$, definiert. Der kritische Wert auf beiden Stufen für einen Test zum globalen Niveau α ist die Konstante $cv_\alpha = \chi^2(\nu_G)_{1-\alpha}$. Nach der Durchführung der ersten Stufe lässt sich folgende bedingte Fehlerfunktion für den zweiten Studienteil in Abhängigkeit des P-Werts auf der zweiten Stufe bestimmen

$$A(p_1) = \alpha_2^* = \begin{cases} 1, & \text{wenn } S_1 \geq cv_\alpha, \\ 1 - F_{\chi^2(\nu_G - \nu_1)}(cv_\alpha - q_1(\nu_1)), & \text{sonst.} \end{cases} \tag{4.1}$$

Dabei ist $S_1 \geq cv_\alpha$ äquivalent zur Bedingung $p_1 \leq 1 - F_{\chi^2(\nu_1)}(cv_\alpha)$.

Nach Müller und Schäfer (2001) können die Anzahl und der Zeitpunkt weiterer Zwischenauswertungen in der Studie unter der Bedingung neu festgesetzt werden, dass die Testprozedur für den weiteren Studienverlauf das bedingte Niveau $A(p_1)$ einhält. Nach jeder Zwischenauswertung ist erneut eine Designänderung erlaubt, vorausgesetzt, der bedingte Fehler erster Art kann bestimmt werden und wird in den folgenden Studienteilen nicht überschritten. Mit der Inverse-χ^2-Methode wird nach Durchführung der ersten Stufe eine weitere zweistufige Studie zum Niveau $A(p_1)$ konstruiert, indem die Freiheitsgrade $\nu_G - \nu_1$ in $\nu_1' = \nu_2$ und $\nu_2' = \nu_G - \nu_\Sigma(2)$ aufgeteilt und Stichprobenumfänge n_2 und n_3^* bestimmt werden. Die Teststatistiken des neuen zweistufigen Plans sind gegeben durch $S_k' = \sum_{i=1}^k q_i'(\nu_i')$, $k = 1, 2$. Als kritischer Wert zur Einhaltung des bedingten Fehlers ergibt sich $cv'_{A(p_1)} = \chi^2(\nu_1' + \nu_2')_{1-A(p_1)} = cv_\alpha - q_1(\nu_1)$. Die Stichprobenumfänge n_i^*, $i \geq 2$, dienen der Beschreibung eines vollständig geplanten gruppensequentiellen Designs, werden jedoch im Verlauf der Studie nach der Interimsanalyse angepasst bzw. ersetzt. Die Stichprobenumfänge können analog zu (3.9) und (3.11) nach Aufteilung der Gewichte bestimmt werden, so dass nach der ersten Stufe des neu geplanten gruppensequentiellen Designs eine Power $1 - \beta_{g,k}$ bzw. für den verbleibenden Teil, bestehend aus einer finalen oder zwei geplanten Stufen, eine Power $1 - \beta$ erzielt wird.

Der bedingte Fehler erster Art für den weiteren Verlauf der Studie nach der Durchführung

von $(k-1)$ Stufen ergibt sich bei konsequenter Verwendung der Inverse-χ^2-Methode aus der bedingten Fehlerfunktion

$$A(p_1,\ldots,p_{k-1}) = \alpha_k^* = \begin{cases} 1, & \text{wenn } S_{k-1} \geq cv_\alpha, \\ 1 - F_{\chi^2(\nu_G-\nu_\Sigma(k-1))}(cv_\alpha - S_{k-1}), & \text{sonst.} \end{cases} \tag{4.2}$$

Die Verwendung der bedingten Ablehnwahrscheinlichkeit des Kombinationstests anstelle der Verwerfungswahrscheinlichkeit, die sich bei Verwendung eines gleichmäßig besten Tests ergibt, führt nach Schäfer et al. (2006) zu einem Powerverlust. Dessen Größe ist aufgrund der Schwierigkeit der Berechnung der bedingten Ablehnwahrscheinlichkeit nach dem Testprinzip von Müller und Schäfer (2004) noch nicht näher bestimmt und muss in Abhängigkeit des jeweils besten Tests zum Testproblem angegeben werden.

Durch die wiederholte Anwendung der Inverse-χ^2-Methode, wenn auch mit unterschiedlichen Gesamtfreiheitsgraden, lässt sich ein zur Testprozedur konsistenter globaler P-Wert gemäß Brannath et al. (2002) rekursiv definieren. Dazu wird für das zweistufige Design bei einer Vorgabe von ν_G Freiheitsgraden insgesamt und ν_1 auf der ersten Stufe ein P-Wert analog zu (2.32) verwendet:

$$\tilde{p}(p_1,p_2) = \begin{cases} p_1, & \text{falls } p_1 \leq \alpha_1 = 1 - F_{\chi^2(\nu_1)}(cv_\alpha), \\ 1 - F_{\chi^2(\nu_G)}(q_1(\nu_1) + q_2(\nu_G - \nu_1)), & \text{sonst.} \end{cases} \tag{4.3}$$

Nach der Durchführung von $(k-1)$ Stufen und der Entscheidung, als Fortsetzung der Studie einen zweistufigen Plan zu wählen, ergibt sich für den P-Wert ab Stufe k, $k > 1$, der Wert

$$\tilde{p}_k(p_k,p_{k+1}) = \begin{cases} p_k, & \text{falls } p_k \leq \alpha_k^* = 1 - F_{\chi^2(\nu_k)}(cv_\alpha - S_{k-1}), \\ 1 - F_{\chi^2(\nu_G-\nu_\Sigma(k-1))}(q_k(\nu_k) + q_{k+1}(\nu_G - \nu_\Sigma(k))), & \text{sonst.} \end{cases} \tag{4.4}$$

Nach (2.33) ergibt sich der globale P-Wert nach Studienende nach der K-ten Stufe als

$$\begin{aligned} p &= \tilde{p}_1(p_1,\tilde{p}_2(p_2,\tilde{p}_3(\ldots,\tilde{p}_{K-2}(p_{K-2},\tilde{p}_{K-1}(p_{K-1},p_K))))) \\ &= 1 - F_{\chi^2(\nu_G)}(S_K). \end{aligned} \tag{4.5}$$

Dabei kann p_K entweder der P-Wert aus der finalen Stufe unter Verwendung der verbleibenden Freiheitsgrade sein oder der P-Wert der ersten Stufe eines zweistufigen Designs, geplant nach Stufe $(K-1)$, der zu einem vorzeitigen Abbruch geführt hat. Nach Definition des kritischen Werts gilt $p \leq \alpha$ genau dann, wenn $S_K \geq cv_\alpha$.

Die Aufteilung des globalen Niveaus α auf die erste und die folgenden Stufen bzw. des verbleibenden Fehlers $A(p_1,\ldots,p_{k-1})$ auf die folgenden Stufen wird über die pro Stufe vergebenen Freiheitsgrade festgelegt. Die Zusammenhänge zwischen den verwendeten Freiheitsgraden und den Fehlerraten werden im folgenden Abschnitt erläutert.

4.2 Zusammenhang von Designparametern und lokalen Fehlerraten

4.2.1 Fehler erster Art

Im Verlauf einer Studie gemäß verallgemeinerter Inverse-χ^2-Methode werden die vor Beginn gewählten Freiheitsgrade ν_G vergeben und im ursprünglichen Design von Hartung und Knapp (2003) durch eine proportionale Gewichtung mit dem Stichprobenumfang verbunden. So wird durch Konstruktion ein Zusammenhang zwischen Gewicht und bedingter Power auf den einzelnen Stufen hergestellt. Die Aufteilung der Freiheitsgrade führt jedoch vor allem dazu, dass der Fehler erster Art vergeben wird. Durch die Inverse-χ^2-Methode muss keine α vergebende Funktion a priori festgelegt werden. Der vergebene Fehler erster Art pro Stufe wird im Verlauf der Studie durch die pro Studienabschnitt gewählte Anzahl Freiheitsgrade bestimmt.

Für die erste Stufe wird durch die Vorgabe von ν_1 und ν_G das Niveau

$$\alpha(1) = \alpha_1 = 1 - F_{\chi^2(\nu_1)}(cv_\alpha) \tag{4.6}$$

festgelegt. Unterschreitet der P-Wert der ersten Stufe diesen Wert, so kann die Studie vorzeitig mit Ablehnung der Nullhypothese abgebrochen werden. Wird die Studie fortgesetzt, ergibt sich eine Fehlerwahrscheinlichkeit für den verbleibenden Studienteil, bedingt auf den P-Wert der ersten Stufe.

Beispiel 1

Hartung und Knapp (2003) schlagen für die erste Stufe eines Versuchs ein Gewicht von $\nu_1 = 2$ *bei einer Gesamtanzahl von Freiheitsgraden* $\nu_G = 10$ *vor. Daraus ergibt sich bei einem globalen Niveau von* $\alpha = 0.025$ *ein lokales Niveau*

$$\alpha_1 = 1 - F_{\chi^2(2)}(\chi^2(10)_{1-0.025}) = 3.5656 \cdot 10^{-05}.$$

Das lokale Niveau der ersten Stufe liegt damit eher in der Nähe des lokalen Niveaus des gruppensequentiellen Plans nach O'Brien und Fleming (1979) mit $2.5374 \cdot 10^{-06}$ *als nach Pocock (1977) mit* 0.0079*, vgl. Tabelle 2.1. Wird auf der ersten Stufe beispielsweise ein P-Wert von 0.2 beobachtet, der im Fortsetzungsbereich der Studie liegt, ergibt sich für den verbleibenden Studienteil gemäß (4.1) ein bedingter Fehler von*

$$1 - F_{\chi^2(\nu_G - \nu_1)}(cv_\alpha - \chi^2(\nu_1)_{1-p_1}) = 1 - F_{\chi^2(8)}(\chi^2(10)_{1-0.025} - \chi^2(2)_{0.8}) = 0.0275.$$

Sofern auf der zweiten Stufe die acht verbleibenden Freiheitsgrade verwendet werden, endet die Studie mit einer Ablehnung der Nullhypothese, wenn der P-Wert der zweiten Stufe kleiner als 0.0275 ausfällt.

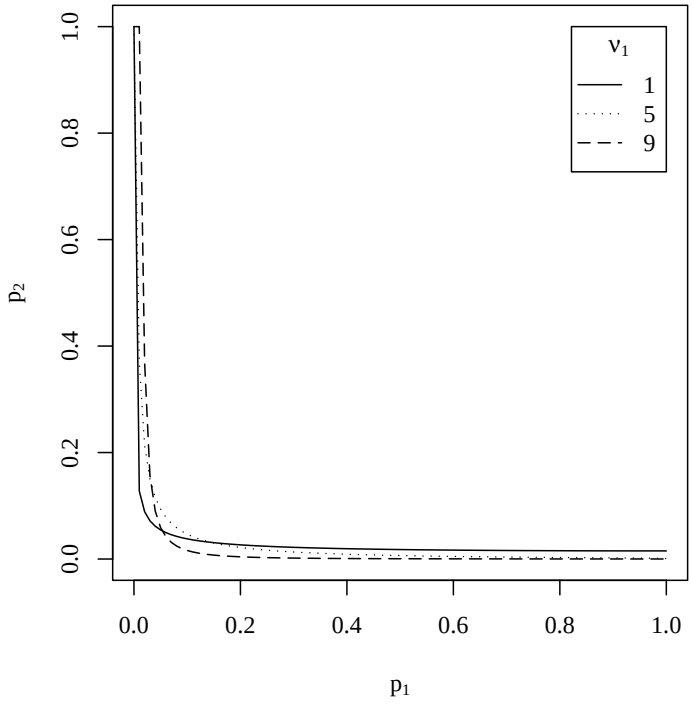

Abbildung 4.1: *Ablehnregionen für unterschiedliche Aufteilungen der Gesamtfreiheitsgrade $\nu_G = 10$ auf zwei Stufen, $\alpha = 0.025$*

Durch die bedingte Fehlerfunktion (4.1) sind Ablehnregionen bezüglich der P-Werte der ersten und zweiten Stufe bei vollständiger Vergabe der verbleibenden Freiheitsgrade auf der zweiten Stufe in Abhängigkeit der Gesamtfreiheitsgrade ν_G und den Freiheitsgraden ν_1 auf der ersten Stufe definiert. Abbildung 4.1 zeigt Ablehnregionen für unterschiedliche Gewichtungen der ersten und zweiten Stufe bei $\nu_G = 10$. Mit dem Anstieg der Anzahl Freiheitsgrade auf der ersten Stufe verschiebt sich die Ablehnregion in den linken Bereich. Das heißt, wenn ein kleiner P-Wert (ungefähr < 0.005) auf der ersten Stufe vorliegt, führt in der zweiten Stufe bei $\nu_1 = 9$ Freiheitsgraden ein wesentlich größerer P-Wert noch zur Ablehnung der Nullhypothese als bei $\nu_1 = 1$. Bei größeren P-Werten auf der ersten Stufe gilt dies umgekehrt.

Da die durch (4.1) und (4.2) definierten bedingten Fehlerfunktionen die Vergabe des lokalen Niveaus gemäß (4.6) und damit den Verlauf der Studie festlegen, muss der Einfluss der Anzahl aufzuteilender Freiheitsgrade ν_G, des Anteils der ersten Stufe $\epsilon_1 = \nu_1/\nu_G$ und des Signifikanzniveaus α näher untersucht werden. Es genügt die Betrachtung von (4.1), da (4.2) durch Umformen der Bedingung $S_{k-1} \geq cv_\alpha$ in $p_{k-1} \leq 1 - F_{\chi^2(\nu_{k-2})}(cv_\alpha - S_{k-2})$ und Betrachtung unterschiedlicher Niveaus eine analoge Form erhält.

Die Abhängigkeit des lokalen Niveaus der ersten Stufe (4.6) von der Gesamtanzahl Freiheitsgrade ν_G und dem gewählten Anteil $\epsilon_1 = \nu_1/\nu_G$ ist in Abbildung 4.2 dargestellt. Es

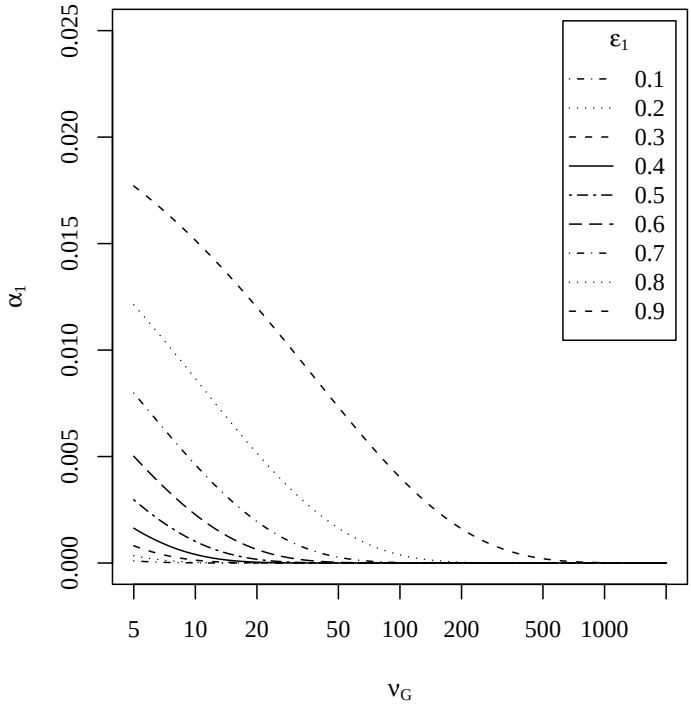

Abbildung 4.2: *Lokales Niveau auf der ersten Stufe in Abhängigkeit der Anzahl Freiheitsgrade ν_G und dem vergebenen Anteil ϵ_1 auf der ersten Stufe, $\alpha = 0.025$*

fällt auf, dass bei einer kleinen Anzahl Freiheitsgrade ($\nu_G \leq 10$) der Anteil des vergebenen Niveaus α_1 am globalen Niveau α annähernd proportional zum Anteil ϵ_1 ist. Bei einer sehr großen Anzahl Freiheitsgrade geht das lokale Niveau auf der ersten Stufe selbst bei einem Anteil von 90% der Freiheitsgrade für ν_1 gegen Null. Für ein vorgegebenes lokales Niveau α_1 auf der ersten Stufe lässt sich aus (4.6) die benötigte Anzahl Freiheitsgrade ν_1 und damit der Anteil ϵ_1 iterativ bestimmen. Bei Freiheitsgraden von $5, 10, 100$ und 1000 ergeben sich z.B. bei $\alpha = 0.025$ und $\alpha_1 = 0.005$, d.h. bei Vergabe von 20% des globalen Niveaus auf der ersten Stufe, Anteile ϵ_1 von 29.97%, 71.21%, 91.06% und 97.22% an den Gesamtfreiheitsgraden für die erste Stufe. Eine intuitive Wahl des vergebenen Niveaus über den Anteil der Freiheitsgrade ist auf Grund dieses Sachverhalts höchstens im Bereich einer sehr kleinen Anzahl Freiheitsgrade möglich.

Nach der Durchführung einer Stufe ergeben sich für den Studienverlauf, wenn die Teststatistik im Fortsetzungsbereich liegt, zwei Möglichkeiten. Zum einen können die verbleibenden Freiheitsgrade in einer finalen Stufe verwendet werden, zum anderen können sie auf zunächst zwei weitere Stufen aufgeteilt werden. Im zweiten Fall ergibt sich in Abhängigkeit von der bedingten Fehlerfunktion (4.2) das Niveau, das analog zu (4.6) auf die zwei Stufen aufgeteilt werden kann. Die Form der Graphen des lokalen Niveaus in Abhängigkeit der Gesamtzahl Freiheitsgrade bleibt auch für andere Werte von α bei entsprechender

Skalierung der y-Achse erhalten (vgl. Abbildungen 4.2 und A.1).

Da in jedem zweistufigen Design nach der Inverse-χ^2-Methode das globale Niveau bei Gleichverteilung der P-Werte ausgeschöpft wird, wird bei Verwendung der restlichen Freiheitsgrade auf der zweiten Stufe im Erwartungswert über die P-Werte der ersten Stufe das verbleibende Niveau $\alpha_2 = \alpha - \alpha_1$ vergeben. Wie hoch das bedingte Niveau ist, das für die Planung des Stichprobenumfangs auf der zweiten Stufe benötigt wird, ergibt sich aus der bedingten Fehlerfunktion. Eine Verschiebung des durch die bedingte Fehlerfunktion definierten Ablehnbereichs nach links wie in Abbildung 4.1 bei steigendem Anteil ϵ_1 ist für jede Anzahl Freiheitsgrade ν_G und jedes Niveau α zu beobachten (siehe Abbildungen A.2 und A.3).

Eine gleichmäßige Aufteilung der Freiheitsgrade auf zwei Stufen führt bei der Wahl von $\nu_G = 4$ zur Kombinationsmethode nach Fisher (siehe 2.8). Aus den Abbildung 4.3, A.2 und A.3 ist ersichtlich, dass sich bzgl. der Verschiebung bei steigendem Anteil an den Gesamtfreiheitsgraden auf der ersten Stufe die Ablehnregionen für alle betrachteten ν_G ähneln. Insgesamt ergibt sich jedoch eine systematische Verschiebung hinsichtlich der Höhe der Anzahl Freiheitsgrade. Bei einer großen Zahl ν_G führen ähnlich große P-Werte auf den beiden Stufen eher zur Ablehnung als bei kleinen, während bei der Wahl von insgesamt vier Freiheitsgraden häufiger extreme Werte auf einer einzelnen Stufe noch zur Ablehnung führen. Die Unterschiede in den Ablehnregionen werden für große (bedingte) Fehlerwahrscheinlichkeiten erster Art α bzw. α_k^*, wie sie nach der Durchführung mehrerer Stufen auftreten können, deutlicher. Insbesondere zeigt sich hier ebenfalls, dass auch bei höheren Niveaus die Wahrscheinlichkeit, die Studie auf der ersten Stufe zu beenden, bei einer großen Anzahl an Freiheitsgraden insgesamt sehr gering ausfällt, während bei 4 und 10 Freiheitsgraden der Bereich, in dem $A(p_1) = 1$ gilt, deutlich wächst.

Die Unterschiede in der bedingten Fehlerfunktion sowie im vergebenen Niveau bei konstantem Anteil, aber unterschiedlichen Anzahlen an Freiheitsgraden zeigen, dass die Gewichtung nicht unabhängig von der Größe der zu vergebenden Varianz beurteilt werden kann. Gewichte gleicher Größe wirken sich für unterschiedliche Gesamtfreiheitsgrade nicht gleich aus. Während die Inverse-Normalmethode im Design nach Lehmacher und Wassmer (1999), Shen und Fisher (1999) oder Hartung (2001) durch Standardisierung auch bei anderen Werten als 1 für die Varianz der finalen Teststatistik bei anteilig gleichem Gewicht zu gleichen bedingten Fehlerraten führt, ist für die Inverse-χ^2-Methode zu berücksichtigen, dass die Freiheitsgrade einer χ^2-Verteilung Erwartungswert und Varianz der Verteilung bestimmen. Daraus ergibt sich ein nahezu linearer Anstieg des $(1-\alpha)$-Quantils in der Anzahl an Freiheitsgraden, der gleichzeitig bei konstantem Anteil an den Freiheitsgraden für die erste Stufe zu einem Absinken des lokalen Niveaus auf der ersten Stufe führt.

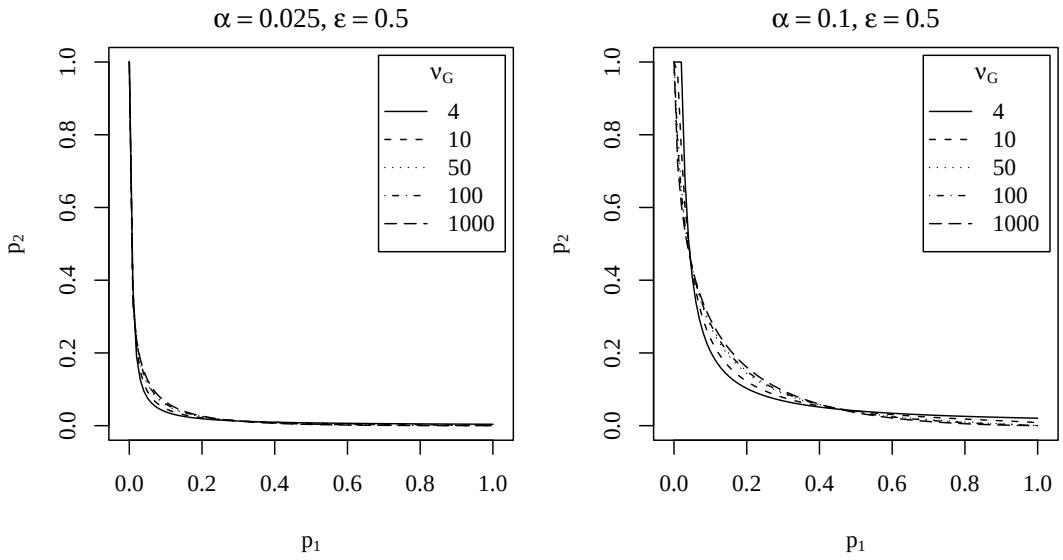

Abbildung 4.3: *Bedingte Fehlerfunktion in Abhängigkeit der Anzahl Freiheitsgrade ν_G bei gleichmäßiger Aufteilung auf zwei Stufen*

Bei einer mehrstufigen Studie gemäß der Inverse-χ^2-Methode verringert sich die Anzahl verbleibender Freiheitsgrade mit jeder Stufe. Daraus folgt, dass mit Fortschreiten der Studie die bedingte Fehlerfunktion (4.2) für den verbleibenden Studienteil nach Stufe $(k-2)$ im Rahmen des rekursiven Kombinationsprinzips eher einen Stopp nach der Durchführung der ersten von zwei folgenden Stufen zulässt im Vergleich zu einer bedingten Fehlerfunktion bei einer größeren Anzahl verbleibender Freiheitsgrade bei gleichem Aufteilungsverhältnis. Das bedingte Niveau der nächsten Stufe, α'_{k-1}, hängt dabei von den Realisierungen der bisherigen Stufen, zusammengefasst in der Statistik S_{k-2}, und den aufgewendeten Freiheitsgraden $\nu_{k-1} \leq \nu_G - \nu_\Sigma(k-2)$ ab:

$$\alpha'_{k-1} = 1 - F_{\chi^2(\nu_{k-1})}(cv_\alpha - S_{k-2}). \tag{4.7}$$

4.2.2 Power

Betrachtungen der Power erfordern die Festlegung der verwendeten Teststatistik T_k bzw. die Kenntnis der Verteilung der beobachteten Zufallsvariable auf jeder Stufe k. Im Folgenden wird davon ausgegangen, dass auf jeder Stufe die Erwartungswerte zweier Normalverteilungen mit bekannter Varianz σ^2 mit einem einseitigen Test verglichen werden.

Wenn θ der wahre Parameterwert ist, gilt für die Teststatistik z_k des Gauß-Tests auf einer einzelnen Stufe k des adaptiven Designs gemäß (2.2) bei einem Stichprobenumfang $n_k/2$ pro Gruppe, dass

$$z_k \sim N\left(\sqrt{\frac{n_k}{4}} \cdot \frac{\theta}{\sigma}, 1\right). \tag{4.8}$$

Mit dem P-Wert der k-ten Stufe $p_k = 1 - \Phi(z_k)$ und der Transformation des P-Werts $q_k(\nu_k) = F^{-1}_{\chi^2(\nu_k)}(1-p_k)$ ergibt sich für den wahren Wert θ auf der ersten Stufe eine Power von

$$\begin{aligned}
\pi_1 &= \mathrm{P}_\theta\big(q_1(\nu_1) \geq cv_\alpha\big) \\
&= \mathrm{P}_\theta\left(F^{-1}_{\chi^2(\nu_1)}(\Phi(z_1)) \geq cv_\alpha\right) \\
&= \mathrm{P}_\theta\left(z_1 \geq \Phi^{-1}(F_{\chi^2(\nu_1)}(cv_\alpha))\right) \\
&= 1 - \Phi\left(\Phi^{-1}(F_{\chi^2(\nu_1)}(cv_\alpha)) - \sqrt{\frac{n_1}{4}} \cdot \frac{\theta}{\sigma}\right). \qquad (4.9)
\end{aligned}$$

Nach Realisierung von $(k-1)$ Stufen und damit der Teststatistik S_{k-1} und der Festlegung der Anzahl Freiheitsgrade $\nu_k \leq \nu_G - \nu_\Sigma(k-1)$ sowie des Stichprobenumfangs n_k für die k-te Stufe berechnet sich die bedingte Power auf der k-ten Stufe gemäß

$$\begin{aligned}
\pi_k^* &= \mathrm{P}_\theta\big(q_k(\nu_k) \geq cv_\alpha - S_{k-1}\big|S_{k-1}\big) \\
&= 1 - \Phi\left(\Phi^{-1}\left(F_{\chi^2(\nu_k)}\left(cv_\alpha - \sum_{i=1}^{k-1} F^{-1}_{\chi^2(\nu_i)}(\Phi(z_i))\right)\right) - \sqrt{\frac{n_k}{4}} \cdot \frac{\theta}{\sigma}\right). \qquad (4.10)
\end{aligned}$$

Die nicht bedingte Berechnung erfordert die Festlegung einer Regel für die Wahl des Stichprobenumfangs und des Gewichts auf der k-ten Stufe. Stichprobenumfänge und Gewichte sind als Funktionen von $z_1, \ldots, z_{k-1}$ oder in Abhängigkeit von externen Informationen Zufallsvariablen. Bei Vorgabe einer solchen Regel ist zu berücksichtigen, dass die Power auf der k-ten Stufe für $k \geq 2$ ein Erwartungswert über Studien ist, deren Verlauf sich in Bezug auf die bis zur $(k-1)$-ten Stufe vergebene und die in Stufe k verwendete Anzahl Freiheitsgrade unterscheidet.

Durch eine a priori Festlegung der Gewichte und der Stichprobenumfänge lässt sich der Powerverlust, der durch die Verwendung eines Kombinationstests entsteht, im Vergleich zum einstufigen Design berechnen. Im Folgenden wird für ein zweistufiges Design gemäß der Inverse-χ^2-Methode die Reduktion der Power bei unterschiedlicher Aufteilung der Freiheitsgrade und proportionaler Aufteilung des Stichprobenumfangs aus dem einstufigen Design, wie sie Hartung und Knapp (2003) vorschlagen, bei einem Zwei-Gruppen-Vergleich beschrieben. Die auf der zweiten Stufe erzielte Power berechnet sich als Integral über die

bedingte Power π_2^* gemäß (4.10)

$$\begin{aligned}
\pi_2 &= \mathrm{P}_\theta\big(q_1(\nu_1) + q_2(\nu_2) \geq cv_\alpha, q_1(\nu_1) < cv_\alpha\big) \\
&= \int_{-\infty}^{\Phi^{-1}(1-\alpha_1)} \mathrm{P}_\theta\big(q_2(\nu_2) \geq cv_\alpha - q_1(\nu_1)\big|Z_1 = z_1\big) \cdot \phi\left(z_1 - \sqrt{\frac{n_1}{4}} \cdot \frac{\theta}{\sigma}\right) \partial z_1 \\
&= 1 - \pi_1 - \int_{-\infty}^{\Phi^{-1}(1-\alpha_1)} \Phi\left(\Phi^{-1}\left(F_{\chi^2(\nu_2)}(cv_\alpha - F^{-1}_{\chi^2(\nu_1)}(\Phi(z_1)))\right) - \sqrt{\frac{n_2}{4}} \cdot \frac{\theta}{\sigma}\right) \\
&\quad \cdot\phi\left(z_1 - \sqrt{\frac{n_1}{4}} \cdot \frac{\theta}{\sigma}\right) \partial z_1 \qquad (4.11)
\end{aligned}$$

wobei α_1 gemäß (4.6) bestimmt wird.

Im einstufigen Design ergibt sich zum Nachweis eines vorgegebenen Unterschieds $\theta_0 > 0$ in den Erwartungswerten bei einer Power von $1-\beta$ ein notwendiger Stichprobenumfang von $M_1 = N_{\text{fix}}$ aus (2.24). In der Praxis resultiert aufgrund von Rundung auf ein Vielfaches von 2 zur Aufteilung von Patienten auf zwei Gruppen für θ_0 eine Power, die mindestens $1-\beta$ beträgt. In den folgenden Berechnungen werden nicht gerundete Stichprobenumfänge betrachtet, um Fallunterscheidungen zu vermeiden. Wird für die erste Stufe des adaptiven Designs ein Anteil ϵ_1 an den Freiheitsgraden und am Umfang M_1 gewählt und auf der zweiten Stufe jeweils der verbleibende Anteil $(1-\epsilon_1)$, lässt sich die Gesamtpower $\pi_{2\text{st}}$ eines zweistufigen Designs für eine feste Wahl von ν_G berechnen. Diese ist unabhängig von θ/σ, da hier geplante und wahre Parameter identisch gewählt werden.

$$\begin{aligned}
\pi_{2\text{st}} &= 1 - \int_{-\infty}^{\Phi^{-1}(F_{\chi^2(\epsilon_1\nu_G)}(cv_\alpha))} \Phi\Bigg(\Phi^{-1}\left(F_{\chi^2((1-\epsilon_1)\nu_G)}(cv_\alpha - F^{-1}_{\chi^2(\epsilon_1\nu_G)}(\Phi(z_1)))\right) \\
&\quad - \sqrt{\frac{(1-\epsilon_1)M_1}{4}} \cdot \frac{\theta}{\sigma}\Bigg) \cdot \phi\left(z_1 - \sqrt{\frac{\epsilon_1 M_1}{4}} \cdot \frac{\theta}{\sigma}\right) \partial z_1 \qquad (4.12) \\
&= 1 - \int_{-\infty}^{\Phi^{-1}(F_{\chi^2(\epsilon_1\nu_G)}(cv_\alpha))} \Phi\Bigg(\Phi^{-1}\left(F_{\chi^2((1-\epsilon_1)\nu_G)}(cv_\alpha - F^{-1}_{\chi^2(\epsilon_1\nu_G)}(\Phi(z_1)))\right) \\
&\quad - \sqrt{(1-\epsilon_1)\cdot(u_{1-\alpha} + u_{1-\beta})^2}\Bigg) \cdot \phi\left(z_1 - \sqrt{\epsilon_1 \cdot (u_{1-\alpha} + u_{1-\beta})^2}\right) \partial z_1
\end{aligned}$$

Tabelle 4.1 enthält die mittels numerischer Integration bestimmte globale Power sowie die Power auf der ersten Stufe an der Stelle des zur Planung verwendeten Werts θ_0/σ für unterschiedliche Kombinationen von ϵ_1 und ν_G für einen Zweistichproben-Gauß-Test zum globalen Niveau von 0.025 und geplanter globaler Power von 0.9. Wie zu erwarten liegen alle Werte unter der Power des einstufigen Designs von 0.9. Die größte Abweichung

mit 0.88352 tritt bei Verwendung von vier Freiheitsgraden und einem Anteil von 0.5 auf. Insgesamt liegt damit die Reduktion der Power, die als Preis für die Verwendung des adaptiven Designs angesehen werden kann, im Vergleich zum einstufigen Design in einem akzeptablen Bereich. Mit steigender Zahl Freiheitsgrade ν_G nähert sich die erreichte Power der Power des einstufigen Verfahrens an. Ebenso werden mit steigender Anzahl die Unterschiede bezüglich des gewählten Anteils ϵ_1 geringer. Die kleinste Power bei konstanten Freiheitsgraden wird jeweils für den Anteil von 0.5 erzielt. Aufgrund des relativ geringen lokalen Niveaus bei einer großen Zahl an Freiheitsgraden sowie bei kleinem Anteil ϵ_1 (vgl. Abschnitt 4.2.1) fällt die Power auf der ersten Stufe in diesen Fällen gering aus. Für $\nu_G = 1000$ und 10000 wird die globale Power fast vollständig in der zweiten Stufe erreicht. Für 4 und 10 Freiheitsgrade ergeben sich bei Aufwendung der Hälfte der Freiheitsgrade und des einstufigen Stichprobenumfangs für die Power auf der ersten Stufe Werte von 0.35 und 0.21, während bei einem Anteil von 0.9 auf der ersten Stufe 0.84 bzw. 0.82 erreicht werden.

Löst man sich von der proportionalen Aufteilung des Stichprobenumfangs und der Freiheitsgrade, ist zu bemerken, dass mit $n_1 \leq M_1$ und $\nu_1 < \nu_G$ und damit vorgegebenem Signifikanzniveau auf der ersten Stufe nicht jede Power $1 - \beta_{g,1} \leq 1 - \beta$ für ein vorgegebenes θ_0 erreicht werden kann. Die maximale Power auf der ersten Stufe ergibt sich für $n_1 = M_1$ beim einseitigen Gauß-Test zu

$$1 - \beta_{g,1}^{\max} = 1 - \Phi\left(\Phi^{-1}\left(F_{\chi^2(\nu_1)}(cv_\alpha) - \sqrt{\frac{M_1}{4}} \cdot \frac{\theta_0}{\sigma}\right)\right). \tag{4.13}$$

Beispiel 2

Geht man von einem einseitigen Zweistichproben-Gauß-Test aus, wobei die Varianz der beobachteten Zufallsvariablen als $\sigma^2 = 1$ angenommen wird, ergibt sich bei den Fehlerraten $\alpha = 0.025$ und $\beta = 0.1$ sowie einem nachzuweisenden Unterschied in den Erwartungswerten von $\theta_0 = 0.5$ ein benötigter Stichprobenumfang von insgesamt $M_k = 170$. Die Verwendung von $\nu_1 = 2, 4, 8$ Freiheitsgraden mit $n_1 = M_1 = 170$ führt bei $\nu_G = 10$ zu einer maximalen Power auf der ersten Stufe bei $\theta = \theta_0$ von 0.2382, 0.4631 bzw. 0.8104.

Wird im zweistufigen Design eine bedingte Power für einen Wert θ/σ auf der zweiten Stufe festgelegt, kann durch Auflösen der Gleichung (4.10) nach n_2 der benötigte Stichprobenumfang auf der zweiten Stufe bestimmt werden.

Der Stichprobenumfang n_2 ist für wenige Freiheitsgrade ν_G eine konvexe Funktion in Abhängigkeit der Teststatistik z_1 wie im Design nach Bauer und Köhne (1994) als Spezialfall der Inverse-χ^2-Methode und in den optimalen Designs nach Brannath und Bauer (2004).

Tabelle 4.1: *Globale Power und Power auf der ersten Stufe (kursiv) eines zweistufigen Testverfahrens nach Inverse-χ^2-Methode bei proportionaler Aufteilung der Freiheitsgrade ν_G und des einstufigen Stichprobenumfangs (einseitiger Gauß-Test mit $\alpha = 0.025$, $1 - \beta = 0.9$) bei Planung mit dem wahren Wert θ/σ*

	ν_G					
ϵ_1	4	10	50	100	1000	10000
0.1	0.88751	0.89293	0.89805	0.89895	0.89988	0.89999
	0.00564	*0.00040*	*<0.00001*	*<0.00001*	*<0.00001*	*<0.00001*
0.2	0.88486	0.89229	0.89802	0.89894	0.89988	0.89999
	0.03521	*0.00583*	*<0.00001*	*<0.00001*	*<0.00001*	*<0.00001*
0.3	0.88396	0.89212	0.89801	0.89894	0.89988	0.89999
	0.10423	*0.03092*	*0.00002*	*<0.00001*	*<0.00001*	*<0.00001*
0.4	0.88361	0.89207	0.89801	0.89894	0.89988	0.89999
	0.21446	*0.09644*	*0.00085*	*<0.00001*	*<0.00001*	*<0.00001*
0.5	0.88352	0.89205	0.89801	0.89894	0.89988	0.89999
	0.35313	*0.21359*	*0.01117*	*0.00031*	*<0.00001*	*<0.00001*
0.6	0.88361	0.89207	0.89801	0.89894	0.89988	0.89999
	0.50015	*0.37183*	*0.06781*	*0.00865*	*<0.00001*	*<0.00001*
0.7	0.88396	0.89212	0.89801	0.89894	0.89988	0.89999
	0.63676	*0.54326*	*0.22746*	*0.08093*	*<0.00001*	*<0.00001*
0.8	0.88486	0.89229	0.89802	0.89894	0.89988	0.89999
	0.75090	*0.69820*	*0.48384*	*0.32046*	*0.00014*	*<0.00001*
0.9	0.88751	0.89293	0.89805	0.89895	0.89988	0.89999
	0.83816	*0.81836*	*0.73702*	*0.66461*	*0.12974*	*<0.00001*

Ab etwa 100 Freiheitsgraden ist die Funktion konkav wie im Design gemäß der gewichteten Inverse-Normalmethode, da mit steigender Anzahl Freiheitsgrade die χ^2-Verteilung gegen eine Normalverteilung konvergiert (vgl. Johnson und Kotz, 1970). Für $\theta/\sigma = 0.5$ und eine bedingte Power von 0.9 sind im Anhang für unterschiedliche ν_G und ϵ_1 die Stichprobenfunktionen abgebildet (Abbildung A.4). Abbildung 4.4 fasst die Graphen des benötigten Stichprobenumfangs n_2 bei gleichmäßiger Aufteilung der Freiheitsgrade auf zwei Stufen aus Abbildung A.4 für verschiedene ν_G zusammen, so dass der Wechsel im Krümmungsverhalten deutlich wird. Bei großen Teststatistiken auf der ersten Stufe führen größere Gewichte auf der ersten Stufe zu kleineren benötigten Stichprobenumfängen auf der zweiten Stufe. Fällt z_1 gering aus, ist es von Vorteil für die Stichprobenplanung, wenn ein großer Anteil des Gewichts noch in der zweiten Stufe zu vergeben ist. Ein hoch

gewichteter schwacher oder negativer Trend aus der ersten Stufe ist nur mit einem großen Stichprobenumfang zu kompensieren.

Gemäß den zugrundeliegenden bedingten Fehlerfunktionen (4.1) ist der benötigte Stichprobenumfang für eine große Anzahl an Freiheitsgraden für sehr große und sehr kleine Werte von z_1 größer als bei wenigen Freiheitsgraden, während sich im mittleren Bereich dieses Verhalten umkehrt. Es ist zu bemerken, dass der beobachtete Effekt in der Nullhypothese liegt, sofern $z_1 \leq 0$ ist. Insbesondere bei einem großen Stichprobenumfang auf der ersten Stufe ist in diesen Fällen ein Stopp wegen Aussichtslosigkeit zu erwägen (siehe auch Abschnitt 4.4).

Der erwartete Stichprobenumfang auf der zweiten Stufe $\mathrm{E}(n_2)$ und insgesamt als Summe $\epsilon_1 \cdot M_1 + \mathrm{E}(n_2)$ sowie die erwartete Power einer zweistufigen Studie lassen sich mittels numerischer Integration oder Simulation bestimmen. Die Berechnung erfolgt für eine vorgegebene bedingte Power von $1-\beta_{g,2}$ für ein festes θ oder eine auf Werte aus der Alternative beschränkte Schätzung $\hat{\theta}_1$ und Vergabe der verbleibenden Freiheitsgrade.

Aufgrund unterschiedlicher lokaler Niveaus auf der ersten Stufe in Abhängigkeit der Freiheitsgrade ν_G und ν_1 und damit unterschiedlicher Power auf der ersten Stufe erweisen

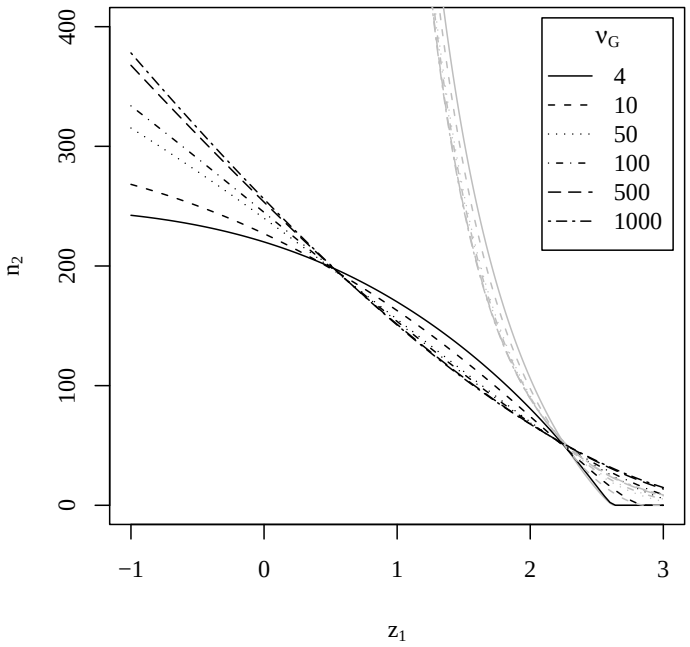

Abbildung 4.4: *Benötigter Stichprobenumfang für bedingte Power von 0.9 bei $\theta/\sigma = 0.5$ bzw. Schätzung (grau) auf der zweiten Stufe in Abhängigkeit der Teststatistik der ersten Stufe z_1 bei gleichmäßiger Aufteilung der Freiheitsgrade auf beide Stufen*

sich Vergleiche zwischen verschiedenen Einstellungen dieser Designparameter als schwierig. Tabelle A.1 fasst für verschiedene Freiheitsgrade und Anteile ϵ_1 den erwarteten Stichprobenumfang und die erwartete Power einer zweistufigen Studie mit Inverse-χ^2-Methode aus jeweils 10000 simulierten Studien zusammen. Der Parameter θ_0 zur Planung der Stichprobe entspricht auf der ersten Stufe dem wahren Parameter $\theta = 0.5$ und auf der zweiten Stufe dem wahren Parameter oder der Schätzung $\hat{\theta}_1 = \max(2\sigma/\sqrt{n_k} \cdot z_1, 0.001)$. Die Stichprobenumfänge M_1 und n_2 werden jeweils für eine (bedingte) Power von 0.9 und $\sigma^2 = 1$ bestimmt. Insbesondere bei der Verwendung des Schätzers ist eine Beschränkung des Stichprobenumfangs n_2 sinnvoll, um durchführbare Stichprobenumfänge zu erhalten. Die obere Grenze für n_2 wird für den folgenden Vergleich auf $10 \cdot M_1$ gesetzt. Bei Verwendung der Schätzung wechselt die Funktion des benötigten Stichprobenumfangs in z_1 nicht das Krümmungsverhalten in Abhängigkeit der Anzahl an Gesamtfreiheitsgraden. Die Funktion ist konkav, da neben dem bedingten Fehler auch der zur Umfangsberechnung verwendete Schätzer mit fallender Teststatistik z_1 abnimmt. Die grauen Linien in Abbildung 4.4 zeigen den steilen Anstieg des Stichprobenumfangs in Abhängigkeit von z_1 bei Verwendung des Schätzers. Die deutliche Trennung der Funktionen des benötigten Stichprobenumfangs n_2 bei Verwendung des Schätzers anstelle der Konstanten zeigt, sich sobald die Parameterschätzung kleiner ausfällt als der wahre Wert von 0.5, d.h. $z_1 < 2.305$.

Bei einer großen Anzahl Freiheitsgrade insgesamt wird gemäß Tabelle 4.1 bei konstantem Stichprobenumfang auf der zweiten Stufe eine etwas größere Power erzielt als bei kleinen Freiheitsgraden. Aufgrund der Unterschiede in den bedingten Fehlerfunktionen ist jedoch bei Adaption des Stichprobenumfangs bei Vorgabe einer bedingten Power für den wahren Parameter θ bei Anteilen $\epsilon_1 \geq 0.5$ in Designs mit vier oder zehn Freiheitsgraden bei etwa gleich großen erwarteten Stichprobenumfängen die erzielte Power etwas größer als in Designs mit mehr Freiheitsgraden insgesamt. Für kleinere Anteile an den Gesamtfreiheitsgraden und dem Umfang M_1 als Gewicht und Umfang auf der ersten Stufe ist bei etwa gleich großer erwarteter Power bei wenigen Freiheitsgraden der erwartete Stichprobenumfang etwas größer als bei vielen Freiheitsgraden.

In allen betrachteten Szenarien wird bei Verwendung des wahren Werts auf der zweiten Stufe eine Power von mindestens 90% erreicht. In Designs mit einem großen Stichprobenumfang auf der ersten Stufe liegt die erwartete Power mit bis zu 98.3% deutlich darüber. Der erwartete Stichprobenumfang liegt nur in den Designs mit einem Anteil von 0.9 auf der ersten Stufe über dem einstufigen Stichprobenumfang $M_1 = 170$. In Studien, in denen der Schätzer aus der ersten Stufe verwendet wird, wird bei kleinen Anteilen trotz eines erheblich erhöhten mittleren Stichprobenumfangs von etwa 600 Patienten, nur eine Power von 0.8 erzielt, was auf die Unsicherheit in der Schätzung des Parameters zurück-

zuführen ist. Der erwartete Stichprobenumfang bei Verwendung des Schätzers ist für alle Anteile deutlich höher als bei Verwendung des konstanten Wertes. Die Beschränkung des Schätzers bis zu einem Wert, der nahe der Nullhypothese liegt, führt zu sehr hohen Stichprobenumfängen bzw. zur Verwendung des Maximums, wie auch in Abschnitt 3.4 an den Häufungspunkten der Vielfachen des maximalen Umfangs zu erkennen ist. Die Beschränkung des Schätzers zur Berechnung des Stichprobenumfangs auf einer Stufe hat einen bedeutenden Einfluss auf den erwarteten Stichprobenumfang, so dass eher ein minimaler klinisch relevanter Wert gewählt werden sollte.

Durch die Gleichheit des Planungsparameters mit dem wahren Wert erscheint die Verwendung eines Schätzwerts als schlechte Alternative zur Verwendung eines konstanten Werts. Ist der wahre Parameter kleiner, jedoch noch klinisch relevant, z.B. $\theta = 0.3$, zeigt sich, dass bei Planung mit einem konstanten Wert von $\theta_0 = 0.5$ in den verschiedenen Designszenarien bei einer kleinen Anzahl an Freiheitsgraden etwa 69% und bei 500 und 1000 Freiheitsgraden etwa 65% der Studien mit einer Ablehnung der Nullhypothese enden, wenn bereits ein großer Stichprobenumfang auf der ersten Stufe gewählt wurde (vgl. Tabelle A.2). Bei einem Anteil von $\epsilon_1 = 0.1$ an den Gesamtfreiheitsgraden und am einstufigen Stichprobenumfang, geplant für $\theta_0 = 0.5$, werden bei konstantem Parameterwert nur etwa 50% Power erreicht. Wird der Parameterschätzer aus der ersten Stufe zur Planung des Umfangs auf der zweiten Stufe verwendet, wird bei kleinen Anteilen eine Power von etwa 73% erzielt, verbunden mit einer mehr als Vervierfachung des erwarteten Stichprobenumfangs im Vergleich zur Planung mit dem konstanten Wert. Der benötigte Stichprobenumfang im einstufigen Design für eine Power von 90% für $\theta = 0.3$ beträgt zum Vergleich 468. Ab einem Anteil von etwa 0.5 bei Verwendung des Schätzers überschreitet die Power des adaptiven Verfahrens 90%. Zur Senkung des erwarteten Stichprobenumfangs ist neben der Absenkung der oberen Schranke des Umfangs, der Erhöhung des minimalen Werts, für den eine Anpassung erfolgt, und der Einführung eines Abbruchs wegen Aussichtslosigkeit die Reduktion der bedingten Power möglich. Ausführliche Überlegungen zur Wahl des Stichprobenumfangs bzw. der bedingten Power sowie der Gewichte, insbesondere bei Erhöhung der Anzahl Interimsanalysen, folgen in Kapitel 5.

4.3 Exakte Berechnung des Stichprobenumfangs pro Stufe bei proportionaler Gewichtung

Hartung und Knapp (2003) wählen die Freiheitsgrade für die folgende Stufe anteilig an den verbleibenden Freiheitsgraden gemäß dem Verhältnis des Stichprobenumfangs m_k auf der folgenden Stufe zum benötigten Stichprobenumfang für eine finale Stufe mit einer

bedingten Power $1-\beta$. Da die Berechnung des Stichprobenumfangs m_k in (3.11) unter der Annahme erfolgt, dass das Gewicht des Studienabschnittes gleich der Anzahl der noch nicht verbrauchten Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ ist, ergibt sich bei Verwendung eines Anteils $\epsilon_k = m_k/M_k$ an den verbleibenden Freiheitsgraden im k-ten Studienabschnitt eine Veränderung für den bedingten Fehler erster Art

$$\alpha'_k = 1 - F_{\chi^2(\epsilon_k \cdot (\nu_G - \nu_\Sigma(k-1)))}(cv_\alpha - S_{k-1}). \tag{4.14}$$

Da die Wahrscheinlichkeitsfunktion der χ^2-Verteilung monoton fallend in der Anzahl Freiheitsgrade ist, fällt der tatsächliche lokale Fehler erster Art im k-ten Studienabschnitt kleiner aus als der bedingte Fehler α^*_k aus (4.2), der in der Berechnung des Umfangs angenommen wird. Dadurch führt der Stichprobenumfang m_k bei Verwendung von $\nu_k = \epsilon_k \cdot (\nu_G - \nu_\Sigma(k-1))$ zu einer geringeren Power als $1-\beta_{g,k}$.

Beispiel 3

Mit den Vorgaben aus Beispiel 2 (siehe S. 50) soll auf der ersten Stufe einer Studie mit der globalen Anzahl von Freiheitsgraden $\nu_G = 10$ *eine Power von* $1-\beta_{g,1} = 0.6$ *für* $\theta = 0.5$ *erzielt werden. Dies führt bei einem Signifikanzniveau von* $\alpha = 0.025$ *zu einem Stichprobenumfang von* $m_k = 80$*. Daraus lässt sich die proportionale Anzahl an Freiheitsgraden für die erste Stufe* $\nu_1 = (m_k/M_k) \cdot \nu_G = 4.7$ *und das tatsächliche Niveau auf der ersten Stufe* $\alpha_1 = 1 - F_{\chi^2(4.7)}(F^{-1}_{\chi^2(10)}(1-0.025)) = 0.00078$ *berechnen. Diese Reduktion des Niveaus führt bei einem wahren Effekt von 0.5 zu einer Senkung der Power auf* $1-\beta_1 = 1 - \Phi(\Phi^{-1}(1-\alpha_1) - \theta/\sigma \cdot \sqrt{m_k/4}) = 0.177$.

Um eine proportionale Wahl der Anzahl Freiheitsgrade und der Umfänge beizubehalten und eine angegebene Power $1-\beta_{g,k} < 1-\beta$ auf der Stufe k zu erreichen, ersetzt man die Anzahl Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ in Formel (3.11) durch $\nu_k = (m_k/M_k) \cdot (\nu_G - \nu_\Sigma(k-1))$. Werden die Stichprobenumfänge m_k und M_k basierend auf einer Schätzung des Parameters $\hat{\theta}_{k-1}$ berechnet, muss diese Schätzung auf den Wertebereich der Alternative beschränkt werden, d.h. $\hat{\theta}_{k-1} > 0$.

Die Gleichung

$$m_k \stackrel{!}{=} f_k\left(1 - F_{\chi^2\left(\frac{m_k}{M_k} \cdot (\nu_G - \nu_\Sigma(k-1))\right)}(cv_\alpha - S_{k-1}), \beta_{g,k}, \hat{\theta}_{k-1}\right). \tag{4.15}$$

stellt ein Fixpunktproblem dar. Es existiert eine Lösung, wenn gemäß der Unverfälschtheit von Tests für die bedingte Fehlerrate erster Art $\alpha^*_k \leq 1-\beta_{g,k} < 1-\beta$ gilt, und die Stichprobenfunktion $f_k(\alpha, \beta, \theta)$ in β und α monoton fallend und in α, β und θ stetig ist. Ist die rechte Seite der Gleichung (4.15) mit $m_k = n_{\min,T_k}$ für den kleinsten Stichprobenumfang $n_{\min,T_k}$, der zur Berechnung der Teststatistik T_k benötigt wird, größer als $n_{\min,T_k}$, existiert eine Lösung im Intervall $[n_{\min,T_k}, M_k]$.

Mit M_k aus (3.9) ist ein Stichprobenumfang für den letzten Studienabschnitt bei Vergabe der verbleibenden Freiheitsgrade für eine bedingte Power von $1-\beta$ an der Stelle $\hat{\theta}_{k-1}$ bestimmt. Die Lösung des Fixpunktproblems (4.15) liefert einen Stichprobenumfang, der für die bedingte Power $1-\beta_{g,k} < 1-\beta$ bei proportionaler Vergabe der Freiheitsgrade und der Stichprobenumfänge auf der k-ten Stufe benötigt wird.

Ein Beweis für die Existenz des Fixpunkts erfolgt durch die Überführung von (4.15) in das Nullstellenproblem

$$0 \stackrel{!}{=} f_k\left(1 - F_{\chi^2\left(\frac{m_k}{M_k}\cdot(\nu_G-\nu_\Sigma(k-1))\right)}\left(cv_\alpha - S_{k-1}\right), \beta_{g,k}, \hat{\theta}_{k-1}\right) - m_k \tag{4.16}$$

und über den Zwischenwertsatz (siehe z.B. Kaballo, 2000).

Zur Verwendung des Zwischenwertsatzes muss jeweils ein Wert auf dem Intervall $[n_{\min,T}, M_k]$ der möglichen Stichprobenumfänge gefunden werden, so dass die rechte Seite von (4.16) größer bzw. kleiner Null wird. Bei Stetigkeit und Monotonie existiert eine eindeutige Nullstelle innerhalb der Grenzen des Intervalls.

Da die Wahrscheinlichkeitsfunktion der χ^2-Verteilung monoton fallend in der Anzahl Freiheitsgrade ist, steigt $1 - F_{\chi^2(m_k/M_k\cdot(\nu_G-\nu_\Sigma(k-1)))}(cv_\alpha - S_{k-1})$ monoton in m_k. Es folgt, dass die rechte Seite der Gleichung (4.15) monoton fallend in m_k ist.

Betrachtet man den Fall $m_k \to 0$, konvergieren die Freiheitsgrade gegen Null und der Wert der Verteilungsfunktion der χ^2-Verteilung gegen Eins. Das Niveau, das in der Stichprobenfunktion verwendet wird, nähert sich der Null, so dass der Wert von f_k sowie die rechte Seite der Gleichung (4.16) unendlich groß und damit größer als Null werden. Ist dies für $m_k = n_{\min,T}$ erfüllt, existiert ein ausführbarer Stichprobenumfang für die folgende Stufe zur Erfüllung der vorgegebenen Power.

Gilt $1-\alpha_k^* \geq \beta_{g,k}$, ergibt sich aufgrund der Monotonie von f_k in β durch Wahl von $m_k = M_k$ auf der rechten Seite der Gleichung (4.16)

$$\begin{aligned}
&f_k(1 - F_{\chi^2(\nu_G-\nu_\Sigma(k-1))}(cv_\alpha - S_{k-1}), \beta_{g,k}, \hat{\theta}_{k-1}) - M_k \\
&= f_k(\alpha_k^*, \beta_{g,k}, \hat{\theta}_{k-1}) - f_k(\alpha_k^*, \beta, \hat{\theta}_{k-1}) \\
&< 0.
\end{aligned}$$

Aufgrund der Monotonie und der Stetigkeit liegt zwischen $n_{\min,T}$ und M_k ein Stichprobenumfang m_k, für den Gleichung (4.16) gilt.

Im Fall einer vollständigen Automatisierung des Studiendesigns ist das Auftreten von $\alpha_k^* > 1-\beta > 1-\beta_{g,k}$ denkbar. Die Wahrscheinlichkeit, im nächsten Studienabschnitt bei Vergabe der verbleibenden Freiheitsgrade unter der Nullhypothese die Nullhypothese zu

verwerfen, ist damit größer als die geforderte Power sowohl für den verbleibenden Studienteil als auch für den nächsten Studienabschnitt, in dem nur ein Teil der Freiheitsgrade verwendet wird. In diesem Fall kann der minimale Stichprobenumfang verwendet und die Studie mit der folgenden Stufe beendet werden. Gilt $1-\beta > \alpha_k^* > 1-\beta_{g,k}$ existiert eine Lösung des Fixpunktproblems, wenn durch die Senkung der Freiheitsgrade auf der k-ten Stufe das erzielte Niveau kleiner als die geforderte Power $1-\beta_{g,k}$ wird. Es wird $\nu_{\beta_{g,k}}$ bestimmt, mit der Eigenschaft $F_{\chi^2(\nu_{\beta_{g,k}})}(cv_\alpha - S_{k-1}) = 1-\beta_{g,k}$. Führt die Wahl von $m_k = m_{\beta_{g,k}} = \nu_{\beta_{g,k}}/(\nu_G - \nu_\Sigma(k-1)) \cdot M_k$ dazu, dass die rechte Seite der Gleichung (4.16) negativ wird und bei Verwendung von $n_{\min,T}$ positiv, so liegt der gesuchte Stichprobenumfang auf dem Intervall $[n_{\min,T}, m_{\beta_{g,k}}]$. Die Beschränkung auf das Intervall ist nötig, um zu garantieren, dass die geforderte Power nicht kleiner als der bedingte vergebene Fehler erster Art auf der entsprechenden Stufe wird und die Stichprobenfunktion gültige Werte liefert.

Das Fixpunktproblem kann mit Suchalgorithmen wie der Regula falsi oder Intervallschachtelung gelöst werden. Als Startpunkte für die Regula falsi können M_k bzw. $m_{\beta_{g,k}}$ und $n_{\min,T}$ oder eine pro Stufe vorgegebene minimale Anzahl $n_{\min}$ verwendet werden. Die Lösung des Fixpunktproblems kann bei sehr kleiner Power $1-\beta_{g,k}$ oder sehr großer bedingter Fehlerrate erster Art kleiner als $n_{\min}$ ausfallen. In diesem Fall ist im Sinne des Mindestumfangs kein gültiger Stichprobenumfang als Lösung des Fixpunktproblems zu erhalten und der vorgegebene minimale Umfang $n_{\min}$ zu wählen. Das Gewicht der k-ten Stufe ν_k wird anteilig gemäß m_k/M_K an den verbleibenden Freiheitsgraden $\nu_G - \nu_\Sigma(k-1)$ gewählt. Die Lösung des Fixpunktproblems kann dabei zu einer Anzahl an Freiheitsgraden führen, die kleiner als ein vorgegebener Minimalwert $\nu_{\min}$ ist. Bei Verwendung von $\nu_{\min}$ liegt die erzielte Power über der ursprünglich geplanten Power $1-\beta_{g,k}$.

Beispiel 3 (Forts.)

Für die Fehlerraten erster und zweiter Art gilt $\alpha = 0.025 < 1-\beta_{g,1} = 0.6 < 1-\beta = 0.9$, *so dass eine eindeutige Lösung des Fixpunktproblems (4.15) gefunden werden kann. Abbildung 4.5 veranschaulicht für ausgewählte Werte der lokalen Power* $1-\beta_{g,1}$ *auf der ersten Stufe das zu lösende Fixpunktproblem, indem die rechte Seite der Gleichung (4.15) in Abhängigkeit des möglichen Umfangs* m_1 *dargestellt ist. Auf der Winkelhalbierenden liegen die Lösungen des Fixpunktproblems. Für eine Power von* 60% *auf der ersten Stufe lässt sich mit den Startwerten* 4 *und* 170 *für die Regula falsi ein Stichprobenumfang von* 124 *für die erste Stufe bei einem proportionalen Gewicht von* $\nu_1 = 10 \cdot 124/170 = 7.29$ *Freiheitsgraden bestimmen. Das heißt, dass auf der ersten Stufe der Studie ca.* 73% *der Gesamtfreiheitsgrade und des Stichprobenumfangs einer konventionellen einstufigen Studie aufgewendet werden müssen, um eine Power von* 60% *zu erhalten. Daraus ergibt sich*

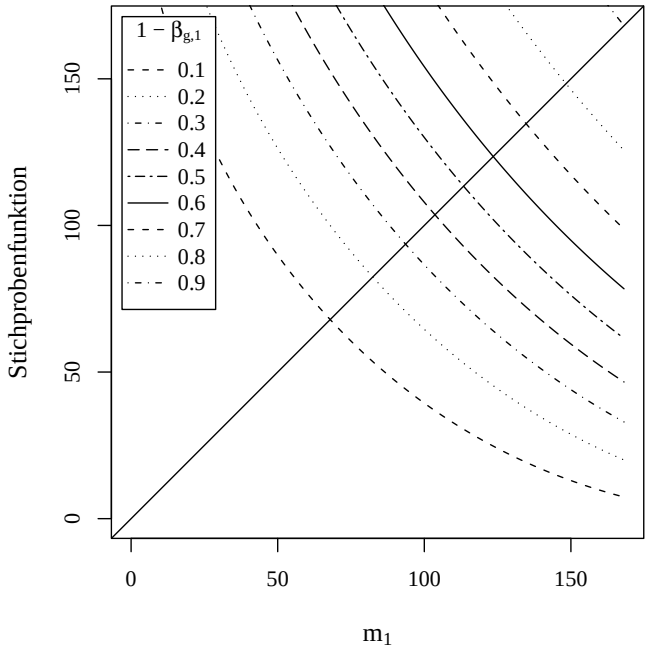

Abbildung 4.5: *Fixpunktproblem auf der ersten Stufe beim einseitigen Zweistichproben-Gauß-Test in Abhängigkeit von der lokalen Power* $1 - \beta_{g,1}$, $\nu_G = 10$, $\alpha = 0.025$, $\beta = 0.1$

auf der ersten Stufe ein Niveau von $\alpha_1 = 1 - F_{\chi^2(\nu_1)(cv_\alpha)} = 0.58\%$.

Aus Abbildung 4.5 folgt, dass etwa 40% *der Freiheitsgrade und des Stichprobenumfangs* M_1 *verwendet werden müssen, um eine Power von* 10% *zu erreichen, während der jeweils zusätzliche Anteil für einen absoluten Zuwachs der lokalen Power auf der ersten Stufe von* 10% *wesentlich geringer und nahezu konstant ausfällt. Dies ist auf den steigenden vergebenen Fehler erster Art auf der ersten Stufe (4.6) bei Erhöhung der Freiheitsgrade* ν_1 *zurückzuführen.*

Tabelle 4.2 enthält für den Zweistichproben-Gauß-Test in Abhängigkeit von der lokalen Power auf der ersten Stufe und der Gesamtanzahl Freiheitsgrade den nach Lösung des Fixpunktproblems sich ergebenden Anteil $\epsilon_1 = m_1/M_1 = \nu_1/\nu_G$ der ersten Stufe am Stichprobenumfang M_1 und an den Freiheitsgraden ν_G sowie das Niveau auf der ersten Stufe. Der Anteil ϵ_1 ist unabhängig vom Wert der Parameter θ und σ, die zur Berechnung von M_k herangezogen werden. M_k und m_k beruhen auf unterschiedlichen Fehlerraten erster und zweiter Art, jedoch auf gleichen Parameterannahmen, so dass das Fixpunktproblem (4.15) für ϵ_k wie folgt formuliert werden kann:

$$\epsilon_k = \left(\frac{u_{1-\alpha^*_{k,\epsilon_k}} + u_{1-\beta_{g,k}}}{u_{1-\alpha^*_k} + u_{1-\beta}} \right)^2 \text{ mit } \alpha^*_{k,\epsilon_k} = 1 - F_{\chi^2(\epsilon_k \cdot (\nu_G - \nu_\Sigma(k-1)))}(cv_\alpha - S_{k-1}). \quad (4.17)$$

Die im Abschnitt 4.2.1 beschriebene Abnahme des lokalen Niveaus auf der ersten Stufe

Tabelle 4.2: *Anteil ϵ_1 der ersten Stufe in Prozent am einstufigen Stichprobenumfang M_1 ($\alpha = 0.025$, $1 - \beta = 0.9$, θ beliebiger fester Wert) und an den Gesamtfreiheitsgraden ν_G sowie Niveau α_1 auf der ersten Stufe bei proportionaler Gewichtung von Stichprobenumfang und Freiheitsgraden in Abhängigkeit von der lokalen Power $1 - \beta_{g,1}$*

	$\nu_G = 10$		$\nu_G = 100$		$\nu_G = 1000$	
$1 - \beta_{g,1}$	$\epsilon_1[\%]$	α_1	$\epsilon_1[\%]$	α_1	$\epsilon_1[\%]$	α_1
0.1	40.39	0.00042	71.23	0.00003	89.37	0.00001
0.2	49.02	0.00093	75.92	0.00012	91.17	0.00004
0.3	55.68	0.00162	79.36	0.00032	92.47	0.00014
0.4	61.65	0.00257	82.35	0.00070	93.59	0.00035
0.5	67.45	0.00388	85.18	0.00139	94.64	0.00081
0.6	73.46	0.00579	88.04	0.00265	95.69	0.00176
0.7	80.13	0.00872	91.14	0.00508	96.82	0.00385
0.8	88.23	0.01379	94.82	0.01031	98.15	0.00890
0.9	100.00	0.025	100.00	0.025	100.00	0.025

bei steigender Anzahl Gesamtfreiheitsgrade und konstantem Anteil ϵ_1 führt dazu, dass bei großem ν_G ein sehr großer Anteil der Freiheitsgrade und des Stichprobenumfangs im einstufigen Design aufgewendet werden muss, um eine relativ geringe lokale Power auf der ersten Stufe zu erreichen. Bei $\nu_G = 10$ Freiheitsgraden werden für eine lokale Power von 10% auf der ersten Stufe 40.39% des Gesamtgewichts und des einstufigen nicht gerundeten Stichprobenumfangs M_1 benötigt. Durch die starke Senkung des lokalen Niveaus muss bei $\nu_G = 1000$ Freiheitsgraden insgesamt ein Anteil von 89.37% aufgewendet werden. Eine große Anzahl Freiheitsgrade führt bei proportionaler Gewichtung von Umfängen und Gewicht im Vergleich zum einstufigen Design, in dem sich für eine Power von 10% bei einem festen θ ein Stichprobenumfang von 4.4% an M_1 ergibt, zu einem starkem Ungleichgewicht zwischen der lokalen Power und dem aufgewendeten Stichprobenumfang. Auf den folgenden Stufen ergibt sich das verbleibende Niveau aus der bedingten Fehlerfunktion (4.2), die neben dem globalen Niveau α und den Gesamtfreiheitsgraden ν_G durch die verbleibenden Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ und die bereits beobachtete Teststatistik S_{k-1} bestimmt ist.

Die bedingten Fehlerfunktionen, die sich als Lösungen des Fixpunktproblems (4.15) ergeben, für $\nu_G = 4, 10, 100, 1000$, $\alpha = 0.025$, $\beta = 0.1$ und $\beta_{g,1} = 0.2$ sind in Abbildung 4.6 dargestellt. Aufgrund der ungleichen Aufteilung der Freiheitsgrade auf die erste und die verbleibenden Stufen sind die Ablehnregionen für 100 und 1000 Freiheitsgrade in den

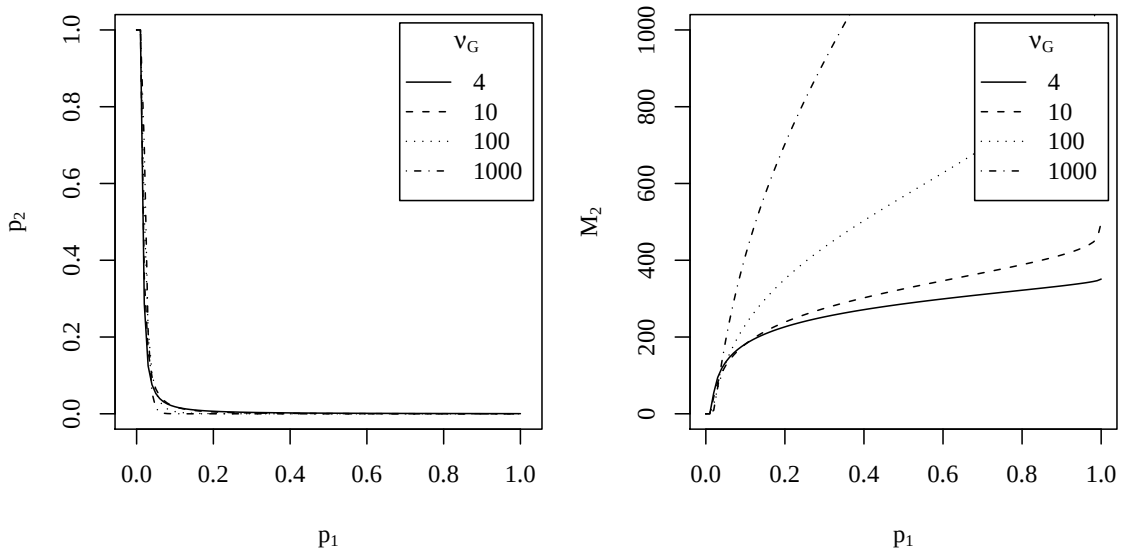

Abbildung 4.6: *Bedingte Fehlerfunktion beim einseitigen Zweistichproben-Gauß-Test für Lösungen des Fixpunktproblems bei lokaler Power* $1 - \beta_{g,1} = 0.8$, $\nu_G = 10$, $\alpha = 0.025$, $\beta = 0.1$, $\theta = 0.5$ *und benötigter Stichprobenumfang für eine bedingte Power von 0.9 unter* θ *bei Vergabe der verbleibenden Freiheitsgrade auf zweiter Stufe*

linken Bereich des positiven Quadranten verschoben. Für 4 und 10 Freiheitsgrade führen bei mittelgroßen und großen P-Werten auf der ersten Stufe höhere P-Werte auf der zweiten Stufe zu einer Ablehnung der Nullhypothese. Nur in einem sehr kleinen Bereich für P-Werte auf der ersten Stufe unter 0.035 liegen die bedingten Fehlerraten bei 100 und 1000 Gesamtfreiheitsgraden über denen bei 4 und 10 Freiheitsgraden. Daraus ergeben sich meist wesentlich höhere benötigte Stichprobenumfänge auf der zweiten Stufe. Abbildung 4.6 zeigt die benötigten Stichprobenumfänge M_2 für eine bedingte Power von 0.9 unter $\theta = 0.5$ bei Vergabe der verbleibenden Freiheitsgrade in Abhängigkeit des P-Werts der ersten Stufe. Durch erneute Lösung eines Fixpunktproblems gemäß (4.15) kann für eine vorgegebene Power $1 - \beta_{g,2} < 1 - \beta$ ein Stichprobenumfang $\epsilon_2 \cdot M_2$ für die zweite Stufe bestimmt werden. Die verbleibenden Freiheitsgrade $\nu_G - \nu_1$ werden proportional auf die zweite und die folgenden Stufen aufgeteilt, um die vorgegebene bedingte Power $1 - \beta_{g,2}$ zu erreichen.

4.4 Abbruch wegen Aussichtslosigkeit

Hartung und Knapp (2003) weisen auf die Möglichkeit hin, in einer Studie gemäß Inverse-χ^2-Methode Grenzen für Stopps wegen Aussichtslosigkeit festzulegen. Sie schlagen untere Schranken $\chi^2(\nu_\Sigma(k))_{1-\alpha_L}$ für die kombinierte Teststatistik S_k auf der k-ten Stufe, $k \geq 1$, mit einem nicht zu konservativ gewählten α_L vor, deren Unterschreitung, $S_k \leq \chi^2(\nu_\Sigma(k))_{1-\alpha_L}$, zu einem Abbruch der Studie mit Annahme der Nullhypothese führt. Dies führt bei Beibehalten des ursprünglichen kritischen Werts dazu, dass die Testprozedur konservativ wird. Um das Gesamtniveau zu erhalten, wird von Hartung und Knapp (2003) eine Korrektur des kritischen Werts vorgeschlagen, sofern das adaptive Design vollständig vom vorgegebenen Algorithmus beschrieben wird (Vollautomat). Der neue kritische Wert wird jeweils in Abhängigkeit von den Regeln zur Wahl der Gewichte und des Stichprobenumfangs sowie den gewählten unteren Schranken durch Simulationen bestimmt.

Hierzu ist zu bemerken, dass die Korrektur des kritischen Werts, bei der Möglichkeit zum Stopp wegen Aussichtslosigkeit auf mehreren Stufen, nur dann global erfolgen kann, sofern die Prozedur als Vollautomat, d.h. ohne Veränderung des festgelegten Algorithmus bei den Interimsanalysen, angewandt wird. Dies kann bei der Einzelbetrachtung einer Studie, die unter Anwendung des Vollautomaten entstanden ist, dazu führen, dass das Signifikanzniveau unter- bzw. überschritten wird. Der angepasste kritische Wert gilt für eine nicht näher definierte mittlere Anzahl Stufen mit entsprechender Vergabe der Freiheitsgrade, jeweils mit der Möglichkeit zum vorzeitigen Abbruch zu Gunsten der Nullhypothese. Wird diese mittlere Anzahl durch eine rasche Vergabe der verbleibenden Freiheitsgrade nicht erreicht, kommt es zu einer Niveauüberschreitung. Bei der Durchführung von sehr vielen Stufen hingegen wird die Wahrscheinlichkeit, zu Gunsten der Nullhypothese abzubrechen, nicht mehr durch die Anpassung des globalen kritischen Werts aufgefangen, und das Niveau wird nicht ausgeschöpft.

Soll für jede geplante Studie und nicht für den Erwartungswert der durch einen Vollautomaten geplanten Studien das globale Niveau eingehalten werden, muss die Korrektur des kritischen Werts für jede Stufe beim Erreichen derselben erfolgen. Erst die Sicherheit, dass eine Stufe mit der Möglichkeit zum Abbruch zu Gunsten der Nullhypothese durchgeführt wird, erlaubt die Korrektur des globalen kritischen Werts ab der entsprechenden Stufe. Eine globale Korrektur ist lediglich für die erste Stufe bzw. bei der Vorgabe fester Freiheitsgrade denkbar.

Insgesamt ist die Möglichkeit des Abbruchs wegen Aussichtslosigkeit lediglich in den frühen Stufen eines Versuchs interessant und dürfte einem Studiensponsor nach der Durch-

führung von mehreren Stufen schwer vermittelbar sein. Die Korrektur des kritischen Werts in einem adaptiven oder auch gruppensequentiellen Design bezieht sich ausschließlich auf die Ergebnisse des Zielparameters. Ein Studienabbruch kann zum Beispiel bei vermutlicher Unterlegenheit des neuen Präparats in das Verfahren integriert werden, während unerwartete Ereignisse, die aus ethischen Gründen zum Abbruch der Studie führen, nicht erfassbar und damit nicht einzubeziehen sind.

Im Folgenden wird die Umsetzung des Stopps wegen Aussichtslosigkeit auf der ersten Stufe und die daraus folgende Anpassung des globalen kritischen Werts bei Verwendung der Inverse-χ^2-Methode dargestellt. Gemäß Abschnitt 4.1 kann der nach der $(k-1)$-ten Stufe verbleibende Teil der Studie als neue Studie gemäß Inverse-χ^2-Methode mit $\nu_G - \nu_\Sigma(k-1)$ Freiheitsgraden und Niveau gemäß bedingter Fehlerfunktion (4.2) angesehen werden. Auf diese Weise lässt sich für die folgenden Stufen analog ein adjustierter kritischer Wert berechnen.

Vor Beginn der Studie werden die Freiheitsgrade für die erste Stufe ν_1 und für den verbleibenden Teil $\nu_2^* = \nu_G - \nu_1$ aufgeteilt. Ein Abbruch wegen Aussichtslosigkeit soll erfolgen, wenn der P-Wert p_1 auf der ersten Stufe größer oder gleich einem vorgegebenen Wert α_0 bzw. $q_1(\nu_1) = F^{-1}_{\chi^2(\nu_1)}(1-p_1) \leq F^{-1}_{\chi^2(\nu_1)}(1-\alpha_0) = cv_0$ ist.

Der globale kritische Wert $cv_{\alpha'}$ bei einem Stopp wegen Aussichtslosigkeit muss folgende Gleichung erfüllen:

$$\begin{aligned} \alpha &= \alpha_1 + \alpha_2 \qquad (4.18) \\ &= \mathrm{P}_{\mathrm{H}_0}(q_1(\nu_1) \geq cv_{\alpha'}) + \mathrm{P}_{\mathrm{H}_0}(S_2 = q_1(\nu_1) + q_2(\nu_2^*) \geq cv_{\alpha'}, cv_0 < q_1(\nu_1) < cv_{\alpha'}). \end{aligned}$$

Die Irrtumswahrscheinlichkeit erster Art auf der ersten Stufe ergibt sich zu

$$\alpha_1 = 1 - F_{\chi^2(\nu_1)}(cv_{\alpha'}). \qquad (4.19)$$

Zur Berechnung von α_2 wird die gemeinsame Dichte von $q_1(\nu_1)$ und $q_2(\nu_2^*)$ unter H_0 benötigt, die sich aufgrund der vorgegebenen Freiheitsgrade ν_1 und ν_2^* als Produkt aus den Dichten der entsprechenden χ^2-Verteilungen ergibt. Die gemeinsame Dichte von S_2 und $q_1(\nu_1)$ ergibt sich durch Ersetzen von $q_2(\nu_2^*)$ durch $S_2 - q_1(\nu_1)$. Für die zweite Stufe gilt unter der Nullhypothese bei verkürzter Darstellung von $q_1(\nu_1)$ als q_1

$$\alpha_2 = \int_{cv_{\alpha'}}^{\infty} \int_{cv_0}^{cv_{\alpha'}} \frac{\left(\frac{1}{2}\right)^{(\nu_1+\nu_2^*)/2}}{\Gamma(\frac{\nu_1}{2}) \cdot \Gamma(\frac{\nu_2^*}{2})} \cdot e^{-\frac{1}{2}S_2} \cdot q_1^{\nu_1/2-1} \cdot (S_2 - q_1)^{\nu_2^*/2-1} \partial q_1 \partial S_2,$$

wobei $\Gamma(\cdot)$ die Gamma-Funktion bezeichnet.

Mit Hilfe der Transformation $q_1 = z \cdot S_2$ ergibt sich

$$\begin{aligned}
\alpha_2 &= \int_{cv_{\alpha'}}^{\infty} \int_{cv_0/S_2}^{cv_{\alpha'}/S_2} \frac{\left(\frac{1}{2}\right)^{(\nu_1+\nu_2^*)/2}}{\Gamma(\frac{\nu_1}{2}) \cdot \Gamma(\frac{\nu_2^*}{2})} \cdot e^{-\frac{1}{2}S_2} \cdot (z \cdot S_2)^{\nu_1/2-1} \\
&\qquad \cdot (S_2 - z \cdot S_2)^{\nu_2^*/2-1} \cdot S_2 \partial z \partial S_2 \qquad (4.20) \\
&= \int_{cv_{\alpha'}}^{\infty} \frac{\left(\frac{1}{2}\right)^{(\nu_1+\nu_2^*)/2}}{\Gamma(\frac{\nu_1}{2}) \cdot \Gamma(\frac{\nu_2^*}{2})} \cdot e^{-\frac{1}{2}S_2} S_2^{(\nu_1+\nu_2^*)/2-1} \int_{cv_0/S_2}^{cv_{\alpha'}/S_2} z^{\nu_1/2-1}(1-z)^{\nu_2^*/2-1} \partial z \partial S_2 \\
&= \int_{cv_{\alpha'}}^{\infty} f_{\chi^2(\nu_1+\nu_2^*)}(S_2) \cdot \left[F_{B(\nu_1/2,\nu_2^*/2)}\left(\frac{cv_{\alpha'}}{S_2}\right) - F_{B(\nu_1/2,\nu_2^*/2)}\left(\frac{cv_0}{S_2}\right) \right] \partial S_2
\end{aligned}$$

dabei steht $F_{B(\nu_1/2,\nu_2^*/2)}(x)$ für die Verteilungsfunktion der Beta-Verteilung mit den Parametern $\nu_1/2$ und $\nu_2^*/2$.

Der globale kritische Wert $cv_{\alpha'}$, der im Intervall $[cv_0, cv_\alpha]$ liegt, kann mittels iterativer Verfahren als Lösung der Gleichung (4.18) unter Verwendung von (4.19) und (4.20) bestimmt werden.

Tabelle 4.3 beinhaltet kritische Werte $cv_{\alpha'}$ für eine Auswahl an Annahmewahrscheinlichkeiten α_0 unter der Nullhypothese auf der ersten Stufe und verschiedene Anzahlen von Gesamtfreiheitsgraden ν_G. Gilt $\alpha_0 = 1$, erhält man das Design ohne Stopp wegen Aussichtslosigkeit mit dem globalen kritischen Wert $cv_\alpha = \chi^2(\nu_G)_{1-\alpha}$. Die Höhe der Senkung des kritischen Werts ist abhängig von der Anzahl Freiheitsgrade auf der ersten Stufe und insgesamt, da sie die Wahrscheinlichkeit des Fehlers erster Art auf der ersten Stufe bestimmen (vgl. Kapitel 4.2).

Ein analoges Vorgehen ist zu Beginn jeder Stufe $k, k > 1$, denkbar. Bedingt auf die Ergebnisse der vorliegenden $(k-1)$ Stufen ergeben sich bei der Aufteilung der verbleibenden Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ auf die Gewichte ν_k und $\nu_{k+1}^* = \nu_G - \nu_\Sigma(k)$ die unter der Nullhypothese χ^2-verteilten Zufallsvariablen $q_k(\nu_k)$ auf der folgenden Stufe und $q_{k+1}(\nu_{k+1}^*)$ für den übrigen Teil der Studie. Dabei ist zu berücksichtigen, dass nach der Durchführung von $(k-1)$ Stufen für die Teststatistik S_k eine untere Schranke $cv_{0,k} > S_{k-1}$ gewählt werden muss. Dies garantiert, dass eine weitere Stufe durchgeführt werden muss, um die Nullhypothese abzulehnen, und führt gleichzeitig zu einer positiven Wahrscheinlichkeit für den Stopp wegen Aussichtslosigkeit. Der neue kritische Wert kann direkt in eine untere Schranke für $q_k(\nu_k)$ umgerechnet werden.

Der ab der k-ten Stufe geltende korrigierte kritische Wert $cv'_{\alpha'}$ ist so zu wählen, dass die verbleibende bedingte Irrtumswahrscheinlichkeit α_k^*, die vom bis zur Stufe $(k-1)$ gültigen

Tabelle 4.3: *Globale kritische Werte $cv_{\alpha'}$ unter Berücksichtigung eines Stopps wegen Aussichtslosigkeit auf der ersten Stufe, in Abhängigkeit der oberen Schranke α_0 für den P-Wert der ersten Stufe und der Anzahl Freiheitsgrade ν_G bei einem Gewicht von $\nu_1 = 0.2\nu_G$ auf der ersten Stufe*

	Anzahl Freiheitsgrade ν_G				
α_0	5	10	50	100	1000
			$\alpha = 0.025$		
1.0	12.8325	20.4831	71.4201	129.5612	1089.5309
0.9	12.7071	20.3704	71.3294	129.4700	1089.3865
0.8	12.5718	20.2415	71.1893	129.3104	1089.0537
0.7	12.4226	20.0915	70.9981	129.0796	1088.5090
0.6	12.2540	19.9128	70.7418	128.7577	1087.6869
0.5	12.0570	19.6933	70.3943	128.3080	1086.4695
0.4	11.8167	19.4112	69.9071	127.6617	1084.6370
0.3	11.5027	19.0227	69.1787	126.6740	1081.7250
			$\alpha = 0.05$		
1.0	11.0704	18.3070	67.5048	124.3421	1074.6794
0.9	10.9322	18.1758	67.3839	124.2142	1074.4519
0.8	10.7819	18.0249	67.2003	123.9965	1073.9574
0.7	10.6150	17.8484	66.9513	123.6856	1073.1725
0.6	10.4246	17.6364	66.6180	123.2551	1072.0101
0.5	10.1998	17.3734	66.1653	122.6548	1070.3084
0.4	9.9212	17.0308	65.5263	121.7889	1067.7568
0.3	9.5493	16.5487	64.5565	120.4493	1063.6777

globalen kritischen Wert $cv_{\alpha'}$ abhängt, bewahrt wird:

$$\underbrace{\mathrm{P}_{\mathrm{H}_0}(S_{k-1} + q_k(\nu_k) \geq cv'_{\alpha'})}_{\alpha_k} +$$

$$\underbrace{\mathrm{P}_{\mathrm{H}_0}(S_{k+1} = S_{k-1} + q_k(\nu_k) + q_{k+1}(\nu^*_{k+1}) \geq cv'_{\alpha'}, cv_{0,k} < S_{k-1} + q_k(\nu_k) < cv'_{\alpha'})}_{\alpha_{k+1}}$$

$$= \quad F^{-1}_{\chi^2(\nu_G - \nu_\Sigma(k-1))}(cv_{\alpha'} - S_{k-1}) = \alpha^*_k. \tag{4.21}$$

Da die Adjustierung der kritischen Werte ausschließlich auf den unter der Nullhypothese vorliegenden χ^2-Verteilungen basiert, erfolgt sie unabhängig von der verwendeten Teststatistik.

Neben der Korrektur des globalen kritischen Werts lässt sich wie bei Bauer und Köhne (1994) auch eine kritische Schranke cv_1 für die Teststatistik der ersten Stufe in Abhängigkeit von der gewählten oberen Schranke α_0 für den P-Wert bzw. unteren Schranke cv_0 für die Teststatistik auf der ersten Stufe bestimmen. Im Vergleich zum Test ohne Abbruch wegen Aussichtslosigkeit verändert sich der globale kritische Wert $cv_\alpha = \chi^2(\nu_G)_{1-\alpha}$ für die folgenden Stufen nicht.

Das globale Niveau setzt sich ähnlich zu (4.18) zusammen aus

$$\underbrace{\mathrm{P}_{\mathrm{H}_0}(q_1(\nu_1) \geq cv_1)}_{\alpha_1} + \underbrace{\mathrm{P}_{\mathrm{H}_0}(S_2 = q_1(\nu_1) + q_2(\nu_2^*) \geq cv_\alpha, cv_0 < q_1(\nu_1) < cv_1)}_{\alpha_2} = \alpha, \quad (4.22)$$

so dass aus der folgenden Gleichung bei Vorgabe von α_0 und α der kritische Wert cv_1 auf der ersten Stufe iterativ bestimmt werden kann:

$$\begin{aligned} \alpha =& 1 - F_{\chi^2(\nu_1)}(cv_1) + \\ & \int_{cv_\alpha}^{\infty} f_{\chi^2(\nu_1+\nu_2^*)}(S_2) \cdot \left[F_{B(\nu_1/2,\nu_2^*/2)}\left(\frac{cv_1}{S_2}\right) - F_{B(\nu_1/2,\nu_2^*/2)}\left(\frac{cv_0}{S_2}\right) \right] \partial S_2 \end{aligned} \quad (4.23)$$

Allgemein ist zu bemerken, dass der Wert α_0 für den Abbruch zu Gunsten der Nullhypothese nicht zu klein gewählt werden sollte. Andernfalls werden aufgrund von geringer Power auf der ersten Stufe viele Versuche mit einem falsch negativen Ergebnis beendet.

Wie in allen sequentiellen Testverfahren muss zum Erhalt des globalen Niveaus bei einer Überschreitung von α_0 durch den P-Wert p_1 die Studie zu Gunsten der Nullhypothese tatsächlich abgebrochen werden (siehe dazu auch: Bauer und Köhne, 1994). Dies ist ein kritischer Aspekt in der Studiendurchführung, bei dessen Missachtung das Niveau überschritten wird. Aus diesem Grund wird von Zulassungsbehörden wie der FDA (U. S. Food and Drug Administration) über kritische Schranken diskutiert (Ellenberg, 2006), die das Niveau auch bei Missachtung des geforderten Abbruchs wegen Aussichtslosigkeit einhalten. Bei Einhaltung der unteren Schranken wird das Verfahren konservativ.

4.5 Aktualisierung der Parameterschätzer

In adaptiven Designs ist die Schätzung des Effekts bei einer Interimsanalyse wesentlich für die Planung des weiteren Studienverlaufs. Eine deutliche Abweichung des Schätzers vom erwarteten bzw. erhofften a priori Parameterwert θ_0 kann zu der Entscheidung führen, den Stichprobenumfang anzupassen, sofern Stichprobenumfänge für die einzelnen Stufen zu Beginn der Studie geplant wurden, z.B. gemäß einem gruppensequentiellen Test (vgl. Abschnitt 2.1). In adaptiven Designs gemäß dem Varianzvergabeprinzip nach Shen

und Fisher (1999), Hartung (2001) oder Hartung und Knapp (2003) bestimmt die Schätzung Gewicht und Stichprobenumfang der folgenden Stufe.

Wie in Kapitel 2.3 beschrieben ist eine unverzerrte Schätzung des Parameters θ bei einer gruppensequentiellen oder adaptiven Studie ab der zweiten Stufe weder am Ende noch bei Interimsanalysen möglich. Eine erwartungstreue Schätzung kann nur mit einem Schätzer $\hat{\theta}(k)$, $k \geq 1$ aus der zuletzt durchgeführten Stufe erreicht werden, sofern $\hat{\theta}(k)$ im einstufigen Design ein unverzerrter Schätzer ist. Korrekturen des Schätzers, wie z.B. von Coburger und Wassmer (2001) vorgeschlagen, sind abhängig vom adaptiven Design, d.h. von der bedingten Fehlerfunktion, der Gewichtung der einzelnen Stufen und den gewählten Stichprobenumfängen. Die Bestimmung korrigierter Schätzer ist zum Teil mit erheblichem Rechenaufwand verbunden und führt zu Schätzern mit großer Varianz im Vergleich zum Mittelwert der gepoolten Daten.

Einfache Schätzer ergeben sich durch die Kombination der Schätzer aus den einzelnen Stufen durch Gewichtung. Ausgehend von der Realisierung von k Stufen eines adaptiven Designs liegt für jede Stufe eine Schätzung $\hat{\theta}(i)$, $i = 1, \ldots, k$, des Parameters θ aus einer Stichprobe vom Umfang n_i, $i = 1, \ldots, k$, vor. Dies kann zum Beispiel die Mittelwertdifferenz zweier Behandlungsarme sein. Ein Kombinationsschätzer aus der Meta-Analyse gewichtet die Schätzung einer Studie mit der Inversen der Studienvarianz. Übertragen auf die Stufen einer adaptiven Studie, in der vorausgesetzt wird, dass die Beobachtungen auf allen Stufen aus der gleichen Verteilung stammen, werden die Schätzer mit dem Stichprobenumfang n_i gewichtet, siehe (3.16). Der kombinierte Schätzer nach k Stufen entspricht dem Schätzer aus den gepoolten Daten bis Stufe k, wenn $\hat{\theta}(i)$, $i = 1, \ldots, k$, Mittelwerte bzw. Mittelwertsdifferenzen sind.

Ein weiterer Kombinationsschätzer ergibt sich als Mittelpunkt der wiederholten Konfidenzintervalle nach Jennison und Turnbull (1989) und gewichtet die Stufenschätzer mit der Wurzel aus dem Stichprobenumfang:

$$\hat{\theta}_k^{\sqrt{n}} = \frac{\sum_{i=1}^{k} \sqrt{n_i} \cdot \hat{\theta}(i)}{\sum_{i=1}^{k} \sqrt{n_i}}. \tag{4.24}$$

Zur Anpassung des Stichprobenumfangs sind neben dem Mittelpunkt eines Konfidenzintervalls auch andere Werte, beschränkt auf Werte aus der Alternative, denkbar. Hartung und Knapp (2006) konstruieren für die verallgemeinerte Inverse-χ^2-Methode geschachtelte wiederholte Konfidenzintervalle für normalverteilte Responsevariablen mit unbekannter Varianz und leiten daraus einen Schätzer ab. Die Teststatistik T_k, $k \geq 1$, wird überführt in eine Teststatistik $T_k(\theta)$, deren Verteilung für das wahre θ der Verteilung von T_k unter der Nullhypothese entspricht. Beim Zweistichproben-Gauß-Test, siehe (2.2), steht θ für die Differenz der Erwartungswerte. Die Statistik $T_k(\theta)$ auf der k-ten Stufe mit

$\hat{\theta}(k) = \bar{X}_{k1} - \bar{X}_{k2}$ ist

$$T_k(\theta) = \frac{\hat{\theta}(k) - \theta}{\sqrt{4\sigma^2/n_k}}, \quad k \geq 1. \tag{4.25}$$

Für den Zweistichproben-t-Test wird in (4.25) die Varianz σ^2 durch die gepoolte Varianz $s^2_{p,k} = \left(\sum_{i=1}^{2}\sum_{j=1}^{n_k/2}(X_{kij} - \bar{X}_{ki})^2\right)/(n_k - 2)$ ersetzt. Die Teststatistik $T_k(\theta)$ ist für das wahre θ t-verteilt mit $(n_k - 2)$ Freiheitsgraden. Für ungleiche Gruppengrößen existieren entsprechende Formulierungen (siehe Brannath et al., 2002; Hartung und Knapp, 2006).

Ausgehend von der Verteilungseigenschaft (3.8) der kombinierten Teststatistik S_k bei Verbrauch aller Freiheitsgrade gilt für den wahren Parameter θ

$$S_k(\theta) = \sum_{i=1}^{k} F^{-1}_{\chi^2(\nu_i)}(F_{i,0}(T_i(\theta))) \sim \chi^2(\nu_G) \text{ mit } \nu_\Sigma(k) = \nu_G. \tag{4.26}$$

Ein einseitiges Konfidenzintervall mit einem Konfidenzkoeffizient von mindestens $(1 - \alpha)$ ist auf jeder Stufe k mit $\nu_\Sigma(k) \leq \nu_G$ definiert durch

$$\text{KI}_{k,L}: \ \{\theta | S_k(\theta) \leq \chi^2(\nu_G)_{1-\alpha}\}. \tag{4.27}$$

Über das Testproblem $\text{H}_0: \ \theta = 0$ vs. $\text{H}_1: \ \theta < 0$ lässt sich mit der Folge

$$R_k(\theta) = \sum_{i=1}^{k} F^{-1}_{\chi^2(\nu_i)}(1 - F_{i,0}(T_i(\theta))) \tag{4.28}$$

analog zu (4.27) ein nach oben begrenztes einseitiges Konfidenzintervall definieren

$$\text{KI}_{k,U}: \ \{\theta | R_k(\theta) \leq \chi^2(\nu_G)_{1-\alpha}\}. \tag{4.29}$$

Durch die Monotonie und Stetigkeit der Folgen $S_k(\theta)$ und $R_k(\theta)$ in θ ergeben sich einseitige geschachtelte Konfidenzintervalle im Verlauf der Studie. Der Schnitt der einseitigen Konfidenzintervalle auf jeder Stufe führt zu geschachtelten zweiseitigen Konfidenzintervallen $\text{KI}_k = [\theta_{k,L}, \theta_{k,U}]$ mit einem Konfidenzkoeffizienten von mindestens $1 - 2\alpha$. Bei starker Heterogenität der Teststatistiken auf den einzelnen Stufen kann der Schnitt der einseitigen Konfidenzintervalle leer sein (siehe Hartung und Knapp, 2006). Anhand eines realen Beispiels einer klinischen Studie, die bei Lehmacher und Wassmer (1999) zitiert wird und auf die Hartung und Knapp (2003) die verallgemeinerte Inverse-χ^2-Methode exemplarisch anwenden, wird die Konstruktion der Konfidenzintervalle verdeutlicht.

Beispiel 4

In einer randomisierten, Placebo-kontrollierten, doppelblinden Studie mit Patienten mit Acne papulopustulosa (Plewigs Stufe II-III) wurde eine Kombination von 1% Chloramphenicol und 0.5% farblosem, schwefelhaltigen Schieferöl mit dem alkoholischen Trägerstoff verglichen. Nach sechs Wochen Behandlung wurde die Reduktion der Bakterien im

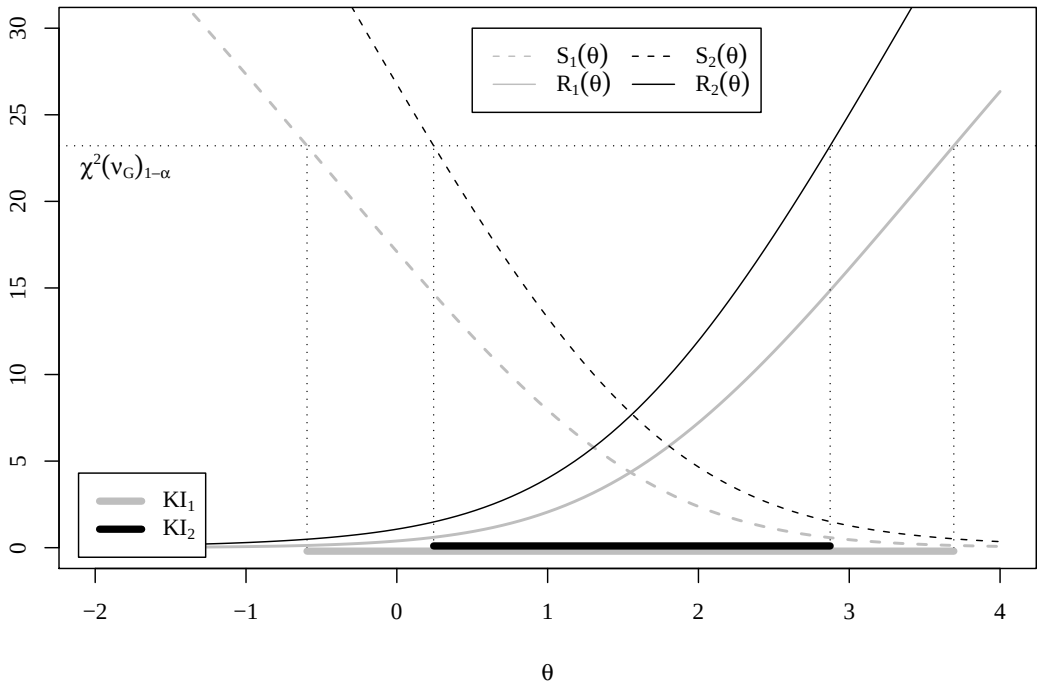

Abbildung 4.7: *Konstruktion der wiederholten Konfidenzintervalle gemäß verallgemeinerter Inverse-χ^2-Methode*

Vergleich zur Baseline, untersucht auf Agar-Platten, bei 24 Patienten in der Behandlungsgruppe und 26 Patienten in der Placebogruppe erhoben. In einem zweiseitigen t-Test wurde mit einem P-Wert von 0.0008 ein signifikanter Unterschied zwischen den Behandlungen festgestellt. Hartung und Knapp (2003) konstruieren anhand der Daten eine Studie gemäß verallgemeinerter Inverse-χ^2-Methode für einen einseitigen t-Test auf den einzelnen Stufen. Das gewählte globale Niveau $\alpha = 0.01$ führt bei insgesamt $\nu_G = 10$ Freiheitsgraden zu einem kritischen Wert von $cv_\alpha = 23.2093$. In der ersten Stufe mit $\nu_1 = 5$ Freiheitsgraden bei 12 Patienten pro Gruppe ($n_1 = 24$) wird eine Differenz der Mittelwerte von $\hat{\theta}(1) = 1.549$ und $s_{p,1} = 1.316$ beobachtet. Eine zweite Stufe mit $\nu_2 = 4$ Freiheitsgraden und einem Stichprobenumfang $n_2 = 12$ führt zu $\hat{\theta}(2) = 1.580$ und $s_{p,2} = 1.472$. Die P-Werte auf den zwei Stufen ergeben sich zu $p_1 = 0.0043$ und $p_2 = 0.0463$. Mit $S_2 = q_1(\nu_1) + q_2(\nu_2) = 17.0989 + 9.6724 = 26.7713 > 23.2093$ kann nach der zweiten Stufe die Studie mit Ablehnung der Nullhypothese beendet werden.

Die Funktionen $S_1(\theta)$, $R_1(\theta)$, $S_2(\theta)$ und $R_2(\theta)$ werden mit der veränderten Teststatistik $T_k(\theta)$ des t-Tests für beide Stufen gebildet. Abbildung 4.7 zeigt die Graphen der Funktionen und den Schnitt mit der Konstanten $\chi^2(\nu_G)_{0.99} = 23.2093$. Der Konfidenzkoeffizient für die Konfidenzintervalle beträgt mindestens $1 - 2 \cdot 0.01 = 0.98$. Auf der ersten Stufe ergibt sich ein breites, die Null einschließendes Konfidenzintervall $KI_1 = [-0.5955; 3.6935]$. Die Funktionen $S_2(\theta)$ und $R_2(\theta)$ definieren ein engeres Konfidenzinter-

Tabelle 4.4: *Punktschätzer auf der zweiten Stufe für Beispiel 4*

$\hat{\theta}_k$	$\hat{\theta}_k^{\sqrt{n}}$	$\hat{\theta}_k^{\text{RCI}}$	$\hat{\theta}_k^{\text{int}}$
1.5593	1.5618	1.5599	1.5588

vall $KI_2 = [0.2443; 2.8734]$*, das entsprechend dem Verwerfen der Nullhypothese die Null nicht umfasst. Durch Verschieben der Konstanten und der Bestimmung der Schnittpunkte der Funktionen* $S_k(\theta)$ *und* $R_k(\theta)$ *mit der Konstanten ergeben sich Konfidenzintervalle mit anderen Konfidenzkoeffizienten.*

Hartung und Knapp (2006) schlagen als Punktschätzer für den Parameter θ zum einen den Mittelpunkt des zweiseitigen Konfidenzintervalls

$$\hat{\theta}_k^{\text{RCI}} = 0.5 \cdot (\theta_{k,U} + \theta_{k,L}) \tag{4.30}$$

vor. Ist der Schnitt der einseitigen Konfidenzintervalle leer, kann durch den Mittelwert aus den Schnittpunkten der Graphen von $S_k(\theta)$ und $R_k(\theta)$ mit der Konstanten $\chi^2(\nu_G)_{1-\alpha}$ dennoch ein Schätzer konstruiert werden. Er ist jedoch in keinem der einseitigen Konfidenzintervalle (4.27) und (4.29) enthalten. Zum anderen bietet der Schnittpunkt der durch die Folgen $S_k(\theta)$ und $R_k(\theta)$ definierten Graphen einen Schätzer. Er ist die Lösung der Gleichung

$$S_k(\hat{\theta}_k^{\text{int}}) = R_k(\hat{\theta}_k^{\text{int}}). \tag{4.31}$$

Für das Beispiel 4 ergeben sich die in Tabelle 4.4 dargestellten Punktschätzer auf der zweiten Stufe. Auf der ersten Stufe entsprechen alle Schätzer dem Wert 1.549. Auch auf der zweiten Stufe unterscheiden sich die Schätzungen wenig. Dies basiert auf den relativ homogenen Schätzern auf den einzelnen Stufen. Im Beispiel fällt die Schätzung, die die einzelnen Schätzer mit der Wurzel des Stufenumfangs gewichtet, am größten aus. Der kleinste Schätzer ergibt sich bei Verwendung des Schnitts der Funktionen $S_2(\theta)$ und $R_2(\theta)$.

Die Eigenschaften dieser Schätzer wie Varianz und Verzerrung sowie die Eignung zur Verwendung in der Stichprobenplanung nach einer Zwischenauswertung werden im folgenden Kapitel näher betrachtet.

Kapitel 5

Strategien der Powerkontrolle

In Kapitel 4 wurde der Einfluss einzelner Designparameter oder spezieller Aufteilungen eines festen Stichprobenumfangs in Designs gemäß verallgemeinerter Inverse-χ^2-Methode betrachtet. Die Beschreibung der Eigenschaften erfolgte dabei vorwiegend für zweistufige Studien, in denen die Aufteilung der Freiheitsgrade zu Beginn festgelegt wird. Weiter wurde der Einfluss einzelner Parameter zum Teil unabhängig von der Variation anderer Parameter betrachtet. In diesem Kapitel wird das Zusammenwirken der Designparameter untersucht, indem Strategien bzgl. der Wahl der Designparameter entwickelt werden, wobei zum Teil eine offene Stufenzahl zugelassen wird. Die verschiedenen Strategien werden hinsichtlich der erzielten Power, dem mittleren benötigten Stichprobenumfang und der mittleren Anzahl durchgeführter Stufen für verschiedene wahre Werte des Parameters θ verglichen.

5.1 Theoretische Überlegungen

Adaptive Designs halten nach Konstruktion ein vorgegebenes globales Niveau α ein, während in den meisten Fällen die Kontrolle der globalen Power sowie des erwarteten benötigten Stichprobenumfangs unberücksichtigt bleibt. Meist wird eine relativ hohe bedingte Power auf den einzelnen Stufen gefordert (siehe Abschnitt 2.4.2), die insgesamt zu einer hohen globalen Power des Verfahrens bei großen Stichprobenumfängen führen kann.

Bei der verallgemeinerten Inverse-χ^2-Methode wird der mittlere benötigte Stichprobenumfang durch den tatsächlichen Stichprobenumfang auf der k-ten Stufe und den Stichprobenumfang M_k bestimmt, der zur Gewichtung herangezogen wird und die Anzahl der Stufen steuert (siehe Abschnitt 3.3). M_k ist der Stichprobenumfang, der benötigt wird, um die Power $1-\beta$ bei Vergabe aller verbleibenden Freiheitsgrade in der nächsten Stufe

zu erreichen. Hartung und Knapp (2003) geben einen konstanten bedingten Fehler zweiter Art $\beta_{g,k} = \beta_g$ bei der adaptiven Bestimmung der Stichprobenumfänge m_k, $k \geq 1$, der einzelnen Stufen vor. Um den erwarteten Stichprobenumfang zu begrenzen und für festgelegte Werte aus der Alternative eine vorgegebene Power $1 - \beta$ zu erreichen, werden außerdem obere Grenzen für die Umfänge m_k und M_k eingeführt. Dies führt zu Häufungen in der Verteilung des benötigten Stichprobenumfangs bei Vielfachen der maximalen Umfänge (siehe Abschnitt 3.4). Durch die offene Anzahl von Stufen, sofern kein minimales Gewicht vorgegeben wird, führt die Festlegung des maximalen Stichprobenumfangs pro Stufe nicht zur Festlegung eines maximalen globalen Stichprobenumfangs.

Betrachtet man den Powerzuwachs gruppensequentieller Designs wie nach Pocock (1977) und O'Brien und Fleming (1979), siehe Tabelle 2.2, wird deutlich, dass die Powervergabe nicht gleichmäßig über die einzelnen Stufen erfolgen muss. Je nach Intention des Studienplaners kann eine langsame oder schnelle Vergabe der Power gewählt werden. Für die Erweiterung möglicher Strategien in adaptiven Designs ist daher die Aufhebung der konstanten bedingten Power in Betracht zu ziehen und eine bedingte Power $1 - \beta_{g,k}$ in Abhängigkeit der Stufe k, $k \geq 1$, vorzugeben. Die bedingte Power kann ab der zweiten Stufe eine Funktion des Parameterschätzers $\hat{\theta}_{k-1}$ sein, so dass beispielsweise bei größeren Parameterwerten eine größere Power vergeben wird als bei kleinen.

In den Strategien nach Hartung und Knapp (2003) wird für den verbleibenden Studienteil jeweils die globale Power $1 - \beta$ zur Berechnung von M_k verwendet. Wird die Studie nicht vorzeitig abgebrochen, wird die letzte Stufe bei Vergabe der verbleibenden Freiheitsgrade mit einer bedingten Power von $1 - \beta$ durchgeführt. Dies führt bei unbeschränkten Stichprobenumfängen zu sehr großen mittleren Stichprobenumfängen und zu einer globalen Power des gesamten Verfahrens, die über $1 - \beta$ liegt. Die Anpassung der Power für den verbleibenden Studienteil kann neben der Aufhebung der konstanten bedingten Power zur Berechnung von m_k den erwarteten Stichprobenumfang regulieren. Anstelle von $1-\beta$ kann eine bedingte Power $\pi_{G,k} = 1 - \beta_{G,k}$ für eine letzte Stufe bei Vergabe der verbleibenden Freiheitsgrade zur Berechnung von M_k verwendet werden. Sie kann in Abhängigkeit des beobachteten Effekts gewählt werden.

Zur Entwicklung von Strategien wird die Zusammensetzung des globalen Fehlers zweiter Art eines gruppensequentiellen Verfahrens betrachtet. Bei der verallgemeinerten Inverse-χ^2-Methode ohne Abbruch wegen Aussichtslosigkeit kann ein Fehler zweiter Art nur in einer finalen Stufe K auftreten, in der die verbleibenden Freiheitsgrade vollständig vergeben werden. In den vorangehenden Stufen entspricht die Fortsetzungswahrscheinlichkeit dem Fehler zweiter Art. Für einen vorgegebenen Wert θ_{H_1} aus der Alternative ergibt sich der globale Fehler zweiter Art aus dem Produkt der bedingten Wahrscheinlichkeiten β'_k,

$k \leq K$, dass die Studie nach Erreichen der k-ten Stufe mit einer weiteren Stufe $k+1$ fortgesetzt, d.h. die Nullhypothese auf Stufe k nicht verworfen wird. Erst auf der K-ten Stufe nach vollständiger Vergabe der Freiheitsgrade ν_G endet die Studie mit der Annahme der Nullhypothese.

Für den globalen Fehler zweiter Art unter θ_{H_1}, formuliert für eine feste Anzahl an Stufen K, gilt

$$\begin{aligned}\beta &= \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(\bigcap_{i=1}^{K}\{S_i \leq cv_\alpha\}\right) = \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(S_1 \leq cv_\alpha\right)\cdot \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(S_2 \leq cv_\alpha \middle| S_1 \leq cv_\alpha\right)\cdot \\ &\qquad \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(S_3 \leq cv_\alpha \middle| \bigcap_{i=1}^{2}\{S_i \leq cv_\alpha\}\right)\cdot \ldots \cdot \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(S_K \leq cv_\alpha \middle| \bigcap_{i=1}^{K-1}\{S_i \leq cv_\alpha\}\right) \\ &:= \prod_{i=1}^{K}\beta_i'. \end{aligned} \tag{5.1}$$

Die Wahrscheinlichkeit, dass bis zur $(k-1)$-ten Stufe keine Ablehnung der Nullhypothese erfolgt ist, wird mit $\beta(k-1)$ bezeichnet und berechnet sich durch $\beta(k-1) = \prod_{i=1}^{k-1}\beta_i'$. Die bedingte Wahrscheinlichkeit, nach $(k-1)$ Stufen im weiteren Verlauf der Studie die Nullhypothese nicht abzulehnen, wird angelehnt an den verbleibenden Fehler erster Art α_k^* mit β_k^* bezeichnet. Der globale Fehler zweiter Art lässt sich damit wie folgt darstellen:

$$\begin{aligned}\beta &= \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(\bigcap_{j=1}^{k-1}\{S_j \leq cv_\alpha\} \cap \left(\bigcap_{i=k}^{K}\{S_i \leq cv_\alpha\}\right)\right) \\ &= \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(\bigcap_{i=1}^{k-1}\{S_i \leq cv_\alpha\}\right)\cdot \mathrm{P}_{\theta_{\mathrm{H}_1}}\left(\bigcap_{j=k}^{K}\{S_j \leq cv_\alpha\} \middle| \bigcap_{i=1}^{k-1}\{S_i \leq cv_\alpha\}\right) \\ &:= \beta(k-1)\cdot\beta_k^*, \quad 2 \leq k \leq K. \end{aligned} \tag{5.2}$$

Bei der proportionalen Wahl der Gewichte und der Stichprobenumfänge in der verallgemeinerten Inverse-χ^2-Methode, d.h. $m_k/M_k = \nu_k/(\nu_G - \nu_\Sigma(k-1))$ (vgl. Abschnitt 4.3), werden die Stichprobenumfänge m_k und M_k zur Regulation des Studienverlaufs verwendet. Bei der Bestimmung dieser Umfänge stehen die bedingten Fehlerwahrscheinlichkeiten zweiter Art $\beta_{g,k}$ und $\beta_{G,k}$ für die theoretischen Werte β_k' und β_k^*. Dabei ist $\beta_{g,k} \geq \beta_{G,k}$ zu wählen.

Aus diesen theoretischen Überlegungen ergibt sich die Möglichkeit über eine entsprechende Wahl von $1-\beta_{g,k}$ und $1-\beta_{G,k}$ den Powerverlauf zu beeinflussen. Zum einen kann eine feste Folge der bedingten Wahrscheinlichkeiten festgelegt werden, die bei der Berechnung von m_k und M_k gemäß (3.11) und (3.9) verwendet werden. Zum anderen kann nach jeder

Stufe eine Anpassung an den beobachteten Parameter erfolgen. Nach der Realisierung der ersten Stufe kann die Parameterschätzung kleiner als der ursprünglich angenommene Wert θ_0 ausfallen. Soll dieser Schätzwert mit einer globalen Power von $1-\beta$ nachgewiesen werden, ist eine ursprünglich geplante bedingte Power $1-\beta_2^* = 1-\beta/\beta_{g,1}$ gemäß (5.2) für den weiteren Studienverlauf nach oben zu korrigieren. Da die Power von $1-\beta_{g,1}$ auf der ersten Stufe für θ_0 gilt, liegt für einen kleineren Parameterwert aus der Alternative ein größerer Fehler zweiter Art vor. Rückblickend kann der unter dem Wert der Parameterschätzung $\hat{\theta}_1$ auftretende Fehler zweiter Art auf der ersten Stufe, $\beta(1)_{\hat{\theta}_1}$, bestimmt werden und die verbleibende Power angepasst werden durch $1-\beta^*_{2,\hat{\theta}_1} = 1-\beta/\beta(1)_{\hat{\theta}_1}$. Eine nachträgliche Berechnung der Power $1-\beta(k)_{\hat{\theta}_k}$ kann nach jeder Stufe k erfolgen. Dazu wird die Wahrscheinlichkeit des vorzeitigen Studienabbruchs mit Ablehnung der Nullhypothese unter dem neuen Parameterwert bei Verwendung der gewählten Freiheitsgrade ν_i, der Stichprobenumfänge n_i und der bedingten Fehlerraten erster Art α_i' gemäß (4.7), $i \leq k$, bestimmt.

Aus Gleichung (5.1) lassen sich für die Wahl der bedingten Power auf der folgenden Stufe $1-\beta_{g,k}$ mehrere mögliche Aufteilungen ableiten. Die Idee einer konstanten bedingten Power pro Stufe, wie in vielen adaptiven Designs verwendet, führt bei einer geplanten Anzahl K an Stufen zur Wahl von $1-\beta_{g,k} = 1-\sqrt[K]{\beta}$. Für den Stichprobenumfang M_k, der zur Bestimmung eines proportional vergebenen Gewichts benötigt wird, können unter anderem folgende Werte für $1-\beta_{G,k}$ verwendet werden:

1. die von Hartung und Knapp (2003) vorgeschlagene Konstante $1-\beta$, d.h. die geplante globale Power,
2. die unter θ_0 erwartete bedingte verbleibende Power ab der k-ten Stufe $1-\beta^{(K-k+1)/K}$ oder
3. eine rückblickend gemäß der Parameterschätzung $\hat{\theta}_{k-1}$ berechnete verbleibende Power $1-\beta/\beta(k-1)_{\hat{\theta}_{k-1}}$.

Die zweite Variante legt die maximale Anzahl durchgeführter Stufen fest. Denn auf der K-ten Stufe gilt $1-\beta_{g,K} = 1-\beta_{G,K}$, was zur vollständigen Vergabe der verbleibenden Freiheitsgrade in der K-ten Stufe führt. Bei der ersten und dritten Variante kann durch die Einführung minimaler Stichprobenumfänge oder Gewichte eine maximale Anzahl an Stufen erzwungen werden. Wird eine Studie bei Verwendung der ersten oder dritten Variante zur Wahl von $1-\beta_{G,k}$ bis zur Stufe K nicht beendet, kann zum Beispiel die bedingte Power der letzten geplanten Stufe $1-\beta_{g,k} = 1-\sqrt[K]{\beta}$ verwendet oder eine beliebige andere Strategie verfolgt werden.

In jeder Variante kann vor Erreichen der K-ten Stufe durch die Vergabe der verbleibenden Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$ auf der k-ten Stufe, $k < K$, das Studienende vorgezogen werden. Eine Situation, in der dies sinnvoll erscheint, liegt vor, wenn die bedingte Irrtumswahrscheinlichkeit erster Art α_k^* größer als $1 - \beta_{G,k}$ oder $1 - \beta_{g,k}$ ist. In der dritten Variante bietet sich die Durchführung einer letzten Stufe an, wenn die durch Rekalkulation bestimmte verbleibende bedingte Power $1 - \beta_{G,k} = 1 - \beta/\beta(k-1)_{\hat{\theta}_{k-1}}$ kleiner als die vorgesehene bedingte Power $1 - \beta_{g,k}$ der folgenden Stufe ausfällt.

Gilt nach der rückblickenden Berechnung der Power $1 - \beta(k-1)_{\hat{\theta}_{k-1}} > 1 - \beta$, ist eine Strategieänderung durchzuführen. Um den geschätzten Parameterwert mit einer Power von $1 - \beta$ nachzuweisen, wurde in der Studie bereits ausreichender Stichprobenumfang aufgewandt. Entweder muss der nachzuweisende Wert des Parameters angepasst oder die Studie wegen Aussichtslosigkeit abgebrochen werden.

Eine weitere Idee, ein adaptives Design gemäß Inverse-χ^2-Methode durch die Vergabe von Power zu konstruieren, greift auf ein gruppensequentielles Verfahren von Bauer (1992) zurück. Die kritischen Schranken werden in diesem Verfahren so bestimmt, dass bei konstanten Stichprobenumfängen der absolute Anteil π_k an der globalen Power $1-\beta = \sum_{i=1}^{K} \pi_i$ auf Stufe k erreicht wird. Ähnlich kann bei Designs gemäß der adaptiven Inverse-χ^2-Methode durch Kopplung von Stichprobenumfang und Gewicht bei der proportionalen Vergabe das adaptive Design über den Powerzuwachs pro Stufe bestimmt werden. Dabei ergibt sich das lokale Niveau gemäß (4.7) über die pro Stufe gewählten Gewichte. Theoretisch setzt sich die globale Power für K geplante Stufen für einen Wert aus der Alternative aus den Powerzuwächsen π_k pro Stufe zusammen:

$$1 - \beta = \sum_{k=1}^{K} \pi_i = \sum_{k=1}^{K} \mathrm{P}_{\theta_{\mathrm{H}_1}} \left(\bigcap_{i=1}^{k-1} \{S_i \leq cv_\alpha\} \cap \{S_k > cv_\alpha\} \right). \tag{5.3}$$

Für die Fehlerrate zweiter Art bis zur k-ten Stufe $\beta(k)$ und die bedingte Wahrscheinlichkeit β_k' des Fortsetzens der Studie nach Erreichen der k-ten Stufe gilt:

$$\beta(k) = 1 - \sum_{i=1}^{k} \pi_i = \mathrm{P}_{\theta_{\mathrm{H}_1}} \left(\bigcap_{i=1}^{k} \{S_i \leq cv_\alpha\} \right), \tag{5.4}$$

$$\beta_k' = \mathrm{P}_{\theta_{\mathrm{H}_1}} \left(\{S_k \leq cv_\alpha\} \middle| \bigcap_{i=1}^{k-1} \{S_i \leq cv_\alpha\} \right) = \frac{\beta(k)}{\beta(k-1)} = \frac{1 - \sum_{i=1}^{k} \pi_i}{1 - \sum_{i=1}^{k-1} \pi_i}. \tag{5.5}$$

Gemäß (5.5) lassen sich für beliebige Verläufe der Power die zu verwendenden bedingten Fehlerraten zweiter Art, $\beta_{g,k}$, bestimmen. Naheliegend ist die Nachahmung des Powerverlaufs bekannter gruppensequentieller Designs wie von Pocock (1977) oder O'Brien und Fleming (1979), siehe Tabelle 2.2. Während in einem fünfstufigen Design nach Pocock

(1977) bei einer Power von insgesamt 90% jeweils mindestens 20% Power auf den ersten drei Stufen erzielt werden, beträgt der Powerzuwachs im Design nach O'Brien und Fleming (1979) auf den ersten drei Stufen 0.1%, 12% und 34%. Der Powerverlauf in diesen gruppensequentiellen Designs ist bei fester globaler Power unabhängig vom Wert des Parameters aus der Alternative. Weiterhin kann u.a. ein theoretisch konstanter Powerzuwachs pro Stufe $\pi_k = (1-\beta)/K$, $k \geq 1$, festgelegt werden, was im fünfstufigen Design bei einer geplanten Power von 90% einem Powerzuwachs von 18% pro Stufe entspricht.

Für die Wahl von $\beta_{G,k}$ bieten sich die entsprechenden bei der Verwendung einer konstanten bedingten Power beschriebenen Möglichkeiten (siehe S. 73) an. Wird gemäß Variante 3 die retrospektiv vergebene Power $1-\beta(k-1)_{\hat{\theta}_{k-1}}$ verwendet, kann der Wert für den Fehler zweiter Art $\beta_{g,k}$ auf der k-ten Stufe, $k \geq 2$, angepasst werden. Eine Möglichkeit der Anpassung der bis zur k-ten Stufe angestrebten Power $1-\beta(k)$ an den geschätzten Parameterwert ergibt sich durch die Wahl von

$$\beta_{g,k} = \frac{\beta(k)}{\beta(k-1)_{\hat{\theta}_{k-1}}}. \tag{5.6}$$

Gilt nach Rekalkulation der Power $\beta(k) > \beta(k-1)_{\hat{\theta}_{k-1}}$, liefert Gleichung (5.6) einen nicht realisierbaren Wert größer 1. Die rückblickend kalkulierte Power bedeutet, dass mit den bisher verwendeten Umfängen und Gewichten unter dem Parameterwert $\hat{\theta}_{k-1}$ die bis zur Stufe $(k-1)$ geplante Power bereits ermöglicht wurde. In diesem Fall kann anstelle von $\beta(k)$ das größte $\beta(i)$, $i > k$, verwendet werden, das $\beta(k-1)_{\hat{\theta}_{k-1}}$ unterschreitet, d.h. $\max_i\{\beta(i)\big|\beta(i) < \hat{\beta}_{k-1}\}$. Ist $\beta(k-1)_{\hat{\theta}_{k-1}} < \beta$ muss wie bereits im Fall der konstanten bedingten Power beschrieben eine Strategieänderung erfolgen.

Eine weitere Variante, einen vorgegebenen Powerverlauf umzusetzen, liegt in der direkten Verwendung des theoretischen Powerzuwachses auf der jeweiligen Stufe. Sofern der bis zur $(k-1)$-ten Stufe nachträglich berechnete Fehler zweiter Art $\beta(k-1)_{\hat{\theta}_{k-1}}$ größer ist als der geplante Powerzuwachs π_k, kann der bedingte Fehler $\beta_{g,k}$ wie folgt gewählt werden:

$$\beta_{g,k} = \frac{\beta(k-1)_{\hat{\theta}_{k-1}} - \pi_k}{\beta(k-1)_{\hat{\theta}_{k-1}}}. \tag{5.7}$$

Verwendet man eine Strategie, die nicht automatisch mit der K-ten Stufe durch die Vergabe der verbleibenden Freiheitsgrade zum Studienende führt, ist keine bedingte Power $1-\beta_{g,K+1}$ für die folgende Stufe festgelegt oder intuitiv vorgegeben. Strategien, die dies umgehen, liegen zum Beispiel in der weiteren Verwendung einer konstanten bedingten Power oder eines konstanten Powerzuwachses gemäß (5.7).

Tabelle 5.1 enthält den theoretischen Powerzuwachs und die zugehörige bedingte Fehlerwahrscheinlichkeit zweiter Art pro Studienabschnitt bei konstanter bedingter Power

Tabelle 5.1: *Powerzuwachs π_k und bedingte Power $1-\beta'_k$ auf Stufe k (in %) in fünfstufigen Studien bei konstantem absoluten Powerzuwachs und bei konstanter bedingter Power pro Stufe, globale Power $1-\beta = 0.9$*

	konstante bedingte Power					konstante Powerzugabe				
k	1	2	3	4	5	1	2	3	4	5
π_k	36.9	23.3	14.7	9.3	5.8	18.0	18.0	18.0	18.0	18.0
$\sum_{i=1}^{k} \pi_i$	36.9	60.2	74.9	84.2	90.0	18.0	36.0	54.0	72.0	90.0
$1-\beta'_k$	36.9	36.9	36.9	36.9	36.9	18.0	22.0	28.1	39.1	64.3

sowie bei konstantem Powerzuwachs in Studien mit fünf geplanten Stufen bei einer globalen Power von 90%. Wird eine konstante bedingte Power $1-\sqrt[K]{\beta}$ pro Stufe angesetzt, wird ein hoher Teil der Gesamtpower bereits auf der ersten Stufe erreicht. Der absolute Zuwachs auf den folgenden Stufen sinkt bis auf 5.8% auf der fünften Stufe. Ist ein konstanter Powerzuwachs pro Stufe gefordert, steigt die bedingte Power pro Stufe von 18% bis zu 64.3% auf der fünften Stufe an.

Allgemein ist bei den Strategien zu beachten, dass der bedingte Fehler erster Art einerseits durch die Kopplung von Umfang und Gewicht durch die proportionale Wahl und andererseits durch die bereits realisierten Stufen beeinflusst wird (vgl. (4.14)). Der benötigte Stichprobenumfang für die gleiche bedingte Power kann daher sehr unterschiedlich ausfallen. Für eine geringe Power auf der ersten Stufe ist aufgrund des sehr geringen lokalen Niveaus (vgl. Tabelle 4.2) ein sehr hoher Stichprobenumfang erforderlich. Hartung und Knapp (2003) wählen in den beschriebenen Beispielen auf der ersten Stufe die Freiheitsgrade ν_1 und den Stichprobenumfang n_1 nicht proportional. Theoretisch kann bei jeder Bestimmung des Gewichts der nachfolgenden Stufe von der proportionalen Gewichtung abgewichen werden. Weiterhin besteht die Möglichkeit, die Kopplungsvorschrift zu verändern, zum Beispiel den Anteil an den Freiheitsgraden $\nu_G - \nu_\Sigma(k)$ höher zu wählen als den Anteil am Stichprobenumfang M_k. Dies kann unter anderem erreicht werden, indem die Freiheitsgrade ν_k anteilig an $\nu_G - \nu_\Sigma(k-1)$ gemäß einer Wurzel aus dem Verhältnis von m_k zu M_k gewählt werden:

$$\epsilon_\nu = \frac{\nu_k}{\nu_G - \nu_\Sigma(k-1)} = \left(\frac{m_k}{M_k}\right)^x = \epsilon_m^x, \quad 0 < x \leq 1. \tag{5.8}$$

Dies führt bei exakter Bestimmung des Stichprobenumfangs für eine bedingte Power $1-\beta_{g,k}$ wie in Abschnitt 4.3 auf ein Fixpunktproblem. Tabelle B.1 enthält die Lösungen des Fixpunktproblems für die erste Stufe bei nichtproportionaler Gewichtung für $x = 0.25, 0.5, 0.75$ analog zu Tabelle 4.2. Es fällt auf, dass der Anteil ϵ_ν der Freiheitsgrade für die drei betrachteten Wurzelfunktionen höher ausfällt als bei der proportionalen Ge-

wichtung. Der Anteil ϵ_m des Stichprobenumfangs auf der ersten Stufe an M_1 ist jedoch erwartungsgemäß deutlich geringer. In Tabelle B.2 ist für $x = 0.5$ die Gesamtpower und die Power auf der ersten Stufe bei Aufteilung von M_1 und ν_G auf zwei Stufen dargestellt. Der Powerverlust fällt im Vergleich zur proportionalen Vergabe etwas größer aus, während die Power der ersten Stufe zum Teil deutlich erhöht ist. Im Vergleich zu Tabelle 4.1, die die erzielte Power bei proportionaler Aufteilung der Freiheitsgrade und des Stichprobenumfangs enthält, geht die Symmetrie der erzielten Power hinsichtlich ϵ und $1 - \epsilon$ verloren.

Neben der Vorgabe der bedingten Power $1 - \beta_{g,k}$ auf der k-ten Stufe und der Wahl von $1 - \beta_{G,k}$ ist für die Bestimmung des Stichprobenumfangs entscheidend, für welchen Wert aus der Alternative die bedingte Power erzielt werden soll. Eine Anpassung für beliebig kleine Werte, die klinisch nicht relevant sind, erscheint nicht sinnvoll bzw. kann nur durch die Festlegung von maximalen Umfängen pro Stufe realisiert werden. Dies führt auf den Stufen, bei denen eine obere Grenze wirksam wird, indirekt zum Nachweis eines größeren Effekts mit der vorgegebenen bedingten Power. Um sehr große Stichprobenumfänge zu vermeiden, sollte in Absprache mit Fachkundigen ein minimaler nachzuweisender Effekt festgelegt werden. Dies kann beispielsweise die Hälfte des zur Planung der ersten Stufe verwendeten Parameterwerts θ_0 sein. Der festgelegte minimale Effekt sollte auch als untere Grenze bei der rückblickenden Berechnung von $\beta(k-1)$ verwendet werden, da für kleinere Werte sehr hohe Fehlerraten zweiter Art $\beta(k-1)_{\hat{\theta}}$ zu hohen Werten für die Power der nachfolgenden Stufen und entsprechend hohen Stichprobenumfängen führen. Anstelle eines geschätzten Werts kann ein a priori festgelegter Parameterwert für die Bestimmung des Stichprobenumfangs auf allen Stufen verwendet werden (vgl. Bauer und König, 2006). Die Anpassung des Stichprobenumfangs erfolgt dadurch nur hinsichtlich der Teststatistiken der durchgeführten Stufen bzw. des sich daraus ergebenden bedingten Fehlers zweiter Art für den weiteren Studienverlauf.

Neben der Auswirkung unterschiedlicher Strategien ist der Einfluss verschiedener Parameterschätzer (siehe 4.5) auf den mittleren Stichprobenumfang und die erzielte Power zu untersuchen. Schätzer, die den wahren Wert unterschätzen, führen zu größeren benötigten Stichprobenumfängen, während eine Überschätzung des wahren Werts nicht immer das Erreichen einer globalen Power $1 - \beta$ gewährleistet.

5.2 Simulationen

Bei der Umsetzung der im voranstehenden Kapitel entwickelten Strategien in einer Simulationsstudie müssen feste Regeln für den Verlauf der simulierten Studien definiert

werden. Es muss festgelegt werden, wann eine Studie mit der Vergabe der verbleibenden Freiheitsgrade beendet wird, falls sich dies nicht durch die Wahl von $\beta_{g,k}$ und $\beta_{G,k}$ ergibt, und welche bedingte Power auf dieser letzten Stufe gewählt wird. Obere und untere Schranken für den Stichprobenumfang bzw. ein minimales Gewicht müssen, sofern vorgegeben, hinsichtlich der Gewichtungsvorschrift berücksichtigt werden. Die folgenden Festlegungen müssen bei der praktischen Durchführung einer klinischen Studie nicht zwingend umgesetzt werden, zur Simulation der Eigenschaften einer Strategie sind sie jedoch notwendig.

Betrachtet wird im Folgenden das einseitige Testproblem (3.1) mit der Teststatistik des Zweistichproben-Gauß-Tests (2.2) auf jeder Stufe des adaptiven Designs, wobei der interessierende Parameter θ für die Erwartungswertdifferenz steht. Die Varianz der Zufallsvariablen beträgt $\sigma^2 = 1$. Für die Planung auf der ersten Stufe wird a priori ein Effekt $\theta_0 = 0.5$ angenommen. Das Niveau wird mit $\alpha = 0.025$ und die angestrebte globale Power mit $1 - \beta = 0.9$ festgelegt. Zur Umsetzung der Teststatistik wird ein minimaler Umfang von $n_{min} = 4$ für m_k und M_k vorgegeben und der Stichprobenumfang auf ein Vielfaches von zwei gerundet. Die Rundung erfolgt nach der Bestimmung des Anteils der Freiheitsgrade, so dass die vorgegebene Power $1 - \beta_{g,k}$ mindestens erreicht wird. Wird der minimale Umfang für m_k gewählt, werden die Freiheitsgrade proportional gewählt, d.h. $\nu_k = (\nu_G - \nu_\Sigma(k)) \cdot 4/M_k$. Ergibt sich für M_k der minimale Stichprobenumfang, werden die verbleibenden Freiheitsgrade in der k-ten Stufe vergeben und die Studie endet nach der k-ten Stufe. Die Simulationen erfolgen unter der Nullhypothese und für $\theta = 0.3, 0.5, 0.7$. Der mittlere benötigte Stichprobenumfang, die mittlere Anzahl an durchgeführten Stufen sowie die erzielte Power werden aus jeweils 10000 Simulationsläufen bestimmt. Die Simulationen werden mit der Software R (R Development Core Team, 2006), Version 2.4.1, durchgeführt.

5.2.1 Adaption für festen Parameterwert bei fester maximaler Stufenzahl

Ein erster Vergleich erfolgt zwischen Strategien, die eine Vergabe der verbleibenden Freiheitsgrade und damit das Studienende nach einer maximalen Zahl an Stufen festlegen. Aufgrund des Aufwands bei der Durchführung von Zwischenauswertungen sowie der geringen zusätzlichen Reduktion des mittleren Stichprobenumfangs bei weiteren Zwischenauswertung in gruppensequentiellen Designs (vgl. McPherson, 1982) wird die maximale Anzahl an Stufen auf $K = 5$ beschränkt. Weiter wird zunächst der a priori Wert $\theta_0 = 0.5$ zur Planung des Stichprobenumfangs auf allen Stufen verwendet.

Tabelle 5.2: *Power, Niveau, Stichprobenumfang und Gewicht auf der ersten Stufe in vier Strategien gemäß S. 79 bei Planung mit* $\theta_0 = 0.5$, $\alpha = 0.025$

Strategie	π_1 [%]	α_1 [%]	n_1	ν_1
I	36.90	0.223	102	5.98
II	18.00	0.080	80	4.73
III	20.26	0.095	82	4.94
IV	0.09	0.001	22	1.27

Folgende Strategien für die Wahl der bedingten Power $1 - \beta_{g,k}$, $k = 1, \ldots, 5$, auf den einzelnen Stufen bei proportionaler Vergabe der Freiheitsgrade und des Stichprobenumfangs werden verglichen:

Strategie I: Konstante bedingte Power $1 - \beta_{g,k} = 1 - \sqrt[5]{0.1}$,

Strategie II: Konstanter Powerzuwachs von $\pi_k = 0.18$,

Strategie III: Powerzuwachs nach Pocock (1977), vgl. Tabelle 2.2,

Strategie IV: Powerzuwachs nach O'Brien und Fleming (1979), vgl. Tabelle 2.2.

Für die Wahl der Gewichte bestimmenden bedingten Power $1 - \beta_{G,k}$, die für den verbleibenden Studienteil festgelegt wird, werden die folgenden zwei Varianten betrachtet:

Strategie A: die theoretische verbleibende Power $1 - \prod_{i=k}^{5} \beta_i'$ und

Strategie B: die konstante globale Power $1 - \beta = 0.9$.

Bei der zweiten Variante wird wie von Hartung und Knapp (2003) vorgeschlagen auf der Stufe k^*, auf der die Freiheitsgrade aufgebraucht werden, spätestens auf der fünften Stufe, die bedingte Power $1 - \beta_{G,k^*} = 1 - \beta$ verwendet. In der ersten Variante gilt hingegen ab der zweiten Stufe $1 - \beta_{G,k} < 1 - \beta$ und auf der fünften Stufe $\beta_{G,5} = \beta_{g,5}$. Nach Lösung des Fixpunktproblems (4.15) ergeben sich die in Tabelle 5.2 dargestellten Startgewichte und Stichprobenumfänge der vier Strategien der Wahl von $\beta_{g,k}$. Die Wahl von $\beta_{G,k}$ führt nicht zu Unterschieden auf der ersten Stufe, da in beiden Varianten $\beta_{G,1} = \beta$ gilt. Die Wahl einer konstanten bedingten Power von 36.9% führt zum größten Stichprobenumfang von 102 auf der ersten Stufe, die Powervergabe nach O'Brien und Fleming (1979) bei einer Power von 0.09% mit 22 zur geringsten Fallzahl auf der ersten Stufe. Die Powervergabe nach Pocock (1977) und die Strategie des konstanten Powerzuwachs unterscheiden sich auf der ersten Stufe wenig und führen zu einem Stichprobenumfang von 82 bzw. 80. Die proportional

zum ungerundeten einstufigen Stichprobenumfang von 168.12 gewählten Gewichte liegen zwischen 1.27 für Strategie IV und 5.98 für Strategie I. Durch die Wahl der Freiheitsgrade ist gemäß (4.6) das lokale Niveau auf der ersten Stufe festgelegt. Es liegt zwischen 0.001% und 0.223% entsprechend der Größenordnung der Freiheitsgrade.

Die vier in Abschnitt 4.5 beschriebenen Schätzer werden jeweils nach dem letzten Studienabschnitt einer simulierten Studie berechnet. Mittelwert und Standardabweichung der finalen Parameterschätzungen werden in Abhängigkeit der gewählten Strategie betrachtet. Der Stichprobenumfang n_k, $k = 1, \ldots, 5$, wird auf jeder Stufe für den konstanten Wert θ_0 bestimmt.

Tabelle 5.3 fasst die simulierten Ablehnungswahrscheinlichkeit, den benötigten mittleren und medianen Stichprobenumfang (ASN und MSN) und die mittlere Anzahl durchgeführter Stufen für die verschiedenen beschriebenen Szenarien und Strategien zusammen. Unter der Nullhypothese sind Abweichungen der Ablehnwahrscheinlichkeit vom Niveau von 2.5% durch Simulationsungenauigkeit zu erklären, da das Verfahren nach Konstruktion das Niveau einhält und bei stetigen Teststatistiken ausschöpft. Bei 10000 Simulationen sind Schwankungen von bis zu 0.3% um den wahren Wert von 2.5% mit einer Wahrscheinlichkeit von 95% zu erwarten.

In den betrachteten Szenarien entspricht die Anzahl durchgeführter Stufen unter der Nullhypothese fast immer der maximal möglichen Anzahl von fünf Stufen unabhängig von der gewählten Strategie. Da in Strategie B auf der letzten Stufe ein Stichprobenumfang für eine bedingte Power von 90% bestimmt wird, liegen die durchschnittlichen und medianen Umfänge deutlich über denen bei Strategie A. 50% der simulierten Studien unter der Nullhypothese benötigen unter Strategie A je nach Strategie zur Wahl von $\beta_{g,k}$, $k = 1, \ldots, K$, höchstens 614-748 Patienten, während bei Strategie B der mediane benötigte Stichprobenumfang zwischen 752 und 944 liegt. Die Simulation unter der Nullhypothese zeigt, dass die Abhängigkeit des Stichprobenumfangs der nächsten Stufe von den bereits realisierten Studienabschnitten auch bei Verwendung eines konstanten Werts aus der Alternativen zur Ermittlung des Stichprobenumfangs zu sehr hohen Patientenzahlen führen kann. Möglichkeiten, den mittleren bzw. medianen Stichprobenumfang zu reduzieren, liegen in der Einführung eines Stopps wegen Aussichtslosigkeit (siehe Abschnitt 4.4) oder der Begrenzung des Umfangs pro Stufe bzw. insgesamt.

Der Unterschied zwischen den Strategien A und B nimmt für steigende Werte des wahren θ ab. Für $\theta = 0.7$ liegen die erzielten Ablehnungswahrscheinlichkeit jeweils um 99% bei ähnlichen mittleren Stichprobenumfängen. Dadurch, dass bei größerem θ häufiger ein Abbruch mit Ablehnung der Nullhypothese erfolgt, bevor alle Freiheitsgrade vergeben sind, wirkt sich der Unterschied zwischen Strategie A und B hinsichtlich der Powervergabe

auf der letzten Stufe bei Vergabe aller Freiheitsgrade wenig aus. Für $\theta_0 = \theta = 0.5$ und $\theta = 0.3$ ergeben sich durch die höheren mittleren Umfänge in Strategie B entsprechend höhere Werte der Power, da auf der Stufe, auf der alle verbleibenden Freiheitsgrade vergeben werden, die globale Power zur Umfangsbestimmung herangezogen wird. Für den wahren Wert $\theta = 0.3$, d.h. einem kleineren Wert als dem Planwert θ_0, wird daher unter Strategie B eine Power von etwa 77% bei einer mittleren Umfangserhöhung von etwa 50 Patienten im Vergleich zu 62% Power unter Strategie A erreicht. Der Unterschied im mittleren Umfang liegt für $\theta = 0.5$ nur bei etwa 7-11 Personen, die erzielte Power liegt unter Strategie A bei etwa 93%, unter Strategie B bei 98%. Für $\theta = 0.5$ und 0.7 liegt der mittlere benötigte Stichprobenumfang mit maximal 144 bzw. 109 für alle Strategien deutlich unter dem einstufigen Stichprobenumfang von 170 Patienten. Durch die konstante Wahl von $\beta_{G,k} = \beta \leq \prod_{i=k}^{5} \beta_i'$, $k = 1, \ldots, 5$, in der zweiten Strategie werden die Freiheitsgrade im Vergleich zu Strategie A langsamer vergeben, vgl. (4.17). Dies führt zu einer geringfügig größeren Anzahl an durchgeführten Stufen in Strategie B.

Bei Verwendung des wahren Werts im Fall $\theta = \theta_0 = 0.5$ zur Bestimmung des Stichprobenumfangs unter Strategie A wird nach (5.1) eine Power von 90% erwartet. Die Erhöhung des simulierten Niveaus um 2-3% erklärt sich zum einen durch die Begrenzung der Stichprobenumfänge m_k und M_k durch $n_{\min} = 4$ nach unten und durch das Runden des verwendeten Stichprobenumfangs auf ein Vielfaches von zwei. Zum anderen führt beim Fortsetzen der Studie die Berechnung eines verbleibenden Niveaus $\alpha_k^* < 1$, das größer als die verbleibende Power $1 - \beta_{G,k}$ ist, gemäß Simulationsregel zum Studienende in der folgenden Stufe. Auf dieser letzten Stufe, in der alle verbleibenden Freiheitsgrade verbraucht werden, ist die Wahrscheinlichkeit des Ablehnens der Nullhypothese unter der Nullhypothese bereits größer als die geplante Power unter θ_0. Die Wahrscheinlichkeit, dass dieses Ereignis nach der ersten Stufe unter $\theta = \theta_0$ eintritt, berechnet sich zu

$$
\begin{aligned}
& P_{\theta_0}\left(\alpha_2^* \geq 1 - \prod_{i=2}^{K} \beta_i',\ q_1(\nu_1) < cv_\alpha\right) \qquad (5.9)\\
&= P_{\theta_0}\left(1 - F_{\chi^2(\nu_G - \nu_1)}(cv_\alpha - q_1(\nu_1)) \geq 1 - \prod_{i=2}^{K} \beta_i',\ q_1(\nu_1) < cv_\alpha\right)\\
&= 1 - \Phi\left(\Phi^{-1}\left(F_{\chi^2(\nu_1)}\big(cv_\alpha - F^{-1}_{\chi^2(\nu_G-\nu_1)}(\prod_{i=2}^{K} \beta_i')\big) - \theta_0 \cdot \sqrt{n_1/4}\right)\right) - (1 - \beta_1).
\end{aligned}
$$

Unter $\theta = 0.5$ liegt die Wahrscheinlichkeit für die Strategien I - III bei etwa 7.5%. In Strategie IV liegt die Wahrscheinlichkeit auf der zweiten Stufe bei 0.36%, erhöht sich jedoch auf der dritten Stufe auf 6.7%. Der simulierte Anteil für das Ereignis $\bigcup_{i=1}^{K}\{\alpha_i^* \geq 1 - \beta_{G,i},\ S_{i-1} < cv_\alpha\}$ insgesamt beträgt für die vier Strategien 10.9, 14.5,

13.8 bzw. 12.9%. Dabei liegt die mittlere Abweichung des Niveaus von der geplanten verbleibenden bedingten Power bei 10.5, 8.6, 9.3 und 8.8%. Eine Simulation ohne Rundung des Stichprobenumfangs, die im Fall $\alpha_k^* \geq 1 - \beta_{G,k}$ gemäß eines nicht fairen Münzwurfs mit Wahrscheinlichkeit $1 - \beta_{G,k}$ über Ablehnung oder Annahme der Nullhypothese entscheidet, führt zu einer simulierten Power bei $\theta_0 = \theta$ gleich der geplanten Power von 90%. Bei Rundung auf ein Vielfaches von zwei und der Verwendung eines minimalen Umfangs kann entsprechend für die letzte Stufe k^* bei Vergabe der verbleibenden Freiheitsgrade aus (5.2) die Wahrscheinlichkeit des Ablehnens für einen nicht fairen Münzwurf bestimmt werden.

Sie ergibt sich aus $1 - \beta / \prod_{k=1}^{k^*-1} \beta_{g,k}^{\text{rnd}}$ mit den rückblickend berechneten tatsächlichen Fortsetzungswahrscheinlichkeiten $\beta_{g,k}^{\text{rnd}}$ auf den bereits durchgeführten Stufen

$$\beta_{g,k}^{\text{rnd}} = 1 - \Phi(\Phi^{-1}(1 - \alpha_k') - \sqrt{n_k/4} \cdot \theta_0). \tag{5.10}$$

Hinsichtlich des Vergleichs von Strategie A und B lässt sich zusammenfassen, dass bei einem großen wahren klinisch relevanten Wert die Strategie der Vergabe einer höheren Power auf einer geplanten letzten Stufe, die durch häufigen vorzeitigen Abbruch unter der Alternative nur selten erreicht wird, zur Steigerung der globalen Power geeignet ist, ohne den mittleren Umfang stark zu erhöhen. Für Parameterwerte nahe der Nullhypothese kann dies jedoch zu einer unnötigen Erhöhung des Stichprobenumfangs führen. Eine andere Option, um die Power bei kleineren klinisch relevanten Werten als dem Planwert θ_0 zu erhöhen, liegt in der Verwendung eines Effektschätzers anstelle eines festen Werts zur Anpassung des Stichprobenumfangs (siehe Abschnitt 5.2.2).

Die vier Strategien hinsichtlich der Wahl der bedingten Power auf den fünf Stufen wirken sich für alle betrachteten wahren Werte des Parameters besonders auf die Anzahl der durchgeführten Stufen aus. Die Powervergabe nach O'Brien und Fleming (1979) (Strategie IV) führt aufgrund der langsamen Powervergabe stets zur höchsten mittleren Anzahl an Stufen $\bar{k}$, die Wahl der konstanten bedingten Power (Strategie I) jeweils zur niedrigsten. Während unter der Nullhypothese alle Strategien eine mittlere Anzahl durchgeführter Stufen von mehr als 4.9 aufweisen, beträgt der Unterschied für die beiden Extreme etwa 0.6 Stufen unter $\theta = 0.3$, 1.3 Stufen unter $\theta = 0.5$ und 1.5 Stufen unter $\theta = 0.7$. Der Unterschied in der mittleren Stufenzahl von Strategie I zur Powervergabe nach Pocock (1977) beträgt für Werte aus der Alternative 0.2 bis 0.3 Stufen, zur Strategie mit konstantem Powerzuwachs zwischen 0.4 und 0.6 Stufen. Unter allen Strategien nimmt mit steigendem wahren Wert die mittlere Anzahl ausgeführter Stufen ab. Die mittleren Stichprobenumfänge (ASN) sind für die Powervergabe nach O'Brien und Fleming (1979) für die Parameterwerte, die nicht dem Planwert $\theta_0 = 0.5$ entsprechen am niedrigsten.

Der höchste mittlere Stichprobenumfang ergibt sich meist für die Strategie der konstanten bedingten Power, d.h. der Strategie mit dem höchsten Umfang auf der ersten Stufe. Insbesondere bei großen Werten des Parameters, wenn häufig nach der ersten Zwischenauswertung die Nullhypothese verworfen werden kann, beträgt der Unterschied etwa 20 Patienten im ASN und 80 im MSN. Die Ergebnisse unter der Powervergabe nach Pocock (1977) und dem konstantem Powerzuwachs unterscheiden sich innerhalb der vier Strategien am wenigsten und liegen bezüglich der beschriebenen Merkmale zwischen den beiden anderen Strategien.

Die vier betrachteten Schätzer für θ nach Studienende überschätzen den wahren Parameter für alle wahren Werte und unter allen Strategien. Es zeigt sich eine Abhängigkeit der Größe der Verzerrung und der Standardabweichung von der gewählten Strategie und vom wahren Parameterwert. Die beobachteten mittleren absoluten Abweichungen vom wahren Wert liegen für $\theta = 0$ zwischen 0.011 und 0.024, für $\theta = 0.3$ zwischen 0.038 und 0.078, für $\theta = 0.5$ zwischen 0.032 und 0.083 und für $\theta = 0.7$ zwischen 0.011 und 0.087. Unter der Nullhypothese und unter $\theta = 0.3$ ergeben sich für den einfachen gepoolten Schätzer im Mittel die größten Abweichungen vom wahren Wert. Für größere Werte von θ gilt dies für den gepoolten Schätzer nur unter Strategie IV, während in den anderen Szenarien der Schätzer $\hat{\theta}_k^{\sqrt{n}}$ die größte mittlere Abweichung vom wahren Wert aufweist. Die geringste Verzerrung liegt für den Schätzer aus dem Schnittpunkt der Funktionen zur Konstruktion des Konfidenzintervalls für alle Strategien und jeden betrachteten Parameterwert vor. Dieser Schätzer besitzt jedoch häufig die größte Standardabweichung. Die Standardabweichung aller Schätzer ist unter der Powervergabe nach O'Brien und Fleming (1979) am größten. Bis auf die Szenarien unter der Nullhypothese und bei $\theta = 0.7$ unter Strategie I, unter denen der Schätzer $\hat{\theta}_k^{\sqrt{n}}$ die geringste Standardabweichung besitzt, ergibt sich für den einfachen gepoolten Schätzer die geringste Standardabweichung bei allen Strategien.

Die Ergebnisse einer entsprechenden Simulation mit einer Gesamtanzahl an Freiheitsgraden $\nu_G = 100$ finden sich in Tabelle B.3. Bei einigen simulierten Studien unter der Nullhypothese wird aufgrund einer sehr geringen verbleibenden Anzahl an Freiheitsgraden das bedingte Niveau α_k^* gemäß (4.2) numerisch null. Dies führt theoretisch zu einem unendlich großen Stichprobenumfang auf der k-ten Stufe. Beim Eintritt einer solchen numerischen Ungenauigkeit wird die simulierte Studie wegen Aussichtslosigkeit abgebrochen. Je nach Strategie, vermehrt unter den Strategien I und III, treten 6 bis 51 Stopps wegen Aussichtslosigkeit auf. Aufgrund des sehr geringen lokalen Niveaus auf der ersten Stufe auch bei großen Freiheitsgraden ν_1 (vgl. Tabelle 4.2), ergeben sich im Vergleich zur Simulation mit $\nu_G = 10$ in allen Szenarien deutlich erhöhte durchschnittliche und mediane Stichprobenumfänge. Die mittleren Stichprobenumfänge liegen auch für $\theta = 0.7$ im

Bereich des Stichprobenumfangs eines einstufigen Designs für $\theta = 0.5$, d.h. nahe 170. Die Power erhöht sich entsprechend, so dass unter der Strategie A für $\theta = 0.3$ eine Power um 72% und unter $\theta = 0.5$ eine Power von 97% erreicht werden. Die mittlere Anzahl durchgeführter Stufen ist für $\theta = 0.3$ um bis zu 0.44 Stufen geringer als unter $\nu_G = 10$, für größere Werte um bis zu 0.6 Stufen.

Durch die Verdoppelung der mittleren Stichprobenumfänge unter der Nullhypothese und die geringe Ersparnis im Vergleich zum einstufigen Design empfiehlt sich eine deutliche Erhöhung der Freiheitsgrade bei den betrachteten Strategien nicht.

Für die verwendeten Schätzer ergibt sich ein ähnliches Bild wie bei $\nu_G = 10$. Die Standardabweichung der Schätzer ist aufgrund des großen Stichprobenumfangs auf der ersten Stufe geringer als für $\nu_G = 10$. Der einfache gepoolte Schätzer besitzt unter der Alternative die geringste Varianz, unter der Nullhypothese ist sie nur etwas größer als die kleinste Varianz, die der mit der Wurzel aus dem Stichprobenumfang gewichtende Schätzer aufweist. Unter der Alternative besitzt der mit der Wurzel des Stichprobenumfangs gewichtende Schätzer jedoch die größte Standardabweichung und ist am stärksten verzerrt. Der Schnittpunkt der Konstruktionsfunktionen des Konfidenzintervalls gehört in allen Szenarien zu den am geringsten verzerrten Schätzern. Unter der Nullhypothese ist der mit der Wurzel der Stichprobenumfänge gewichtete Schätzer ebenfalls wenig verzerrt und unter $\theta = 0.7$ der einfache gepoolte Schätzer. Für $\theta \geq 0.5$ treten unter der Powervergabe nach O'Brien und Fleming (1979) die stärksten mittleren Abweichungen vom wahren Parameterwert auf.

Die Bedeutung der Kopplung von Stichprobenumfang und Gewicht der Studie zeigt sich, wenn die Freiheitsgrade auf den fünf Stufen konstant auf $\nu_k = 2$, $k = 1, \ldots, 5$, gesetzt werden, was der Kombinationsmethode nach Fisher entspricht, siehe (2.11). Tabelle B.4 enthält die Simulationsergebnisse unter den acht Strategien hinsichtlich der Wahl von $\beta_{g,k}$ und $\beta_{G,k}$ bei konstantem Gewicht von zwei Freiheitsgraden pro Stufe. Um eine festgelegte bedingte Power wie in den vier Strategien zur Wahl von $\beta_{g,k}$ auf den einzelnen Stufen zu erzielen, ergeben sich vor allem auf den ersten Stufen hohe Stichprobenumfänge. Dies beruht auf den sehr geringen bedingten Niveaus, die von den Freiheitsgraden auf den einzelnen Stufen abhängen (vgl. 4.2). Um z.B. eine Power von 36.90% auf der ersten Stufe bei $\nu_1 = 2$ Freiheitsgraden zu erzielen, wird mit 212 ein Stichprobenumfang benötigt, der höher ist als der des einstufigen Designs. Nur unter Strategie IV ist eine Reduktion des Stichprobenumfangs unter $\theta \geq 0.5$ zu beobachten, die jedoch kleiner ausfällt als bei proportionaler Wahl der Gewichte.

Tabelle 5.3: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$*, ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien für Wahl von* $\beta_{G,k}$ *und* $\beta_{g,k}$ *in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für konstanten Wert* $\theta_0 = 0.5$ *auf allen Stufen, Erwartungswert und Standardabweichung (kursiv) für vier Effektschätzer nach finaler Stufe (*$\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$		$\hat{\theta}_k^{\sqrt{n}}$		$\hat{\theta}_k^{\mathrm{int}}$		$\hat{\theta}_k^{\mathrm{RCI}}$	
0.0	A	I	0.0250	720	748	4.94	0.024	(*0.106*)	0.015	(*0.104*)	0.013	(*0.117*)	0.019	(*0.109*)
		II	0.0230	681	710	4.97	0.022	(*0.106*)	0.013	(*0.103*)	0.012	(*0.112*)	0.017	(*0.107*)
		III	0.0248	699	732	4.95	0.024	(*0.110*)	0.015	(*0.106*)	0.014	(*0.118*)	0.019	(*0.112*)
		IV	0.0260	589	614	4.97	0.024	(*0.117*)	0.015	(*0.113*)	0.014	(*0.118*)	0.019	(*0.115*)
	B	I	0.0259	906	944	4.95	0.018	(*0.097*)	0.012	(*0.096*)	0.012	(*0.109*)	0.015	(*0.100*)
		II	0.0247	793	832	4.97	0.020	(*0.102*)	0.014	(*0.102*)	0.014	(*0.109*)	0.017	(*0.103*)
		III	0.0236	871	908	4.97	0.018	(*0.096*)	0.011	(*0.094*)	0.011	(*0.106*)	0.015	(*0.099*)
		IV	0.0270	718	752	4.97	0.020	(*0.108*)	0.013	(*0.106*)	0.013	(*0.111*)	0.016	(*0.107*)
0.3	A	I	0.6197	246	206	3.60	0.366	(*0.146*)	0.365	(*0.170*)	0.338	(*0.164*)	0.349	(*0.155*)
		II	0.6385	248	218	4.04	0.371	(*0.160*)	0.363	(*0.182*)	0.343	(*0.175*)	0.355	(*0.166*)
		III	0.6235	247	214	3.83	0.371	(*0.159*)	0.363	(*0.179*)	0.342	(*0.175*)	0.354	(*0.166*)
		IV	0.6301	240	218	4.30	0.373	(*0.182*)	0.351	(*0.185*)	0.344	(*0.185*)	0.357	(*0.182*)
	B	I	0.7894	301	248	3.72	0.366	(*0.140*)	0.362	(*0.161*)	0.336	(*0.155*)	0.349	(*0.147*)
		II	0.7665	288	250	4.10	0.373	(*0.157*)	0.362	(*0.178*)	0.343	(*0.170*)	0.356	(*0.162*)
		III	0.7889	294	250	3.90	0.373	(*0.153*)	0.364	(*0.171*)	0.343	(*0.167*)	0.356	(*0.159*)
		IV	0.7672	281	252	4.32	0.378	(*0.183*)	0.354	(*0.184*)	0.347	(*0.184*)	0.361	(*0.182*)

Fortsetzung nächste Seite

Fortsetzung Tabelle 5.3

θ	Strategie		$\hat{P}_\theta(\not{H}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$		$\hat{\theta}_k^{\sqrt{n}}$		$\hat{\theta}_k^{\text{int}}$		$\hat{\theta}_k^{\text{RCI}}$	
0.5	A	I	0.9283	136	106	2.17	0.540	(*0.156*)	0.554	(*0.173*)	0.532	(*0.176*)	0.535	(*0.166*)
		II	0.9365	131	112	2.82	0.557	(*0.170*)	0.566	(*0.199*)	0.543	(*0.199*)	0.548	(*0.183*)
		III	0.9313	131	112	2.54	0.558	(*0.170*)	0.566	(*0.193*)	0.545	(*0.196*)	0.550	(*0.182*)
		IV	0.9267	131	116	3.47	0.582	(*0.221*)	0.559	(*0.228*)	0.558	(*0.234*)	0.569	(*0.224*)
	B	I	0.9860	144	110	2.20	0.549	(*0.151*)	0.560	(*0.166*)	0.535	(*0.172*)	0.540	(*0.162*)
		II	0.9785	139	116	2.85	0.564	(*0.165*)	0.572	(*0.193*)	0.547	(*0.194*)	0.554	(*0.178*)
		III	0.9816	140	116	2.56	0.562	(*0.162*)	0.570	(*0.185*)	0.547	(*0.189*)	0.553	(*0.175*)
		IV	0.9796	143	124	3.50	0.583	(*0.214*)	0.559	(*0.218*)	0.555	(*0.223*)	0.569	(*0.215*)
0.7	A	I	0.9911	108	102	1.34	0.710	(*0.176*)	0.721	(*0.176*)	0.712	(*0.182*)	0.711	(*0.179*)
		II	0.9871	93	84	1.82	0.733	(*0.185*)	0.748	(*0.198*)	0.734	(*0.205*)	0.733	(*0.194*)
		III	0.9871	95	82	1.67	0.731	(*0.186*)	0.744	(*0.196*)	0.732	(*0.203*)	0.731	(*0.193*)
		IV	0.9820	87	84	2.86	0.789	(*0.261*)	0.763	(*0.267*)	0.767	(*0.280*)	0.777	(*0.264*)
	B	I	0.9981	109	102	1.35	0.716	(*0.175*)	0.727	(*0.173*)	0.716	(*0.181*)	0.715	(*0.178*)
		II	0.9931	95	84	1.84	0.733	(*0.182*)	0.749	(*0.195*)	0.732	(*0.203*)	0.732	(*0.191*)
		III	0.9958	98	84	1.68	0.731	(*0.184*)	0.744	(*0.193*)	0.731	(*0.200*)	0.730	(*0.191*)
		IV	0.9947	90	86	2.89	0.787	(*0.262*)	0.761	(*0.267*)	0.763	(*0.280*)	0.775	(*0.264*)

5.2.2 Adaption an Parameterschätzung bei fester maximaler Stufenzahl

Im folgenden Abschnitt werden die vier Parameterschätzer zur Anpassung des Stichprobenumfangs anstelle des festen Werts θ_0 verwendet. Um sehr hohe Stichprobenumfänge für nicht klinisch relevante Parameterwerte zu verhindern, erfolgt die Berechnung des Umfangs für das Maximum aus dem Parameterschätzer und einer unteren Schranke θ_{u}.

Ohne Rekalkulation der Power

Die untere Schranke, für die eine Anpassung des Stichprobenumfangs erfolgen soll, wird im Folgenden auf $\theta_{\mathrm{u}} = 0.3$ gesetzt. Der Parameterwert für die Planung der ersten Stufe beträgt wie in den Simulationen des vorangegangenen Abschnitts $\theta_0 = 0.5$. Für jeden der vier beschriebenen Schätzer werden für die Nullhypothese und drei Werte aus der Alternative, 0.3, 0.5 und 0.7, 10000 Studien simuliert. Der Stichprobenumfang und das Gewicht auf den einzelnen Stufen werden jeweils adaptiv gemäß einer der acht Strategien bestimmt, die sich aus der Kombination der unterschiedlichen Vergabe der bedingten Power auf den einzelnen Stufen und der Wahl der bedingten verbleibenden Power für den weiteren Verlauf der Studie insgesamt ergeben, siehe Abschnitt 5.1. Die Tabellen B.5 bis B.8 beinhalten die erzielte Power, den mittleren und medianen benötigten Stichprobenumfang, die mittlere Anzahl durchgeführter Stufen sowie Erwartungswert und Standardabweichung des gewählten Effektschätzers der jeweils letzten Stufe.

Die Verwendung einer Schätzung anstelle eines festen Werts führt insbesondere bei $\theta = 0.3$ zu einer Erhöhung der Power unter allen Strategien (vgl. Tabelle 5.3). Bei Verwendung der theoretischen verbleibenden Power (Strategie A) bleibt die erreichte Power für $\theta = 0.3$ mit 87% bis 89% unterhalb von 90%. Der Stichprobenumfang auf der ersten Stufe, der für den Planwert $\theta_0 = 0.5$ bestimmt wird, führt unter kleineren wahren Werten zu einer niedrigeren Power auf der ersten Stufe als $1-\beta_{g,1}$. Auch bei Schätzungen nahe des wahren Werts für die Stichprobenbestimmung auf den folgenden Stufen erfolgt kein Ausgleich für die zu geringe Power auf der ersten Stufe, so dass unter Strategie A die geplante Power nicht erreicht wird. Daher fällt unter Strategie I mit der größten geplanten Power auf der ersten Stufe die Power jeweils um etwa 1.5% geringer aus als unter den Strategien II bis IV. Unter Strategie B ergeben sich durch die bedingte Power von 90% bei Vergabe der verbleibenden Freiheitsgrade für alle vier Strategien bei der Wahl von $\beta_{g,k}$ und bei jedem Schätzer simulierte Werte der Power von über 95%. Bei $\theta = 0.3$ führt dies zu mittleren und medianen Stichprobenumfängen von etwa 380 und 340, die damit etwa um

30 bis 40 größer sind als unter Strategie A. Für $\theta = 0.5$ ergeben sich unter Strategie A etwa um 4.5% größere Werte der Power als bei einem festen Wert zur Bestimmung des Stichprobenumfangs. Der mittlere Stichprobenumfang ist bei den Strategien I - IV unter $\theta = 0.5$ um etwa 30 Patienten höher, während der mediane Stichprobenumfang im Vergleich zur Anpassung an einen festen Wert vor allem unter der Powervergabe nach O'Brien und Fleming (1979) erhöht ist. Der Unterschied in der Anzahl durchgeführter Stufen durch die langsamere Gewichtsvergabe unter Strategie B, der sich bei Anpassung des Umfangs an einen festen Wert unter $\theta = 0.3$ zeigt (siehe Tabelle 5.3), fällt bei der Anpassung von Gewicht und Stichprobenumfang an einen Schätzwert geringer aus.

Die vier Schätzer führen hinsichtlich der benötigten Fallzahlen, der durchgeführten Stufen sowie der erzielten Power zu ähnlichen Ergebnissen. Die Schätzung aus dem Schnitt der Konfidenzintervallfunktionen $\hat{\theta}_k^{\text{int}}$ besitzt bei Betrachtung der Schätzung nach der finalen Stufe meist die geringste Verzerrung, führt für die langsame Powervergabe nach O'Brien und Fleming (1979) jedoch unter $\theta = 0.5$ zum höchsten mittleren Stichprobenumfang. Der gepoolte Schätzer weist bei jeder Strategie und jedem betrachteten Wert aus der Alternative die niedrigste Varianz für die Parameterschätzung auf der finalen Stufe auf. Die größten Verzerrungen der finalen Schätzer treten wie in Abschnitt 5.2.1 bei der Powervergabe nach O'Brien und Fleming (1979) auf.

Im Vergleich zur Simulation von Hartung und Knapp (2003), beschrieben im Abschnitt 3.4, in der für die Stichprobenumfänge m_k und M_k obere Schranken festgelegt werden, steigen in der vorliegenden Simulation unter $\theta = 0.5$ die mittleren Stichprobenumfänge um 25-30 Patienten an und liegen damit näher am Umfang eines einstufigen Designs mit $\alpha = 0.025$, $\beta = 0.1$ und $\theta_0 = 0.5$. Die hier betrachteten Strategien I - III ohne Beschränkung des Stichprobenumfangs, jedoch mit unterer Grenze $\theta_u = 0.3$, führen zu ähnlichen oder teilweise geringeren medianen Umfängen. Für die Vergabe nach O'Brien und Fleming (1979) liegt der mediane Stichprobenumfang über dem des einstufigen Designs. Die erzielte Power liegt jedoch für alle Strategien unter $\theta = 0.5$ mehr als 6% über der erzielten Power aus der ursprünglichen Simulationsstudie. Die mittlere Anzahl Stufen fällt in der ursprünglichen Simulationsstudie höher aus, da dort keine Beschränkung auf fünf Stufen erfolgte. Eine Anhebung der unteren Schranke für den Parameterwert, an den die Anpassung erfolgt, würde zu einer Reduktion des mittleren Stichprobenumfangs sowie des Medians führen. Dies zeigen die Ergebnisse bei Konstanthaltung des Parameters zur Planung der Stichprobenumfänge in Tabelle 5.3, in der bei leicht erhöhter Power die medianen und mittleren Stichprobenumfänge unter $\theta = 0.5$ im Bereich der Simulationsstudie mit oberen Schranken für die gewichtsbestimmenden Stichprobenumfänge liegen. Bei konstantem Parameter von 0.5 bei der Anpassung liegt die Power beim wahren Wert

von 0.3 je nach Wahl von $1 - \beta_{G,k}$ 5 - 15% über der Power aus der Simulationsstudie von Hartung und Knapp (2003). Eine Power in der Nähe von 90% für einen wahren Wert von $\theta = 0.3$ wird jedoch erst für Strategien mit einer Anpassung des Stichprobenumfangs an eine Schätzung erreicht.

Die Ergebnisse der vorliegenden Simulationen lassen sich hinsichtlich der erzielten Power wie folgt zusammenfassen. In allen Strategien wird die vorgegebene globale Power von 90% überschritten, wenn der wahre Parameter dem Planwert auf der ersten Stufe $\theta_0 = 0.5$ entspricht oder ihn übersteigt. Für den kleineren wahren Wert $\theta = 0.3$ führt die Verwendung des Planwerts in den weiteren Stufen zu einer geringeren Power als 90%, die durch die Verwendung des Effektschätzers erhöht wird. Unter Strategie B führt die bedingte Power von 90% auf der fünften Stufe dazu, dass auch unter $\theta = 0.3$ mindestens eine Power von 90% insgesamt erzielt wird, während sie unter Strategie A zwischen 87% und 89% liegt, da kein Ausgleich für die zu geringe Power auf der ersten Stufe erfolgt.

Um unter Strategie A die zu geringe Power auf der ersten Stufe auszugleichen, bietet sich eine Powerrekalkulation an (siehe Seite 73, Punkt 3). Dies führt zu einer Erhöhung der verbleibenden bedingten Power, sofern der Schätzwert der ersten Stufe unter θ_0 liegt. Entsprechend kann für größere Werte als θ_0 eine Verringerung der verbleibenden Power erfolgen, um die Fallzahl zu reduzieren.

Mit Rekalkulation der Power

Aufgrund des ähnlichen Verhaltens der finalen Schätzer werden im Folgenden die Simulationen auf die Verwendung des gepoolten Schätzers, der in den meisten Szenarien die kleinste Varianz aufweist, und des Schätzers aus dem Schnitt der Konstruktionsfunktionen des Konfidenzintervalls beschränkt, der meist die kleinste Verzerrung besitzt. Liegt nach $(k-1)$ Stufen eine Schätzung $\hat{\theta}_{k-1}$ für den Parameter θ vor, kann die bis zur $(k-1)$-ten Stufe vergebene Power für diesen Parameterwert geschätzt werden. Nach Durchführung der ersten Stufe lässt sich die vergebene Power gemäß (4.9) durch Ersetzen von θ/σ durch $\hat{\theta}_1/\sigma$ berechnen. Analog lässt sich für alle bereits durchgeführten Studienabschnitte mittels Stichprobenumfang und bedingtem Niveau α'_k die unter dem Schätzwert als angenommenem wahren Wert vergebene bedingte Power bzw. der bedingte Fehler zweiter Art bestimmen. Der bis zur $(k-1)$-ten Stufe vorliegende Fehler zweiter Art ergibt sich aus dem Produkt der bedingten Fehlerwahrscheinlichkeiten zweiter Art auf den $(k-1)$ Stufen gemäß (5.1) und (5.2). Daraus wird die verbleibende Power geschätzt, welche anstelle der geplanten Power für $1 - \beta_{G,k}$ verwendet wird.

Nach Schätzung der verbleibenden Power ergeben sich mehrere Möglichkeiten, diese zu

vergeben:

Variante a: Die bedingte Power $1 - \beta_{g,k}$ für die folgende Stufe kann zum einen gemäß (5.6) angepasst werden, dass nach Durchführung der k-ten Stufe für den geschätzten Parameterwert die theoretische geplante Power $1 - \beta(k)$ erreicht wird.

Variante b: Zum anderen kann gemäß (5.7) der geplante Powerzuwachs π_k in der folgenden Stufe umgesetzt werden.

Variante c: Die Verwendung der ursprünglich geplanten bedingten Power $1 - \beta_i'$ bietet eine weitere Möglichkeit.

Die simulierten Studien enden spätestens nach der fünften Stufe, in der für die Power $1 - \beta_{G,5} = 1 - \beta_{g,5} = 1 - \beta/\beta(4)_{\hat{\theta}_4}$ der Stichprobenumfang bei Vergabe der verbleibenden Freiheitsgrade $\nu_G - \nu_\Sigma(4)$ für $\hat{\theta}_4$ bestimmt wird. Die Rekalkulation der Power erfolgt bei negativen oder kleineren Werten aus der Alternative zur unteren Grenze θ_u.

In den Tabellen B.9 und B.10 sind die Simulationsergebnisse für die zwölf Strategien enthalten, die sich aus den vier Strategien des Poweranstiegs (siehe S. 79) und den drei Varianten der Umsetzung ergeben. Dabei basiert die Rekalkulation der Power in der ersten Ergebnistabelle auf dem gepoolten Schätzer, in der zweiten auf dem Schätzer $\hat{\theta}_k^{\mathrm{int}}$, jeweils mit $\theta_\mathrm{u} = 0.3$ als unterer Grenze.

Wie in den vorangehenden Simulationen besitzt der gepoolte finale Schätzer die kleinere Varianz, während der Schätzer $\hat{\theta}_k^{\mathrm{int}}$ die geringere Verzerrung aufweist. Der Erwartungswert des finalen Schätzers liegt für beide Schätzer in allen Szenarien über dem wahren Wert. Dabei beträgt die Verzerrung unter $\theta = 0.3$ für den gepoolten Schätzer zum Teil über 20% des wahren Werts, für $\hat{\theta}_k^{\mathrm{int}}$ über 10% des wahren Werts. Die prozentualen Abweichungen, die von der Anzahl durchgeführter Stufen abhängen, sind für die beiden größeren Parameterwerte für die Strategien I - III deutlich geringer, während für die langsame Powervergabe mit erhöhter Anzahl an Stufen auch größere Verzerrungen auftreten. Der Unterschied in der Anzahl durchgeführter Stufen ist bzgl. der Wahl des Schätzers gering. Die Power bei Verwendung des Schätzers $\hat{\theta}_k^{\mathrm{int}}$ ist meist etwas höher als bei Verwendung des gepoolten Schätzers, wobei die größten Abweichungen von bis zu 2.45% unter der Powervergabe nach O'Brien und Fleming (1979) auftreten. Mit der erhöhten Power ergibt sich ein bis zu 8 Patienten größerer mittlerer und bis zu 18 Patienten größerer medianer Stichprobenumfang. Die medianen Stichprobenumfänge unter den Strategien I - III unterscheiden sich hinsichtlich der Wahl des Schätzers deutlich weniger als unter Strategie IV.

Vergleicht man die Ergebnisse der zwei Schätzer mit und ohne Powerrekalkulation (Tabellen B.5 und B.7), zeigen sich insbesondere unter Variante a bei Verwendung der Powerrekalkulation erhöhte mittlere Stufenzahlen um bis zu 0.94 Stufen unter Strategie IV. In Variante c, der Verwendung der unveränderten bedingten Power, ergeben sich ähnliche mittlere Stufenzahlen wie bei den Simulationen ohne Powerrekalkulation. Die Power unter $\theta = 0.3$ hat sich im Vergleich zu den Simulationen ohne Powerrekalkulation unter fast allen Strategien erhöht. Insbesondere unter Strategie I, der Verwendung der konstanten bedingten Power, beträgt die Erhöhung zwischen 2.7% und 4.3%. Bis auf Variante a IV führen die Strategien mit Powerrekalkulation zu einer Power um 90% unter $\theta = 0.3$. Die Erhöhung der Power ist im Vergleich zu den Strategien ohne Powerrekalkulation mit einer Erhöhung des mittleren Stichprobenumfangs zwischen 8 und 22 Patienten verbunden. Die mittleren Stichprobenumfänge von 343 bis 370 sowie die medianen Umfänge zwischen 310 und 342 liegen damit deutlich unter dem benötigten Stichprobenumfang von 468 Patienten einer klassischen Studie mit entsprechendem Niveau und einer Power von 90% für $\theta = 0.3$. Eine Erklärung für die niedrigere Power unter Strategie IV liefert der im Vergleich zu den anderen Strategien deutlich geringere Stichprobenumfang auf der ersten Stufe, der zu weniger stabilen Schätzungen auf der ersten Stufe führt. Fällt der Schätzer größer als $\theta_u = 0.3$ aus, wird er bei der Rekalkulation und bei der Bestimmung des Stichprobenumfangs verwendet, was zu einer zu geringen Power für den wahren Wert $\theta = 0.3$ führt. Aufgrund der Beschränkung durch $\theta_u = 0.3$ erfolgt kein Ausgleich der zu geringen Power durch kleinere Werte. Eine Senkung der unteren Schranke würde zu einer Erhöhung der Power führen, jedoch auch zu einer Erhöhung des mittleren Stichprobenumfangs für größere Parameterwerte. Die Anpassung des Stichprobenumfangs erfolgt zwar für $\theta = 0.3$ für die Strategien I bis III ebenso nur maximal bis zum wahren Wert, durch die stabilere Schätzung auf der ersten Stufe weicht die erzielte Power im Mittel jedoch weniger von der für den wahren Wert gewünschten ab. In Kombination mit dem bereits beschriebenen möglichen Ereignis (vgl. S. 81), dass das bedingte Niveau die geplante Power übersteigt, führen diese Strategien zur angestrebten globalen Power von 90%.

Bis auf die Strategievariante a IV, in der mittlerer und medianer Stichprobenumfang bei Powerrekalkulation im Vergleich zur Strategie ohne Powerrekalkulation bis zu 20 bzw. 48 Patienten kleiner sind, fallen für $\theta = 0.5$ und 0.7 die Unterschiede in Power und Stichprobenumfang geringer aus als unter $\theta = 0.3$. Die erwünschte Reduktion des Stichprobenumfangs für $\theta = 0.5$ und 0.7 durch die nachträgliche Anpassung der verbleibenden Power tritt nicht ein. Dies liegt zum einen an der Varianz des Schätzers, zum anderen an der auftretenden Überschreitung der bedingten Power durch das bedingte Niveau. Bis auf die erhöhte mittlere Stufenzahl unter Strategie a IV verursacht die Rekalkulation jedoch keine zusätzlichen Kosten im Vergleich zu den Strategien ohne Powerrekalkulation, führt

jedoch zur erwünschten Erhöhung der Power für einen kleineren Parameterwert als den Ausgangswert, der zur Planung des Stichprobenumfangs auf der ersten Stufe herangezogen wird.

Im folgenden Abschnitt wird die Beschränkung der Anzahl durchgeführter Stufen aufgehoben und die Auswirkung unterschiedlicher Strategien auf den mittleren Stichprobenumfang und die mittlere sowie maximale Stufenzahl untersucht.

5.2.3 Adaption an Parameterschätzung bei offener Stufenzahl

Wird die Anzahl der Studienabschnitte nicht mehr wie in den vorangegangen Abschnitten auf fünf beschränkt, ergibt sich insbesondere bei der Nachahmung des Powerverlaufs eines gruppensequentiellen Designs eine beliebige Anzahl an Fortsetzungsmöglichkeiten ab der fünften Stufe. Die Änderung des Vorgehens im Vergleich zu den auf fünf Stufen beschränkten Studien erfolgt nach Abschluss der vierten Stufe, sofern die Teststatistik im Fortsetzungsbereich liegt. Der bedingte verbleibende Fehler erster Art (4.7) wird aufgeteilt auf die fünfte und mögliche weitere Stufen, indem nur ein Teil der verbleibenden Freiheitsgrade in der fünften Stufe vergeben wird. Im Folgenden werden Strategien betrachtet, die aufgrund ihrer Konstruktion keine neuen Überlegungen zur Vergabe der Power vor bzw. nach der fünften Stufe verlangen, sondern das Vorgehen auf den ersten vier Stufen auf den folgenden Stufen fortsetzen. Dies lässt sich zum einen durch die Wahl einer konstanten bedingten Power verwirklichen, zum anderen durch einen konstanten Powerzuwachs beim Vorgehen nach Variante c I bzw. b II im vorangehenden Abschnitt (vgl. S. 90). Sofern nicht vorzeitig eine Ablehnung der Nullhypothese erfolgt, wird eine letzte Stufe mit Vergabe der verbleibenden Freiheitsgrade durchgeführt, wenn sich unter dem minimalen Stichprobenumfang $n_{\min}$ bereits eine größere bedingte Power ergibt als die nach Powerrekalkulation berechnete Power $1 - \beta_{G,k}$. Dies schließt den Spezialfall $1 - \beta_{G,k} < \alpha_k^*$ ein. Die untere Schranke für die Parameterschätzung zur Powerrekalkulation von $\theta = 0.3$ wird beibehalten. Da die bisher gewählte bedingte konstante Power von $1 - \sqrt[5]{\beta} = 36.9\%$ sowie der konstante Powerzuwachs von 18% auf eine fünfstufige Studie abzielen, wird zusätzlich jeweils eine verlangsamte Powervergabe betrachtet. Mit einer konstanten bedingten Power von $1 - \sqrt[9]{\beta} = 22.6\%$ bzw. einem konstanten Powerzuwachs von 10% wird die Vergabe der Power deutlich verzögert und damit die erwartete Anzahl Zwischenauswertungen erhöht.

Die Tabellen B.11 und B.12 beinhalten die Simulationsergebnisse bei unbeschränkter Stufenzahl bei Verwendung des gepoolten Schätzers bzw. des Schätzers aus dem Schnitt der Konstruktionsfunktionen des Konfidenzintervalls. Die beiden Schätzer führen unter der

Alternative nur zu sehr geringen Unterschieden hinsichtlich des mittleren und medianen benötigten Stichprobenumfangs sowie der mittleren, medianen und maximalen Anzahl Stufen. Unter der Nullhypothese treten etwas größere mediane Stichprobenumfänge bei Verwendung des gepoolten Schätzers unter den Strategien des konstanten Powerzuwachses auf als bei Verwendung der konstanten bedingten Power. Die Aufhebung der Beschränkung der Stufenzahl führt unter der Nullhypothese bei der Powervergabe, ausgerichtet auf fünf Stufen, bei mittlerer und medianer Anzahl zu sechs Stufen. Die a priori Planung von neun Stufen liefert sowohl bei konstanter bedingter Power als auch bei konstantem Powerzuwachs im Mittel und Median zehn Studienabschnitte unter der Nullhypothese. Die maximale Anzahl simulierter Stufen liegt bei sieben bzw. bei der langsameren Powervergabe je nach Strategie bei zwölf oder dreizehn Stufen.

Mit steigendem Wert aus der Alternative nähern sich mittlere und mediane Anzahl Studienabschnitte unter der langsamen Powervergabe den Werten aus der schnelleren Powervergabe an. Unter $\theta = 0.7$ fallen die Mediane von einer Stufe bei der konstanten bedingten Power und zwei Stufen bei der Strategie des konstanten Powerzuwachs für die langsame und die schnellere Powervergabe zusammen. Der Unterschied in der mittleren Anzahl durchgeführter Stufen liegt unter $\theta = 0.7$ bei 0.3 bzw. 0.45 Stufen. Die beobachtete maximale Anzahl an Stufen sinkt mit wachsendem wahren Parameter deutlich langsamer als medianer und mittlerer Umfang. Unter $\theta = 0.7$ werden je nach Powervergabe noch sieben bzw. zehn Stufen maximal beobachtet. Unter der langsamen Vergabe der Power erhöht sich der mittlere und mediane Stichprobenumfang unter der Nullhypothese um 1500 Patienten im Vergleich zur fünfstufigen Planung, während der mittlere Umfang unter den betrachteten Werten aus der Alternative ähnlich groß, der mediane zum Teil sogar geringfügig kleiner ist.

Bei beiden Schätzern führt sowohl die verlangsamte Powervergabe sowie die Strategie des konstanten Powerzuwachses zu größeren Verzerrungen beim finalen Schätzer und zu einer größeren Varianz. Dies ergibt sich durch die mit diesen Strategien verbundene höhere Anzahl durchgeführter Stufen.

Im Vergleich zu den Strategien c I und b II aus dem vorangehenden Abschnitt 5.2.2 mit Beschränkung auf maximal fünf Stufen erhöht sich unter der schnellen Powervergabe die mittlere Anzahl an Studienabschnitten unter der Nullhypothese um eine Stufe, unter $\theta = 0.3$ um maximal 0.2 Stufen. Für die größeren Parameterwerte entsprechen sich die mittleren Anzahlen an Stufen. Die schnellere Powervergabe bei offener Stufenzahl führt im Vergleich zu den entsprechenden Strategien mit beschränkter Stufenzahl nur unter der Nullhypothese zu einer deutlichen Erhöhung des Stichprobenumfangs, während unter den betrachteten Werten aus der Alternative die Stichprobenumfänge gleiche Größe besitzen.

Kapitel 6

Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurden mögliche Strategien in adaptiven Designs am Beispiel der Inverse-χ^2-Methode von Hartung und Knapp (2003) untersucht. Im ersten Teil lag der Fokus auf der Darstellung der theoretischen Eigenschaften und der Erweiterung des Verfahrens. Im zweiten Teil wurde der Einfluss der Art der Powervergabe und der Verwendung unterschiedlicher Schätzer auf Stichprobenumfang, Anzahl durchgeführter Stufen und erzielte Power durch umfangreiche Simulationen beschrieben.

Es erfolgte zunächst eine Einordnung der Inverse-χ^2-Methode in die bestehenden adaptiven Designs. Sie kann zum einen als rekursiver Kombinationstest aufgefasst werden, zum anderen bietet sie die Möglichkeit, Studiendesigns gemäß dem Prinzip von Müller und Schäfer (2001) zu konstruieren. Dies lieferte die theoretische Grundlage für die Erweiterung des Designs um Regeln, die optional auf jeder Stufe des Verfahrens die Anpassung des kritischen Werts bei der Einführung einer Schwelle zum Abbruch der Studie wegen Aussichtslosigkeit erlauben. Die Anpassung ist unabhängig von der verwendeten Teststatistik auf den einzelnen Stufen. Weiterhin konnte die Forderung an die Konvergenz der Summe der gewählten Gewichte aufgehoben werden, so dass die Umsetzung des Designs durch das Wegfallen einer entsprechenden Prüfung erheblich erleichtert wird. Gemäß dem Vorgehen in rekursiven Kombinationstests konnte ein P-Wert nach jeder Stufe des Verfahrens konstruiert werden, der nach Ende der Studie genau dann kleiner als das globale Signifikanzniveau ist, wenn die Studie mit einer Ablehnung der Nullhypothese endet.

Aus den aufgezeigten Zusammenhängen zwischen Gewicht und bedingter Fehlerrate erster Art ergibt sich die Empfehlung, die Anzahl an Freiheitsgraden, die der maximalen Summe der Stufengewichte entspricht, nicht beliebig groß zu wählen. Es zeigte sich, dass bei vier oder zehn Freiheitsgraden mit der Wahl eines hohen Gewichts auch ein großer Teil des Niveaus vergeben wird. Bei einer großen Anzahl von Freiheitsgraden ist hingegen keine intuitive Interpretation möglich, da auch die Vergabe hoher Anteile an den Freiheitsgraden

nur zu sehr geringen lokalen Niveaus auf der ersten Stufe führt. Daraus ergeben sich bei vorgegebener Power für eine einzelne Stufe unverhältnismäßig hohe Stichprobenumfänge.

Für die von Hartung und Knapp (2003) vorgeschlagene proportionale Gewichtung der Freiheitsgrade und des Stichprobenumfangs der folgenden Stufe im Verhältnis zum gesamten verbleibenden Studienverlauf wurde eine Berechnungsmethode entwickelt, um eine vorgegebene bedingte Power bei Erhaltung der Proportionalität exakt zu erreichen. Durch die Lösung des dabei entstehenden Fixpunktproblems kann bei Vorgabe des verbleibenden Fehlers erster Art sowie der bedingten Power für den verbleibenden Studienverlauf für jede Power, die maximal der geforderten verbleibenden Power insgesamt entspricht, ein Stichprobenumfang und ein proportionales Gewicht für die nachfolgende Stufe bestimmt werden. Wird die proportionale Gewichtung aufgehoben, kann sich für eine einzelne Stufe ein Stichprobenumfang ergeben, der größer als der Stichprobenumfang für eine höhere Power für den gesamten verbleibenden Studienverlauf ist.

Die Untersuchungen waren beschränkt auf den einseitigen Zwei-Stichproben-Gauß-Test. Prinzipiell gelten die beschriebenen Eigenschaften auch beim Test einer zweiseitigen Hypothese. Die Power auf den einzelnen Stufen sollte jedoch gewährleisten, dass keine widersprüchlichen Richtungen auf einzelnen Stufen insgesamt zur Ablehnung der Nullhypothese führen, um die Studie interpretieren zu können. Das Verfahren ist auf alle stetigen Teststatistiken anwendbar. Bei zusätzlicher Schätzung der Varianz ist ein Anstieg des Stichprobenumfangs zu erwarten, insbesondere wenn auf der ersten Stufe ein geringer Stichprobenumfang aufgewendet wird. Lösungen für das Fixpunktproblem müssen je nach Stichprobenfunktion mit erhöhtem Aufwand iterativ bestimmt werden. Die Anwendung des Designs auf Teststatistiken mit nicht auf [0,1] gleichverteilten P-Werten führt gemäß Bauer (1989a) im Allgemeinen zu konservativen Verfahren. Das vorgegebene Niveau wird nicht überschritten, sofern unabhängige Stichprobeneinheiten auf den verschiedenen Stufen rekrutiert werden und die angewendeten Tests zu jedem vorgegebenen Signifikanzniveau den Fehler erster Art einhalten (Brannath et al., 2002).

Bei der Entwicklung unterschiedlicher Strategien stellte die Aufhebung der Festlegung einer konstanten bedingten Power für alle Stufen wie in vielen in der Literatur vorgeschlagenen Designs einen wichtigen Schritt dar. Dabei wurden bei der Inverse-χ^2-Methode sowohl die Power auf den einzelnen Stufen als auch die Power jeweils für den verbleibenden Studienteil insgesamt nicht mehr zwingend konstant gewählt. In den betrachteten Strategien wurden für die verbleibende Power insgesamt neben der globalen Power entweder die erwartete bedingte Power bei vorgegebener globaler Power oder eine unter dem Parameterschätzer rückblickend berechnete verbleibende Power verwendet.

Aus der Vielzahl möglicher Strategien wurden zum einen die Powerverläufe der zwei be-

kanntesten gruppensequentiellen Verfahren nach Pocock (1977) und O'Brien und Fleming (1979) herangezogen, zum anderen die in der Literatur beschriebene Strategie der konstanten bedingten Power sowie eine Strategie des konstanten Powerzuwachses. Anstelle oberer Schranken für die gewichtsbestimmenden Stichprobenumfänge wie bei Hartung und Knapp (2003) wurde eine untere Schranke für den Parameterschätzer gewählt. Es wurden vier unterschiedliche Schätzer zur Anpassung des Stichprobenumfangs betrachtet, die entweder aus der Gewichtung der Einzelschätzer oder mit Hilfe der Konstruktionsfunktionen für wiederholte Konfidenzintervalle bestimmt wurden.

Die betrachteten vier Schätzer zur Anpassung des Stichprobenumfangs hatten im Vergleich zur gewählten Strategie wenig Einfluss auf Stichprobenumfang und Gewicht. Der einfache gepoolte Schätzer wies häufig die geringste Varianz auf, während der Schnittpunkt der Konstruktionsfunktionen des Konfidenzintervalls meist die geringste Verzerrung bei erhöhter Varianz aufwies. Die Verzerrung der finalen Schätzung stieg mit der mittleren Anzahl der durchgeführten Stufen.

In allen betrachteten Strategien wurde die vorgegebene globale Power von 90% überschritten, wenn der wahre Parameter dem Planwert auf der ersten Stufe entsprach oder ihn überstieg. Für den betrachteten kleineren wahren Wert führte die wiederholte Verwendung des Planwerts zur Bestimmung des Stichprobenumfangs in den weiteren Stufen zu einer geringeren Power als 90%. Durch die Verwendung des Effektschätzers wurde die Power für alle betrachteten wahren Werte aus der Alternative deutlich erhöht. Bei Verwendung der globalen Power zur Planung des Stichprobenumfangs auf der letzten Stufe wurde auch für den kleineren Wert eine Power von mindestens 90% insgesamt erzielt. Bei Strategien mit der Verwendung einer theoretischen verbleibenden Power, die kleiner als die globale Power war, wurde die geplante globale Power nicht ganz erreicht.

Durch Verwendung einer geschätzten verbleibenden Power für den verbleibenden Studienteil konnte schließlich die globale Power für den kleineren wahren Parameterwert erreicht werden. Der Stichprobenumfang für die höheren Parameterwerte konnte durch die Powerrekalkulation nicht reduziert werden. Die Rekalkulation verursacht keine zusätzlichen Kosten hinsichtlich Stichprobenumfang und Anzahl durchgeführter Stufen im Vergleich zu den Strategien ohne Powerrekalkulation, führt jedoch zur erwünschten Erhöhung der Power für einen kleineren Parameterwert als den Ausgangswert. In anderen adaptiven Verfahren, die keine Steuerung des Studienverlaufs über Gewichte zulassen, bietet sie ebenfalls eine Möglichkeit einen vorgegebenen Powerverlauf umzusetzen und dabei an veränderte Planungsparameter anzupassen. Die Auswirkung auf die globale Power und den benötigten mittleren Stichprobenumfang ist in weiteren Simulationen zu untersuchen.

Die Komplexität des Verfahrens und das Fehlen fest definierter Optimaltätskriterien er-

schweren die Entwicklung von universell einsetzbaren Strategien. Die hier aufgezeigten Strategien zeigen einen Ausschnitt der sich bietenden Möglichkeiten bei der Umsetzung adaptiver Verfahren. Je nach Intention der Studienplaner kann bei pessimistischer Sicht auf den anfänglich gewählten Startparameter eine Strategie mit Rekalkulation der Power verwendet werden, um auch kleinere wahre Werte zu entdecken. Dabei muss auch bei größeren wahren Parameterwerten ein mittlerer Stichprobenumfang nahe des einstufigen Designs akzeptiert werden. Bei optimistischer Sicht bietet sich die Verwendung eines konstanten Parameterwerts zur Anpassung des Stichprobenumfangs an, um mit möglichst wenigen Studienteilnehmern die Wirksamkeit eines Präparats nachzuweisen. Zur Senkung der hohen Stichprobenumfänge unter der Nullhypothese, ein Problem in allen adaptiven Verfahren ohne Beschränkung des Stichprobenumfangs, ist die Aufnahme eines Stopps wegen Aussichtslosigkeit nach einer ausreichend stabilen Schätzung zu empfehlen.

In einer Diskussion zum Nutzen adaptiver Designs (Burman und Sonesson, 2006) sagt Bauer (2006), der große Vorteil der adaptiven Designs liege in der Möglichkeit auf externe Einflüsse zu reagieren. Dies ist in Simulationen nicht erfassbar, aber in jede der betrachteten Strategien integrierbar.

Anhang A

Eigenschaften und Erweiterungen

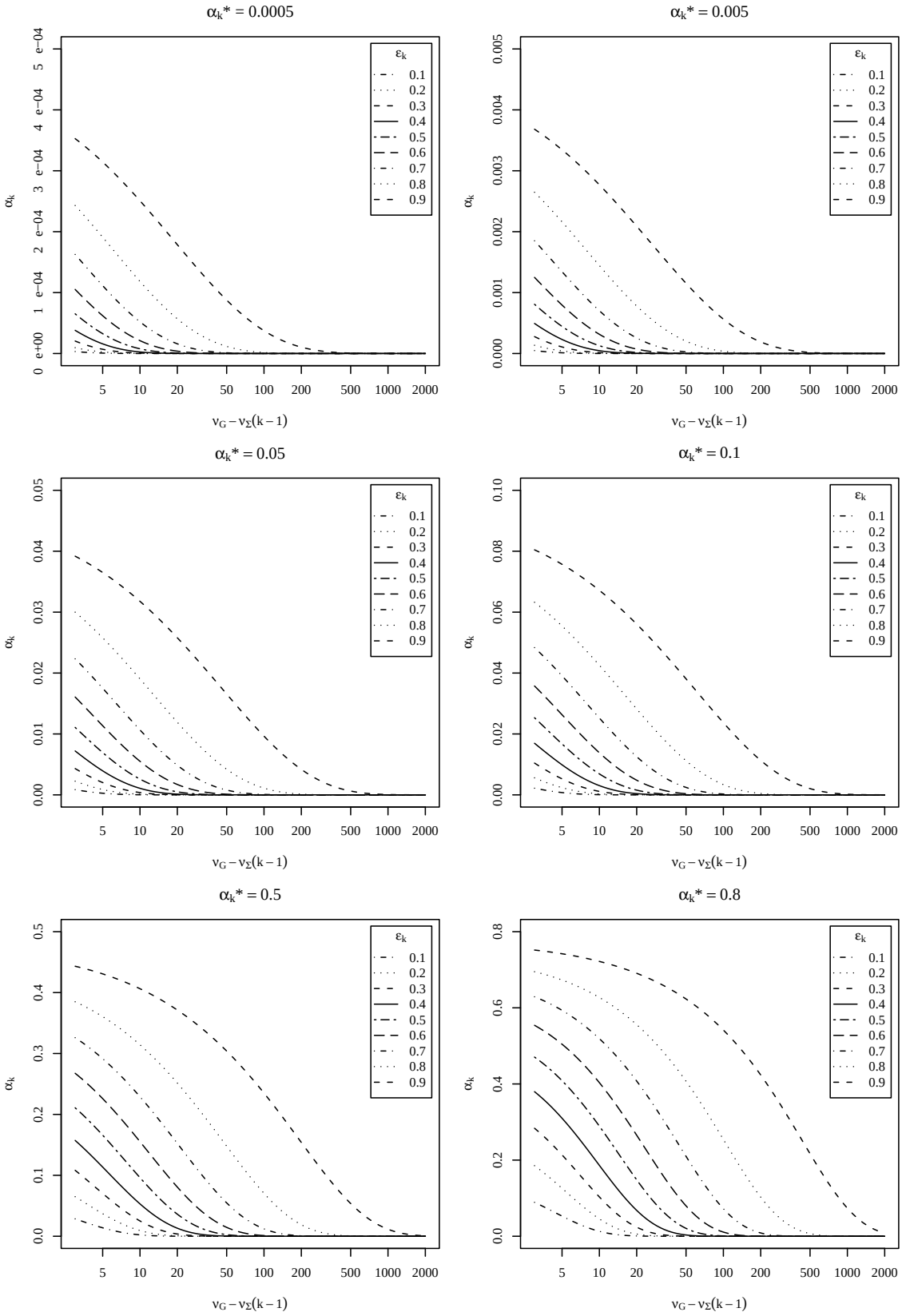

Abbildung A.1: *Lokales bedingtes Niveau gemäß Inverse-χ^2-Methode für Stufe k in Abhängigkeit der Anzahl Freiheitsgrade $\nu_G - \nu_\Sigma(k-1)$, des vergebenen Anteils ϵ_k auf der k-ten Stufe und des verbleibenden Niveaus α_k^**

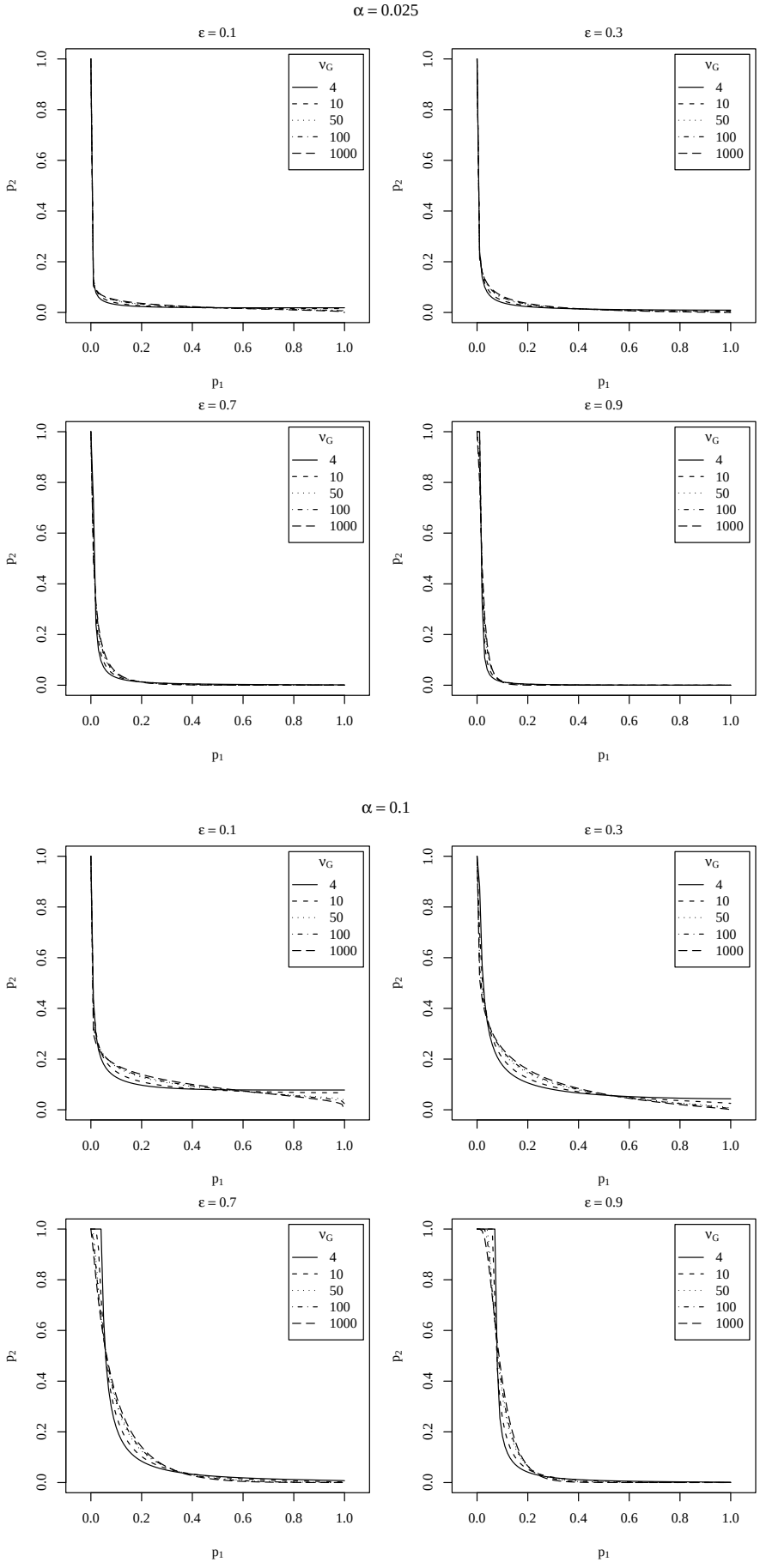

Abbildung A.2: *Bedingte Fehlerfunktion in Abhängigkeit der Anzahl Freiheitsgrade ν_G bei unterschiedlicher Aufteilung auf zwei Stufen, $\alpha = 0.025, 0.1$*

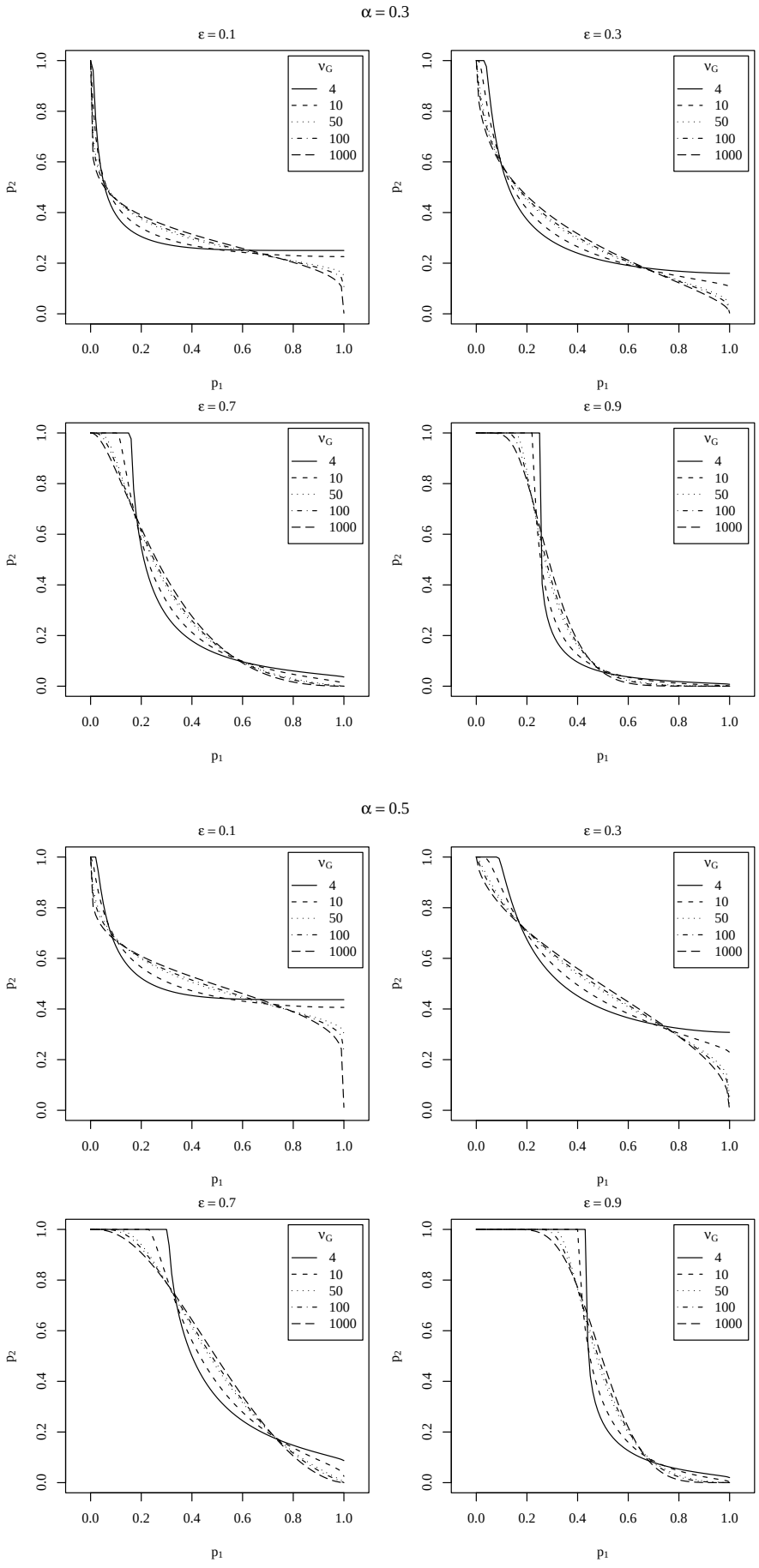

Abbildung A.3: *Bedingte Fehlerfunktion in Abhängigkeit der Anzahl Freiheitsgrade ν_G bei unterschiedlicher Aufteilung auf zwei Stufen, $\alpha = 0.3, 0.5$*

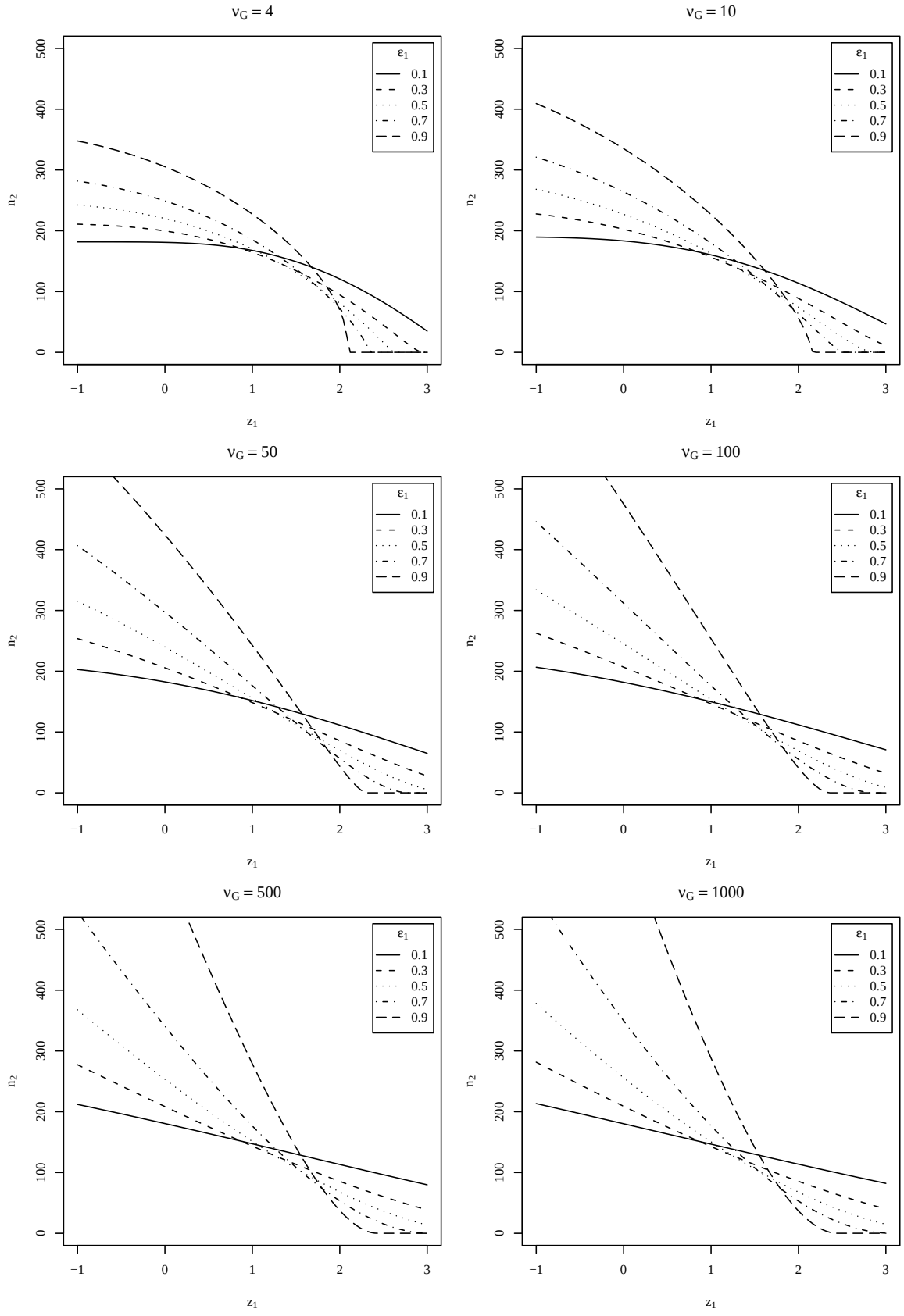

Abbildung A.4: *Benötigter Stichprobenumfang für eine bedingte Power von 0.9 bei $\theta/\sigma = 0.5$ auf der zweiten Stufe in Abhängigkeit der Teststatistik der ersten Stufe z_1 bei Vergabe der verbleibenden Freiheitsgrade $(1 - \epsilon_1) \cdot \nu_G$*

Tabelle A.1: *Erwartete Stichprobenumfänge und Power bei einer bedingten Power* $\pi_2^* = 0.9$ *auf der zweiten Stufe für* $\theta_0 = 0.5$ *bzw. Schätzung* $\hat{\theta}_1 = \max(2\sigma/\sqrt{n_k} \cdot z_1, 0.001)$, *maximaler Umfang* $n_{2,\max} = 10 \cdot M_1$, *wahrer Wert* $\theta = 0.5$, $\sigma^2 = 1$, *10000 Simulationsläufe*

		$\theta_0 = 0.5$			$\hat{\theta}_1$		
ν_G	ϵ_1	$\mathrm{E}(n_2)$	$n_1 + \mathrm{E}(n_2)$	π_{ges}	$\mathrm{E}(n_2)$	$n_1 + \mathrm{E}(n_2)$	π_{ges}
4	0.1	151.3	168.1	0.901	559.41	576.23	0.807
	0.3	103.7	154.2	0.910	372.11	422.54	0.924
	0.5	64.4	148.4	0.935	245.87	329.93	0.972
	0.7	37.0	154.7	0.965	158.54	276.22	0.994
	0.9	23.7	175.0	0.983	110.37	261.67	0.999
10	0.1	149.2	166.1	0.901	578.64	595.45	0.806
	0.3	101.9	152.3	0.907	364.67	415.11	0.913
	0.5	63.1	147.2	0.926	243.10	327.16	0.966
	0.7	37.6	155.3	0.957	163.42	281.11	0.988
	0.9	22.3	173.6	0.981	105.77	257.08	0.999
50	0.1	146.6	163.4	0.905	569.11	585.92	0.805
	0.3	101.9	152.4	0.904	360.23	410.66	0.904
	0.5	62.7	146.7	0.911	231.37	315.43	0.942
	0.7	36.4	154.1	0.947	149.34	267.02	0.978
	0.9	21.9	173.2	0.976	99.57	250.88	0.996
100	0.1	146.3	163.1	0.898	564.31	581.12	0.803
	0.3	101.7	152.1	0.902	356.57	407.01	0.897
	0.5	64.5	148.6	0.910	243.01	327.07	0.941
	0.7	38.2	155.9	0.945	156.15	273.83	0.978
	0.9	22.4	173.7	0.976	100.09	251.39	0.993
500	0.1	145.9	162.7	0.901	556.13	572.94	0.798
	0.3	102.6	153.0	0.896	359.97	410.41	0.889
	0.5	63.6	147.6	0.909	224.91	308.97	0.930
	0.7	38.6	156.3	0.943	153.43	271.11	0.972
	0.9	24.9	176.2	0.974	102.04	253.34	0.992
1000	0.1	146.3	163.1	0.902	563.98	580.79	0.800
	0.3	103.0	153.4	0.899	351.97	402.40	0.882
	0.5	64.7	148.8	0.911	234.05	318.11	0.936
	0.7	38.9	156.5	0.935	156.51	274.19	0.967
	0.9	26.5	177.8	0.973	105.11	256.42	0.991

Tabelle A.2: *Erwartete Stichprobenumfänge und Power bei einer bedingten Power* $\pi_2^* = 0.9$ *auf der zweiten Stufe für* $\theta_0 = 0.5$ *bzw. Schätzung* $\hat{\theta}_1 = \max(2\sigma/\sqrt{n_k} \cdot z_1, 0.001)$, *maximaler Umfang* $n_{2,\max} = 10 \cdot M_1$, *wahrer Wert* $\theta = 0.3$, $\sigma^2 = 1$, *10000 Simulationsläufe*

		$\theta_0 = 0.5$			$\hat{\theta}_1$		
ν_G	ϵ_1	$\mathrm{E}(n_2)$	$n_1 + \mathrm{E}(n_2)$	π_{ges}	$\mathrm{E}(n_2)$	$n_1 + \mathrm{E}(n_2)$	π_{ges}
4	0.1	163.1	179.9	0.510	817.1	833.9	0.729
	0.3	145.3	195.7	0.546	762.4	812.8	0.836
	0.5	128.5	212.6	0.591	701.9	786.0	0.902
	0.7	113.6	231.2	0.650	638.8	756.5	0.948
	0.9	114.1	265.4	0.684	624.0	775.3	0.978
10	0.1	162.0	178.9	0.502	829.4	846.2	0.732
	0.3	143.4	193.8	0.541	755.2	805.6	0.833
	0.5	126.0	210.0	0.587	691.4	775.5	0.899
	0.7	114.9	232.5	0.633	646.8	764.5	0.942
	0.9	111.7	263.0	0.688	598.3	749.7	0.976
50	0.1	160.2	177.0	0.511	823.0	839.8	0.735
	0.3	142.8	193.2	0.543	747.4	797.9	0.840
	0.5	125.2	209.2	0.590	669.0	753.0	0.886
	0.7	114.9	232.6	0.641	620.6	738.2	0.935
	0.9	121.9	273.2	0.670	599.9	751.2	0.967
100	0.1	159.9	176.7	0.512	821.6	838.4	0.738
	0.3	142.3	192.7	0.548	739.6	790.1	0.830
	0.5	127.6	211.6	0.579	681.4	765.5	0.888
	0.7	118.4	236.1	0.622	627.6	745.3	0.929
	0.9	126.8	278.1	0.670	590.5	741.8	0.964
500	0.1	159.5	176.3	0.510	805.8	822.6	0.733
	0.3	142.8	193.2	0.551	737.1	787.6	0.831
	0.5	127.2	211.2	0.575	665.8	749.8	0.882
	0.7	121.1	238.8	0.626	618.1	735.8	0.928
	0.9	145.2	296.5	0.651	596.0	747.3	0.950
1000	0.1	159.8	176.6	0.518	816.7	833.5	0.732
	0.3	143.2	193.6	0.537	737.4	787.9	0.827
	0.5	128.6	212.7	0.585	668.1	752.1	0.886
	0.7	123.5	241.1	0.618	620.4	738.1	0.927
	0.9	153.8	305.1	0.649	602.6	753.9	0.945

Anhang B

Strategien

Tabelle B.1: *Prozentuale Anteile ϵ_m und $\epsilon_\nu = \epsilon_m^x$ der ersten Stufe am Stichprobenumfang M_1 bzw. an den Gesamtfreiheitsgraden ν_G bei exakter Bestimmung des Stichprobenumfangs für lokale Power $1 - \beta_{g,1}$ bei globaler Power $1 - \beta = 0.9$*

		$\nu_G = 10$		$\nu_G = 100$		$\nu_G = 1000$	
x	$1 - \beta_{g,1}$	$\epsilon_m[\%]$	$\epsilon_\nu[\%]$	$\epsilon_m[\%]$	$\epsilon_\nu[\%]$	$\epsilon_m[\%]$	$\epsilon_\nu[\%]$
0.25	0.1	18.99	66.01	41.14	80.09	68.68	91.03
	0.2	27.68	72.53	48.79	83.58	73.42	92.57
	0.3	35.15	77.00	54.94	86.09	77.00	93.68
	0.4	42.40	80.69	60.55	88.21	80.18	94.63
	0.5	49.97	84.08	66.18	90.20	83.22	95.51
	0.6	58.28	87.37	72.12	92.16	86.36	96.40
	0.7	68.04	90.82	78.87	94.24	89.80	97.35
	0.8	80.55	94.74	87.34	96.67	93.99	98.46
0.5	0.1	28.30	53.20	56.54	75.19	81.03	90.02
	0.2	37.09	60.90	62.96	79.35	84.11	91.71
	0.3	44.34	66.59	67.85	82.37	86.36	92.93
	0.4	51.08	71.47	72.17	84.95	88.34	93.99
	0.5	57.87	76.07	76.41	87.41	90.19	94.97
	0.6	65.18	80.74	80.76	89.87	92.08	95.96
	0.7	73.55	85.76	85.58	92.51	94.13	97.02
	0.8	84.08	91.70	91.47	95.64	96.56	98.27
0.75	0.1	35.09	45.59	65.43	72.75	86.38	89.60
	0.2	43.83	53.87	70.83	77.21	88.65	91.36
	0.3	50.79	60.17	74.89	80.50	90.30	92.63
	0.4	57.11	65.70	78.42	83.33	91.72	93.73
	0.5	63.34	71.00	81.82	86.03	93.07	94.75
	0.6	69.93	76.47	85.26	88.73	94.41	95.78
	0.7	77.34	82.47	89.03	91.65	95.88	96.89
	0.8	86.49	89.68	93.56	95.13	97.59	98.18

Tabelle B.2: *Globale Power und Power auf der ersten Stufe (kursiv) eines zweistufigen Testverfahrens nach Inverse-χ^2-Methode bei Aufteilung der Freiheitsgrade ν_G und des Stichprobenumfangs M_1 gemäß $\nu_1/\nu_G = \sqrt{m_1/M_1}) = \sqrt{\epsilon_m}$ (einseitiger Zwei-Stichproben-Gauß-Test mit $\alpha = 0.025$, $1 - \beta = 0.9$) bei Planung mit dem wahren Wert θ/σ*

	ν_G			
ϵ_m	4	10	100	1000
0.1	0.88108	0.88360	0.88188	0.87892
	0.02412	*0.00511*	*<0.00001*	*<0.00001*
0.2	0.87802	0.88262	0.88257	0.87996
	0.09540	*0.03800*	*<0.00001*	*<0.00001*
0.3	0.87658	0.88276	0.88435	0.88227
	0.20712	*0.11697*	*0.00020*	*<0.00001*
0.4	0.87587	0.88330	0.88642	0.88489
	0.34044	*0.23876*	*0.00489*	*<0.00001*
0.5	0.87564	0.88404	0.88854	0.88757
	0.47611	*0.38397*	*0.03880*	*<0.00001*
0.6	0.87588	0.88491	0.89065	0.89020
	0.60023	*0.52990*	*0.14890*	*<0.00001*
0.7	0.87679	0.88592	0.89271	0.89275
	0.70533	*0.65962*	*0.34896*	*0.00157*
0.8	0.87900	0.88724	0.89470	0.89520
	0.78923	*0.76471*	*0.58318*	*0.07624*
0.9	0.88433	0.88970	0.89664	0.89756
	0.85314	*0.84384*	*0.77758*	*0.48962*

Tabelle B.3: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) für Wahl von* $\beta_{G,k}$ *und* $\beta_{g,k}$ *in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für konstanten Wert* $\theta_0 = 0.5$ *auf allen Stufen, Erwartungswert und Standardabweichung (kursiv) für vier Effektschätzer nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 100$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$		$\hat{\theta}_k^{\sqrt{n}}$		$\hat{\theta}_k^{\mathrm{int}}$		$\hat{\theta}_k^{\mathrm{RCI}}$	
0.0	A	I	0.0272	1640	1646	4.93	0.023	(*0.088*)	0.016	(*0.087*)	0.015	(*0.107*)	0.019	(*0.097*)
		II	0.0246	1591	1596	4.95	0.022	(*0.086*)	0.015	(*0.084*)	0.014	(*0.101*)	0.018	(*0.093*)
		III	0.0254	1621	1632	4.94	0.022	(*0.087*)	0.015	(*0.086*)	0.015	(*0.104*)	0.018	(*0.095*)
		IV	0.0271	1450	1456	4.96	0.024	(*0.092*)	0.016	(*0.089*)	0.016	(*0.102*)	0.019	(*0.096*)
	B	I	0.0271	1905	1932	4.94	0.019	(*0.082*)	0.013	(*0.080*)	0.015	(*0.101*)	0.018	(*0.090*)
		II	0.0272	1757	1780	4.96	0.021	(*0.083*)	0.014	(*0.082*)	0.014	(*0.099*)	0.017	(*0.090*)
		III	0.0226	1853	1874	4.96	0.018	(*0.078*)	0.012	(*0.075*)	0.014	(*0.096*)	0.016	(*0.086*)
		IV	0.0265	1639	1648	4.97	0.021	(*0.087*)	0.015	(*0.084*)	0.015	(*0.098*)	0.017	(*0.092*)
0.3	A	I	0.6930	303	212	3.30	0.359	(*0.120*)	0.357	(*0.151*)	0.327	(*0.143*)	0.334	(*0.137*)
		II	0.7401	299	220	3.60	0.361	(*0.124*)	0.357	(*0.157*)	0.330	(*0.145*)	0.337	(*0.139*)
		III	0.7138	302	222	3.48	0.360	(*0.125*)	0.356	(*0.155*)	0.328	(*0.146*)	0.335	(*0.141*)
		IV	0.7501	288	228	3.89	0.369	(*0.143*)	0.355	(*0.163*)	0.336	(*0.158*)	0.343	(*0.154*)
	B	I	0.8386	347	242	3.45	0.359	(*0.118*)	0.357	(*0.147*)	0.326	(*0.140*)	0.333	(*0.134*)
		II	0.8312	329	240	3.69	0.361	(*0.124*)	0.357	(*0.154*)	0.329	(*0.143*)	0.336	(*0.138*)
		III	0.8462	331	240	3.55	0.363	(*0.121*)	0.360	(*0.151*)	0.331	(*0.142*)	0.338	(*0.136*)
		IV	0.8391	318	246	3.93	0.369	(*0.140*)	0.354	(*0.159*)	0.334	(*0.155*)	0.341	(*0.151*)

Fortsetzung nächste Seite

Fortsetzung Tabelle B.3

θ	Strategie		$\hat{P}_\theta(\not{H}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$		$\hat{\theta}_k^{\sqrt{n}}$		$\hat{\theta}_k^{\text{int}}$		$\hat{\theta}_k^{\text{RCI}}$	
0.5	A	I	0.9699	159	142	1.89	0.522	(*0.141*)	0.534	(*0.164*)	0.515	(*0.157*)	0.517	(*0.154*)
		II	0.9714	155	132	2.25	0.528	(*0.144*)	0.537	(*0.178*)	0.518	(*0.165*)	0.520	(*0.161*)
		III	0.9665	155	132	2.13	0.528	(*0.146*)	0.537	(*0.177*)	0.519	(*0.166*)	0.521	(*0.161*)
		IV	0.9640	143	122	2.88	0.551	(*0.162*)	0.546	(*0.195*)	0.531	(*0.189*)	0.534	(*0.181*)
	B	I	0.9896	163	142	1.91	0.526	(*0.140*)	0.536	(*0.162*)	0.517	(*0.157*)	0.519	(*0.153*)
		II	0.9859	159	132	2.28	0.530	(*0.145*)	0.540	(*0.177*)	0.520	(*0.167*)	0.522	(*0.162*)
		III	0.9871	158	132	2.16	0.531	(*0.141*)	0.540	(*0.173*)	0.521	(*0.163*)	0.523	(*0.158*)
		IV	0.9808	145	124	2.89	0.554	(*0.159*)	0.551	(*0.194*)	0.535	(*0.188*)	0.539	(*0.180*)
0.7	A	I	0.9977	140	138	1.22	0.704	(*0.164*)	0.710	(*0.166*)	0.705	(*0.165*)	0.705	(*0.165*)
		II	0.9965	132	128	1.45	0.710	(*0.164*)	0.718	(*0.178*)	0.711	(*0.171*)	0.711	(*0.169*)
		III	0.9957	132	128	1.40	0.709	(*0.167*)	0.718	(*0.177*)	0.710	(*0.172*)	0.710	(*0.171*)
		IV	0.9886	105	92	2.31	0.726	(*0.179*)	0.729	(*0.219*)	0.719	(*0.215*)	0.720	(*0.205*)
	B	I	0.9988	140	138	1.23	0.704	(*0.162*)	0.711	(*0.164*)	0.705	(*0.164*)	0.705	(*0.164*)
		II	0.9981	132	128	1.47	0.706	(*0.163*)	0.716	(*0.175*)	0.708	(*0.170*)	0.708	(*0.168*)
		III	0.9988	132	128	1.43	0.707	(*0.165*)	0.716	(*0.176*)	0.708	(*0.171*)	0.708	(*0.169*)
		IV	0.9921	107	94	2.30	0.725	(*0.176*)	0.730	(*0.217*)	0.719	(*0.214*)	0.720	(*0.203*)

Tabelle B.4: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für konstanten Wert* $\theta_0 = 0.5$ *auf allen Stufen,* $\nu_k = 2$, $k = 1, \ldots, 5$ *(*$\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$
0.0	A	I	0.0257	796	816	4.99
		II	0.0255	738	764	4.99
		III	0.0264	766	792	4.99
		IV	0.0267	612	638	4.99
	B	I	0.0250	967	998	4.99
		II	0.0240	842	872	4.99
		III	0.0228	919	948	4.99
		IV	0.0273	732	764	4.99
0.3	A	I	0.7883	362	350	3.83
		II	0.7452	310	294	4.16
		III	0.7726	334	322	4.01
		IV	0.6833	276	266	4.43
	B	I	0.8850	384	362	3.86
		II	0.8387	334	308	4.17
		III	0.8724	357	336	4.01
		IV	0.8060	308	290	4.42

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$
0.5	A	I	0.9836	244	220	2.20
		II	0.9652	196	178	2.88
		III	0.9767	210	192	2.60
		IV	0.9480	160	148	3.55
	B	I	0.9958	245	218	2.17
		II	0.9895	198	178	2.86
		III	0.9921	211	190	2.59
		IV	0.9829	167	152	3.57
0.7	A	I	0.9998	215	212	1.18
		II	0.9971	158	150	1.64
		III	0.9987	167	158	1.51
		IV	0.9923	107	106	2.87
	B	I	0.9998	215	212	1.19
		II	0.9990	158	150	1.65
		III	0.9998	167	158	1.52
		IV	0.9981	109	108	2.89

Tabelle B.5: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte, gepoolte Schätzung* $\max(0.3, \hat{\theta}_k)$, *Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$	$\sigma(\hat{\theta}_k)$
0.0	A	I	0.0250	1811	1892	4.94	0.019	0.083
		II	0.0235	1737	1818	4.97	0.017	0.082
		III	0.0259	1790	1880	4.95	0.018	0.087
		IV	0.0277	1565	1640	4.97	0.021	0.108
	B	I	0.0259	2325	2432	4.95	0.014	0.076
		II	0.0252	2048	2154	4.97	0.015	0.080
		III	0.0235	2260	2366	4.97	0.013	0.074
		IV	0.0270	1923	2024	4.97	0.018	0.099
0.3	A	I	0.8669	332	300	2.97	0.368	0.139
		II	0.8816	336	304	3.47	0.375	0.151
		III	0.8724	336	316	3.16	0.372	0.151
		IV	0.8707	345	322	3.75	0.382	0.198
	B	I	0.9720	381	336	3.03	0.371	0.136
		II	0.9599	380	328	3.52	0.375	0.149
		III	0.9722	380	342	3.21	0.373	0.146
		IV	0.9562	389	350	3.77	0.380	0.191
0.5	A	I	0.9747	161	106	1.90	0.550	0.149
		II	0.9742	156	116	2.43	0.569	0.166
		III	0.9721	160	114	2.23	0.566	0.162
		IV	0.9510	176	184	3.09	0.604	0.243
	B	I	0.9940	169	110	1.92	0.554	0.147
		II	0.9884	164	122	2.46	0.569	0.162
		III	0.9919	169	120	2.23	0.569	0.161
		IV	0.9771	183	190	3.11	0.608	0.241
0.7	A	I	0.9950	112	102	1.30	0.718	0.173
		II	0.9901	99	84	1.73	0.739	0.179
		III	0.9900	102	82	1.60	0.738	0.177
		IV	0.9636	111	82	2.87	0.822	0.270
	B	I	0.9985	114	102	1.32	0.719	0.169
		II	0.9941	101	84	1.75	0.739	0.177
		III	0.9959	105	84	1.61	0.737	0.174
		IV	0.9797	112	84	2.88	0.827	0.266

Tabelle B.6: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte Schätzung* $\max(0.3, \hat{\theta}_k^{\sqrt{n}})$, *Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k^{\sqrt{n}}$	$\sigma(\hat{\theta}_k^{\sqrt{n}})$
0.0	A	I	0.0263	1813	1900	4.94	0.013	0.084
		II	0.0250	1728	1812	4.96	0.013	0.083
		III	0.0268	1789	1880	4.95	0.013	0.085
		IV	0.0258	1559	1638	4.97	0.016	0.101
	B	I	0.0243	2324	2434	4.96	0.011	0.073
		II	0.0225	2056	2150	4.97	0.010	0.074
		III	0.0241	2259	2364	4.96	0.010	0.074
		IV	0.0248	1922	2022	4.97	0.013	0.096
0.3	A	I	0.8631	332	300	2.95	0.360	0.158
		II	0.8905	339	306	3.46	0.362	0.172
		III	0.8772	338	314	3.16	0.358	0.169
		IV	0.8776	345	322	3.73	0.355	0.196
	B	I	0.9761	387	342	3.03	0.360	0.151
		II	0.9657	375	328	3.50	0.361	0.163
		III	0.9711	383	340	3.20	0.359	0.163
		IV	0.9567	389	348	3.77	0.352	0.188
0.5	A	I	0.9778	159	106	1.89	0.552	0.169
		II	0.9753	159	120	2.42	0.561	0.194
		III	0.9739	163	116	2.21	0.561	0.193
		IV	0.9554	176	184	3.08	0.564	0.254
	B	I	0.9950	170	112	1.94	0.554	0.165
		II	0.9906	166	124	2.44	0.565	0.193
		III	0.9906	171	126	2.22	0.562	0.187
		IV	0.9812	186	196	3.11	0.564	0.248
0.7	A	I	0.9944	112	102	1.29	0.725	0.172
		II	0.9910	102	84	1.74	0.746	0.196
		III	0.9907	103	82	1.61	0.746	0.193
		IV	0.9683	112	84	2.85	0.776	0.291
	B	I	0.9978	115	102	1.33	0.724	0.172
		II	0.9931	102	84	1.75	0.749	0.195
		III	0.9957	106	84	1.61	0.743	0.190
		IV	0.9826	114	88	2.87	0.783	0.293

Tabelle B.7: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte Schätzung* $\max(0.3, \hat{\theta}_k^{\mathrm{int}})$, *Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k^{\mathrm{int}}$	$\sigma(\hat{\theta}_k^{\mathrm{int}})$
0.0	A	I	0.0230	1820	1906	4.95	0.012	0.092
		II	0.0223	1737	1822	4.97	0.012	0.087
		III	0.0268	1787	1872	4.95	0.014	0.094
		IV	0.0246	1558	1638	4.97	0.015	0.097
	B	I	0.0229	2327	2431	4.96	0.011	0.083
		II	0.0250	2053	2152	4.97	0.012	0.085
		III	0.0228	2253	2362	4.97	0.011	0.083
		IV	0.0266	1919	2014	4.97	0.013	0.097
0.3	A	I	0.8616	332	296	2.97	0.341	0.156
		II	0.8938	336	304	3.46	0.347	0.167
		III	0.8825	337	310	3.16	0.344	0.165
		IV	0.8867	348	320	3.72	0.353	0.198
	B	I	0.9717	388	342	3.04	0.338	0.150
		II	0.9573	377	330	3.53	0.343	0.160
		III	0.9715	383	344	3.23	0.342	0.159
		IV	0.9600	387	346	3.74	0.351	0.188
0.5	A	I	0.9762	160	106	1.88	0.537	0.169
		II	0.9720	156	118	2.45	0.545	0.194
		III	0.9728	161	116	2.21	0.545	0.191
		IV	0.9606	183	202	3.04	0.560	0.253
	B	I	0.9949	168	108	1.93	0.535	0.168
		II	0.9898	165	124	2.46	0.544	0.194
		III	0.9912	170	126	2.24	0.543	0.187
		IV	0.9865	190	212	3.04	0.564	0.251
0.7	A	I	0.9937	112	102	1.30	0.718	0.181
		II	0.9908	100	80	1.72	0.738	0.204
		III	0.9915	103	82	1.60	0.735	0.198
		IV	0.9721	117	90	2.80	0.776	0.318
	B	I	0.9979	113	102	1.30	0.718	0.177
		II	0.9948	102	84	1.73	0.736	0.202
		III	0.9963	106	84	1.62	0.731	0.195
		IV	0.9817	121	96	2.83	0.773	0.316

Tabelle B.8: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei acht Strategien (siehe S. 79) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte Schätzung* $\max(0.3, \hat{\theta}_k^{\mathrm{RCI}})$, *Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k^{\mathrm{RCI}}$	$\sigma(\hat{\theta}_k^{\mathrm{RCI}})$
0.0	A	I	0.0229	1821	1912	4.95	0.015	0.082
		II	0.0282	1730	1813	4.96	0.017	0.089
		III	0.0240	1797	1876	4.96	0.015	0.084
		IV	0.0245	1563	1639	4.97	0.016	0.099
	B	I	0.0268	2306	2424	4.96	0.015	0.078
		II	0.0292	2042	2146	4.96	0.015	0.085
		III	0.0225	2261	2370	4.96	0.012	0.076
		IV	0.0252	1916	2012	4.97	0.014	0.093
0.3	A	I	0.8687	335	304	2.96	0.353	0.146
		II	0.8937	339	306	3.46	0.359	0.158
		III	0.8748	335	312	3.17	0.358	0.156
		IV	0.8806	351	328	3.75	0.363	0.191
	B	I	0.9731	385	342	3.01	0.355	0.143
		II	0.9583	377	330	3.51	0.360	0.153
		III	0.9713	377	340	3.21	0.360	0.151
		IV	0.9585	388	350	3.76	0.368	0.189
0.5	A	I	0.9749	161	106	1.90	0.541	0.159
		II	0.9704	157	114	2.41	0.559	0.179
		III	0.9749	162	114	2.21	0.557	0.176
		IV	0.9599	179	192	3.06	0.587	0.248
	B	I	0.9951	167	108	1.92	0.546	0.157
		II	0.9905	163	122	2.47	0.558	0.172
		III	0.9899	168	118	2.23	0.558	0.171
		IV	0.9796	187	206	3.11	0.587	0.246
0.7	A	I	0.9952	112	102	1.31	0.715	0.174
		II	0.9892	100	84	1.74	0.735	0.190
		III	0.9907	103	82	1.62	0.735	0.187
		IV	0.9686	112	84	2.85	0.805	0.282
	B	I	0.9983	115	102	1.31	0.718	0.174
		II	0.9946	102	84	1.75	0.734	0.188
		III	0.9941	106	84	1.62	0.733	0.185
		IV	0.9805	115	88	2.86	0.807	0.278

Tabelle B.9: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$*, ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei zwölf Strategien (siehe S. 79/90) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte, gepoolte Schätzung* $\max(0.3, \hat{\theta}_k)$ *nach Powerrekalkulation, Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe (*$\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$	$\sigma(\hat{\theta}_k)$
0.0	a	I	0.0268	1888	1987	4.95	0.018	0.082
		II	0.0250	1604	1642	4.97	0.017	0.085
		III	0.0249	1700	1754	4.97	0.018	0.085
		IV	0.0261	1209	1288	4.98	0.022	0.115
	b	I	0.0252	1766	1846	4.95	0.017	0.083
		II	0.0244	1714	1798	4.96	0.018	0.086
		III	0.0233	1789	1870	4.96	0.016	0.082
		IV	0.0292	1579	1658	4.97	0.020	0.104
	c	I	0.0246	1924	2024	4.95	0.018	0.082
		II	0.0235	1780	1868	4.96	0.017	0.083
		III	0.0263	1853	1940	4.95	0.019	0.086
		IV	0.0250	1592	1664	4.97	0.019	0.100
0.3	a	I	0.8939	348	318	3.26	0.369	0.138
		II	0.8895	355	312	3.85	0.376	0.153
		III	0.8974	350	320	3.69	0.374	0.150
		IV	0.8499	365	326	4.53	0.387	0.205
	b	I	0.9036	350	308	3.32	0.370	0.137
		II	0.9017	349	314	3.59	0.376	0.152
		III	0.9052	344	312	3.35	0.375	0.148
		IV	0.8958	355	328	3.77	0.382	0.199
	c	I	0.9061	348	318	2.96	0.369	0.138
		II	0.8975	348	316	3.48	0.371	0.148
		III	0.9057	343	324	3.15	0.374	0.149
		IV	0.8934	353	332	3.74	0.382	0.199

Fortsetzung nächste Seite

Fortsetzung Tabelle B.9

θ	Strategie		$\hat{P}_\theta(\not{H}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k$	$\sigma(\hat{\theta}_k)$
0.5	a	I	0.9778	162	106	2.10	0.550	0.146
		II	0.9584	159	114	2.76	0.568	0.164
		III	0.9673	157	110	2.57	0.566	0.161
		IV	0.9233	168	136	4.03	0.618	0.247
	b	I	0.9778	158	106	2.00	0.551	0.148
		II	0.9679	154	112	2.45	0.571	0.163
		III	0.9730	159	112	2.25	0.565	0.159
		IV	0.9503	178	186	3.10	0.609	0.248
	c	I	0.9808	162	106	1.86	0.554	0.146
		II	0.9702	158	116	2.42	0.566	0.160
		III	0.9743	163	114	2.19	0.566	0.159
		IV	0.9575	180	188	3.09	0.604	0.243
0.7	a	I	0.9937	113	102	1.37	0.718	0.173
		II	0.9869	101	84	1.87	0.736	0.180
		III	0.9855	101	84	1.75	0.732	0.178
		IV	0.9419	91	76	3.60	0.833	0.276
	b	I	0.9937	112	102	1.32	0.718	0.173
		II	0.9863	100	84	1.73	0.737	0.181
		III	0.9909	104	84	1.61	0.733	0.180
		IV	0.9625	111	84	2.85	0.825	0.277
	c	I	0.9934	113	102	1.29	0.719	0.172
		II	0.9855	100	84	1.73	0.738	0.181
		III	0.9892	104	84	1.59	0.737	0.179
		IV	0.9607	112	82	2.86	0.826	0.270

Tabelle B.10: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN und mittlere Stufenzahl* $\bar{k}$ *bei zwölf Strategien (siehe S. 79/90) in fünfstufiger Studie bei Anpassung des Stichprobenumfangs für begrenzte Schätzung* $\max(0.3, \hat{\theta}_k^{\mathrm{int}})$ *nach Powerrekalkulation, Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe (*$\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k^{\mathrm{int}}$	$\sigma(\hat{\theta}_k^{\mathrm{int}})$
0.0	a	I	0.0258	1886	1986	4.96	0.014	0.091
		II	0.0268	1594	1638	4.97	0.013	0.092
		III	0.0236	1697	1740	4.97	0.012	0.086
		IV	0.0263	1216	1298	4.98	0.013	0.106
	b	I	0.0254	1762	1845	4.95	0.013	0.092
		II	0.0241	1717	1802	4.97	0.012	0.089
		III	0.0234	1780	1860	4.96	0.012	0.088
		IV	0.0243	1591	1672	4.97	0.011	0.093
	c	I	0.0270	1932	2026	4.95	0.013	0.092
		II	0.0232	1780	1864	4.97	0.011	0.084
		III	0.0221	1868	1952	4.96	0.012	0.088
		IV	0.0215	1585	1658	4.97	0.012	0.093
0.3	a	I	0.9013	344	310	3.23	0.342	0.155
		II	0.8937	350	312	3.86	0.346	0.165
		III	0.8977	346	312	3.67	0.345	0.162
		IV	0.8744	369	342	4.49	0.351	0.199
	b	I	0.9043	351	310	3.34	0.340	0.155
		II	0.8989	346	308	3.61	0.346	0.165
		III	0.9051	345	314	3.35	0.345	0.162
		IV	0.8977	353	324	3.74	0.353	0.198
	c	I	0.9021	350	318	2.98	0.341	0.156
		II	0.9045	345	313	3.46	0.346	0.166
		III	0.9096	347	324	3.16	0.344	0.163
		IV	0.8986	357	332	3.72	0.349	0.187

Fortsetzung nächste Seite

Fortsetzung Tabelle B.10

θ	Strategie		$\hat{P}_\theta(\not{H}_0)$	ASN	MSN	$\bar{k}$	$\hat{\theta}_k^{\text{int}}$	$\sigma(\hat{\theta}_k^{\text{int}})$
0.5	a	I	0.9794	161	106	2.09	0.532	0.170
		II	0.9695	160	117	2.76	0.542	0.195
		III	0.9719	156	110	2.58	0.546	0.191
		IV	0.9504	176	154	3.95	0.571	0.266
	b	I	0.9795	159	106	2.02	0.530	0.169
		II	0.9713	156	116	2.46	0.545	0.197
		III	0.9757	162	116	2.26	0.544	0.193
		IV	0.9635	183	200	3.05	0.562	0.255
	c	I	0.9817	165	106	1.90	0.533	0.171
		II	0.9716	160	118	2.41	0.542	0.196
		III	0.9732	163	116	2.18	0.546	0.189
		IV	0.9686	185	206	3.02	0.566	0.263
0.7	a	I	0.9946	112	102	1.34	0.719	0.179
		II	0.9875	100	84	1.85	0.735	0.206
		III	0.9876	102	84	1.73	0.728	0.201
		IV	0.9598	97	80	3.51	0.779	0.328
	b	I	0.9941	112	102	1.31	0.713	0.180
		II	0.9870	100	84	1.74	0.732	0.208
		III	0.9899	103	84	1.60	0.731	0.198
		IV	0.9709	118	90	2.81	0.773	0.314
	c	I	0.9942	113	102	1.29	0.716	0.176
		II	0.9871	101	84	1.72	0.732	0.204
		III	0.9904	105	84	1.58	0.734	0.198
		IV	0.9730	117	90	2.80	0.783	0.323

Tabelle B.11: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$, *ASN, MSN, mittlere, mediane und maximale Stufenzahl* ($\bar{k}$, $\tilde{k}_{0.5}$, $k_{\max}$) *bei konstanter bedingter Power (Strategie I) bzw. konstantem bedingten Powerzuwachs (Strategie II) bei Anpassung des Stichprobenumfangs für begrenzte gepoolte Schätzung* $\max(0.3, \hat{\theta}_k)$ *nach Powerrekalkulation, Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe* ($\alpha = 0.025$, $\beta = 0.1$, $\nu_G = 10$, 10000 *Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\tilde{k}_{0.5}$	$k_{\max}$	$\hat{\theta}_k$	$\sigma(\hat{\theta}_k)$
0.0	I	0.37	0.0238	2361	2456	5.96	6	7	0.016	0.076
		0.23	0.0246	3885	4082	9.89	10	12	0.015	0.075
	II	0.18	0.0237	2160	2258	5.99	6	7	0.017	0.083
		0.10	0.0233	3472	3618	9.95	10	12	0.015	0.074
0.3	I	0.37	0.9058	354	322	3.12	3	7	0.367	0.136
		0.23	0.9077	353	312	4.31	4	13	0.376	0.144
	II	0.18	0.9052	343	308	3.78	4	7	0.377	0.150
		0.10	0.9076	356	308	5.39	5	12	0.383	0.157
0.5	I	0.37	0.9806	165	106	1.90	2	7	0.550	0.147
		0.23	0.9715	157	112	2.45	2	12	0.563	0.156
	II	0.18	0.9714	153	112	2.47	2	7	0.570	0.163
		0.10	0.9610	151	120	3.25	3	11	0.575	0.172
0.7	I	0.37	0.9954	113	102	1.30	1	6	0.719	0.172
		0.23	0.9892	103	86	1.63	1	10	0.732	0.180
	II	0.18	0.9870	99	84	1.74	2	7	0.736	0.179
		0.10	0.9789	93	74	2.16	2	10	0.752	0.187

Tabelle B.12: *Ablehnwahrscheinlichkeit* $\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$*, ASN, MSN, mittlere, mediane und maximale Stufenzahl (*$\bar{k}$*,* $\tilde{k}_{0.5}$*,* $k_{\max}$*) bei konstanter bedingter Power (Strategie I) bzw. konstantem bedingten Powerzuwachs (Strategie II) bei Anpassung des Stichprobenumfangs für begrenzte Schätzung* $\max(0.3, \hat{\theta}_k^{\mathrm{int}})$ *nach Powerrekalkulation, Erwartungswert und Standardabweichung des Effektschätzers nach finaler Stufe (*$\alpha = 0.025$*,* $\beta = 0.1$*,* $\nu_G = 10$*, 10000 Simulationsläufe)*

θ	Strategie		$\hat{\mathrm{P}}_\theta(\not{\mathrm{H}}_0)$	ASN	MSN	$\bar{k}$	$\tilde{k}_{0.5}$	$k_{\max}$	$\hat{\theta}_k^{\mathrm{int}}$	$\sigma(\hat{\theta}_k^{\mathrm{int}})$
0.0	I	0.37	0.0258	2366	2482	5.94	6	7	0.012	0.091
		0.23	0.0250	3903	4082	9.89	10	13	0.012	0.089
	II	0.18	0.0232	2157	2244	6.00	6	7	0.013	0.087
		0.10	0.0218	3451	3572	9.96	10	12	0.013	0.082
0.3	I	0.37	0.9047	348	318	3.07	3	7	0.341	0.156
		0.23	0.9064	351	310	4.30	4	12	0.344	0.163
	II	0.18	0.9098	350	312	3.82	4	7	0.343	0.167
		0.10	0.9081	351	304	5.38	5	12	0.348	0.177
0.5	I	0.37	0.9814	164	106	1.90	2	7	0.534	0.170
		0.23	0.9746	155	110	2.44	2	11	0.545	0.189
	II	0.18	0.9731	154	116	2.49	2	7	0.547	0.196
		0.10	0.9658	151	120	3.24	3	11	0.551	0.217
0.7	I	0.37	0.9960	113	102	1.29	1	6	0.719	0.178
		0.23	0.9919	103	86	1.60	1	10	0.729	0.195
	II	0.18	0.9884	100	84	1.71	2	6	0.733	0.207
		0.10	0.9855	94	74	2.17	2	10	0.744	0.229

Literaturverzeichnis

Armitage, P., McPherson, C.K., Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* 132, 2: 235–244.

Banik, N., Köhne, K., Bauer, P. (1996). On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical Journal* 38, 1: 25–37.

Bauer, P. (1989a). Multistage testing with adaptive designs + discussion. *Biometrie und Informatik in Medizin und Biologie* 20, 4: 130–148.

Bauer, P. (1989b). Sequential tests of hypotheses in consecutive trials. *Biometrical Journal* 31, 6: 663–676.

Bauer, P. (1992). The choice of sequential boundaries based on the concept of power spending. *Biometrie und Informatik in Medizin und Biologie* 23, 1: 3–15.

Bauer, P. (2006). Discussions. *Biometrics* 62, 3: 676–678.

Bauer, P., Brannath, W., Posch, M. (2001). Flexible two-stage designs: An overview. *Methods of Information in Medicine* 40, 2: 117–121.

Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 18, 14: 1833–1848.

Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 50, 4: 1029–1041.

Bauer, P., Köhne, K. (1996). Erratum: Evaluation of experiments with adaptive interim analysis. *Biometrics* 52, 1: 380.

Bauer, P., König, F. (2006). The reassessment of trial perspectives from interim data – a critical view. *Statistics in Medicine* 25, 1: 23–36.

Bauer, P., Röhmel, J. (1995). An adaptive method for establishing a dose response relationship. *Statistics in Medicine* 14, 14: 1595–1607.

Bischoff, W., Miller, F. (2005). Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika* 92, 1: 197–212.

Brannath, W., Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. *Biometrics* 60, 3: 715–723.

Brannath, W., König, F., Bauer, P. (2006a). Estimation in flexible two stage designs. *Statistics in Medicine* 25, 19: 3366–3381.

Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* 97, 457: 236–244.

Brannath, W., Posch, M., Bauer, P. (2006b). On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference* 136, 6: 1956–1961.

Burman, C.F., Sonesson, C. (2006). Are flexible designs sound? *Biometrics* 62, 3: 664–669.

Cheng, Y., Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics* 60, 4: 910–918.

Coburger, S., Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal* 43, 7: 821–833.

Cui, L., Hung, H.M., Wang, S.J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* 55, 3: 853–857.

DeMets, D.L., Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69, 3: 661–663.

Denne, J.S. (2000). Estimation following extension of a study on the basis of conditional power. *Journal of Biopharmaceutical Statistics* 10, 2: 131–144.

Ellenberg, S.S. (2006). Group discussion. *Statistics in Medicine* 25, 19: 3326–3347.

Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine* 17, 14: 1551–1562.

Hartung, J. (2000). A new class of self-designing clinical trials. In: Hasman, A., Blobel, B., Dudeck, J., Engelbrecht, R., Gell, G., Prokosch, H.U. (Hrsg.), *Medical Infobahn for Europe. Proceedings of MIE 2000 and GMDS 2000.* Amsterdam: IOS Press, 310–314.

Hartung, J. (2001). A self-designing rule for clinical trials with arbitrary response variables. *Controlled Clinical Trials* 22, 2: 111–116.

Hartung, J. (2006). Flexible designs by adaptive plans of generalized Pocock- and O'Brien-Fleming-type and by self-designing clinical trials. *Biometrical Journal* 48, 4: 521–536.

Hartung, J., Knapp, G. (2003). A new class of completely self-designing clinical trials. *Biometrical Journal* 45, 1: 3–19.

Hartung, J., Knapp, G. (2006). Repeated confidence intervals in self-designing clinical trials and switching between noninferiority and superiority. *Biometrical Journal* 48, 4: 697–709.

Hedges, L.V., Olkin, I. *Statistical Methods for Meta-Analysis*. New York: Academic Press, 1985.

Jennison, C., Turnbull, B.W. (1989). Interim analyses: The repeated confidence interval approach. *Journal of the Royal Statistical Society, Series B* 51, 3: 305–361.

Jennison, C., Turnbull, B.W. (1991). Exact calculations for sequential t, χ^2 and F tests. *Biometrika* 78, 1: 133–141.

Jennison, C., Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton und London: Chapman and Hall/CRC, 2000.

Jennison, C., Turnbull, B.W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 22, 6: 971–993.

Johnson, N.L., Kotz, S. *Continuous Univariate Distributions*, Band 1. New York: Wiley, 1970.

Kaballo, W. *Einführung in die Analysis*, Band I. Heidelberg - Berlin: Spektrum Akademischer Verlag GmbH, 2000, 2. Auflage.

Kim, K., DeMets, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 74, 1: 149–154.

Lan, K.K.G., DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70, 3: 659–663.

Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* 55, 4: 1286–1290.

Li, G., Shih, W.J., Xie, T., Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics* 3, 2: 227–287.

Li, G., Shih, W.J., Wang, Y. (2005). Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics* 15, 4: 707–718.

Liu, A., Hall, W.J. (1999). Unbiased estimation following a group sequential test. *Biometrika* 86, 1: 71–78.

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine* 1, 1: 25–36.

Mehta, C.R., Patel, N.R. (2006). Adaptive, group sequential and decision theoretic approaches to sample size determination. *Statistics in Medicine* 25, 19: 3250–3269.

Müller, H.H., Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 57, 3: 886–891.

Müller, H.H., Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 23, 16: 2497–2508.

O'Brien, P.C., Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 3: 549–556.

Pampallona, S., Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* 42, 1: 19–35.

Pocock, S.J. (1977). Group sequential methods in the design and analyses of clinical trials. *Biometrika* 64, 2: 191–199.

Pocock, S.J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* 38, 1: 153–162.

Posch, M., Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal* 41, 6: 689–696.

Posch, M., Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics* 56, 4: 1170–1176.

Posch, M., Timmesfeld, N., König, F., Müller, H.H. (2004). Conditional rejection probabilities of student's t-test and design adaptations. *Biometrical Journal* 46, 4: 389–403.

Posch, M., Bauer, P., Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* 22, 6: 953–969.

Proschan, M.A. (2003). The geometry of two-stage tests. *Statistica Sinica* 13: 163–177.

Proschan, M.A., Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* 51, 4: 1315–1324.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL http://www.R-project.org. ISBN 3-900051-07-0.

Schäfer, H., Müller, H.H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 20, 24: 3741–3751.

Schäfer, H., Timmesfeld, N., Müller, H.H. (2006). An overview of statistical approaches for adaptive designs and design modifications. *Biometrical Journal* 48, 4: 507–520.

Shen, Y., Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 55, 1: 190–197.

Shih, W.C.J., Quan, H., Li, G. (2004). Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 23, 18: 2781–2798.

Shih, W.J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. *Statistics in Medicine* 25, 6: 933–941.

Thach, C.T., Fisher, L.D. (2002). Self-designing two-stage trials to minimize expected costs. *Biometrics* 58, 2: 432–438.

Tsiatis, A.A., Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 90, 2: 367–378.

Wang, S.J., Hung, H.M., Tsong, Y., Cui, L. (2001). Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine* 20, 13: 1903–1912.

Wang, S.K., Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43, 1: 193–199.

Wassmer, G. (1999). Multistage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal* 41, 3: 279–293.

Wassmer, G. *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien. Theoretische Konzepte und deren praktische Umsetzung mit SAS.* Schriftenreihe des IMSIE. München: Verlag Alexander Mönch, 2001, 2. überarbeitete Auflage.

Wassmer, G., Eisebitt, R., Coburger, S. (2001). Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Information Journal* 35, 4: 1131–1146.

Yin, G., Shen, Y. (2005). Self-designing trial combined with classical group sequential monitoring. *Journal of Biopharmaceutical Statistics* 15, 4: 667–675.