Vorwort

Die vorliegende Arbeit stellt die überarbeitete Fassung meiner Habilitationsschrift dar, die der Medizinischen Fakultät der Universität zu Köln im März 1999 vorgelegt wurde. Sie entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Medizinische Statistik, Informatik und Epidemiologie der Universität zu Köln.

Mein besonderer Dank gilt dem Institutsdirektor Prof. Dr. Walter Lehmacher, der mich zu dem in dieser Arbeit behandelten Thema führte und mich mit großer Unterstützung begleitete. Ohne ihn wäre die Arbeit in der vorliegenden Form nicht zustande gekommen. Danken möchte ich auch allen Kollegen, von denen ich durch Diskussionen methodischer Aspekte, Verbesserungsvorschläge und Korrekturen zahlreiche Anregungen bekommen habe. Besonders zu nennen sind Frau Dipl.-Stat. Silke Coburger, Frau Dipl.-Stat. Karin Funke und Herr Dipl.-Math. Dr. Martin Hellmich, die auch im Rahmen des von der DFG geförderten Projekts „Weiterentwicklung von gruppensequentiellen und adaptiven Plänen in klinischen Studien“ wertvolle methodische Beiträge leisten.

Gernot Wassmer Köln, Oktober 1999

Vorwort zur 2. Auflage

In der vorliegenden zweiten Auflage sind Substanz und Umfang unverändert geblieben, an einigen Stellen konnten Fehler beseitigt werden. Ich danke allen Kolleginnen und Kollegen, die mich auf Unstimmigkeiten im Text der ersten Auflage aufmerksam gemacht haben. Die rasante Entwicklung insbesondere der adaptiven Verfahren spiegelt sich in den neu hinzugekommenen Literaturangaben wider. Ich habe versucht, die Literaturstellen so aktuell wie möglich anzugeben und auch neue Entwicklungen einzubeziehen.

Gernot Wassmer Köln, März 2001

Inhaltsverzeichnis

1 Einführung **1**

2 Gruppensequentielle Testverfahren **11**

2.1 Die allgemeine Konstruktionsmethode 13

2.2 Verfahren mit gleichen Sequenzgrößen 22

2.2.1 Klassische Verfahren 23

2.2.2 Abbruch mit der Annahme der Nullhypothese 31

2.2.3 Einseitige Pläne 35

2.3 Verfahren mit ungleichen Sequenzgrößen 43

2.3.1 Der Fall fest vorgegebener Sequenzgrößen 44

2.3.2 Eine *worst case scenario*-Lösung 47

2.3.3 Der α-*spending function*-Ansatz 59

2.3.4 Vergleich der Verfahren 63

2.4 Andere Studientypen 68

3 Adaptive Testverfahren **71**

3.1 Zweistufige Designs 73

3.1.1 Der Ansatz von Bauer und Köhne 74

3.1.2 Der Ansatz von Proschan und Hunsberger 88

3.1.3 *Power*-Gewinn bei adaptiver Planung 97

3.1.4 Ein allgemeines Prinzip für die Konstruktion adaptiver Pläne 104

3.1.5 Zweiseitige Pläne 114

3.2 Mehrstufige Designs 118

3.2.1 Mehrstufige Designs auf der Basis von Fishers Kombinationstest 119

3.2.2 Mehrstufige Designs auf der Basis der *inverse normal method* 129

3.2.3 Adaptive Verfahren ohne Vorgabe der Stufenanzahl . . . 135

4 Zusammenfassung **137**

Literaturverzeichnis **140**

A SAS/IML-Implementierung gruppensequentieller Pläne **153**

A.1 Kritische Werte nach Wang und Tsiatis 154

A.2 Schranken mit optimalem *ASN* 156

A.3 Der Ansatz nach Pampallona und Tsiatis 158

A.4 Ungleiche Sequenzgrößen 163

A.5 Implementierung des *worst case scenario*-Ansatzes 165

A.6 Implementierung des α-*spending*-Ansatzes 168

A.7 Bestimmung der *power* bei Fishers Produktregel 171

A.8 Kritische Werte der Prozedur nach Proschan und Hunsberger . . 173

A.9 Berechnung der kritischen Werte bei Fishers Produktregel 174

Tabellenverzeichnis

2.1 Kritische Werte nach O'Brien und Fleming bzw. Pocock im zweiseitigen Testproblem 24

2.2 Kritische Werte der Δ-Klasse von Fortsetzungsbereichen nach Wang und Tsiatis im zweiseitigen Testproblem 28

2.3 Optimales Δ und kritische Werte in der Δ-Klasse von Fortsetzungsbereichen im zweiseitigen Testproblem 29

2.4 Optimales Δ und kritische Werte des zweiseitigen Verfahrens von Pampallona und Tsiatis 34

2.5 Kritische Werte im einseitigen Pocock-Design nach DeMets und Ware 38

2.6 Optimales Δ und kritische Werte des einseitigen Verfahrens von Pampallona und Tsiatis 41

2.7 Kritische Werte nach O'Brien und Fleming bzw. Pocock im zweiseitigen Testproblem bei ungleichen Sequenzgrößen 45

2.8 Kritische Werte im zweiseitigen bzw. einseitigen Testproblem unter der Voraussetzung unabhängiger Teststatistiken 51

2.9 Kritische Werte im zweiseitigen Testproblem nach Pocock mit gemeinsamer oberer Schranke für die Sequenzgrößen 55

2.10 Kritische Werte im zweiseitigen Testproblem nach O'Brien und Fleming mit gemeinsamer oberer und unterer Schranke für die Sequenzgrößen 57

3.1 Signifikanzniveau α_1 und Schranke c_α im zweistufigen Verfahren nach Bauer und Köhne 77

3.2 Signifikanzniveau α_1 und c_{α_2} im zweistufigen Verfahren mit $\alpha_1 = \alpha_2$ 79

3.3 Benötigter maximaler Stichprobenumfang $n_1 + n_2$ und ASN beim zweistufigen Verfahren nach Bauer und Köhne 87

3.4 Kritische Werte der Proschan und Hunsberger-Prozedur 93

3.5 Vergleich von $power$ und ASN im adaptiven Design nach Proschan und Hunsberger-Prozedur und Test mit festem Stichprobenumfang 101

3.6 Vergleich von $power$ und ASN im adaptiven Design nach Proschan und Hunsberger-Prozedur und gruppensequentiellen Testverfahren 103

3.7 Kritische Werte für das Produkt der p-Werte basierend auf konstanten lokalen Signifikanzniveaus 123

3.8 Kritische Werte für das Produkt der p-Werte und lokale Signifikanzniveaus mit voller Ausschöpfung des Niveaus auf der letzten Stufe und konstanten lokalen Niveaus 125

3.9 Kritische Werte für das Produkt der p-Werte und lokale Signifikanzniveaus mit voller Ausschöpfung des Niveaus auf der letzten Stufe und Vermeidung von Interaktionen 127

3.10 Vergleich von $power$ und ASN im adaptiven Verfahren nach Lehmacher und Wassmer mit gruppensequentiellen Test bei fest vorgegebenen Stichprobenumfängen 132

Abbildungsverzeichnis

2.1 Fortsetzungs- bzw. Annahmebereiche für O'Brien und Flemings und Pococks Design 25

2.2 Fortsetzungs- bzw. Entscheidungsbereiche für das optimale zweiseitige Design nach Pampallona und Tsiatis 36

2.3 Fortsetzungs- bzw. Entscheidungsbereiche für das optimale einseitige Design nach Pampallona und Tsiatis 42

2.4 Fortsetzungs- bzw. Annahmebereiche für O'Brien und Flemings Design für den Fall ungleicher, fest vorgegebener Sequenzgrößen 46

2.5 Fortsetzungs- bzw. Annahmebereiche für O'Brien und Flemings und Pococks Design mit beliebigen Stichprobenumfängen . . . 52

2.6 Fehlerwahrscheinlichkeit 1. Art im zweiseitigen Design nach Pocock in Abhängigkeit von τ_2 und τ_3 56

2.7 Fehlerwahrscheinlichkeit 1. Art im zweiseitigen Design nach O'Brien und Fleming in Abhängigkeit von τ_2 und τ_3 58

2.8 Beispiele von α-$spending$-Funktionen 61

3.1 Ablehnregion für das Bauer und Köhne-Verfahren 78

3.2 Dichtefunktion von (p_1, p_2) unter der Alternativhypothese . . . 84

3.3 Vergleich der Ablehnregionen der adaptiven zweistufigen Verfahren 105

3.4 Vergleich der Ablehnregionen der adaptiven zweistufigen Verfahren 109

3.5 Vergleich der globalen $power$ der adaptiv zweistufigen Verfahren 112

3.6 Vergleich des ASN der adaptiv zweistufigen Verfahren 113

3.7 Ablehnregion für das zweiseitige zweistufige Testverfahren mit Fishers Produktregel 116

1 Einführung

Bei den meisten klinischen Studien werden in der Planungsphase feste Stichprobenumfänge festgelegt. Die statistische Auswertung wird bei diesen Studien erst nach der vollständigen Beobachtung aller zur Stichprobe gehörenden Patienten durchgeführt. Gruppensequentielle Verfahren erlauben die Durchführung klinischer Studien, in denen jeweils nach einer bestimmten Anzahl von rekrutierten Patienten in einer Zwischenauswertung (*interim analysis*) eine Entscheidung über den Abbruch oder die Weiterführung der Studie entschieden werden kann. Im Gegensatz zu statistischen Verfahren mit einem fest vorgegebenen Stichprobenumfang ist bei diesen Verfahren der Gesamt-Stichprobenumfang somit nicht mehr fest, sondern zufällig. In klinischen Studien wird durch den Gebrauch dieser Pläne in den meisten Fällen eine Verminderung der benötigten Anzahl von Patienten erreicht. Dies hat sowohl ökonomische wie ethische Konsequenzen, deren Stellenwert in der heutigen Zeit als hochrangig anzusehen ist.

Gruppensequentielle Verfahren sind hauptsächlich in den 70er und 80er Jahren eingeführt und entwickelt worden. Es handelt sich dabei um Test- und Schätzverfahren, die – besonders in den U.S.A. und Großbritannien – eine relativ weite Verbreitung gefunden haben. Sie stellen eine Erweiterung der klassischen sequentiellen Pläne dar, die eine Entscheidung nach jeder einzelnen Beobachtung vorsehen. Das letztere kommt der Praxis klinischer Studien nicht entgegen. Nicht nur aus organisatorischen Gründen ist es zweckmäßiger, erst nach Erreichen einer bestimmten Anzahl von Beobachtungen sequentiell zu testen. Bei gruppensequentiellen Verfahren ist eine statistische Entscheidung nach jeder Gruppe („Sequenz“) von Beobachtungen möglich. Nach erfolgter Wirksamkeitsmessung ist zu entscheiden, ob weitere Sequenzen in die Studie aufzunehmen sind. Zur Zeit werden gruppensequentielle Pläne hauptsächlich bei onkologischen Studien angewandt, da man hier insbesondere aus ethischen Gründen an einem frühzeitigen Abbruch der Studie interessiert ist.

In der klinischen Forschung ist man aus ethischen, ökonomischen und auch organisatorischen Gründen sehr an Zwischenauswertungen interessiert. Prinzipiell ermöglichen diese ein statistisch signifikantes Studienergebnis mit einer kleineren Anzahl von in die Studie aufzunehmenden Patienten. Eine Therapie, die sich als überlegen erwies, kann somit früher verwendet werden und die unterlegene(n) Therapie(n) ersetzen. Zwischenauswertungen ermöglichen auch, die Qualität der Dokumentation einer Studie zu bewerten und notwendigenfalls zu verbessern; die Einhaltung des Studienprotokolls kann überprüft werden; unerwünschte Nebenwirkungen können früher aufgedeckt werden, etc. Die Relevanz und die generellen Probleme bei der Durchführung von Zwischenauswertungen sind ausführlich in Köpcke (1984) beschrieben. In dieser Arbeit sind auch typische Beispiele von Therapiestudien angegeben, die mit gruppensequentiellen Verfahren geplant und durchgeführt wurden. Es wird darüber hinaus dargestellt, welche Faktoren zu einem frühzeitigen Abbruch einer Studie führen können (vgl. auch Armitage, 1991; McPherson, 1990).

Ein gängiges Problem bei klinischen Studien ist die Fehlspezifikation des für die Fallzahlplanung zugrundegelegten nachzuweisenden Effekts. Ein zu klein oder zu groß angenommener Effekt führt dazu, daß die Studie *over-* bzw. *underpowered* ist. Bei der sequentiellen Durchführung einer Studie ist es wünschenswert, daß der in einer Zwischenauswertung beobachtete Effekt in die Fallzahlplanung der weiteren Sequenzen mit einfließen kann und eine Neufestlegung der Fallzahlen prinzipiell möglich ist. Man möchte also Adaptionen am Prüfplan vornehmen können, um Informationen der Zwischenauswertungen in die Stichprobenkalkulation der restlichen Studie eingehen zu lassen. Besonders interessant sind daher Auswertungsstrategien, die die Resultate der Zwischenauswertung für die weitere Planung „verwerten“ dürfen. Diese adaptiven Pläne ermöglichen eine flexiblere Planung und Durchführung von Studien als dies bei klassischen gruppensequentiellen Plänen vorgesehen ist. Dadurch sind sie auch aus ökonomischen Gründen interessant, da sie insgesamt den Prozeß der Medikamentenentwicklung und -zulassung beschleunigen können. Adaptive Pläne finden in der heutigen Zeit immer mehr Verwendung in der klinischen und auch (bzw. vor allem) pharmazeutischen Forschung.

Eine weitere Anwendung adaptiver Pläne ergibt sich aus der Möglichkeit der datenabhängigen Wahl der zu testenden Hypothesen. Dadurch können beispiels-

weise in einer mehrarmigen Studie solche Studienarme von einer weiteren Überprüfung ausgeschlossen werden, die in einer Zwischenauswertung einen klaren (signifikanten) Effekt zeigten. Das verfügbare Patientenkollektiv kann damit auf die verbleibenden Arme aufgeteilt werden. Adaptive Pläne bieten auch für solche Fälle eine allgemeine theoretische Lösung für das Problem der datenabhängigen Bestimmung und Planung der Sequenzen in einem gruppensequentiellen Design.

Eine kleine Tour d'Horizon

Die im medizinischen Bereich bahnbrechende Arbeit für die Anwendung gruppensequentieller Pläne stammt von Pocock (1977), der klare Richtlinien für die Anwendung dieser Pläne gegeben hat. Auf Armitage, McPherson und Rowe (1969) geht die der Arbeit von Pocock zugrundeliegende Bestimmungsmethode der Entscheidungsbereiche zurück, die für normalverteilte Merkmale vorgesehen ist. Unter der Voraussetzung gleicher Gruppengrößen lassen sich entsprechend adjustierte kritische Schranken ableiten, die dann in praxi auch für ungleiche Gruppengrößen verwendet werden können, da dies keinen nennenswerten Einfluß auf das nominelle Niveau der Testverfahrens hat (Pocock, 1982). Aus dem gleichen Grund sind die kritischen Werte auch für andere Verteilungsannahmen verwendbar.

Ein großer Teil der nachfolgenden Literatur befaßt sich mit entsprechend exakten Verfahren, die nicht die Voraussetzung gleicher Gruppengrößen mit normalverteilten Merkmalen benötigen. Eine sehr übersichtliche Darstellung findet sich in Jennison und Turnbull (1991b) und Turnbull (1997), die das breite Spektrum der heute bekannten Verfahren zusammenfassen (vgl. auch Whitehead, 1997; Jennison und Turnbull, 2000). Besonders hervorzuheben ist der sogenannte *use function*-Ansatz (Lan und DeMets, 1983), der ein attraktives Instrument für ein relativ flexibles Datenhandling darstellt. Vor dessen Implementation muß der maximale (gesamte) Stichprobenumfang spezifiziert werden. Während im ursprünglichen Ansatz die Anzahl der Zwischenauswertungen vorab festgelegt werden muß, muß bei diesem Ansatz lediglich eine *use function* oder α-*spending function* vorgegeben werden, die die Aufteilung der Wahrscheinlichkeit des Fehlers 1. Art auf die durch die Zwischenauswertungen festgelegten Stufen der Studie festlegt. Dieser Ansatz ist besonders in solchen Fällen zu verwenden, in de-

nen die Anzahl der Beobachtungseinheiten pro Gruppe unvorhersehbar ist. Dies ist beispielsweise dann der Fall, wenn ein Studienprotokoll die Zwischenauswertungen an bestimmten Zeitpunkten und nicht nach einer bestimmten Anzahl von Patienten vorschreibt. Ein gruppensequentieller Ansatz für beliebige Fallzahlen in den Gruppen bzw. Sequenzen der Studie wird auch in Wassmer (1999a) vorgeschlagen. Dieser Ansatz kann als Alternative zum *use function*-Ansatz betrachtet werden.

Ist man an einer Schätzung des Effekts interessiert, so ist zu unterscheiden, ob dies nur für den Fall, daß die Studie tatsächlich beendet worden ist, oder in jedem Fall (auch bei Nichtbeendigung der Studie) zu geschehen hat. Entsprechende Punktschätzer und Konfidenzintervalle wurden in der Literatur vorgeschlagen. Durch die Verwendung einer Stoppregel, d.h. der Möglichkeit der frühzeitigen Beendigung der Studie bei der Durchführung eines gruppensequentiellen Plans, sind Punktschätzer für interessierende Parameter im statistischen Sinne verfälscht. Dies bedeutet, daß auf lange Sicht mit einer Unter- oder Überschätzung der entsprechenden Parameter zu rechnen ist. Es gibt Versuche, unverfälschte Schätzer zu konstruieren, die die Stoppregeln des gruppensequentiellen Plans berücksichtigen, bzw. mittels numerischer Methoden mit dem Grad der Verfälschtheit (*estimation bias*) zu korrigieren (z. B. Coburger und Wassmer, 2000; Emerson, 1993; Emerson und Fleming, 1990; Emerson und Kittelson, 1997; Kim, 1989; Liu und Hall, 1999; Pinheiro und DeMets, 1997; Skovlund und Walløe, 1989; Todd, Whitehead und Facey, 1996; Todd und Whitehead, 1997; Whitehead, 1986). Vergleichende Untersuchungen existieren nicht bzw. in nur eingeschränktem Maße.

Die Konstruktion von Konfidenzintervallen wird durch die Tatsache erschwert, daß keine eindeutigen Regeln für die Konstruktion entsprechender Punktschätzer existieren. In der Literatur wurden zwei Methoden vorgeschlagen, die sich wesentlich unterscheiden. Die eine Methode ermöglicht die Bestimmung eines Konfidenzintervalls erst nach erfolgter Durchführung der Studie, d.h. nach Beenden des Experiments gemäß einer spezifizierten Stoppregel (z. B. Chang, 1989; Chang und O'Brien, 1986; Coad und Woodroofe, 1996; Cohen und Sackrowitz, 1996; Duffy und Santner, 1987; Kim und DeMets, 1987a; Rosner und Tsiatis, 1988; Todd et al., 1996; Tsiatis, Rosner und Mehta, 1984). Diese Verfahren hängen von der Ordnung im Stichprobenraum ab, d.h. der Frage, ob z.B.

ein bestimmter in der 1. Sequenz beobachteter Effekt „extremer“ als ein doppelt starker Effekt nach Beobachtung der 3. Sequenz zu gelten hat oder nicht. Eine alternative Vorgehensweise wurde von Jennison und Turnbull (1989) vorgeschlagen. Diese zweite Methode ergibt Konfidenzintervalle, die unabhängig von der Stoppregel verwendet werden und auch ohne Abbruch der Studie ermittelt und beispielsweise dem Studienkomitee in einem Zwischenbericht vorgelegt werden können. Die resultierenden sog. *Repeated Confidence Intervals* (RCI's) sind relativ einfach zu bestimmen, da lediglich die in den Testverfahren bereits berücksichtigte Multiplizität Eingang in die Konstruktion der Intervalle findet.

Eine weitere Möglichkeit für die Planung und Auswertung von gruppensequentiell konzipierten Studien ergibt sich aus Verfahren, die auf der *bootstrap*-Methode basieren. Dies wird aktuell von Scharfstein und Tsiatis (1998) vorgeschlagen. *Bootstrap*-Methoden zur Bestimmung von Konfidenzintervallen in gruppensequentiellen Designs wurden von Chuang und Lai (1998) eingeführt. Weitere Untersuchungen über diese Art der gruppensequentiellen Planung und Auswertung sind Gegenstand der aktuellen Forschungstätigkeiten.

Die grundlegende Schwierigkeit in der Berechnung der kritischen Schranken bei gruppensequentiellen Verfahren besteht in der Komplexität der Berechnung von Wahrscheinlichkeiten abhängiger Ereignisse. Die normalverteilte Daten voraussetzende Rekursionsformel von Armitage et al. (1969) benötigt entsprechende numerische Integrationsmethoden. Man ist auf nur kommerziell erwerbbare Spezialsoftware angewiesen, z.B die Softwarepakete „EaSt“ oder „PEST“, die bei Drucklegung der vorliegenden Auflage in den Versionen „EaSt 2000“ bzw. „PEST 4“ erhältlich sind. Ein ausführlicher Vergleich der Pakete PEST 3 sowie der Vorversion von EaSt 2000 findet sich in Emerson (1996). Neben diesen Produkten sind SAS/IML Routinen ab Version 6.11 verfügbar (SAS Institute Inc., 1989, 1995), mit deren Hilfe sich Programme für die Anwendung gruppensequentieller Pläne entwickeln lassen. In Wassmer (1999a) ist bereits beschrieben, wie der *use function*-Ansatz mit Hilfe dieser Routinen implementiert werden kann. Mit dem vom Autor dieser Arbeit entwickelten Programm ADDPLAN (Wassmer und Eisebitt, 2001) ist die Planung und Analyse insbesondere von adaptiv gruppensequentiellen Designs durchführbar.

Den in der Literatur im gruppensequentiellen Zusammenhang beschriebenen Verfahren ist gemein, daß die Ergebnisse der Zwischenauswertung(en) nicht in

die Planung der weiteren Studie einfließen dürfen. Insbesondere dürfen auch die Auswertungszeitpunkte nicht abhängig vom beobachteten Effekt gewählt werden. Die verfälschenden Effekte, die bei Nichtbeachtung der Regel der datenunabhängigen Weiterplanung resultieren, wurden in Proschan, Follmann und Waclawiw (1992) beschrieben. In den in dieser Arbeit betrachteten Fällen war kein übermäßig starker Effekt beobachtet worden. Allerdings bleibt unklar, ob sich dies in allgemeineren Fällen bestätigt.

Das adaptive Planen von klinischen Studien wird von deren theoretischer Konzeption her durch Verfahren ermöglicht, die erst kürzlich in der Literatur vorgeschlagen wurden. Im wesentlichen sind hier die in den Arbeiten von Bauer (1989a), Bauer und Köhne (1994), Bauer und Röhmel (1995), Bauer, Bauer und Budde (1998), Bauer und Kieser (1999) sowie von Proschan und Hunsberger (1995) beschriebenen Vorgehensweisen zu nennen. In Bauers Konzept wird durch die Kombination der separaten p-Werte der einzelnen Sequenzen eine *overall*-Teststatistik erzeugt, die die adaptive Planung ermöglicht. Als Kombinationsregel wird die auf Fisher zurückgehende Produktregel verwendet. Dies ist im wesentlichen ein nichtparametrischer Ansatz, der immer dann zu exakten Testverfahren führt, solange die zugrundeliegenden p-Werte auf exakten Verfahren beruhen. Im Gegensatz dazu entwickelten Proschan und Hunsberger ein Verfahren, das normalverteilte Zielvariablen voraussetzt. Die Anwendung auf andere Verteilungssituationen beruht auf asymptotischen Überlegungen. Die kritischen Werte werden gemäß einer Methode gefunden, die auf der Betrachtung des für die Wahrscheinlichkeit für den Fehler 1. Art „ungünstigsten“ Falls basiert. Darauf aufbauend bietet deren Konzept der Spezifikation einer *conditional error function* eine allgemeine Möglichkeit zur Konstruktion von adaptiven Plänen. Wassmer (1998) konnte zeigen, daß die beiden Verfahren im Fall normalverteilter Zielvariablen zu nahezu identischen Entscheidungsregeln mit vergleichbarer *power* und durchschnittlichen Stichprobenumfängen führt. Bauers Konzept ist für bis zu höchstens zwei adaptive Zwischenauswertungen vorgeschlagen, im Ansatz von Proschan und Hunsberger ist lediglich eine einzige Zwischenauswertung vorgesehen. In der praktischen Anwendung sind jedoch durchaus mehr adaptive Planungsmöglichkeiten denkbar, was zu adaptiven Designs mit mehreren (> 3) Sequenzen führt. Erweiterungsmöglichkeiten des Produktregel-Ansatzes auf beliebig viele Zwischenauswertungen werden in Wassmer (1999b) beschrieben. Einige Resultate über die *power* von Fishers Kombinationstest und deren

numerische Berechenbarkeit finden sich in Wassmer (1997). Die Bestimmung von Konfidenzintervallen in adaptiv geplanten Studien wird in Wassmer und Lehmacher (1997) behandelt (vgl. auch Frick, 2000; Wassmer, Eisebitt und Coburger, 2001). In Lehmacher und Wassmer (1999) ist eine Methode der Durchführung adaptiv geplanter Studien beschrieben, in der die klassischen gruppensequentiellen Designvarianten Verwendung finden können (vgl. auch Wassmer et al., 2001). Dieser Ansatz ist daher als eine Verknüpfung adaptiver Methoden mit rein gruppensequentiellen Verfahren zu sehen.

Neben diesen Ansätzen sind im Zusammenhang mit adaptiven Plänen die Verfahren zu nennen, die auf der Behauptung beruhen, daß die Wahrscheinlichkeit für den Fehler 1. Art nicht wesentlich durch die adaptive Planung der Studie beeinflußt wird. Erste Untersuchungen wurden von Hayre (1985) durchgeführt. Wie oben bereits erwähnt, ist diese Vorgehensweise vom theoretischen Standpunkt her unbefriedigend, und weitere Resultate sind erstrebenswert. Es sind auch die Arbeiten von Gould zu nennen (vgl. Gould und Shih, 1998), die explizite Verfahrensweisen für die Schätzung der Variabilität und damit der Bestimmung der Stichprobenumfänge in den Zwischenauswertungen beinhalten. Kieser und Friede (2000b) schlagen ein Verfahren vor, das bei der Schätzung der Variabilität das Niveau exakt einhält (vgl. auch Denne und Jennison, 1999; Proschan und Wittes, 2000; Wittes, Schabenberger, Zucker, Brittain und Proschan, 1999; Zucker, Wittes, Schabenberger und Brittain, 1999).

Die Idee der adaptiven Planung und Auswertung von klinischen Studien wird in der neuesten Literatur auch in Form von sogenannten „*self designing clinical trials*“ (Fisher, 1998; Shen und Fisher, 1999) vorgeschlagen. Sämtliche zur Verfügung stehende Information kann dazu benutzt werden, die Art der Gewichtung der Daten der folgenden Sequenzen der Studie festzulegen. Dabei ist allerdings ein „positives“ Studienresultat (Ablehnung von H_0) erst am Ende der Studie vorgesehen. Dieses Verfahren ist erweiterbar. Insbesondere ist zu klären, inwieweit die vorgeschlagenen Methoden mit den oben erwähnten Verfahren der adaptiven Planung und Auswertung klinischer Studien zu verbinden sind (vgl. Posch und Bauer, 1999).

Überblick über die Arbeit

Diese Arbeit enthält eine Beschreibung der statistischen Testverfahren, die eine

gruppensequentielle und/oder adaptive Planung von klinischen Studien ermöglichen. Diese werden auf einem einheitlichen und soliden theoretischen Fundament dargestellt, so daß das Rationale und die Herleitung der Verfahren klar nachvollziehbar sind. Besonderer Wert wird auf die Darstellung der vom Autor dieser Arbeit vorgeschlagenen Verfahren und Untersuchungen gelegt, die im Rahmen einer mehrjährigen Forschungstätigkeit am Institut für Medizinische Statistik, Informatik und Epidemiologie (IMSIE) der Universität zu Köln entstanden und teilweise bereits publiziert wurden. Auf eine Behandlung von Schätzverfahren wird verzichtet, da sie den Rahmen der Arbeit sprengen würde.

Die Arbeit gliedert sich in zwei Blöcke. Der erste Block behandelt Testverfahren für gruppensequentielle Pläne. Nach der Darstellung der allgemeinen Konstruktionsmethode zur Herleitung der kritischen Schranken für den Parallelgruppenvergleich der Mittelwerte normalverteilter Merkmale werden Verfahren vorgestellt, die auf der Annahme identischer Stichprobenumfänge in den Sequenzen der Studie beruhen. Für den allgemeineren Fall ungleicher Sequenzgrößen werden neben dem Fall, daß die Sequenzgrößen fest vorgegeben sind, zwei Verfahren für den Fall beliebiger Sequenzgrößen vorgestellt. Zuerst wird eine vom Autor dieser Arbeit vorgeschlagene *worst case scenario*-Lösung beschrieben, die auf der Betrachtung des für die Fehlerwahrscheinlichkeit 1. Art ungünstigsten Falles beruht. Dieser Ansatz wird daraufhin mit dem inzwischen recht populär gewordenen *use function*-Ansatz verglichen. Ein Literaturüberblick über die Verwendung dieser Designs in anderen Studientypen als dem Parallelgruppenvergleich normalverteilter Merkmale beendet dieses Kapitel.

Der zweite Block beinhaltet die Beschreibung der in der Literatur vorgeschlagenen adaptiven Testverfahren und einige neue bis dato nicht publizierten Ergebnisse. Für den zweistufigen Fall, d.h. die Durchführung von einer einzigen Zwischenauswertung, werden die Verfahren von Bauer und Kollegen sowie von Proschan und Hunsberger vorgestellt und verglichen. Eine vom Autor der vorliegenden Arbeit durchgeführte Verallgemeinerung dieser Verfahren auf mehr als zwei (bzw. drei) Stufen wird dargestellt und diskutiert. Schließlich wird die von Lehmacher und Wassmer (1999) vorgeschlagene Lösung beschrieben und deren Anwendbarkeit in bezug auf Gütebetrachtungen und praktische Durchführbarkeit behandelt. Dieses Verfahren stellt gewissermaßen das „Sahnehäubchen“ dieser Schrift dar, da es die beiden Blöcke in sehr eindrucksvoller Weise

kombiniert und zur Anwendung empfehlen kann.

Bei der Beschreibung der Verfahren wird auf die praktische Anwendung und die konkrete Umsetzung in klinischen Studien Wert gelegt. Die numerische Berechenbarkeit der Verfahren wird transparent gehalten, indem die zu den Ergebnissen dieser Arbeit führenden SAS-Programme im Anhang dokumentiert sind. Neben den schon recht ausführlichen und für die meisten praktische Fälle ausreichenden Tabellen im Hauptteil des Textes können diese Programme beispielsweise dazu benutzt werden, für beliebige Vorgaben des Signifikanzniveaus α die Entscheidungsbereiche eines gruppensequentiellen oder adaptiven Testverfahrens zu bestimmen.

2 Gruppensequentielle Testverfahren

In der klassischen, auf Neyman und Pearson (1928) zurückgehenden Testtheorie ist der Stichprobenumfang fest vorgegeben. Zum Testen einer Nullhypothese H_0 gegen eine Alternativhypothese H_1 ermittelt man die Entscheidungsbereiche derart, daß die irrtümliche Ablehnung von H_0 (Fehler 1. Art) mit einer a priori festgelegten Irrtumswahrscheinlichkeit α unter Kontrolle ist. Unter den so gegebenen Niveau-α-Tests ist derjenige Test besser als ein anderer, der eine geringere Wahrscheinlichkeit, die Hypothese H_0 fälschlicherweise beizubehalten, d.h. eine höhere *power* besitzt. Dies ist das klassische frequentistische Testprinzip, das zur Durchführung von Signifikanztests führt, bei der die irrtümliche Ablehnung von H_0 mit dem *Signifikanzniveau* α unter Kontrolle ist. Dieses Testprinzip hat eine weite Verbreitung gefunden und findet in randomisierten und kontrollierten klinischen Studien fast ausnahmslos Verwendung. Insbesondere läßt sich durch die in Statistik-Softwarepaketen bestimmten p-Werte (Überschreitungswahrscheinlicheiten) eine sehr einfach durchführbare Entscheidungsregel angeben, die lautet: Lehne H_0 ab, falls $p \leq \alpha$. Nicht nur wegen dieser Eigenschaft, sondern auch wegen dem den p-Werten innewohnenden *explorativen* Charakter werden diese – auch in nicht kontrollierten Studien – so häufig angewendet.

Ist der Stichprobenumfang nicht fest gegeben und eine sequentielle Studie geplant, so führt eine Verwendung der nicht-sequentiellen Testverfahren bzw. der damit zusammenhängenden Entscheidungsbereiche zu einer Erhöhung der tatsächlichen Fehlerwahrscheinlichkeiten. Das Niveau α wird nicht mehr eingehalten, da eine irrtümliche Ablehnung von H_0 nicht mehr nur einmal, sondern prinzipiell in jeder „Sequenz“ der Studie geschehen kann. Somit muß eine Adjustierung der Entscheidungsbereiche durchgeführt werden. Die gängigen gruppensequentiellen Pläne beruhen prinzipiell auf dieser Adjustierung und unterscheiden sich u.a. in der Aufteilung der Fehlerwahrscheinlichkeit 1. Art auf die Sequenzen der Studie. Bei gruppensequentiellen Testverfahren werden wiederholte Signifikanztests durchgeführt, und in der Regel geht lediglich die Kontrolle

des Fehlers 1. Art in die Konstruktion dieser Tests ein. Die Güte der Verfahren wird neben der *power* an dem durchschnittlich benötigten Stichprobenumfang beurteilt. Dies wird deshalb betont, da dies im Gegensatz zu alternativen sequentiellen Testverfahren steht, bei denen per definitionem *beide* Fehlerwahrscheinlichkeiten in die Bestimmung der Testprozedur eingehen und diese daher nicht mehr als reine Signifikanztests zu bezeichnen sind. Diese Verfahren sind auch häufig dadurch charakterisiert, daß nach jeder Beobachtungseinheit eine Testentscheidung getroffen werden kann. Das bekannteste Verfahren dieser Art geht auf den „Gründervater“ der sequentiellen Verfahren Abraham Wald (Wald, 1947) zurück. Es handelt sich dabei um den *Sequential Probability Ratio Test (SPRT)*. Der *SPRT* besitzt zwar mathematische Optimalitätseigenschaften, er wird in der Praxis aber nur selten verwendet. Insbesondere gilt für die nach dieser Vorschrift ermittelten Tests, daß sie den kleinstmöglichen durchschnittlichen Stichprobenumfang besitzen und dies sowohl bei Gültigkeit von H_0 wie bei Gültigkeit von H_1. Darüber hinaus ist bei Verwendung sogenannter offener Sequentialpläne, bei denen kein maximaler Stichprobenumfang vorgegeben wird, der tatsächliche Stichprobenumfang mit Wahrscheinlichkeit 1 endlich. Eine ausführliche Behandlung dieses Verfahrens und die Darstellung der weit ausgebauten mathematischen Theorie der sequentiellen statistischen Tests finden sich in den einschlägigen Monographien (z.B. Bauer, Scheiber und Wohlzogen, 1986; Ghosh, 1970; Siegmund, 1985; Wetherill, 1975; Whitehead, 1997).

In den nächsten Abschnitten wird eine Übersicht über gruppensequentielle Pläne gegeben, die seit der Arbeit von Armitage et al. (1969) in der statistischen Literatur vorgeschlagen und diskutiert wurden. Bedingt durch die Tatsache, daß es sich dabei um relativ „moderne“ Verfahren handelt, sind diese in den oben erwähnten Monographien über sequentielle Verfahren eher nur am Rande oder gar nicht erwähnt. Kurz vor Drucklegung der vorliegenden Arbeit erschien das Buch von Jennison und Turnbull (2000), das ein umfangreiches Kompendium über gruppensequentielle Verfahren und deren Anwendung in klinischen Studien darstellt. Auch Whitehead (1997) behandelt gruppensequentielle Pläne in einem eigenem Kapitel und beschreibt neuere Entwicklungen. Die aus medizinischer Sicht relevanten Aspekte zur Durchführung von wiederholten Signifikanztests werden ebenso in Armitage (1975) beschrieben. Köpcke (1984) enthält explizit die heute als klassisch geltenden gruppensequentiellen Pläne und eigene Weiterentwicklungen. Dieses Buch ist nach wie vor speziell für den deutschsprachigen

Raum eine empfehlenswerte Referenz.

2.1 Die allgemeine Konstruktionsmethode

Dieser Abschnitt beinhaltet die allgemeine Konstruktionsmethode und behandelt die von Armitage et al. (1969) und McPherson und Armitage (1971) eingeführte Vorgehensweise zur Bestimmung der Entscheidungsbereiche eines gruppensequentiellen Planes. Die Methode zur Konstruktion dieser Verfahren wird hier wie in den folgenden Abschnitten für den Parallelgruppenvergleich der Mittelwerte normalverteilter Merkmale beschrieben, in dem beispielsweise zwei Therapien bzgl. ihrer Effektivität miteinander verglichen werden. Im nichtsequentiellen Design wird dabei eine unabhängige Stichprobe $X_{11}, X_{12}, \ldots, X_{1n_1}$ eines $N(\mu_1, \sigma_1^2)$-verteilten Merkmals und davon unabhängig von einem $N(\mu_2, \sigma_2^2)$-verteilten Merkmal eine unabhängige Stichprobe $X_{21}, X_{22}, \ldots, X_{2n_2}$ gezogen. Die Stichprobenumfänge n_1 und n_2 der beiden Gruppen sind dabei fest vorgegeben, und eine Entscheidung geschieht nach der vollständigen Beobachtung der Erhebung. In der Regel interessiert man sich für den Vergleich der Mittelwerte μ_1 und μ_2, d.h. es soll entschieden werden, ob man die Hypothese

$$\mathrm{H}_0 : \ \mu_1 - \mu_2 = 0 \tag{2.1}$$

aufgrund einer vorliegenden Stichprobe zum Signifikanzniveau α verwerfen kann. Die Varianzen σ_1^2 und σ_2^2 seien als bekannt vorausgesetzt. Je nachdem, ob man gegen eine zweiseitige Alternative

$$\mathrm{H}_1 : \ \mu_1 - \mu_2 \neq 0$$

oder gegen eine einseitige Alternative

$$\mathrm{H}_1 : \ \mu_1 - \mu_2 \overset{(<)}{>} 0$$

testet, sind die Bedingungen an die kritischen Werte durch

$$\begin{aligned} P_{\mathrm{H}_0}(|\,Z\,| \geq u) &= \alpha \quad \text{(im zweiseitigen Fall)} \quad \text{bzw.} \\ P_{\mathrm{H}_0}(Z \overset{(\leq)}{\geq} u) &= \alpha \quad \text{(im einseitigen Fall)} \end{aligned} \tag{2.2}$$

gegeben, wobei

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \tag{2.3}$$

die unter H_0 standardnormalverteilte Prüfgröße des Tests bezeichnet. Die kritischen Werte u ergeben sich aus der Quantilfunktion $\Phi^{-1}(\cdot)$ der Standardnormalverteilung und sind gemäß (2.2) durch $u = \Phi^{-1}(1 - \alpha/2)$ im zweiseitigen und $u = (-)\Phi^{-1}(1 - \alpha)$ im einseitigen Fall gegeben.

Im gruppensequentiellen Design wird zugelassen, daß nach der Erhebung einer bestimmten Anzahl von Probanden pro Therapiegruppe die Möglichkeit besteht zu entscheiden, ob die Studie weitergeführt wird oder nicht. O.B.d.A. wird dies für den Fall erläutert, daß pro Therapiegruppe gleich viele Beobachtungen zugrundeliegen und die Varianzen gleich groß sind, d.h. $n_1 = n_2$ und $\sigma_1^2 = \sigma_2^2 =: \sigma^2$. Werden pro Sequenz („Stufe“) k des gruppensequentiellen Plans n_k Beobachtungen pro Therapiegruppe vereinbart, so kann pro Stufe k in Analogie zu (2.3) die Statistik

$$Z_k = \frac{\bar{X}_{1k} - \bar{X}_{2k}}{\sigma}\sqrt{\frac{n_k}{2}}$$

berechnet werden. Dabei wird n_k, eine maximale Anzahl K von Sequenzen und damit die Folge der Stichprobenumfänge $n_1, n_2, \ldots, n_K$ und der maximale Stichprobenumfang $N = \sum_{k=1}^{K} n_k$ pro Therapiegruppe als gegeben vorausgesetzt. Da die Stichproben der einzelnen Sequenzen voneinander unabhängig sind, sind dies auch die Statistiken $Z_1, Z_2, \ldots, Z_K$. Die Entscheidungsregel basiert auf der bis zu einer bestimmten Stufe k verfügbaren Information. Wird mit

$$Z_k^* = \frac{\frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \sum_{\tilde{k}=1}^{k} n_{\tilde{k}} (\bar{X}_{1\tilde{k}} - \bar{X}_{2\tilde{k}})}{\sigma}\sqrt{\frac{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}{2}}$$

die standardisierte *overall* Statistik bezeichnet, die die gesamte Stichprobenin-

formation bis zur Sequenz k zusammenfaßt, so gilt:

$$Z_k^* = \frac{\sum_{\bar{k}=1}^{k} \sqrt{n_{\bar{k}}} Z_{\bar{k}}}{\sqrt{\sum_{\bar{k}=1}^{k} n_{\bar{k}}}} \tag{2.4}$$

und

$$E(Z_k^*) = \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\sum_{\bar{k}=1}^{k} \frac{n_{\bar{k}}}{2}} \, . \tag{2.5}$$

Offensichtlich sind die Statistiken $Z_1^*, Z_2^*, \ldots, Z_K^*$ voneinander abhängig. Die Kovarianz zwischen Z_k^* und $Z_{k'}^*$ ergibt sich zu

$$\begin{aligned} & Cov(Z_k^*, Z_{k'}^*) = \\ & \frac{1}{\sqrt{\sum_{\bar{k}=1}^{k} n_{\bar{k}}} \sqrt{\sum_{\bar{k}=1}^{k'} n_{\bar{k}}}} Cov(\sum_{\bar{k}=1}^{k} \sqrt{n_{\bar{k}}} Z_{\bar{k}}, \sum_{\bar{k}=1}^{k'} \sqrt{n_{\bar{k}}} Z_{\bar{k}}) = \\ & \frac{1}{\sqrt{\sum_{\bar{k}=1}^{k} n_{\bar{k}}} \sqrt{\sum_{\bar{k}=1}^{k'} n_{\bar{k}}}} \sum_{\bar{k}=1}^{k} n_{\bar{k}} = \\ & \frac{\sqrt{\sum_{\bar{k}=1}^{k} n_{\bar{k}}}}{\sqrt{\sum_{\bar{k}=1}^{k'} n_{\bar{k}}}}, \text{ falls } k \leq k' \, . \end{aligned} \tag{2.6}$$

Die gemeinsame Verteilung von $Z_1^*, Z_2^*, \ldots, Z_K^*$ ist die multivariate Normalverteilung mit dem durch (2.5) spezifizierten Erwartungswertvektor und der sich durch (2.6) ergebenden (regulären) Kovarianz- bzw. Korrelationsmatrix $\mathbf{\Sigma}$ mit

Elementen

$$\mathbf{\Sigma}_{kk'} = Cov(Z_k^*, Z_{k'}^*) = \frac{\sqrt{\sum_{\bar{k}=1}^{\min\{k,k'\}} n_{\bar{k}}}}{\sqrt{\sum_{\bar{k}=1}^{\max\{k,k'\}} n_{\bar{k}}}} .$$

In einer Studie könnte ein zweistufiges gruppensequentielles Verfahren mit $n_1 = n_2$ beispielsweise so geplant sein, daß in der ersten Stufe die Hypothese H_0 verworfen werden soll, falls $|Z_1^*| \geq 1.96$. Ist $|Z_1^*| < 1.96$, so wird die zweite Sequenz erhoben und H_0 nach deren Durchführung abgelehnt, falls $|Z_2^*| \geq 1.96$. Die Wahrscheinlichkeit für den Fehler 1. Art dieser Testprozedur ist gegeben durch

$$\begin{aligned} &P_{\mathrm{H}_0}(|Z_1^*| \geq 1.96 \text{ oder } (|Z_1^*| < 1.96 \text{ und } |Z_2^*| \geq 1.96)) = \\ &\underbrace{P_{\mathrm{H}_0}(|Z_1^*| \geq 1.96)}_{=0.05} + \underbrace{P_{\mathrm{H}_0}(|Z_1^*| < 1.96 \text{ und } |Z_2^*| \geq 1.96)}_{(*)} = \\ &0.0831 > 0.05 , \end{aligned} \tag{2.7}$$

wobei die Wahrscheinlichkeit $(*)$ mit der Verteilungsfunktion der bivariaten Normalverteilung mit Korrelation $1/\sqrt{2}$ berechnet werden kann. Dies kann beispielsweise in SAS mit der seit Version 6.11 zur Verfügung stehenden Funktion PROBBNRM geschehen.

Bei einem Signifikanzniveau von $\alpha = 0.05$ wird damit die Niveaubedingung nicht erfüllt. Deshalb muß eine entsprechend größere kritische Grenze verwendet werden, die als 2.178 gewählt werden kann. Denn dann ist

$$P_{\mathrm{H}_0}(|Z_1^*| \geq 2.178 \text{ oder } (|Z_1^*| < 2.178 \text{ und } |Z_2^*| \geq 2.178)) = 0.05 ,$$

und der Test schöpft das Niveau α voll aus. Dieser gruppensequentielle Plan mit konstanten kritischen Schranken $u_1 = u_2$ für Z_1^* und Z_2^* wurde von Pocock (1977) vorgeschlagen. Prinzipiell ist diese „konstante“ Aufteilung der kritischen Werte jedoch willkürlich, denn z.B. auch der von O'Brien und Fleming (1979) vorgeschlagene sequentielle Plan mit $u_1 = 2.797$ und $u_2 = 1.977$ und beliebig viele andere Aufteilungen erfüllen die Niveaubedingung.

Allgemein ist ein sequentielles Verfahren durch Bereiche $\mathscr{C}_k$, $k = 1, 2, \ldots, K-1$, im Stichprobenraum spezifiziert, in denen die Studie fortgesetzt wird. Bezeichnet zusätzlich $\mathscr{C}_K$ den Annahmebereich für H_0 in der K-ten Sequenz, so ist die Wahrscheinlichkeit für den Fehler 1. Art gegeben durch

$$1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{Z_k^* \in \mathscr{C}_k\}) \; . \tag{2.8}$$

Alle sequentiellen Pläne, für die (2.8) höchstens gleich α ist, sind valide Niveau-α-Tests. Entsprechend ist die Güte oder *power* des Testverfahrens bei geeignet spezifizierter Alternativhypothese H_1 gegeben durch

$$1 - P_{\mathrm{H}_1}(\bigcap_{k=1}^{K} \{Z_k^* \in \mathscr{C}_k\}) \; . \tag{2.9}$$

Der bis zu der Testentscheidung benötigte Stichprobenumfang ist nicht fest, sondern zufällig, und die Güte eines sequentiellen Verfahrens wird neben der *power* am durchschnittlich benötigten Stichprobenumfang (*Average Sample Number ASN)* beurteilt. Der ASN berechnet sich aus einem der folgenden alternativ zu verwendenden Ausdrücke:

$$\begin{aligned} ASN \;\; = \;\; & n_1 + \sum_{k=2}^{K} n_k P(\bigcap_{\bar{k}=1}^{k-1} \{Z_{\bar{k}}^* \in \mathscr{C}_{\bar{k}}\}) = \\ & N - \sum_{k=1}^{K-1} (N - \sum_{\bar{k}=1}^{k} n_{\bar{k}}) \, P(Z_k^* \in \overline{\mathscr{C}}_k \cap \bigcap_{\bar{k}=1}^{k-1} \{Z_{\bar{k}}^* \in \mathscr{C}_{\bar{k}}\}) \; , \end{aligned} \tag{2.10}$$

wobei $\overline{\mathscr{C}}_k$ das Komplement des Fortsetzungsbereichs $\mathscr{C}_k$ bezeichnet, und die Wahrscheinlichkeiten sowohl unter H_0 wie unter H_1 berechnet werden können.

Zur Bestimmung eines gruppensequentiellen Plans müssen die Größen (2.8) – (2.10) prinzipiell mit der Berechnung von Wahrscheinlichkeiten der multivariaten Normalverteilung ermittelt werden. Armitage et al. (1969) gaben jedoch eine bei der Durchführung wiederholter Signifikanztests spezifische Darstellungsform an, die die Berechenbarkeit des multivariaten Normalintegrals entscheidend vereinfacht. Im allgemeinen Fall (d.h. $n_1, n_2, \ldots, n_K$ und μ_1, μ_2 beliebig)

ergibt sich diese Darstellung zweckmäßigerweise aus der Verteilung der Statistik

$$S_k = \frac{\sum_{\bar{k}=1}^{k} \sqrt{n_{\bar{k}}} Z_{\bar{k}}}{\sqrt{n_1}} \ . \tag{2.11}$$

Man erhält hierdurch $Var(S_1) = 1$ und

$$Var(S_k) = Var(S_{k-1}) + \tau_k \ , \tag{2.12}$$

wobei $\tau_k := \frac{n_k}{n_1}$, $k = 2, 3, \ldots, K$, das auf n_1 „standardisierte Zeitintervall“ zwischen $(k-1)$-ter und k-ter Stufe bezeichnet und damit zur Zeit $\tau_1 \equiv 1$ die erste Zwischenauswertung durchgeführt wird. Durch

$$\begin{aligned} E(S_k) \quad = \quad & \frac{\mu_1 - \mu_2}{\sigma\sqrt{2}} \frac{1}{\sqrt{n_1}} \sum_{\bar{k}=1}^{k} n_{\bar{k}} = \\ & \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\sum_{\bar{k}=1}^{k} \frac{n_{\bar{k}}}{2}} \sqrt{\sum_{\bar{k}=1}^{k} \tau_{\bar{k}}} \ = \\ & \frac{\mu_1 - \mu_2}{\sigma} \sqrt{\frac{n_1}{2}} \sum_{\bar{k}=1}^{k} \tau_{\bar{k}} \ =: \ \vartheta_k \end{aligned} \tag{2.13}$$

ist der Nichtzentralitätsparameter ϑ_k für die Verteilung von S_k unter einer spezifizierten Alternative H_1 gegeben. Die Kovarianz zwischen S_k und $S_{k'}$ ergibt sich zu

$$\mathbf{\Sigma}_{kk'} = Cov(S_k, S_{k'}) = \sum_{\bar{k}=1}^{k^*} \tau_{\bar{k}} \ , \text{ wobei } k^* = \min\{k, k'\} \ .$$

Es ist offensichtlich, daß für die Fortsetzungsbereiche $\mathscr{C}_k$ für Z_k^* und die Fortsetzungsbereiche $\breve{\mathscr{C}}_k$ für S_k die Beziehung

$$\breve{\mathscr{C}}_k = \sqrt{\frac{\sum_{\bar{k}=1}^{k} n_{\bar{k}}}{n_1}} \mathscr{C}_k = \sqrt{\sum_{\bar{k}=1}^{k} \tau_{\bar{k}}} \, \mathscr{C}_k \tag{2.14}$$

gilt, wobei die lineare Transformation des Intervalls $\mathscr{C}_k$ elementweise zu verstehen ist.

Mit $f_k(s_k, \vartheta_k)$, $k = 1, 2, \ldots, K$, wird die Dichte der Verteilung von S_k im gruppensequentiellen Design verstanden. $f_k(s_k, \vartheta_k)$ bezeichnet also – etwas grob formuliert – die Dichte der gemeinsamen Verteilung von S_k und einer Zufallsgröße, deren Verteilung die Wahrscheinlichkeit angibt, bis zur Stufe $k-1$ zu keiner Testentscheidung gelangt zu sein. Bezeichnet $\phi(\cdot)$ die Dichtefunktion der Standardnormalverteilung, so gilt für $f_k(s_k, \vartheta_k)$ die folgende rekursive Form (vgl. auch SAS Institute Inc., 1995, S. 54):

$$f_k(s_k, \vartheta_k) =$$

$$\begin{cases} \displaystyle\int_{\breve{\mathscr{C}}_{k-1}} f_{k-1}(s_{k-1}, \vartheta_{k-1}) \; \frac{1}{\sqrt{\tau_k}} \phi\left(\frac{s_k - s_{k-1} - \vartheta_k + \vartheta_{k-1}}{\sqrt{\tau_k}} \right) \, ds_{k-1} & \\ & \text{falls } k = K \text{ oder } s_k \notin \breve{\mathscr{C}}_k \text{ für } k = 2, \ldots, K-1 \\ 0 & \text{sonst,} \end{cases} \tag{2.15}$$

$k = 2, 3, \ldots, K$, wobei ϑ_k den in (2.13) definierten Nichtzentralitätsparameter bezeichnet. $f_1(s_1, \vartheta_1)$ ist durch

$$f_1(s_1, \vartheta_1) = \phi(s_1 - \vartheta_1), \text{ falls } s_1 \notin \breve{\mathscr{C}}_1$$

gegeben. Man beachte, daß sich aus (2.15)

$$\sum_{k=1}^{K} \int_{-\infty}^{\infty} f_k(s_k, \vartheta_k) \, ds_k = 1$$

für alle ϑ_k ergibt.

Die Gültigkeit von (2.15) folgt aus der wiederholten Verwendung des Satzes von Fubini und des Transformationssatzes für Dichten (vgl. z.B. Pfanzagl, 1991, S. 51ff.). Bezeichnet man mit $g(s_1, s_2, \ldots, s_K)$ die Dichte der gemeinsamen Verteilung von $S_1, S_2, \ldots, S_K$ unter H_0 (d.h. unter der Annahme $\vartheta_k = 0$, $k = 1, 2, \ldots, K$), so gilt:

$$P\Big(\bigcap_{k=1}^{K} \{S_k \in \breve{\mathscr{C}}_k\}\Big) =$$

$$\int_{\breve{\mathscr{C}}_K} \int_{\breve{\mathscr{C}}_{K-1}} \cdots \int_{\breve{\mathscr{C}}_1} g(s_1, \ldots, s_{K-1}, s_K) \, ds_1 \cdots ds_{K-1} \, ds_K =$$

$$\int\limits_{\{z_K:\sqrt{\tau_K}z_K\in\breve{\mathscr{C}}_K-s_{K-1}\}}\int\limits_{\breve{\mathscr{C}}_{K-1}}\cdots\int\limits_{\breve{\mathscr{C}}_1} g(s_1,\ldots,s_{K-1},z_K)\,ds_1\cdots ds_{K-1}dz_K\ , \tag{2.16}$$

da

$$s_K\in\breve{\mathscr{C}}_K\Leftrightarrow\sqrt{n_K}z_K\in\sqrt{n_1}\breve{\mathscr{C}}_K-\sum_{k=1}^{K-1}\sqrt{n_k}z_k\Leftrightarrow\sqrt{\tau_K}z_K\in\breve{\mathscr{C}}_K-s_{K-1}\ .$$

Das Integral (2.16) ist wegen der Unabhängigkeit von Z_K und $S_1,\ldots,S_{K-1}$ identisch

$$\int\limits_{\breve{\mathscr{C}}_K}\int\limits_{\breve{\mathscr{C}}_{K-1}}\cdots\int\limits_{\breve{\mathscr{C}}_1} g(s_1,\ldots,s_{K-1})\,\frac{1}{\sqrt{\tau_K}}\phi(\frac{z_K-s_{K-1}}{\sqrt{\tau_K}})\,ds_1\cdots ds_{K-1}dz_K\ .$$

Eine sukzessive Wiederholung dieses Vorgehens und die aus der Normalverteilungseigenschaft folgende Identität

$$f(s_k,\vartheta_k)=f(s_k-\vartheta_k,0),\ k=1,2,\ldots,K\,,$$

validiert die Gültigkeit von (2.15).

□

Setzt man $\boldsymbol{\vartheta}=(\vartheta_1,\vartheta_2,\ldots,\vartheta_K)=\mathbf{0}$, so ist die Wahrscheinlichkeit für den Fehler 1. Art gegeben durch

$$1-\int_{\breve{\mathscr{C}}_K} f_K(s_K,0)\,ds_K\ , \tag{2.17}$$

und entsprechend die *power* für gegebenes ϑ durch

$$1-\int_{\breve{\mathscr{C}}_K} f_K(s_K,\vartheta_K)\,ds_K\ . \tag{2.18}$$

Der ASN ist durch (2.10) gegeben, indem man

$$P(\bigcap_{\bar{k}=1}^{k-1}\{Z_{\bar{k}}^*\in\mathscr{C}_{\bar{k}}\})=\int_{\breve{\mathscr{C}}_{k-1}} f_{k-1}(s_{k-1},\vartheta_{k-1})\,ds_{k-1} \tag{2.19}$$

berechnet. Für ungleiche Stichprobenumfänge muß dabei bei gegebenen τ_k, $k = 2, 3, \ldots, K$, das entsprechende Integral ermittelt werden.

Die Dichte (2.15) sowie deren „finale“ Integration lassen sich durch geeignete numerische Integrationsroutinen (wie z.B. die Newton-Cotes-Methode) berechnen, was von den Protagonisten gruppensequentieller Pläne in geeigneten Programmiersprachen (z.B. FORTRAN, vgl. Reboussin, DeMets, Kim und Lan, 1995) umgesetzt wurde. Da die Implementierung der Verfahren typischerweise sehr zeit- und rechenintensiv ist, ist man als Anwender auf nur kommerziell erwerbbare Spezialsoftware angewiesen. Im wesentlichen sind zwei Softwarepakete für die Planung und Durchführung einer gruppensequentiell konzipierten Studie verfügbar. Zum einen „EaSt 2000“, entwickelt von K. Kim, C. Mehta, P. Pampallona, A. Tsiatis und S. Vasundhra (erhältlich von Cytel Software Cooperation); zum anderen „PEST, Version 4“, entwickelt von H. Brunier und J. Whitehead (erhältlich von MPS Research Unit, University of Reading). Neben diesen Produkten stellt SAS in SAS/IML ab Version 6.11 (SAS Institute Inc., 1995) das Modul SEQ zur Verfügung, mit dem die Integrale ebenso berechnet werden können und mit dessen Hilfe sich Programme für die Anwendung und Untersuchung gruppensequentieller Pläne entwickeln lassen (vgl. Wassmer und Biller, 1998, für eine deutschsprachige Einführung in SAS/IML). Alle in dieser Arbeit durchgeführten Berechnungen wurden damit durchgeführt (vgl. auch Wassmer, 1999a). Für die Planung und Auswertung von adaptiv gruppensequentiellen Testdesigns (vgl. Kapitel 3) wurde darüber hinaus das Paket ADDPLAN entwickelt (Wassmer und Eisebitt, 2001), das vom Autor dieser Arbeit erhältlich ist. Dieses Programm ermöglicht ebenso die Planung von klassisch gruppensequentiellen Plänen.

2.2 Verfahren mit gleichen Sequenzgrößen

Sind die Stichprobenumfänge $n_1, n_2, \ldots, n_K$ gleich groß und identisch n, so vereinfachen sich die Ausdrücke (2.4) – (2.6) zu

$$Z_k^* = \frac{1}{\sqrt{k}} \sum_{\bar{k}=1}^{k} Z_{\bar{k}} \ ,$$

$$
\begin{aligned}
E(Z_k^*) &= \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{kn}{2}}\,, \\
Cov(Z_k^*, Z_{k'}^*) &= \sqrt{\frac{k}{k'}}\,,\ \text{falls } k \le k'\,.
\end{aligned}
$$

Die Formel für die Dichten $f_k(s_k, \vartheta_k)$ der sich aus (2.11) ergebenden Statistiken

$$
S_k = \sum_{\bar{k}=1}^{k} Z_{\bar{k}}\,, k = 1, 2, \ldots, K\,, \tag{2.20}
$$

ergibt sich aus (2.15), indem man $\tau_2 = \tau_3 = \ldots = \tau_K = 1$ setzt. Für $k = 2, 3, \ldots, K$ gilt (vgl. (2.15)):

$f_k(s_k, \vartheta_k) =$

$$
\begin{cases}
\displaystyle\int_{\breve{\mathscr{C}}_{k-1}} f_{k-1}(s_{k-1}, \vartheta_{k-1})\ \phi\left(s_k - s_{k-1} - \vartheta_1\right)\ ds_{k-1} & \\
\qquad\qquad \text{falls } k = K \text{ oder } s_k \notin \breve{\mathscr{C}}_k \text{ für } k = 2, \ldots, K-1 & \\
0 \qquad\qquad \text{sonst,} &
\end{cases}
$$

und $f_1(s_1, \vartheta_1) = \phi(s_1 - \vartheta_1)$. Man beachte, daß in dieser Formulierung die Fortsetzungs- bzw. Annahmebereiche $\breve{\mathscr{C}}_k$, $k = 1, 2, \ldots, K$, wiederum für die Summenscores (2.20) und nicht für die standardisierten Summenscores Z_k^* spezifiziert sind. Entsprechend (2.12) – (2.14) gilt

$$
\begin{aligned}
Var(S_k) &= k\,, \\
E(S_k) = \vartheta_k &= \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{n}{2}}\,k\,, \\
\breve{\mathscr{C}}_k &= \sqrt{k}\,\mathscr{C}_k,\ k = 1, 2, \ldots, K\,.
\end{aligned} \tag{2.21}
$$

2.2.1 Klassische Verfahren

Pocock (1977) und O'Brien und Fleming (1979) schlugen für das zweiseitige Testproblem gruppensequentielle Pläne vor, die eine breite Akzeptanz für die medizinische Anwendung gefunden haben. Diese Verfahren sind dementsprechend populär und in diesem Sinne als „klassisch“ zu verstehen. Das von Pocock

vorgeschlagene Verfahren beruht auf einer konstanten Aufteilung der kritischen Schranken u_k für Z_k^*, während das von O'Brien und Fleming angegebene Verfahren monoton (in k) fallende kritische Werte verwendet. Die Entscheidungsbereiche für Z_k^* lauten:

- für Pococks Design:
  $\mathscr{C}_k = (-u;\ u)$ mit $u = c_{\mathrm{P}}(K, \alpha)$, $k = 1, 2, \ldots, K$,
- für O'Brien und Flemings Design:
  $\mathscr{C}_k = (-u_k;\ u_k)$ mit $u_k = c_{\mathrm{OBF}}(K, \alpha)/\sqrt{k}$, $k = 1, 2, \ldots, K$.

Entsprechend lauten die Entscheidungsbereiche für S_k:

- für Pococks Design:
  $\breve{\mathscr{C}}_k = (-\breve{u}_k;\ \breve{u}_k)$ mit $\breve{u}_k = c_{\mathrm{P}}(K, \alpha)\sqrt{k}$, $k = 1, 2, \ldots, K$,
- für O'Brien und Flemings Design:
  $\breve{\mathscr{C}}_k = (-\breve{u};\ \breve{u})$ mit $\breve{u} = c_{\mathrm{OBF}}(K, \alpha)$, $k = 1, 2, \ldots, K$.

Je nachdem, welches Verfahren verwendet werden soll, ist $c(K, \alpha)$ bei gegebenem K und α numerisch zu bestimmen, indem (2.8) identisch α gesetzt wird. Tabelle 2.1 enthält die Werte u_k, $k = 1, 2, \ldots, K$, und die für den Abbruch der Studie benötigten p-Werte nach O'Brien und Fleming bzw. Pocock für $\alpha = 0.05$ und $K = 2, 3, 4, 5$. Ausführlichere Tabellen werden in den folgenden Ausführungen präsentiert (vgl. z.B. Tabelle 2.2). Die Entscheidungsbereiche für die beiden Designs sind für $K = 5$ und $\alpha = 0.05$ in Abbildung 2.1 illustriert.

Bei gegebenem α, standardisiertem Effekt $\frac{\mu_1 - \mu_2}{\sigma}$, gegebener Stichprobengröße n pro Stufe des gruppensequentiellen Plans sowie maximaler Anzahl K der Stufen kann sowohl die Güte (*power*) $1-\beta$ wie der pro Therapiegruppe benötigte ASN berechnet werden. Umgekehrt kann bei Vorgabe von α, $1 - \beta$ und K aus (2.18) der Wert

$$\vartheta^* = \vartheta^*(\alpha, \beta, K) := \frac{\mu_1 - \mu_2}{\sigma}\sqrt{\frac{n}{2}} \tag{2.22}$$

bestimmt werden, für den bei Verwendung eines spezifischen K-stufigen gruppensequentiellen Plans zum Niveau α die Güte $1 - \beta$ beträgt. Man beachte, daß hieraus gemäß (2.21) der komplette Vektor $\boldsymbol{\vartheta}$ resultiert. Aus (2.22) ergibt sich

Tabelle 2.1: *Kritische Werte u_k für Z_k^* nach O'Brien und Fleming (1979) bzw. Pocock (1977) im zweiseitigen Testproblem; $\alpha = 0.05$. Die für den Abbruch der Studie benötigten p-Werte sind in Klammern angegeben.*

	K	u_1	u_2	u_3	u_4	u_5
O'Brien/	2	2.797	1.977			
Fleming		(0.0052)	(0.0480)			
	3	3.471	2.454	2.004		
		(0.0005)	(0.0141)	(0.0451)		
	4	4.049	2.863	2.337	2.024	
		(0.0001)	(0.0042)	(0.0194)	(0.0429)	
	5	4.562	3.226	2.634	2.281	2.040
		(<0.0001)	(0.0013)	(0.0084)	(0.0226)	(0.0413)
Pocock	2	2.178	2.178			
		(0.0294)	(0.0294)			
	3	2.289	2.289	2.289		
		(0.0221)	(0.0221)	(0.0221)		
	4	2.361	2.361	2.361	2.361	
		(0.0182)	(0.0182)	(0.0182)	(0.0182)	
	5	2.413	2.413	2.413	2.413	2.413
		(0.0158)	(0.0158)	(0.0158)	(0.0158)	(0.0158)

für die benötigte Fallzahl n pro Gruppe:

$$n = n(\alpha, \beta, K) = 2\left(\vartheta^* \frac{\sigma}{\mu_1 - \mu_2}\right)^2, \tag{2.23}$$

woraus sich aus (2.10), (2.19) und (2.21)

$$ASN = ASN(\alpha, \beta, K) = 2\left(\vartheta^* \frac{\sigma}{\mu_1 - \mu_2}\right)^2 \Big(1 + \sum_{k=1}^{K-1} \int_{\check{\mathscr{C}}_k} f_k(s_k, \vartheta^* k)\, ds_k\Big) \tag{2.24}$$

ergibt. Der ASN ist wie bei nichtsequentiellen Plänen (d.h. bei fest vorgegebenem Gesamt-Stichprobenumfang) proportional in $\left(\frac{\sigma}{\mu_1 - \mu_2}\right)^2$. Es genügt daher, den Fall $\frac{\sigma}{\mu_1 - \mu_2} = 1$ zu betrachten, um Eigenschaften eines gruppensequen-

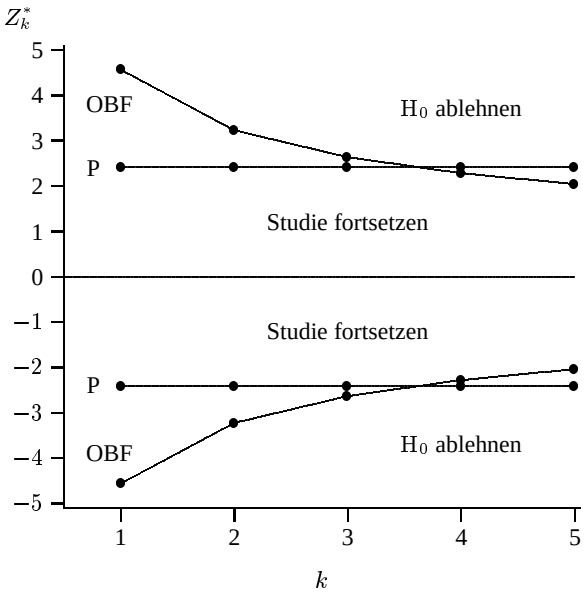

Abbildung 2.1: *Fortsetzungs- bzw. Annahmebereiche $\mathscr{C}_k$ für O'Brien und Flemings (OBF) bzw. Pococks (P) Design für den Fall gleicher Sequenzgrößen; $K = 5$, $\alpha = 0.05$.*

tiellen Testverfahren bzgl. des zu erwartenden Stichprobenumfangs zu beurteilen.

Offensichtlich ist der ASN bei gegebenem α und β monoton fallend in K, während der maximal benötigte Stichprobenumfang $K \cdot n(\alpha, \beta, K)$ in K steigt. Bei gruppensequentiellen Plänen mit $K = 5$ ist allerdings die maßgebliche Reduktion des ASN verglichen mit gruppensequentiellen Plänen mit $K < 5$ erreicht (Pocock, 1982; McPherson, 1982). Weitere Verkleinerungen des ASN für $K > 5$ sind für praktisch relevante Fälle eher marginal und nicht von Bedeutung. Hieraus ist zu schließen, daß nicht nur aus logistischen Überlegungen und der praktischen Durchführbarkeit wegen ein gruppensequentielles Design mit $K = 5$ zu empfehlen ist.

Pocock (1982) gab darüber hinaus optimale Pläne an, die bei vorgegebenem α, β und K den ASN unter der Gültigkeit der durch (2.22) spezifizierten Alternativhypothese H_1 minimieren. Dabei sind beliebige Bereiche $\mathscr{C}_k$ bzw. $\breve{\mathscr{C}}_k$, $k = 1, 2, \ldots, K$, zugelassen. Entsprechende numerische Methoden sind erforderlich, um dieses K-dimensionale Optimierungsproblem zu lösen. In Pocock (1982) sind die Ergebnisse von *grid searches* enthalten, die mit den in SAS/IML seit Version 6.11 enthaltenen Optimierungsroutinen (z.B. dem Double Dogleg Optimierungsalgorithmus, vgl. SAS Institute Inc., 1995) einfach und elegant reproduziert werden können.

Beispielsweise ergeben sich für $\alpha = 0.05$, $1 - \beta = 0.95$ und $K = 5$ die optimalen Schranken

$$(u_1, u_2, u_3, u_4, u_5) = (2.446, 2.404, 2.404, 2.404, 2.396)$$

für Z_k^* mit minimierenden $ASN = 15.65$.

Für $\alpha = 0.05$, $1 - \beta = 0.50$ und $K = 5$ ergibt sich

$$(u_1, u_2, u_3, u_4, u_5) = (3.671, 2.884, 2.573, 2.375, 2.037)$$

mit minimierenden $ASN = 7.14$.

Dies bedeutet, daß bei einer Gütevorgabe von 95% das Pocock-Verfahren (d.h. eine konstante Aufteilung der kritischen Schranken für Z_k^*) nahezu optimal ist, während bei einer Gütevorgabe von 50% das optimale Verfahren eher vom Typ eines O'Brien und Fleming-Designs ist (d.h. ein Ausschöpfen des Niveaus α möglichst am Ende der Studie). Dies ist auch intuitiv naheliegend, denn ein moderater Effekt (*power* = 0.50) ist – wenn überhaupt – erst bei einer großer Stichprobe und damit in späteren Stufen zu erkennen, während ein starker Effekt (*power* = 0.95) eher früh erkannt werden kann und die Studie zum Abbruch führt.

In Köpcke (1984, 1989) wurden sogenannte „gemischte Strategien" vorgeschlagen und untersucht, die eine Zwischenlösung der Verfahren nach O'Brien und Fleming bzw. Pocock darstellen. Wang und Tsiatis (1987) schlugen eine (allgemeinere) Klasse von zweiseitigen Schranken vor, die durch einem Parameter Δ spezifiziert sind. Die Fortsetzungs- bzw. Annahmebereichen $\mathscr{C}_k$ sind definiert durch

$$\begin{aligned} \mathscr{C}_k &= (-u_k;\, u_k), \text{ wobei } u_k = c(K,\alpha,\Delta)\, k^{\Delta-0.5} \text{ bzw.} \\ \breve{\mathscr{C}}_k &= (-\breve{u}_k;\, \breve{u}_k), \text{ wobei } \breve{u}_k = c(K,\alpha,\Delta)\, k^{\Delta}\,. \end{aligned} \tag{2.25}$$

$\Delta = 0$ ergibt das Design nach O'Brien und Fleming (1979); $\Delta = 0.5$ führt zu konstanten Fortsetzungsbereichen nach Pocock (1977). In Tabelle 2.2 sind die Werte für $\alpha = 0.10$, 0.05, 0.025, 0.01, 0.005, $\Delta = 0.0, 0.1, \ldots, 0.7$ und $K = 2, 3, 4, 5$ mit einer Genauigkeit von 4 Nachkommastellen tabelliert. In Anhang A.1 ist das für die Berechnung dieser Werte benötigte SAS-Programm angegeben. Die in Tabelle 2.2 angegebenen Werte sind mit der höchsten Genauigkeit, die SAS zur Verfügung stellt, berechnet. Sie unterscheiden sich von den in Table 1 von Wang und Tsiatis (1987) angegebenen Werten, die in der in dieser Arbeit angegebenen Genauigkeit nicht korrekt sind (Wassmer und Bock, 1999).

Wang und Tsiatis (1987) stellten dar, daß diese Pläne approximativ optimal in dem Sinne sind, als daß der für Δ minimierte ASN mit dem optimalen ASN unter Zugrundelegung beliebiger Bereiche $\mathscr{C}_k$ praktisch übereinstimmt, d.h. der letztere nur marginal kleiner ausfällt. Dies folgt aus der Tatsache, daß die Entscheidungsbereiche praktisch identisch sind.

Beispielsweise ergeben sich für die in der Δ-Klasse optimalen Pläne in den beiden obigen Fällen ($\alpha = 0.05$, $K = 5$):

$\Delta = 0.506$, d.h. $(u_1, u_2, u_3, u_4, u_5) = (2.400, 2.410, 2.417, 2.421, 2.424)$
für $1 - \beta = 0.95$

mit minimierenden $ASN = 15.65$ sowie

$\Delta = 0.077$, d.h. $(u_1, u_2, u_3, u_4, u_5) = (4.071, 3.036, 2.558, 2.265, 2.061)$
für $1 - \beta = 0.50$

mit minimierenden $ASN = 7.16$ (vgl. Table 2 in Wang und Tsiatis, 1987). Praktisch entspricht dies dem obigen Resultat. Weitere Ergebnisse für praktisch relevante Fälle sind Tabelle 2.3 zu entnehmen, die die optimalen Δ und die Werte $c(K, \alpha, \Delta)$ für $\alpha = 0.05$, 0.01 und $1 - \beta = 0.80$, 0.90, 0.95 enthält. Tabelle 2.3 enthält ebenso das Verhältnis der optimalen ASN nach Pocock (1982) bzw. Wang und Tsiatis (1987). Die Berechnung dieser Tabelle mit SAS ist in Anhang A.2 dokumentiert.

Tabelle 2.2: *Kritische Werte $c(K, \alpha, \Delta)$ der Δ-Klasse von Fortsetzungsbereichen nach Wang und Tsiatis (1987) im zweiseitigen Testproblem. $\Delta = 0$ ergibt das Design nach O'Brien und Fleming (1979); $\Delta = 0.5$ führt zu konstanten Fortsetzungsbereichen nach Pocock (1977).*

		α				
Δ	K	0.10	0.05	0.025	0.01	0.005
0.00	2	2.3730	2.7965	3.1826	3.6481	3.9724
	3	2.9611	3.4711	3.9352	4.4945	4.8851
	4	3.4662	4.0486	4.5787	5.2182	5.6651
	5	3.9151	4.5617	5.1506	5.8611	6.3578
0.10	2	2.2425	2.6314	2.9856	3.4136	3.7129
	3	2.6943	3.1442	3.5543	4.0496	4.3958
	4	3.0690	3.5692	4.0250	4.5752	4.9601
	5	3.3936	3.9371	4.4323	5.0304	5.4488
0.20	2	2.1287	2.4877	2.8132	3.2058	3.4805
	3	2.4670	2.8639	3.2254	3.6622	3.9681
	4	2.7356	3.1643	3.5550	4.0273	4.3580
	5	2.9615	3.4174	3.8329	4.3351	4.6868
0.30	2	2.0305	2.3651	2.6667	3.0284	3.2807
	3	2.2766	2.6297	2.9494	3.3345	3.6039
	4	2.4616	2.8307	3.1657	3.5701	3.8532
	5	2.6114	2.9943	3.3425	3.7631	4.0577
0.40	2	1.9465	2.2625	2.5457	2.8837	3.1183
	3	2.1197	2.4395	2.7270	3.0709	3.3101
	4	2.2412	2.5651	2.8567	3.2062	3.4498
	5	2.3349	2.6624	2.9579	3.3124	3.5598
0.50	2	1.8754	2.1783	2.4492	2.7718	2.9952
	3	1.9922	2.2895	2.5557	2.8730	3.0929
	4	2.0674	2.3613	2.6246	2.9387	3.1564
	5	2.1217	2.4132	2.6745	2.9863	3.2026
0.60	2	1.8161	2.1110	2.3753	2.6906	2.9095
	3	1.8904	2.1751	2.4308	2.7367	2.9496
	4	1.9332	2.2113	2.4616	2.7615	2.9707
	5	1.9613	2.2347	2.4812	2.7770	2.9835
0.70	2	1.7676	2.0590	2.3215	2.6364	2.8560
	3	1.8113	2.0917	2.3460	2.6529	2.8681
	4	1.8327	2.1068	2.3564	2.6592	2.8724
	5	1.8449	2.1149	2.3618	2.6622	2.8742

Tabelle 2.3: *Optimales Δ (unter H_1) und kritische Werte $c = c(K, \alpha, \Delta)$ in der Δ-Klasse von Fortsetzungsbereichen im zweiseitigen Testproblem. Der benötigte Stichprobenumfang n pro Stufe ist für $\frac{\mu_1 - \mu_2}{\sigma} = 1$ angegeben. Ebenso angegeben ist das Verhältnis von optimalen ASN nach Pocock (1982) (=ASN_1) und dem approximativ optimalen ASN nach Wang und Tsiatis (1987) (=ASN_2).*

		$\alpha = 0.05$				$\alpha = 0.01$			
$1-\beta$	K	Δ	c	n	$\frac{ASN_1}{ASN_2}[\%]$	Δ	c	n	$\frac{ASN_1}{ASN_2}[\%]$
0.80	2	0.418	2.246	8.48	99.99	0.451	2.822	12.50	99.98
	3	0.389	2.458	5.77	99.99	0.408	3.052	8.41	99.98
	4	0.367	2.646	4.36	99.92	0.385	3.256	6.32	99.89
	5	0.353	2.806	3.50	99.85	0.375	3.415	5.08	99.82
0.90	2	0.486	2.189	11.50	99.99	0.512	2.760	16.20	99.99
	3	0.481	2.315	7.98	99.99	0.483	2.902	11.05	99.98
	4	0.461	2.434	6.06	99.92	0.458	3.041	8.32	99.86
	5	0.445	2.542	4.87	99.83	0.441	3.163	6.66	99.71
0.95	2	0.507	2.173	14.23	99.99	0.536	2.739	19.47	99.99
	3	0.533	2.248	10.04	99.95	0.530	2.825	13.51	99.98
	4	0.522	2.324	7.71	99.91	0.507	2.923	10.22	99.85
	5	0.506	2.400	6.22	99.80	0.488	3.019	8.18	99.63

Tabelle 2.3 zeigt, daß die Δ-Klasse tatsächlich in allen Fällen eine nahezu identische Optimalität bzgl. des zu minimierenden ASN aufweist. Bei der Suche nach einem optimalen Plan kann man sich aufgrund dieser Ergebnisse auf die Δ-Klasse beschränken. Für die in der Praxis häufiger auftretenden Fälle (*power* = 0.80 oder 0.90) entspricht die konstante Aufteilung der kritischen Werte auf die Stufen (d.h. Δ nahe 0.5) dem optimalen Design eher als stark monoton fallende kritische Schranken (d.h. Δ nahe 0). Bei gewähltem α, β und K kann der hierfür optimale Plan verwendet werden. Dieser Plan hängt explizit von der gewünschten Güte des Testverfahrens ab und stellt deshalb ein alternatives Vorgehen zu dem Ansatz der als eigentlich „klassisch“ bezeichneten Pläne dar.

Die optimalen Δ-Werte unterscheiden sich von den in Table 4 von Wang und Tsiatis (1987) angegebenen Werten für Δ. Dies ist wiederum auf die in der an-

gegebenen Exaktheit nicht korrekten Berechnungen in Wang und Tsiatis (1987) zurückzuführen, und in Wassmer und Bock (1999) berichtigt.

Beispiel
Aus Tabelle 2.3 ergibt die Wahl von $K = 5$, $\alpha = 0.05$ und $1 - \beta = 0.80$ die Werte $\Delta = 0.353$ und $c = 2.806$, woraus sich die kritischen Schranken
$(u_1, u_2, u_3, u_4, u_5) = (2.806, 2.535, 2.389, 2.290, 2.216)$
für Z_k^* ergeben. Die für den Abbruch der Studie benötigten (zweiseitigen) p-Werte, die zum Abbruch der Studie führen, sind gegeben durch
$(p_1, p_2, p_3, p_4, p_5) = (0.0050, 0.0112, 0.0169, 0.0220, 0.0267)$.
In Tabelle 2.3 ist der Stichprobenumfang n pro Stufe k des gruppensequentiellen Plans für den standardisierten Effekt $\frac{\mu_1 - \mu_2}{\sigma} = 1$ angegeben, der für diesen Fall (auf 3 Nachkommastellen angegeben) $n = 3.496$ beträgt. Die Zwischenauswertungen sind damit nach jeweils 4 Beobachtungen pro Therapiegruppe durchzuführen, womit die Bedingung an die *power* des Verfahrens leicht übererfüllt, aber damit garantiert ist. Geht man von einem standardisierten Effekt $\frac{\mu_1 - \mu_2}{\sigma} = 0.5$ aus, so ist der Stichprobenumfang gemäß (2.23) gleich $4 \cdot 3.496 = 13.98$, und die Güteforderung für 14 Beobachtungen pro Therapiegruppe garantiert.

In die Konstruktion des beschriebenen Testverfahrens geht neben dem Signifikanzniveau α auch die *power* $1 - \beta$ ein. Man beachte, daß durch die Berücksichtigung der *power* eine Kontrolle des Fehlers 2. Art bei dem durch (2.22) spezifiziertem Wert der Alternative erfolgt, falls am Ende der Studie (und nur in diesem Fall) die Teststatistik Z_K^* in den Annahmebereich $\mathscr{C}_K$ fällt. Eine komplexere Vorgehensweise ergibt sich, wenn auch der frühzeitige Abbruch der Studie mit der Annahme der Nullhypothese in Stufen $k < K$ berücksichtigt werden soll. Der nächste Abschnitt beinhaltet dieses Vorgehen mit der Darstellung des Ansatzes von Pampallona und Tsiatis (1994).

2.2.2 Abbruch mit der Annahme der Nullhypothese

Ein Design, in dem der kontrollierte Abbruch der Studie mit der Annahme der Nullhypothese H_0 möglich ist, wurde von Pampallona und Tsiatis (1994) vorgeschlagen (vgl. auch Gould und Pecore, 1982). Dieser Ansatz stellt eine Erweiterung des von Emerson und Fleming (1989) beschriebenen Vorgehens dar, deren

Ziel es war, eine symmetrische Behandlung der Ablehnung und der Annahme von H_0 zu erzielen. Die der allgemeineren Formulierung zugrundeliegende Idee ist die folgende. Wird in einer Stufe k des gruppensequentiellen Plans ein betragsmäßig nur geringer Effekt beobachtet, so ist es sinnvoll, die Studie mit der Annahme von H_0 zu beenden, da es nicht mehr sehr wahrscheinlich ist, aufgrund der vorliegenden Beobachtung zu einem positiven Studienresultat (im Sinne der Ablehnung von H_0) zu kommen. Ein Design, das dieses Studienergebnis berücksichtigt, kann für die zweiseitige Alternative in der Weise formuliert werden, daß aus zwei Intervallen bestehende Fortsetzungsbereiche angegeben werden. Demgemäß wird die Studie fortgesetzt, falls mit $0 < u_k^0 < u_k^1$, $k = 1, 2, \ldots, K-1$,

$$Z_k^* \in (-u_k^1;\, -u_k^0) \,\cup\, (u_k^0;\, u_k^1) \quad \Leftrightarrow \quad |\, Z_k^* \,| \in (u_k^0;\, u_k^1)\ .$$

Die korrespondierenden Bereiche für die Teststatistiken S_k lassen sich völlig analog formulieren.

Zusätzlich wird angenommen, daß man auf der Stufe K auf jeden Fall eine Entscheidung (zugunsten von H_0 oder H_1) zu treffen hat, d.h. man setzt

$$u_K^0 = u_K^1\ . \tag{2.26}$$

Die kritischen Werte u_k^0 und u_k^1, $k = 1, 2, \ldots, K$, werden in der Δ-Klasse von Wang und Tsiatis (1987) gesucht. Dementsprechend sind die Werte $u_k^{\,0}$ und u_k^1 gegeben durch

$$u_k^0 = \sqrt{k}\vartheta^* - c^0(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5} \tag{2.27}$$

und

$$u_k^1 = c^1(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5}\ , \tag{2.28}$$

wobei ϑ^* gemäß (2.22) gegeben ist und damit durch (2.27) die Werte $u_k^{\,0}$ spezifiziert sind, wie weit Z_k^* zugunsten H_0 von seinem Erwartungswert $\sqrt{k}\vartheta^*$ abweichen muß, um zu einer Annahme von H_0 zu gelangen. Die zu ermittelnden Werte $c^0 = c^0(K, \alpha, \beta, \Delta)$ und $c^1 = c^1(K, \alpha, \beta, \Delta)$ hängen dabei wegen der Berücksichtigung der *power* des Testverfahrens zusätzlich von β ab.

Aus (2.26) – (2.28) ergibt sich

$$\vartheta^* = (c^0 + c^1)\, K^{\Delta - 1} \tag{2.29}$$

und gemäß (2.23) ergibt sich hieraus der pro Stufe k benötigte Stichprobenumfang

$$n = 2(c^0 + c^1)^2 \, K^{2(\Delta-1)} (\frac{\sigma}{\mu_1 - \mu_2})^2 \, . \tag{2.30}$$

Der ASN läßt sich analog (2.24) berechnen und ist wie im letzten Abschnitt o.B.d.A. unter $\frac{\sigma}{\mu_1-\mu_2} = 1$ zu ermitteln.

Die Werte c^0 und c^1 werden bei gegebenem α, β, K und Δ aus

$$\begin{aligned}
\sum_{k=1}^{K} P_{\mathrm{H}_0}(\mid Z_k^* \mid \geq u_k^1 \cap \bigcap_{\bar{k}=1}^{k-1} \{\mid Z_{\bar{k}}^* \mid \in (u_{\bar{k}}^0;\, u_{\bar{k}}^1)\}) &= \alpha \quad \text{und} \\
\sum_{k=1}^{K} P_{\mathrm{H}_1}(\mid Z_k^* \mid \geq u_k^1 \cap \bigcap_{\bar{k}=1}^{k-1} \{\mid Z_{\bar{k}}^* \mid \in (u_{\bar{k}}^0;\, u_{\bar{k}}^1)\}) &= 1 - \beta
\end{aligned}$$

ermittelt, wobei die Alternative H_1 durch (2.29) spezifiziert ist. Dies kann durch die numerische Berechnung der Integrale und durch geeignete Suchprogramme geschehen. In Pampallona und Tsiatis (1994) sind c^0 und c^1 für α = 0.05, 0.01, β = 0.20, 0.10, 0.05, Δ = 0.0, 0.1,. . .,0.5 und $K = 1, 2, \ldots, 5, 10$ umfassend (und exakt) tabelliert. Anhang A.3 beschreibt den (etwas komplexeren) SAS-Programmcode zur Reproduzierung dieser Werte. Das Verfahren von Emerson und Fleming (1989) ergibt sich für den Spezialfall $\alpha = \beta$. Der ASN der resultierenden Verfahren wurde von Pampallona und Tsiatis (1994) unter der Annahme von H_0, unter der Annahme von H_1 sowie unter der Annahme, daß der Parameter in der Mitte zwischen H_0 und H_1 liegt, berechnet. Dabei ergibt sich der kleinste ASN unter allen Parameterkonstellationen in den meisten Fällen für $\Delta = 0.40$ oder 0.50, in seltenen Fällen für $\Delta = 0.30$. Dies entspricht dem Resultat des letzten Abschnitts, daß ein optimales Verfahren „eher dem Pocock-Design“ entspricht. Analog den von Wang und Tsiatis (1987) durchgeführten Berechnungen können auch hier die Werte Δ ermittelt werden, die den ASN minimieren. Die unter H_1 optimalen Δ-Werte mit c^0, c^1, und die benötigten Stichprobenumfänge n pro Stufe sind für $\alpha = 0.05$, 0.01 und $1 - \beta = 0.80$, 0.90 und 0.95 Tabelle 2.4 zu entnehmen.

Der Vergleich von Tabelle 2.3 und 2.4 zeigt, daß bei der Berücksichtigung des frühen Abbruchs mit der Annahme von H_0 die optimalen Δ-Werte zu kleineren

Tabelle 2.4: *Optimales Δ (unter H_1) und Werte $c^1 = c^1(K, \alpha, \beta, \Delta)$ und $c^0 = c^0(K, \alpha, \beta, \Delta)$ des zweiseitigen Verfahrens von Pampallona und Tsiatis (1994) mit frühzeitiger Annahme von H_0. Der benötigte Stichprobenumfang n pro Stufe ist für $\frac{\mu_1 - \mu_2}{\sigma} = 1$ angegeben.*

		$\alpha = 0.05$				$\alpha = 0.01$			
$1-\beta$	K	Δ	c^1	c^0	n	Δ	c^1	c^0	n
0.80	2	0.382	2.215	1.117	9.43	0.443	2.779	1.104	13.93
	3	0.414	2.333	1.245	7.07	0.438	2.920	1.248	10.11
	4	0.428	2.412	1.319	5.70	0.434	3.028	1.348	7.97
	5	0.424	2.501	1.366	4.68	0.433	3.110	1.416	6.60
0.90	2	0.396	2.220	1.554	12.34	0.466	2.768	1.523	17.56
	3	0.444	2.310	1.654	9.26	0.480	2.861	1.638	12.92
	4	0.454	2.386	1.706	7.37	0.478	2.940	1.721	10.22
	5	0.445	2.479	1.716	5.90	0.470	3.019	1.779	8.36
0.95	2	0.385	2.243	1.930	14.85	0.463	2.779	1.881	20.63
	3	0.447	2.319	2.004	11.09	0.497	2.843	1.968	15.33
	4	0.437	2.437	2.032	8.39	0.501	2.900	2.031	12.19
	5	0.499	2.372	1.960	7.48	0.488	2.981	2.073	9.83

Werten tendieren. Generell ist der Unterschied aber nicht sehr groß. Die Empfehlung, für einen gruppensequentiellen Plan eher einen Plan mit konstanten kritischen Schranken zu benutzen, gilt auch hier. Vergleicht man bei gegebenem Δ und β die kritischen Schranken für die Ablehnung von H_0, so sind diese für das Design nach Pampallona und Tsiatis (1994) kleiner, d.h. eine Ablehnung wahrscheinlicher. Dies resultiert aus der Tatsache, daß sich die Ablehnungsbereiche vergrößern, falls zusätzliche Abbruchkriterien gefordert werden. Geht die geforderte *power* gegen 1, so geht das beschriebene Verfahren in das „klassische“ Verfahren über.

Das folgende Beispiel illustriert die Verwendung eines optimalen Plans nach Pampallona und Tsiatis (1994).

Beispiel

Bei Wahl von $K = 5$, $\alpha = 0.05$ und $1 - \beta = 0.80$ ergibt sich aus Tabelle 2.4

$\Delta = 0.424$, $c^0 = 1.366$ und $c^1 = 2.501$, woraus sich die kritischen Schranken
$(u_1^0, u_2^0, u_3^0, u_4^0, u_5^0) = (0.164, 0.868, 1.394, 1.831, 2.213)$ und
$(u_1^1, u_2^1, u_3^1, u_4^1, u_5^1) = (2.501, 2.373, 2.301, 2.251, 2.213)$
für Z_k^* ergeben. Entsprechende p-Werte für die Ablehnung bzw. Annahme von H_0 lassen sich ebenso berechnen. Der Stichprobenumfang n pro Stufe k des gruppensequentiellen Plans für $\frac{\mu_1-\mu_2}{\sigma} = 1$ ist $n = 4.68$, womit nach jeweils 5 Beobachtungen pro Therapiegruppe die Zwischenauswertungen durchzuführen sind. Dies garantiert 80% *power*.
Für $K = 5$, $\alpha = 0.05$ und $1 - \beta = 0.95$ ergibt sich aus Tabelle 2.4: $\Delta = 0.499$, $c^0 = 1.960$ und $c^1 = 2.372$, woraus sich die kritischen Schranken
$(u_2^0, u_3^0, u_4^0, u_5^0) = (0.777, 1.393, 1.911, 2.369)$ und
$(u_1^1, u_2^1, u_3^1, u_4^1, u_5^1) = (2.372, 2.371, 2.370, 2.369, 2.369)$
ergeben bei einem Mindeststichprobenumfang von $n = 7.48$ (d.h. Zwischenauswertungen nach 8 Beobachtungen). Man beachte, daß ein Abbruch mit Annahme von H_0 erst ab der zweiten Zwischenauswertung vorgesehen ist, was rechnerisch aus der Bedingung $u_k^0 > 0$ folgt. In Abbildung 2.2 sind die Fortsetzungs- und Entscheidungsbereiche der beiden sich ergebenden Testverfahren illustriert.

2.2.3 Einseitige Pläne

Eine einseitige Formulierung des Testproblem, d.h. die Überprüfung der Hypothese (2.1) gegen eine einseitige Alternative

$$\mathrm{H}_1 : \ \mu_1 - \mu_2 > 0$$

wurde von DeMets und Ware (1980, 1982) vorgeschlagen und diskutiert (die Behandlung der Alternative $\mathrm{H}_1 : \ \mu_1 - \mu_2 < 0$ geschieht völlig analog). Die einfachste Möglichkeit zur Bereitstellung eines gruppensequentiellen Plans besteht in der Bestimmung der kritischen Werte, die der Bedingung

$$1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{Z_k^* \in \mathscr{C}_k\}) = \alpha$$

genügen, wobei die Fortsetzungsbereiche $\mathscr{C}_k$ durch

$$\mathscr{C}_k = (-\infty;\, u_k), \ k = 1, 2, \ldots, K, \tag{2.31}$$

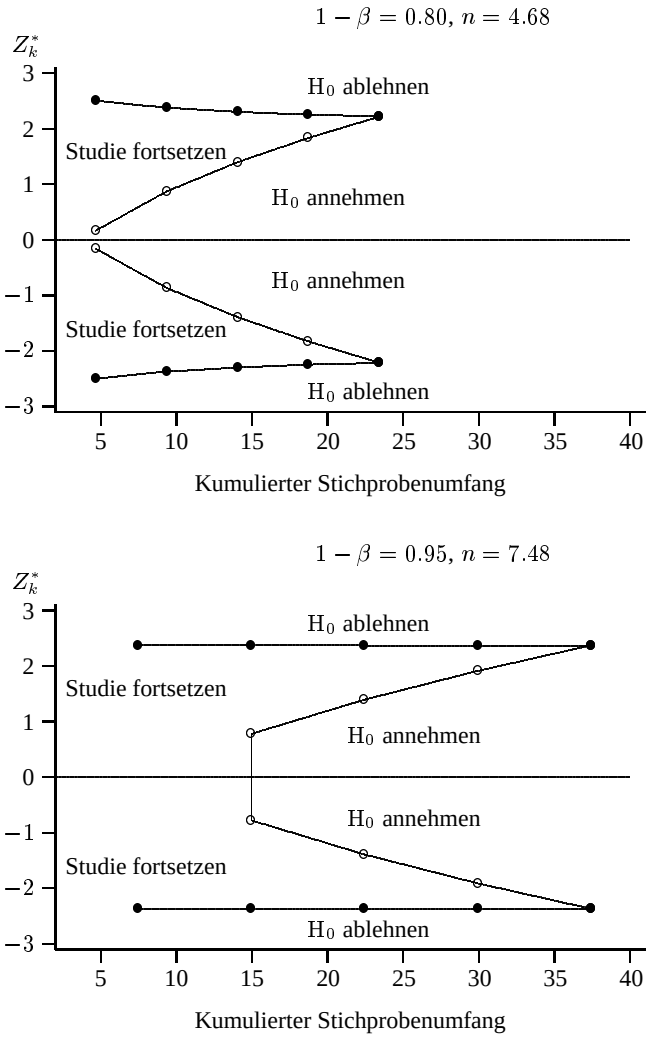

Abbildung 2.2: *Fortsetzungs- bzw. Entscheidungsbereiche für das optimale zweiseitige Design nach Pampallona und Tsiatis (1994) mit minimalem ASN unter H_1; $K = 5$, $\alpha = 0.05$. $1-\beta = 0.80$ ergibt $\Delta = 0.424$ und $1-\beta = 0.95$ ergibt $\Delta = 0.499$ (vgl. Tabelle 2.4).*

gegeben sind. Dies bedeutet, daß H_0 nur dann abgelehnt wird, falls die Teststatistik Z_k^* genügend große Werte annimmt. Hier wird der Wunsch nach der Berücksichtigung eines frühen Abbruchs mit Annahme von H_0 besonders deutlich. Ist der beobachtete Wert der Teststatistik in einer Stufe der Studie sehr klein oder gar negativ, so ist die Chance, die Studie zu einem positiven Ergebnis zu führen, sehr gering. Die Fortsetzungsbereiche sind dem entsprechend zu modifizieren. Hier bieten sich eine Reihe von Möglichkeiten an, von denen DeMets und Ware (1980, 1982) zwei Verfahren diskutierten: Die *asymmetric method* verwendet konstante Werte u^L als Abbruchkriterium für die Annahme von H_0, d.h. $\mathscr{C}_k$ ist gegeben durch

$$\mathscr{C}_k = (u^L;\, u_k),\ k = 1, 2, \ldots, K.$$

Die *constant likelihood method* benutzt von der *power* $1 - \beta$ abhängige und an den *SPRT* (Wald, 1947) angelehnte Schranken u_k^L. Diese Methoden wurden für das Design nach Pocock (DeMets und Ware, 1980) wie für das Design nach O'Brien und Fleming (DeMets und Ware, 1982) vorgeschlagen.

In Erweiterung zu Table 1 und 2 in DeMets und Ware (1980) sind in Tabelle 2.5 die kritischen Werte für $\alpha = 0.05$, 0.025, 0.01, 0.005, $u^L = 0.5$, 0, –1.0, $-\infty$ und den Fall konstanter Abbruchschranken $u_k = u,\ k = 1, 2, \ldots, K$, nach Pocock angegeben. Man beachte, daß $u^L = 0$ dem Fall entspricht, in dem auf einer Stufe k die Studie mit der Annahme von H_0 abgebrochen wird, falls ein gegenläufiger Effekt (d.h. $p_k \geq 0.50$) beobachtet wird. In Tabelle 2.5 sind ebenso die durchschnittliche Anzahl durchgeführter Stufen $\overline{K}$ unter Gültigkeit von H_0 angegeben, die sich aus dem (unter H_0 berechneten) ASN mit $n = 1$ ergibt.

Tabelle 2.5 zeigt, daß wie im letzten Abschnitt die kritischen Werte mit zunehmend starker Bedingung an Z_k^* (d.h. wachsendem u^L) kleiner werden, womit eine Ablehnung von H_0 eher geschehen kann. Man beachte jedoch, daß der hierdurch erzielte „Gewinn" nicht besonders groß ist. Beispielsweise ergibt sich durch die Bedingung, H_0 bei $Z_k^* \leq u^L$ abbrechen zu müssen(!), bei $\alpha = 0.05$, $K = 5$ und $u^L = -\infty$ der kritische Wert $u = 2.122$, während sich ceteris paribus bei $u^L = -1.0$ der kaum nennenswert unterschiedliche Wert $u = 2.120$ und bei $u^L = 0$ der Wert $u = 2.109$ ergibt. Erst bei $u^L = 0.5$ zeigt sich ein etwas größerer „Sprung" auf $u = 2.034$. Eine Implementation des Verfahrens

Tabelle 2.5: *Kritische Werte u im einseitigen Pocock-Design nach DeMets und Ware (1980) mit Abbruch für H_0, falls $Z_k^* \leq u^L$. Die durchschnittliche Anzahl durchgeführter Stufen $\overline{K}$ ist unter Gültigkeit von H_0 berechnet.*

		$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$		$\alpha = 0.005$	
u^L	K	u	$\overline{K}$	u	$\overline{K}$	u	$\overline{K}$	u	$\overline{K}$
0.5	2	1.855	1.28	2.168	1.29	2.527	1.30	2.770	1.31
	3	1.947	1.44	2.264	1.48	2.623	1.50	2.865	1.50
	4	2.000	1.56	2.320	1.61	2.681	1.64	2.923	1.65
	5	2.034	1.65	2.358	1.72	2.720	1.76	2.963	1.77
0.0	2	1.871	1.47	2.176	1.49	2.531	1.49	2.772	1.50
	3	1.979	1.81	2.283	1.84	2.633	1.86	2.871	1.87
	4	2.045	2.09	2.348	2.14	2.697	2.17	2.934	2.18
	5	2.091	2.33	2.394	2.40	2.743	2.44	2.979	2.45
-1.0	2	1.875	1.81	2.178	1.83	2.531	1.84	2.772	1.84
	3	1.992	2.55	2.289	2.58	2.636	2.60	2.873	2.60
	4	2.066	3.24	2.361	3.28	2.704	3.31	2.938	3.32
	5	2.120	3.89	2.412	3.96	2.753	3.99	2.986	4.01
$-\infty$	2	1.875	1.97	2.178	1.99	2.531	1.99	2.772	2.00
	3	1.992	2.94	2.289	2.97	2.636	2.99	2.873	2.99
	4	2.067	3.91	2.361	3.95	2.704	3.98	2.939	3.99
	5	2.122	4.87	2.413	4.94	2.754	4.98	2.986	4.99

mit positivem u^L ist jedoch in der Regel für die Praxis nicht zu empfehlen, da eine Nichtablehnung von H_0 hierdurch auch bei Gültigkeit der Alternative sehr häufig ist.

Die Berücksichtigung des Abbruchs mit der Annahme von H_0 macht sich deutlicher bei der durchschnittlichen Anzahl $\overline{K}$ durchgeführter Tests bemerkbar. Der Unterschied ist mit zunehmenden K immer stärker ausgeprägt, falls die Gültigkeit von H_0 unterstellt wird. Allerdings muß bemerkt werden, daß dies kein allzu großer Gewinn ist, da ein Abbruch mit der Annahme von H_0 generell immer geschehen kann, weil damit keine inferenzstatistische Aussage verbunden ist.

Eine weitere bemerkenswerte Eigenschaft ergibt sich aus dem Vergleich der zweiseitigen und einseitigen kritischen Schranken in Tabelle 2.2 bzw. Tabelle 2.5. Die einseitigen kritischen Werte mit $u^L = -\infty$ (d.h. keine Berücksichtigung eines frühen Abbruchs mit der Annahme von H_0) zum Niveau α stimmen mit den zweiseitigen kritischen Werten für $\Delta = 0.5$ zum Niveau $2\,\alpha$ überein. Man beachte jedoch, daß diese Übereinstimmung lediglich numerisch für die in der Tabelle angegebene Genauigkeit gilt. Dies läßt sich im Fall $K = 2$ verdeutlichen. Damit exakte Übereinstimmung gilt, müßten die Wahrscheinlichkeiten

$$\begin{aligned} P_1 &= 1 - P_{\mathrm{H}_0}(|\,Z_1^*\,| < u,\, |\,Z_2^*\,| < u) \quad \text{und} \\ P_2 &= 2(1 - P_{\mathrm{H}_0}(Z_1^* < u,\, Z_2^* < u)) \end{aligned}$$

übereinstimmen, was jedoch nur für genügend große u der Fall ist, wie die folgende kleine Tabelle zeigt (die Werte der binormalen Normalverteilung mit Korrelation $1/\sqrt{2}$ (vgl. (2.6)) wurden mit der SAS-Funktion PROBBNRM berechnet):

u	0.4	0.8	1.2	1.6	2.0	2.4
P_1	0.8699	0.5978	0.34704	0.174531	0.07597316	0.0285025575
P_2	0.9189	0.6021	0.34720	0.174533	0.07597317	0.0285025575

Im wesentlichen entspricht dies der Tatsache, daß „signifikant widersprüchliche Effekte“ unter der Normalverteilungsannahme extrem unwahrscheinlich sind und praktisch nicht auftauchen. Von praktischer Bedeutung ist dies, da wie im nicht-sequentiellen Fall die Benutzung einer einzigen Tabelle (z.B. für den zweiseitigen Fall) ausreicht. Die Abbruchschranken u_k können überdies wie im zweiseitigen Testproblem in der Δ-Klasse angegeben werden. Dies bedeutet: Für das einseitige Testen zum Niveau α mit Fortsetzungsbereichen (2.31) können die in Tabelle 2.2 angegebenen Werte mit $2\,\alpha$ verwendet werden. Der Unterschied in den sich ergebenden kritischen Schranken ist für gebräuchliche Werte von α verschwindend klein und fällt mit kleiner werdendem α. Die Berechnungen für bis zu $K = 5$ zeigten, daß der maximale Unterschied für $\alpha = 0.10$ kleiner als 0.00005 und somit von keiner praktischen Bedeutung ist (vgl. auch Proschan, 1999b).

Pampallona und Tsiatis (1994) schlugen das Verfahren, das einen kontrollierten Abbruch mit der Annahme von H_0 vorsieht, auch für das einseitige Testproblem

vor. Die Vorgehensweise ist analog der des zweiseitigen Testproblems. Die Studie wird fortgesetzt, falls

$$Z_k^* \in (u_k^0;\, u_k^1)\ , \tag{2.32}$$

wobei $u_k^0 < u_k^1$, $k = 1, 2, \ldots, K-1$. Man beachte, daß hier die Forderung $u_k^0 > 0$ nicht getroffen wird. Wie im zweiseitigen Fall wird zusätzlich angenommen, daß man auf der Stufe K auf jeden Fall eine Entscheidung (zugunsten von H_0 oder H_1) zu treffen hat, d.h. man setzt die Gültigkeit von (2.26) voraus. Die kritischen Werte u_k^0 und u_k^1, $k = 1, 2, \ldots, K$, sind durch (2.27) und (2.28) gegeben, wobei ϑ^* gemäß (2.22) für die durch (2.32) gegebenen Entscheidungsbereiche spezifiziert ist. Wie im zweiseitigen Fall ist der pro Stufe k benötigte Stichprobenumfang durch (2.30) gegeben, womit sich der ASN o.B.d.A. unter $\frac{\sigma}{\mu_1 - \mu_2} = 1$ ermitteln läßt.

Die kritischen Werte c^0 und c^1 des einseitigen Plans werden bei gegebenem α, β, K und Δ aus

$$\begin{aligned}
\sum_{k=1}^{K} P_{\mathrm{H}_0}(Z_k^* \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{Z_{\tilde{k}}^* \in (u_{\tilde{k}}^0;\, u_{\tilde{k}}^1)\}) &= \alpha \quad \text{und} \\
\sum_{k=1}^{K} P_{\mathrm{H}_1}(Z_k^* \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{Z_{\tilde{k}}^* \in (u_{\tilde{k}}^0;\, u_{\tilde{k}}^1)\}) &= 1 - \beta
\end{aligned}$$

berechnet, wobei die Alternative H_1 durch (2.29) spezifiziert ist. Auch für diesen Fall sind c^0, c^1 und der ASN in Pampallona und Tsiatis (1994) für α = 0.05, 0.01, β = 0.20, 0.10, 0.05, Δ = 0.0, 0.1,...,0.5 und $K = 1, 2, \ldots, 5, 10$ tabelliert (vgl. auch Anhang A.3). Wie im zweiseitigen Fall ergibt sich als allgemeine Richtlinie, daß eine Wahl von Δ nahe bei 0.50 einen niedrigeren ASN sowohl unter H_0 wie unter H_1 ergibt. Zur Verdeutlichung sind analog Tabelle 2.4 in Tabelle 2.6 die unter H_1 optimalen Δ-Werte mit c^0, c^1 und benötigten Stichprobenumfang n pro Stufe für $\alpha = 0.05$, 0.01 und $1 - \beta = 0.80$, 0.90 und 0.95 angegeben. In Abbildung 2.3 sind die resultierenden optimalen Entscheidungsbereiche für $K = 5$, $\alpha = 0.05$ und $1 - \beta = 0.80$ illustriert.

Durch das Abbruchkriterium für die Annahme von H_0 wird wiederum eine Verkleinerung der kritischen Werte erreicht. Der „Gewinn“ fällt generell größer aus als in der *asymmetric method* von DeMets und Ware (1980, 1982). Dies ist durch

Tabelle 2.6: *Optimales Δ (unter H_1) und Werte $c^1 = c^1(K, \alpha, \Delta)$ und $c^0 = c^0(K, \alpha, \Delta)$ des einseitigen Verfahrens von Pampallona und Tsiatis (1994) mit frühzeitiger Annahme von H_0. Der benötigte Stichprobenumfang n pro Stufe ist für $\frac{\mu_1 - \mu_2}{\sigma} = 1$ angegeben.*

		$\alpha = 0.05$				$\alpha = 0.01$			
$1-\beta$	K	Δ	c^1	c^0	n	Δ	c^1	c^0	n
0.80	2	0.323	1.940	1.137	7.40	0.423	2.549	1.108	12.01
	3	0.366	2.073	1.274	5.56	0.427	2.687	1.250	8.80
	4	0.384	2.164	1.367	4.52	0.424	2.796	1.353	6.97
	5	0.393	2.234	1.438	3.82	0.420	2.889	1.434	5.78
0.90	2	0.336	1.951	1.586	9.97	0.442	2.544	1.533	15.34
	3	0.400	2.052	1.695	7.52	0.467	2.638	1.645	11.38
	4	0.427	2.114	1.763	6.14	0.470	2.713	1.728	9.07
	5	0.440	2.163	1.816	5.22	0.467	2.781	1.796	7.53
0.95	2	0.324	1.976	1.976	12.23	0.437	2.559	1.896	18.19
	3	0.407	2.058	2.058	9.21	0.482	2.625	1.981	13.59
	4	0.442	2.105	2.105	7.55	0.494	2.675	2.041	10.94
	5	0.461	2.138	2.138	6.45	0.495	2.724	2.094	9.14

den größeren Entscheidungsbereich für die Annahme von H_0 verursacht (vgl. Abbildung 2.3). Wie im zweiseitigen Fall ist im Ansatz von Pampallona und Tsiatis (1994) im Gegensatz zum Ansatz von DeMets und Ware die Annahme von H_0 bei der durch ϑ^* spezifizierten Alternative unter Kontrolle. Pampallona und Tsiatis verglichen ihr Verfahren auch mit dem sich durch die *constant likelihood method* ergebenden Verfahren (das auch von der *power* abhängt). Bzgl. des ASN fällt ihr Verfahren besser aus, womit der modernere Ansatz vorzuziehen ist.

Für die konkrete Planung einer Studie ergeben sich einige Diskussionspunkte, die im folgenden erörtert werden. Zum Einen erscheint es höchst fraglich, ob ein frühzeitiger Abbruch der Studie mit der Annahme von H_0 oft geschehen soll, oder ob nur für „extreme Fälle“ ein derartiger Abbruch erfolgen soll. In der Planung klinischer Studien ist die durch ϑ^* spezifizierte Alternative meistens nur als Anhaltspunkt für die Fallzahlplanung gedacht, um zu einem statistisch signi-

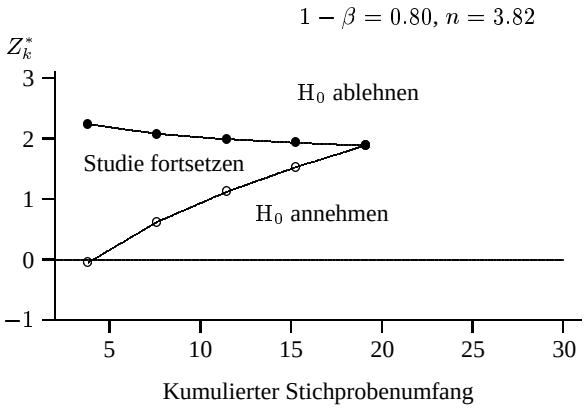

Abbildung 2.3: *Fortsetzungs- bzw. Entscheidungsbereiche für das optimale einseitige Design nach Pampallona und Tsiatis (1994) mit minimalem ASN unter H_1; $K = 5$, $\alpha = 0.05$. $1 - \beta = 0.80$ ergibt $\Delta = 0.393$ (vgl. Tabelle 2.6).*

fikanten Resultat zu gelangen. In der klinischen Praxis ist damit ϑ^* nicht der im strengen Sinne zu spezifizierende „klinisch relevante Unterschied“, sondern eine Größe, die auch von Aspekten der Durchführbarkeit einer Studie bestimmt ist. Ein erweiterter Entscheidungsraum für die Annahme von H_0 beinhaltet somit die Gefahr von häufigen falsch negativen Resultaten. Damit zusammenhängend soll auf die relative Willkürlichkeit hingewiesen werden, als Optimalitätskriterium den unter H_1 spezifizierten ASN zu verwenden. Zwar ergeben sich unter H_0 bzw. zwischen H_0 und H_1 sehr ähnliche Resultate (vgl. Table 2 in Pampallona und Tsiatis, 1994), doch konkurrieren der maximal benötigte Stichprobenumfang mit dem ASN als Entscheidungskriterium: der maximal benötigte Stichprobenumfang ist für kleines Δ geringer, während der ASN für größeres Δ optimal ist. Ein tatsächlicher Konflikt entsteht hieraus jedoch nicht, da in der Durchführung einer gruppensequentiellen klinischen Studie im Vordergrund stehen sollte, die Studie möglichst früh zu einem positiven Resultat zu bringen. Dies ist für Werte Δ nahe 0.40 bzw. 0.50 der Fall. Allerdings sollte beachtet werden, daß der ASN für verschiedene Δ-Werte – insbesondere im einseitigen

Fall – nicht sehr stark variiert, während dies für den maximalen Stichprobenumfang eher der Fall ist (vgl. Table 1 und Table 2 in Pampallona und Tsiatis, 1994). Dies macht die Entscheidung zwischen den verschiedenen Designvarianten noch schwerer. Monoton in k abfallende Schranken können insbesondere dann verwendet werden, falls ein früher Abbruch der Studie nur in solchen Fällen geschehen soll, in denen schon früh ein sehr starker Effekt beobachtet werden kann.

Beim Verfahren von O'Brien und Fleming (1979) ist diese letzte Eigenschaft besonders deutlich ausgeprägt: die Studie wird beispielsweise bei zweiseitigen $\alpha = 0.05$ und $K = 5$ abgebrochen, falls $| Z_1^* | \geq 4.562$ (vgl. Tabelle 2.1). Dies entspricht einem zweiseitigen p-Wert von ≤ 0.000005, der in der Praxis kaum anzutreffen ist. Ein solcher Plan erscheint also eher „widersinnig" bzw. unrealistisch. Viel realistischer wäre es, in einem solchen Fall auf diese, Aufwand und Kosten verursachende Auswertung einfach zu verzichten, und die erste Zwischenauswertung nach $2 \cdot n$ Patienten (pro Therapiegruppe) durchzuführen. Dies ist ein einfaches Beispiel eines Verfahrens mit ungleichen Sequenzgrößen. Einige hierfür konzipierte Verfahren werden im folgenden Abschnitt behandelt.

2.3 Verfahren mit ungleichen Sequenzgrößen

Die Annahme gleicher Sequenzgrößen bei gruppensequentiellen Plänen ist in vielen Fällen unrealistisch bzw. nicht durchführbar. So ist es z.B. in vielen Situationen weit zweckmäßiger, bei der Planung einer Studie die *Zeitpunkte*, an denen die Zwischenauswertungen durchgeführt werden, festzulegen. Dieses Vorgehen wird typischerweise zu verschieden großen Sequenzgrößen führen. Ebenso kann es sinnvoll sein, von vornherein eine größere Stichprobe bis zur ersten Auswertung zu verlangen (vgl. die Bemerkung am Ende des letzten Abschnitts); *drop out*-Fälle können zur Ungleichheit der Sequenzgrößen führen, etc.

Pocock (1977, 1982) stellte dar, daß die Verwendung der für den Fall gleich großer Sequenzen bestimmten kritischen Werte keinen sehr großen Einfluß auf die Fehlerwahrscheinlichkeit 1. Art hat, wenn diese auch für den Fall ungleicher Sequenzgrößen herangezogen werden. Insbesondere für den Fall kleinerer Abweichungen von der Bedingung sind die für den Fall gleicher Sequenzgrößen

bestimmten Schranken verwendbar. Vom theoretischen Standpunkt ist dies natürlich unbefriedigend. Darüber hinaus sind durchaus Fälle zu berücksichtigen, in denen sich die Sequenzgrößen so stark unterscheiden, daß eine unbedenkliche Anwendung der „einfachen“ Schranken nicht geboten ist.

In den folgenden Abschnitten werden Verfahren vorgestellt, die speziell für den Fall ungleicher Sequenzgrößen zugeschnitten sind und damit der Problemstellung adäquatere Vorgehensweisen liefern. Zuerst wird der Fall betrachtet, daß die Sequenzgrößen fest vorgegeben sind, womit der Fall gleich großer Sequenzgrößen als Spezialfall resultiert. Der von Wassmer (1999a) vorgeschlagene Ansatz beruht auf der Betrachtung des für den Fehler 1. Art ungünstigsten Falles und einer dementsprechenden Adjustierung der kritischen Werte. Dem gegenüber ist der α-$spending\ function$- oder $use\ function$-Ansatz ein auf Lan und DeMets (1983) und Kim und DeMets (1987b) zurückgehendes Verfahren, das die auf den Analysezeitpunkt bezogene interaktive Berechnung der kritischen Schranken vorsieht.

2.3.1 Der Fall fest vorgegebener Sequenzgrößen

Die in Abschnitt 2.1 dargestellte allgemeine Konstruktionsmethode für die Bestimmung gruppensequentieller Pläne läßt sich für den Fall beliebiger, fest vorgegebener Sequenzgrößen verwenden. Wie in diesem Abschnitt bereits beschrieben wurde, hängt die Fehlerwahrscheinlichkeit 1. Art lediglich von den Analysezeitpunkten ab, die durch die Angabe der auf n_1 standardisierten Zeitintervalle $\tau_k = \frac{n_k}{n_1}$ zwischen Sequenz $k-1$ und k, $k = 2, 3, \ldots, K$, gegeben sind. Bei spezifiziertem $\boldsymbol{\tau} = (1, \tau_2, \ldots, \tau_K)$ lassen sich so die kritischen Schranken berechnen, falls eine „Gestalt“ der Ablehnbereiche vorgegeben wird. Beispielsweise lassen sich für ungleiche Sequenzgrößen zweiseitige kritische Schranken u nach Pocock durch die in (2.17) angegebene Bedingung bestimmen, falls $\breve{\mathscr{C}}_k$ durch (2.14) mit $\mathscr{C}_k = (-u;\ u)$ gegeben ist. Entsprechend lassen sich Schranken nach O'Brien und Fleming oder allgemein Schranken in der Δ-Klasse nach Wang und Tsiatis formulieren, indem

$$\mathscr{C}_k = (-u_k;\ u_k) \quad \text{mit} \quad u_k = c(K, \alpha, \Delta, \tau)\, k^{\Delta - 0.5}$$

gesetzt wird. Einseitige kritische Schranken lassen sich völlig analog formulieren. In Tabelle 2.7 sind einige kritische Werte für $\alpha = 0.05$, 0.025, 0.01, $\Delta = 0$ und $\Delta = 0.5$ (d.h. O'Brien und Flemings bzw. Pococks Design) angegeben. Diese können für beliebige Fälle mit dem in Anhang A.4 angegebenen SAS-Programm berechnet werden.

Tabelle 2.7: *Kritische Werte $c = c(K, \alpha, \Delta, \tau)$ bei ungleichen Sequenzgrößen im zweiseitigen Testproblem.*

	O'Brien/Fleming			Pocock		
	α			α		
$\boldsymbol{\tau}$	0.05	0.025	0.01	0.05	0.025	0.01
(1, 1, 1, 1, 1)	4.562	5.151	5.861	2.413	2.674	2.987
(1, 2, 1, 1, 1)	4.531	5.126	5.843	2.405	2.666	2.977
(1, 4, 1, 1, 1)	4.492	5.094	5.818	2.386	2.647	2.958
(1, 2, 2, 2, 2)	4.581	5.166	5.873	2.441	2.699	3.007
(1, 4, 4, 4, 4)	4.593	5.175	5.880	2.458	2.713	3.018
(1, 1, 1, 1, 2)	4.624	5.206	5.908	2.429	2.689	2.999
(1, 1, 1, 1, 4)	4.682	5.257	5.950	2.445	2.703	3.011
(1, 0.5, 0.5, 0.5)	4.021	4.557	5.202	2.319	2.587	2.906
(1, 0.25)	2.776	3.171	3.643	2.111	2.389	2.718

Der Vergleich mit den Werten für gleiche Sequenzgrößen (d.h. $\boldsymbol{\tau} = (1, 1, 1, 1, 1)$) verdeutlicht den Effekt der ungleichen Sequenzgrößen. Es gibt Fälle, in denen eine stärkere Bedingung für den Abbruch der Studie zu fordern ist und Fälle, die eine weniger starke Bedingung mit sich bringen. Prinzipiell lassen sich also exakte Verfahren angeben, falls die Sequenzgrößen vorgegeben und bekannt sind. Wird beispielsweise ein vierstufiges Verfahren gewünscht, das erst nach $2 \cdot n$ Beobachtungen eine Zwischenauswertung vorsieht, so sind die kritischen Schranken gemäß Tabelle 2.7 für $\boldsymbol{\tau} = (1, 0.5, 0.5, 0.5)$ zu verwenden. Entsprechend kann auch ein zweistufiges Verfahren mit $4 \cdot n$ Beobachtungen bis zur ersten Zwischenauswertung erwünscht sein. Diese Varianten sind für $\alpha = 0.05$ in Abbildung 2.4 für O'Brien und Flemings Design graphisch dargestellt. Dabei wurde willkürlich der maximale Stichprobenumfang $N = 100$ pro Behandlungsgruppe gewählt und das entsprechende fünfstufige Design mit gleichen Sequenzgrößen gegenübergestellt.

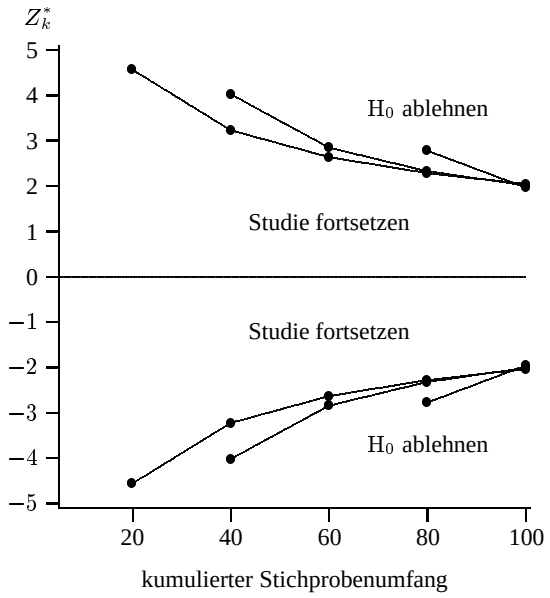

Abbildung 2.4: *Fortsetzungs- bzw. Annahmebereiche $\mathscr{C}_k$ für O'Brien und Flemings Design für den Fall ungleicher, fest vorgegebener Sequenzgrößen mit maximalen Stichprobenumfang $N = 100$; $\alpha = 0.05$.*

Interessant ist die Frage, wie sich die Verfahren bei ungleichen Sequenzgrößen im Vergleich zum Fall gleicher Sequenzgrößen bzgl. des ASN und der $power$ verhalten. Diese Größen sind berechenbar (siehe Anhang A.4). Beispielsweise ergibt sich für die in Abbildung 2.4 illustrierten Designvarianten nach O'Brien und Fleming im Vergleich zu konstanten Sequenzgrößen ein größerer (d.h. „schlechterer") ASN mit einer größeren (d.h. „besseren") $power$, falls für jedes Design der maximale Stichprobenumfang $N = 100$ und die fest vorgegebene Effektgröße $\frac{\mu_1 - \mu_2}{\sigma}$ zugrundegelegt werden. So ist z.B. bei einem Effekt $\frac{\mu_1 - \mu_2}{\sigma} = 0.4018$ die $power$ des O'Brien und Fleming-Designs bei gleichen Sequenzgrößen $= 0.80$ und der $ASN = 79.5$. Bei $\boldsymbol{\tau} = (1, 0.5, 0.5, 0.5)$ ist $power = 0.804$ und $ASN = 82.7$, bei $\boldsymbol{\tau} = (1, 0.25)$ ist $power = 0.811$

und $ASN = 91.9$. Bei Vorgabe dieser Effektgröße ergibt sich für Pococks Design bei gleichen Sequenzgrößen $power = 0.722$ und $ASN = 70.0$; bei $\boldsymbol{\tau} = (1, 0.5, 0.5, 0.5)$ ist hier $power = 0.748$ und $ASN = 71.5$, bei $\boldsymbol{\tau} = (1, 0.25)$ ist $power = 0.791$ und $ASN = 86.7$. Das letztere Design hat also einen durchweg geringeren ASN mit den Kosten einer niedrigeren $power$.

Gruppensequentielle Verfahren lassen sich für den Fall beliebig vorgegebener Sequenzgrößen bestimmen und deren Verhalten läßt sich bzgl. adäquater Gütekriterien beurteilen. Prinzipiell bietet sich auch die Möglichkeit an, die Pläne nach einem bestimmten Zielkriterium optimal zu gestalten. Die Ausführungen in den letzten Abschnitten haben aber schon gezeigt, daß die Wahl eines gruppensequentiellen Verfahrens relativ „sensitiv“ auf Änderung der Optimalitätsforderung reagiert und beispielsweise die Forderung, den maximalen Stichprobenumfang zu minimieren in Konflikt damit steht, den ASN zu minimieren. Eine somit noch komplexere Optimalität ist darüber hinaus dem Anwender oft schwer zu vermitteln. Die Idee des in diesem Sinne optimalen gruppensequentiellen Versuchsplans wird hier deshalb nicht weiter verfolgt (man vergleiche dazu Brittain und Bailey, 1993; Müller und Schäfer, 1999), da sie auch nicht im Zentrum der Untersuchungen des Verfassers dieser Arbeit steht.

2.3.2 Eine *worst case scenario*-Lösung

Wie bereits erwähnt, schlug Pocock (1977, 1982) die Verwendung der für gleiche Sequenzgrößen ermittelten Schranken auch für den Fall ungleicher Sequenzgrößen vor. Der hierdurch entstehende Effekt auf die Fehlerwahrscheinlichkeit 1. Art wurde von Proschan et al. (1992) untersucht. Wassmer (1999a) schlug ein Verfahren vor, das auf einer Adjustierung der kritischen Werte bzgl. dieses Effekts beruht. Dieser Ansatz wird im folgenden dargestellt.

Proschan et al. (1992) beschrieben unter Zugrundelegung des einseitigen Testproblems einen eher moderaten Effekt auf die Fehlerwahrscheinlichkeit 1. Art, solange die maximale Anzahl K der Stufen nicht zu groß ist. Mit Hilfe eines *grid search* ermittelten sie die Konfigurationen der Sequenzgrößen, die zu einer maximalen Abweichung vom vorgegebenen α führen. Allgemein kann bei Vorgabe beliebiger Bereiche $\mathscr{C}_k$ unter Voraussetzung beliebiger Sequenzgrößen die

Größe

$$\sup_{\tau_k>0,\, k=2,3,\ldots,K} 1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{Z_k^* \in \mathscr{C}_k\}) \tag{2.33}$$

betrachtet und der Vektor $\boldsymbol{\tau}$ ermittelt werden, für den die Fehlerwahrscheinlichkeit 1. Art maximal ist. Zweckmäßigerweise wird dies im folgenden durch den *Vektor der Analysezeitpunkte* ausgedrückt. Dieser ist durch $\boldsymbol{t} = (t_1, t_2, \ldots, t_K)$, $0 < t_1 < t_2 < \ldots < t_K = 1$, gegeben, wobei $t_k = \sum_{\bar{k}=1}^{k} n_{\bar{k}}/N$ und damit der letzte Analysezeitpunkt auf den Wert 1 standardisiert ist. Zwischen τ_k und t_k, $k = 1, 2, \ldots, K$, besteht die Beziehung

$$t_k = \frac{n_1}{N} \sum_{\bar{k}=1}^{k} \tau_{\bar{k}},$$

und die für den allgemeinen Fall gültige Kovarianz zwischen Z_k^* und $Z_{k'}^*$, $k \leq k'$, ist in dieser Terminologie gegeben durch (vgl. (2.6))

$$Cov(Z_k^*, Z_{k'}^*) = \sqrt{\frac{t_k}{t_{k'}}}\,.$$

In Proschan et al. (1992) ist für den einseitigen Fall gezeigt, daß die Fehlerwahrscheinlichkeit 1. Art maximal ist, falls der Grenzfall der Unabhängigkeit zwischen den Statistiken Z_k^*, $k = 1, 2, \ldots, K$, vorliegt. Dies folgt aus der auf Slepian (1962) zurückgehenden Ungleichung

$$1 - P(\bigcap_{k=1}^{K} \{Z_k^* < u_k\}) \leq 1 - \prod_{k=1}^{K} P(Z_k^* < u_k) \tag{2.34}$$

für multivariat normalverteilte Zufallsgrößen mit Erwartungswertvektor $\mathbf{0}$ und Korrelationsmatrix $\boldsymbol{R}$ mit Elementen $\rho_{kk'} \geq 0$ ($u_1, u_2, \ldots, u_K$ beliebig). Setzt man $t_k = \epsilon^{K-k}$ und betrachtet man den Grenzfall $\epsilon \to 0$ (d.h. den Grenzfall $t_k = 0$ für $k < K$ und $t_K = 1$), so gilt

$$\lim_{\epsilon \to 0} Cov(Z_k^*, Z_{k'}^*) = \lim_{\epsilon \to 0} \sqrt{\frac{\epsilon^{K-k}}{\epsilon^{K-k'}}} = \lim_{\epsilon \to 0} \sqrt{\epsilon^{k'-k}} = 0 \tag{2.35}$$

für $k < k'$. Somit ist die rechte Seite in (2.34) auch die kleinste obere Schranke für die Fehlerwahrscheinlichkeit 1. Art, da dieser Fall den durch (2.35) beschriebenen Unabhängigkeitsfall beschreibt.

Analog läßt sich diese Überlegung auf den zweiseitigen Fall anwenden. Aufgrund der Gültigkeit der von Šidák (1967) hergeleiteten Ungleichung

$$1 - P(\bigcap_{k=1}^{K} \{|\, Z_k^* \,| < u_k\}) \leq 1 - \prod_{k=1}^{K} P(|\, Z_k^* \,| < u_k) \tag{2.36}$$

für multivariat normalverteilte Zufallsgrößen mit Erwartungswertvektor $\mathbf{0}$ und positiv definiter Korrelationsmatrix $\boldsymbol{R}$ folgt wie oben, daß die rechte Seite in (2.36) eine kleinste obere Schranke für das Niveau des Testverfahrens ist, falls die kritischen Schranken $u_1, u_2, \ldots, u_K$ verwendet werden. Man erkennt hieraus, daß beispielsweise bei der Verwendung von nicht adjustierten kritischen Werten der Standardnormalverteilung (vgl. das einführende Beispiel (2.7) auf S. 16) zum Niveau $\alpha = 0.05$ sogar

$$\lim_{K\to\infty} \sup_{0<t_1<\ldots<t_K=1} 1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{|\, Z_k^* \,| < 1.96\}) = \lim_{K\to\infty} 1 - (1-\alpha)^K = 1$$

gilt, und somit die Fehlerwahrscheinlichkeit beliebig wächst, falls auch die Anzahl der Zwischenauswertungen beliebig groß ist (in Proschan et al., 1992, wird gezeigt, daß sich dieses Ergebnis auch bei Verwendung der Schranken nach Pocock bzw. O'Brien und Fleming ergibt).

Legt man K fest, so ist durch (2.34) und (2.36) eine einfache Möglichkeit für die Adjustierung der gruppensequentiellen Schranken für beliebige Sequenzgrößen gegeben. Da für den Fall beliebiger Sequenzgrößen

$$\sup_{0<t_1<t_2<\ldots<t_K=1} 1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{Z_k^* < u_k\}) = 1 - \prod_{k=1}^{K} \Phi(u_k)$$

im einseitigen Fall bzw.

$$\sup_{0<t_1<t_2<\ldots<t_K=1} 1 - P_{\mathrm{H}_0}(\bigcap_{k=1}^{K} \{|\, Z_k^* \,| < u_k\}) = 1 - \prod_{k=1}^{K} (2\Phi(u_k) - 1)$$

im zweiseitigen Fall gilt, lassen sich entsprechend adjustierte Werte durch die Bedingung

$$1 - \prod_{k=1}^{K} \Phi(u_k) = \alpha \tag{2.37}$$

im einseitigen Fall bzw.

$$1 - \prod_{k=1}^{K} (2\Phi(u_k) - 1) = \alpha \tag{2.38}$$

im zweiseitigen Fall angeben (Wassmer, 1999a). Die hierfür adjustierten Werte werden im folgenden mit $\tilde{u}_k$, $k = 1, 2, \ldots, K$, bezeichnet.

Wie man leicht erkennt, lassen sich diese Werte in Pococks Design (d.h. $\tilde{u}_k = \tilde{u}$, $k = 1, 2, \ldots, K$) besonders leicht bestimmen. Sie sind im einseitigen Fall durch

$$\tilde{u} = \tilde{c}_{\mathrm{P}}(K, \alpha) = \Phi^{-1}(\sqrt[K]{1-\alpha})$$

und im zweiseitigen Fall durch

$$\tilde{u} = \tilde{c}_{\mathrm{P}}(K, \alpha) = \Phi^{-1}((\sqrt[K]{1-\alpha} + 1)/2)$$

gegeben. Im allgemeinen Fall sind sie durch (2.38) bzw. (2.37) spezifiziert, indem z.B. $\tilde{u}_k$ in der Δ-Klasse nach Wang und Tsiatis (1987) festgelegt wird (vgl. (2.25)). Die dann durch iterative Berechnung ermittelten kritischen Werte $\tilde{c}(K, \alpha, \Delta)$ sind in Tabelle 2.8 für den zweiseitigen bzw. den einseitigen Fall für $\alpha = 0.10, 0.05, 0.025, 0.01, 0.005$, $\Delta = 0$ (d.h. O'Brien und Fleming-Schranken) und $\Delta = 0.5$ (d.h. Pocock-Schranken) für $K = 2, 3, 4, 5$ und 10 tabelliert. Zum leichteren Vergleich sind auch die jeweiligen gruppensequentiellen Schranken für den Fall gleicher Sequenzgrößen angegeben.

In Abbildung 2.5 sind im zweiseitigen Testproblem für $K = 5$ und $\alpha = 0.05$ die Entscheidungsbereiche des adjustierten Falles denen des nicht adjustierten Falles gegenübergestellt.

Der Vergleich mit den nicht adjustierten Schranken zeigt keinen „dramatischen" Unterschied. Erst bei größer werdendem K ($K > 5$) ist ein entscheidender Nachteil bei Verwendung der adjustierten Werte im Vergleich zur Verwendung der Werte für gleiche Sequenzgrößen zu erkennen. Diese Eigenschaft ist beim Design nach O'Brien und Fleming etwas stärker ausgeprägt als bei Pococks Design. Beispielsweise lauten im Fall $\alpha = 0.05$ die zweiseitigen Schranken für die für den Abbruch der Studie benötigten p-Werte im nicht-adjustierten Pocock-Design bei $K = 2$ $p_1 = p_2 = 0.0294$, während die adjustierten Grenzen für die zweiseitigen p-Werte bei $K = 2$ durch $\tilde{p}_1 = \tilde{p}_2 = 0.0253$ gegeben sind.

Tabelle 2.8: *Kritische Werte $\tilde{c}(K, \alpha, \Delta)$ nach Pocock sowie O'Brien und Fleming im zweiseitigen bzw. einseitigen Testproblem unter der Voraussetzung unabhängiger Teststatistiken. In Klammern die Werte für den Fall gleicher Sequenzgrößen.*

Zweiseitiges Testproblem

		α			
	K	0.05	0.025	0.01	0.005
O'Brien/	2	2.828 (2.797)	3.200 (3.183)	3.655 (3.648)	3.976 (3.972)
Fleming	3	3.580 (3.471)	4.013 (3.935)	4.545 (4.494)	4.921 (4.885)
	4	4.247 (4.049)	4.731 (4.579)	5.326 (5.218)	5.748 (5.665)
	5	4.856 (4.562)	5.384 (5.151)	6.035 (5.861)	6.497 (6.358)
	10	7.394 (6.598)	8.093 (7.420)	8.956 (8.411)	9.573 (9.103)
Pocock	2	2.236 (2.178)	2.495 (2.449)	2.806 (2.772)	3.023 (2.995)
	3	2.388 (2.289)	2.635 (2.556)	2.934 (2.873)	3.143 (3.093)
	4	2.491 (2.361)	2.731 (2.625)	3.022 (2.939)	3.227 (3.156)
	5	2.569 (2.413)	2.804 (2.674)	3.089 (2.987)	3.290 (3.203)
	10	2.800 (2.555)	3.020 (2.811)	3.289 (3.117)	3.480 (3.329)

Einseitiges Testproblem

		α			
	K	0.05	0.025	0.01	0.005
O'Brien/	2	2.431 (2.373)	2.829 (2.797)	3.315 (3.300)	3.655 (3.648)
Fleming	3	3.118 (2.961)	3.583 (3.471)	4.148 (4.077)	4.545 (4.495)
	4	3.735 (3.466)	4.251 (4.049)	4.882 (4.740)	5.326 (5.218)
	5	4.299 (3.915)	4.862 (4.562)	5.549 (5.330)	6.035 (5.861)
	10	6.665 (5.696)	7.404 (6.598)	8.312 (7.670)	8.958 (8.411)
Pocock	2	1.955 (1.875)	2.239 (2.178)	2.575 (2.531)	2.807 (2.772)
	3	2.121 (1.992)	2.391 (2.289)	2.712 (2.636)	2.935 (2.873)
	4	2.234 (2.067)	2.494 (2.361)	2.806 (2.704)	3.023 (2.939)
	5	2.319 (2.122)	2.572 (2.413)	2.877 (2.754)	3.090 (2.986)
	10	2.568 (2.270)	2.803 (2.555)	3.089 (2.889)	3.290 (3.117)

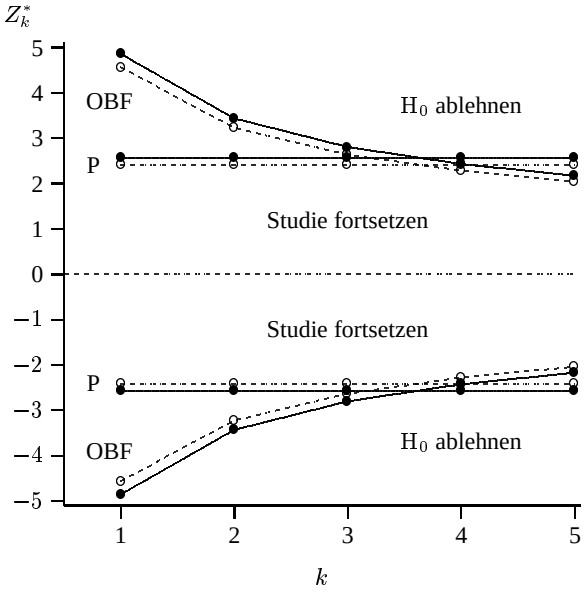

Abbildung 2.5: *Fortsetzungs- bzw. Annahmebereiche für O'Brien und Flemings (OBF) bzw. Pococks (P) Design mit beliebigen Stichprobenumfängen (durchgezogene Linie) im Vergleich mit den Bereichen für gleiche Sequenzgrößen (gestrichelte Linie);* $K = 5, \alpha = 0.05$.

Die benötigten p-Werte in O'Brien und Flemings Design lauten ceteris paribus $(p_1, p_2) = (0.0052, 0.0480)$ bzw. $(\tilde{p}_1, \tilde{p}_2) = (0.0047, 0.0455)$. Für $K = 5$ ergibt sich

– in Pococks Design:

$$p_1 = p_2 = p_3 = p_4 = p_5 = 0.0158 \qquad \text{bzw.}$$
$$\tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3 = \tilde{p}_4 = \tilde{p}_5 = 0.0102$$

– in O'Brien und Flemings Design:

$$(p_1, p_2, p_3, p_4, p_5) = (0.000005, 0.0013, 0.0084, 0.0226, 0.0413) \quad \text{bzw.}$$
$$(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4, \tilde{p}_5) = (0.000001, 0.0006, 0.0051, 0.0152, 0.0299)$$

O'Brien und Flemings Design reagiert damit etwas empfindlicher auf diese Art der Adjustierung, was für noch größere K zu recht konservativen Testverfahren führen würde. Welchen Effekt dies auf die $power$ der Verfahren hat, wird in Abschnitt 2.3.4 dargestellt. Man beachte auch, daß die numerische Übereinstimmung der zweiseitigen kritischen Schranken zum Niveau α mit den einseitigen Schranken zum Niveau $\alpha/2$ bei den adjustierten kritischen Werten nicht mehr gegeben ist.

Dem Nachteil, daß diese Art der Adjustierung kritischer Werte zu konservativen Testverfahren führen kann, steht ein großer Vorteil gegenüber: Unabhängig von den Sequenzgrößen führen die Verfahren zu validen Niveau-α-Testverfahren. Insbesondere brauchen die Stichprobenumfänge der Sequenzen des gruppensequentiellen Plans bei Studienbeginn nicht bekannt zu sein. So läßt sich beispielsweise das in den einführenden Bemerkungen angedeutete Vorgehen durchführen, nach spezifizierten Zeitpunkten die Zwischenauswertungen zu planen. Solange die weitere Planung der Zwischenauswertungen datenunabhängig geschieht, ist bei Verwendung der in Tabelle 2.8 angegebenen kritischen Werte ein zulässiges Vorgehen garantiert.

Die Verwendung der auf diese Art adjustierten Werte ist extrem „pessimistisch“ in bezug auf die Stichprobenumfänge der Sequenzen des gruppensequentiellen Plans. Viel realistischer ist es, eine obere Grenze für die Größe der Sequenzen zu fordern. Beispielsweise könnte es vernünftig sein, höchstens eine Verdoppelung der Sequenzgrößen im Vergleich zur ersten Sequenzgröße n_1 zuzulassen. Diese

obere Grenze τ^o für τ_k, $k = 2, 3, \ldots, K$, muß vor Beginn der Studie festgelegt werden. Die unter dieser Nebenbedingung spezifizierten kritischen Werte sind gemäß (2.33) gegeben durch

$$\sup_{0<\tau_k\leq\tau^o,\, k=2,3,\ldots,K} P_{\mathrm{H}_0}(\bigcap_{k=1}^{K}\{Z_k^* \in \mathscr{C}_k\}) = 1-\alpha \ ,$$

was sich bei Verwendung des zweiseitigen Designs nach Pocock zu

$$\sup_{0<\tau_k\leq\tau^o,\, k=2,3,\ldots,K} P_{\mathrm{H}_0}(\bigcap_{k=1}^{K}\{\mid Z_k^* \mid < \tilde{u}\}) = 1-\alpha \qquad (2.39)$$

ergibt. Die auf diese Art bestimmten kritischen Werte können mit Hilfe nichtlinearer Optimierungsmethoden bestimmt werden (vgl. Wassmer, 1999a, und Anhang A.5). Es ergeben sich bis auf zwei Nachkommastellen angebbare numerisch stabile Resultate. In Tabelle 2.9 sind die sich so ergebenden kritischen Werte für $\alpha = 0.005$, 0.01, 0.025, 0.05, 0.10, $\tau^o = 2$, 4, 8, und $K = 2$, 3, 4, 5 angegeben, was praktisch relevante Fälle widerspiegeln soll. Entsprechend einseitige kritische Werte mit $u^L = -\infty$ (vgl. Abschnitt 2.2.3) zum Niveau α ergeben sich hier durch die in der Tabelle zum Niveau $2\,\alpha$ angegebenen Werte.

Die *worst case scenarios*, d.h. die Konstellationen der Sequenzgrößen, die die Fehlerwahrscheinlichkeiten 1. Art in Pococks Design maximieren, sind wie folgt charakterisiert: In allen betrachteten Fällen ist die Wahrscheinlichkeit maximal für $\tau_K = \tau^o$ und es gibt (für $K > 2$) einen maximierenden Wert $\tau_2, \ldots, \tau_{K-1}$, der in der überwiegenden Anzahl von Fällen durch ein $0 < \tau_2 < \tau^o$ und $\tau_3 = \ldots = \tau_{K-1} = \tau^o$ gegeben ist. Ist z.B. $K = 4$, $\alpha = 0.01$ und $\tau^o = 4$, dann gilt (2.39) für den Vektor (1.50, 4, 4) mit $\tilde{u} = 2.98$ (vgl. Tabelle 2.9). Die Berechnungen zeigen jedoch, daß dies nicht immer der Fall ist. Ist z.B. $\tau^o = 8$, dann gilt (2.39) für (1.73, 5.14, 8) mit $\tilde{u} = 2.99$ (vgl. Tabelle 2.9).

Dieses Verhalten ist in Abbildung 2.6 für $K = 3$, $\alpha = 0.05$ und $\tau^o = 4$ illustriert. In Abhängigkeit von τ_2 und τ_3 ergibt sich ein nicht-monotones Verhalten der Fehlerwahrscheinlichkeit 1. Art in Abhängigkeit von τ_2. Man erkennt darüber hinaus, daß das so entstandene Testverfahren sehr konservativ sein kann. Dies ergibt sich (in diesem Beispiel) für kleine Werte von τ_2 und τ_3, was dem Fall von kleinen Stichproben in späteren Sequenzen im Vergleich zur ersten Sequenz entspricht. Man beachte jedoch, daß dies nicht maßgeblich durch die hier

Tabelle 2.9: *Kritische Werte $\tilde{u} = \tilde{c}(K, \alpha)$ nach Pocock im zweiseitigen Testproblem; $n_1, \ldots, n_K$ beliebig, aber $\tau_k = n_k/n_1$ nach oben beschränkt durch τ^o, $k = 2, 3, \ldots, K$.*

		α				
τ^o	K	0.10	0.05	0.025	0.01	0.005
2	2	1.90	2.20	2.47	2.79	3.01
	3	2.03	2.32	2.58	2.89	3.11
	4	2.10	2.39	2.65	2.96	3.18
	5	2.16	2.45	2.70	3.01	3.23
4	2	1.92	2.22	2.48	2.80	3.02
	3	2.05	2.34	2.60	2.91	3.12
	4	2.13	2.42	2.67	2.98	3.19
	5	2.19	2.47	2.73	3.03	3.24
8	2	1.94	2.23	2.49	2.80	3.02
	3	2.07	2.35	2.61	2.92	3.13
	4	2.15	2.43	2.69	2.99	3.20
	5	2.21	2.49	2.74	3.04	3.25

durchgeführte Adjustierung der kritischen Werte verursacht ist, sondern eine generelle Eigenschaft von Pococks Design darstellt.

Eine etwas andere Situation ergibt sich für das Design nach O'Brien und Fleming. Die Berechnungen zeigen, daß die Fehlerwahrscheinlichkeit mit fallendem τ_2 wächst. Da der Fall $\tau_2 = 0$ nur als Grenzfall existiert, wird das Maximum nicht angenommen. Aus praktischen Erwägungen heraus bietet sich deshalb hier an, auch eine untere Schranke τ^u für τ_k, $k = 2, 3, \ldots, K$, anzugeben, und die kritischen Werte im zweiseitigen Fall durch die Bedingung

$$\sup_{\tau^u \le \tau_k \le \tau^o,\, k=2,3,\ldots,K} P_{\mathrm{H}_0}\Big(\bigcap_{k=1}^{K} \{\mid Z_k^* \mid < \tilde{c}(K,\alpha)/\sqrt{k}\}\Big) = 1 - \alpha \qquad (2.40)$$

zu definieren. Analog Tabelle 2.9 sind in Tabelle 2.10 die sich durch (2.40) ergebenden adjustierten kritischen Werte für $\tau^u = 0.10$ angegeben. Die einseitigen Niveau-α-Werte ergeben sich analog Tabelle 2.9 durch Betrachtung des Falles $2\,\alpha$.

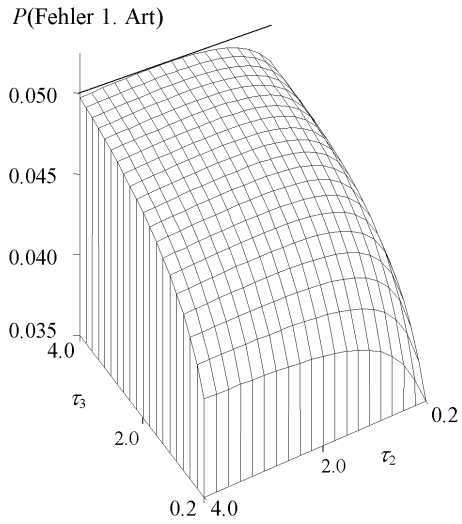

Abbildung 2.6: *Wahrscheinlichkeit für den Fehler 1. Art im zweiseitigen Design nach Pocock in Abhängigkeit von τ_2, τ_3 bei Benutzung des kritischen Wertes $\tilde{u} = 2.34$ im Fall $\tau^o = 4$, $K = 3$, $\alpha = 0.05$ (vgl. Tabelle 2.9). Das Maximum wird erreicht für $\tau_2 = 2.47$ und $\tau_3 = 4$.*

In allen in Tabelle 2.10 betrachteten Fällen ist die Fehlerwahrscheinlichkeit 1. Art maximal für $\tau_K = \tau^o$ und $\tau_2 = \tau^u$ bzw. für $\tau_2 = \tau^o$, falls $K = 2$. Für $K > 3$ existiert typischerweise ein maximierender Wert $\tau_3, \ldots, \tau_{K-1}$ mit $\tau^u < \tau_k < \tau^o$, $k = 3, \ldots, K-1$. Beispielsweise gilt im zweiseitigen Design mit $K = 4$, $\alpha = 0.05$, $\tau^u = 0.10$ und $\tau^o = 4$ (2.40) für den Vektor (0.10, 0.55, 4) mit $\tilde{c}(K, \alpha) = 4.17$ (vgl. Tabelle 2.10). Dieser Fall ist in Abbildung 2.7 in der τ_2-τ_3-Ebene illustriert. Hieraus ist zu erkennen, daß die Konservativität im Vergleich zu Pococks Design nicht so groß ist, was durch die in τ_2 bestehende Antitonie der Fehlerwahrscheinlichkeit verursacht ist. Ist $\tau^o = 2$, dann gilt (2.40) ceteris paribus für den Vektor (0.10, 0.10, 2) mit $\tilde{c}(K, \alpha) = 4.14$. Dies bedeutet, daß der maximierende Wert auch am Rand des zulässigen Bereichs für τ_k liegen kann.

Tabelle 2.10: *Kritische Werte $\tilde{c}(K, \alpha)$ nach O'Brien und Fleming im zweiseitigen Testproblem; $n_1, \ldots, n_K$ beliebig, aber $\tau_k = n_k / n_1$ nach oben beschränkt durch τ^o und nach unten beschränkt durch $\tau^u = 0.10$, $k = 2, 3, \ldots, K$.*

		α				
τ^o	K	0.10	0.05	0.025	0.01	0.005
2	2	2.39	2.81	3.19	3.65	3.97
	3	3.03	3.53	3.98	4.53	4.91
	4	3.57	4.14	4.66	5.28	5.72
	5	4.07	4.70	5.27	5.96	6.44
4	2	2.41	2.82	3.19	3.65	3.98
	3	3.06	3.55	4.00	4.54	4.91
	4	3.62	4.17	4.68	5.30	5.73
	5	4.13	4.74	5.31	5.98	6.46
8	2	2.42	2.82	3.20	3.65	3.98
	3	3.08	3.56	4.00	4.54	4.92
	4	3.65	4.20	4.70	5.31	5.74
	5	4.17	4.78	5.33	6.00	6.48

Der Vergleich der in den Tabellen 2.9 und 2.10 angegebenen kritischen Werte mit den kritischen Werten in Tabelle 2.8 verdeutlicht die sich durch die unteren und oberen Schranken für τ_k ergebenden Effekte. Die kritischen Werte unterscheiden sich nur dann in stärkerem Ausmaß, falls eine deutliche Nebenbedingung an τ_k gestellt werden kann ($\tau^o = 2$). In allen anderen Fällen ist der Effekt eher marginal. Die adjustierten Werte für $\tau^o = 2$ unterscheiden sich nur wenig von den nicht adjustierten Werten. Umgekehrt präzisiert dies die von Pocock (1977, 1982) vorgeschlagene Verwendung der nicht adjustierten kritischen Werte, falls eine nur „geringe“ Abweichung von der Voraussetzung gleicher Sequenzgrößen vorliegt. Die trotzdem notwendige (geringfügige) Adjustierung wird durch die angegebene numerische Methode definiert.

Der vorgeschlagene Ansatz für gruppensequentielle Pläne mit beliebigen und auch nicht vorhersagbaren Sequenzgrößen beruht auf der Betrachtung des für die Fehlerwahrscheinlichkeit 1. Art ungünstigsten Falles. Das resultierende Testverfahren hält damit das vorgegebene Niveau α ein. Das Niveau α wird voll

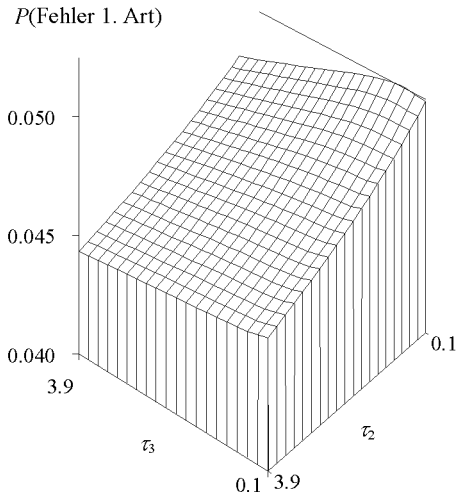

Abbildung 2.7: *Wahrscheinlichkeit für den Fehler 1. Art im zweiseitigen Design nach O'Brien und Fleming in Abhängigkeit von τ_2, τ_3 bei Benutzung des kritischen Wertes $\tilde{u} = 4.17$ im Fall $\tau^u = 0.10$, $\tau^o = 4$, $K = 4$, $\alpha = 0.05$ (vgl. Tabelle 2.10). Das Maximum wird erreicht für $\tau_2 = 0.10$, $\tau_3 = 0.55$ und $\tau_4 = 4$.*

ausgeschöpft, falls die ungünstigste Sequenzkonfiguration eintritt. In allen anderen Fällen ist das Verfahren konservativ. Welchen Effekt die Konservativität auf die $power$ des Verfahrens hat, wird in Abschnitt 2.3.4 besprochen. Eine praktische Anwendung des vorgeschlagenen Verfahrens mit Nebenbedingungen ist dann gegeben, falls man bei der Studienplanung die Größe der Sequenzen nicht strikt vorgeben, aber in einem gewissen Rahmen halten will. Die Rahmenvorgaben ergeben sich meist aus ökonomischen Erwägungen. Wie bereits erwähnt, läßt dieses Verfahren eine sich aus den Ergebnissen der Studie durchgeführte Weiterplanung der Sequenzen *nicht* zu. Trotzdem können studienunabhängige Faktoren diese Planung leiten und dementsprechend Anpassungen am Studienprotokoll vorgenommen werden (vgl. Pocock, 1996, für die Rolle von „exter-

nen“ Faktoren im *monitoring* einer Studie). Im nächsten Abschnitt wird auf ein inzwischen sehr bekanntes Verfahren eingegangen, das ähnlich wie der eben besprochene Ansatz für unvorhersagbare und beliebige Sequenzgrößen konzipiert ist.

2.3.3 Der α-*spending function*-Ansatz

Der auf der α-$spending\ function$ oder $use\ function$ basierende Ansatz geht auf die Arbeiten von Lan und DeMets (1983) und Kim und DeMets (1987b) zurück (vgl. DeMets und Lan, 1994). Verwandte Verfahren wurden von Fleming, Harrington und O'Brien (1984) und Slud und Wei (1982) vorgeschlagen. Der α-$spending\ function$-Ansatz oder kurz α-$spending$-Ansatz ist für den Fall konzipiert, Zwischenauswertungen an beliebigen Zeitpunkten der Datenerhebung durchzuführen. Die Stichprobenumfänge der Sequenzen sind dabei beliebig und auch unvorhersagbar. Dies wird durch die Angabe einer streng monoton wachsenden Funktion $\alpha^*(t_k)$ erreicht, die die bis zum Zeitpunkt t_k „verbrauchte“ Fehlerwahrscheinlichkeit 1. Art spezifiziert. An $\alpha^*(t_k)$ wird die Nebenbedingung $0 \leq \alpha^*(t_k) \leq \alpha$ gestellt. t_k ist wie in Abschnitt 2.3.2 definiert. Der gesamte Analysezeitraum ist damit durch das Einheitsintervall $[0;\ 1]$ gegeben, und t_k wird in diesem Zusammenhang als „Informationsrate“ oder „Informationszeit“ der k-ten Zwischenauswertung bezeichnet (im Unterschied zur *Kalenderzeit*; man vgl. dazu Lan und DeMets, 1989; Lan, Reboussin und DeMets, 1994).

Die Anwendung dieses Ansatzes beruht auf der folgenden Vorgehensweise. Bei Vorgabe des maximalen Stichprobenumfangs N und der Funktion $\alpha^*(t_k)$ läßt sich das bis zur ersten Zwischenauswertung „zu verbrauchende“ Signifikanzniveau $\alpha^*(t_1)$ angeben, wobei $t_1 = n_1/N$ den ersten Informationszeitpunkt definiert. Man beachte, daß n_1 nicht vorgegeben zu sein braucht. Die zur Ablehnung von H_0 und zum Abbruch der Studie nach der ersten Sequenz benötigte kritische Schranke läßt sich im zweiseitigen Fall durch die Bedingung

$$P_{\mathrm{H}_0}(|\,Z_1^*\,| \geq u_1) = \alpha^*(t_1)$$

ermitteln, was zu $u_1 = \Phi^{-1}(1 - \alpha^*(t_1)/2)$ führt. Zum Zeitpunkt $t_2 = (n_1 +$

$n_2)/N$ lautet die Bedingung für die kritische Schranke u_2:

$$P_{\mathrm{H}_0}(|\,Z_1^*\,| < u_1 \cap |\,Z_2^*\,| \geq u_2) = \alpha^*(t_2) - \alpha^*(t_1)\ ,$$

deren Auflösung durch die in den letzten Abschnitten beschriebene numerische Methode mit ungleichen Sequenzgrößen zu geschehen hat. Damit ist bis zum Zeitpunkt t_2 das Signifikanzniveau $\alpha^*(t_2)$ erreicht. Diese Spezifikation der kritischen Schranken wird fortgeführt, bis das gesamte Signifikanzniveau α ausgeschöpft ist. Die kritischen Schranken $u_1, u_2, \ldots, u_K$ sind damit rekursiv durch die Bedingung

$$P_{\mathrm{H}_0}\Big(\bigcap_{\bar{k}=1}^{k-1}\{|\,Z_{\bar{k}}^*\,| < u_{\bar{k}}\} \cap |\,Z_k^*\,| \geq u_k\Big) = \alpha^*(t_k) - \alpha^*(t_{k-1}) \tag{2.41}$$

gegeben, was einer Aufteilung des Signifikanzniveaus α auf die einzelnen Stufen des Experiments entspricht. Der einseitige Fall wird völlig analog behandelt. Man beachte, daß die maximale Anzahl K der Zwischenauswertungen nicht spezifiziert ist. Durch die Vorgabe des maximalen Stichprobenumfangs N ist diese zwar implizit gegeben, die Ermittlung der kritischen Schranken geschieht aber allein durch die zum Zeitpunkt der Zwischenauswertung beobachtete Informationszeit t_k. Bei Vorgabe der kalendarischen Zeitpunkte der Zwischenauswertungen sind diese durch den bis zur k-ten Zwischenauswertung erhobenen Anteil des Stichprobenumfangs am maximalen Stichprobenumfang gegeben. Dieser Anteil ist aber nicht vorgegeben, sondern ergibt sich aus dem konkreten Verlauf der Studie.

Es wurden verschiedene Vorschläge für die Wahl von $\alpha^*(t_k)$ gemacht. Beispielsweise ergibt sich durch die Wahl von

$$\begin{aligned} \alpha_1^*(t_k) &= \alpha\,\ln(1+(e-1)t_k) \quad \text{bzw.} \\ \alpha_2^*(t_k) &= 4(1-\Phi(\Phi^{-1}(1-\alpha/4)/\sqrt{t_k})) \end{aligned}$$

im Fall von gleichen Sequenzgrößen approximativ das Design von Pocock (1977) bzw. O'Brien und Fleming (1979). Kim und DeMets (1987b) schlug eine Klasse von Funktionen vor, die durch

$$\alpha_3^*(\delta, t_k) = \alpha\, t_k^\delta$$

gegeben ist. Beispiele von α-$spending$-Funktionen sind in Abbildung 2.8 illustriert. Die von Kim und DeMets (1987b) vorgeschlagenen Funktionen sind für $\delta = 1$, 1.5 und 2 angegeben und als zwischen α_1^* und α_2^* liegende Alternativen erkennbar.

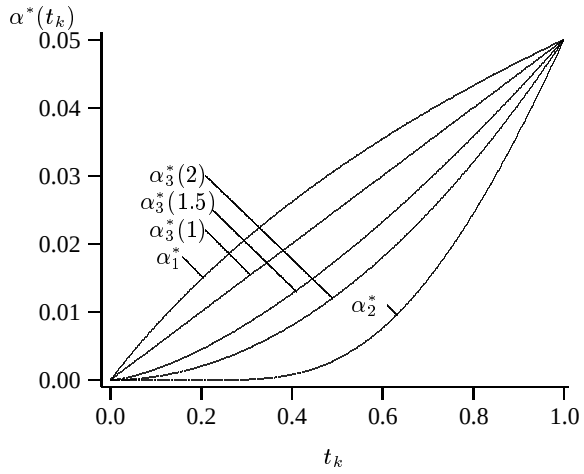

Abbildung 2.8: *Beispiele von α-$spending$-Funktionen. α_1^* und α_2^* ergeben approximativ kritische Werte nach Pocock bzw. O'Brien und Fleming. $\alpha_3^*(\delta, \cdot)$ nach Kim und DeMets (1987b) sind für $\delta = 1, 1.5$ und 2 angegeben; $\alpha = 0.05$.*

Noch allgemeiner schlugen Hwang, Shih und DeCani (1990) die Klasse

$$\alpha_4^*(\gamma, t_k) = \begin{cases} \alpha \, \frac{1-e^{-\gamma t_k}}{1-e^{-\gamma}} & \text{für} \quad \gamma \neq 0 \\ \alpha \, t_k & \text{für} \quad \gamma = 0 \end{cases}$$

vor und zeigten, daß die so ermittelten Pläne approximativ den Plänen in der Δ-Klasse von Wang und Tsiatis (1987) entsprechen und damit „nahezu optimal" sind. Alternativ schlug Jennison (1987) für das einseitige Testproblem eine innerhalb der Klasse aller gruppensequentieller Designs approximativ optimale vierparametrige Familie von α-$spending$-Funktionen vor. Li und Geller (1991) untersuchten allgemeine für α-$spending$-Funktionen zu postulierende Eigen-

schaften und schlugen stückweise linear konvexe Funktionen vor (vgl. auch Geller, 1994).

Am Beispiel der α-$spending$-Funktion α_2^* (O'Brien und Flemings Design) sei das Vorgehen durch die konkrete Angabe kritischer Werte illustriert. Angenommen, bei gegebenem Signifikanzniveau $\alpha = 0.05$ ist die erste Auswertung zur Informationszeit $t_1 = 0.4$ durchzuführen. Dies ergibt $\alpha_2^*(0.4) = 0.00079$ und $u_1 = 3.357$ als kritischen Wert der ersten Sequenz. Ist $t_2 = 0.80$, so ergibt sich durch die Berechnungsmethode der kritischen Schranken bei ungleichen Sequenzgrößen der Wert $u_2 = 2.255$. Bei $t_3 = 1$ (d.h. bei Erreichen der letzten Sequenz) kann die kritische Schranke $u_3 = 2.026$ berechnet werden. Die Werte u_1, u_2 und u_3 werden dabei sukzessive ermittelt, d.h. insbesondere, daß die Informationsraten t_k bei Studienbeginn nicht feststehen. Ist beispielsweise nach Durchführung der dritten Sequenz erst $t_3' = 0.9$ erreicht, so errechnet sich $u_3' = 2.179$ und bei $t_4' = 1$ ist $u_4' = 2.073$ die kritische Grenze. Dies bedeutet, daß man zwar an beliebigen Zeitpunkten die Zwischenauswertungen durchführen darf, daß jedoch – was selbstverständlich ist – die Anzahl der tatsächlich durchgeführten Zwischenauswertungen Einfluß auf die kritischen Schranken hat. In Anhang A.6 ist ein SAS-Programm angegeben, das die (sukzessive) Berechnung der kritischen Schranken durchführt.

Ein Problem für die Anwendung des beschriebenen Verfahrens ergibt sich, falls der maximale Stichprobenumfang N am Ende der Studie nicht erreicht, d.h. entweder über- oder unterschritten wird. Dies wird in der Praxis häufig geschehen, da eine Anwendung des Ansatzes besonders in solchen Fällen geeignet ist, in denen die Sequenzgrößen als Realisationen einer Zufallsgröße aufzufassen sind (beispielsweise bei Vorgabe der Analysezeitpunkte) und damit sog. *random underrunning* bzw. *random overrunning* mit großer Wahrscheinlichkeit auftritt. In diesen Fällen erreicht man jedoch durch einen einfachen „Trick" eine effiziente Anwendung des Verfahrens. Im Fall von *random overrunning* verhindert man eine Antikonservativität des Vefahrens dadurch, indem für $t_K > 1$ die α-$spending$-Funktion identisch α gesetzt wird. Allgemein wird dies durch die Definition

$$\tilde{\alpha}^*(t_K) := \min\{\alpha, \alpha^*(t_K)\}$$

erreicht. Ein *random underrunning* bei Beendigung der Studie tritt auf, falls ex-

terne Gründe für die Beendigung der Studie nach einer festgelegten Zeit sprechen. Beispielsweise können ökonomische Erwägungen für den erzwungenen Abbruch nach $t_K < 1$ führen, falls eine Weiterführung der Studie zu viele Ressourcen benötigt oder/und die Patienten-Rekrutierung eine längere Zeit als vorgesehen benötigt. Eine hier erlaubte Designvariante ergibt sich durch die Festlegung

$$\tilde{\alpha}^*(t_K) := \alpha \ ,$$

falls $t_K < 1$ die festgesetzte letzte Informationszeit bezeichnet (Kim, Boucher und Tsiatis, 1995). Man beachte, daß insbesondere *random underrunning* in vielen Fällen durch eine wenig sorgfältige Studienplan verursacht ist, und daher durch eine detailliertere Planung vermieden werden kann. Nichtsdestoweniger ist der auf der modifizierten α-$spending$-Funktion $\tilde{\alpha}^*(t_k)$ basierende Ansatz so flexibel, daß auch diese Abweichungen vom Studienprotokoll berücksichtigt werden können. Eine entsprechende Bestimmung der kritischen Schranken mit der modifizierten α-$spending$-Funktion geschieht wie oben beschrieben durch die Bedingung (2.41).

2.3.4 Vergleich der Verfahren

Wassmer (1999a) verglich das auf der α-$spending$-Funktion beruhende Verfahren mit dem von ihm vorgeschlagenen Ansatz in bezug auf die Güte ($power$) und den ASN. Diese Größen wurden berechnet, indem verschiedene Konfigurationen der Informationszeiten $t_1, t_2, \ldots, t_K$ unterstellt wurden. Beispielsweise wurde der Fall der Verdoppelung der Information (d.h. des Stichprobenumfangs) nach der ersten und vor der letzten Sequenz betrachtet und mit dem Design mit gleichem Stichprobenumfang verglichen (vgl. Table 5 und 6 in Wassmer, 1999a). Als Designvarianten wurden Pococks und O'Brien und Flemings Design betrachtet (d.h. im α-$spending$-Ansatz die Funktionen $\tilde{\alpha}_1^*(t_k)$ bzw. $\tilde{\alpha}_2^*(t_k)$). Wie man leicht erkennt, bringt bei $K > 2$ eine Verdoppelung des Stichprobenumfangs nach der ersten Sequenz eine Verringerung der Anzahl der Sequenzen beim α-$spending$-Ansatz mit sich. Der *worst case scenario*-Ansatz bleibt davon unberührt. Der Vergleich der Verfahren ist deshalb in manchen Fällen „ungerecht", da ein unterschiedliches Design zugrundegelegt wird.

Der Vergleich der $power$ und des ASN zeigt, daß der α-$spending$-Ansatz durchweg bessere $power$ besitzt, solange das für diesen Ansatz zugrundegelegte Design eingehalten oder „fast“ eingehalten wird. Dies folgt aus der Tatsache, daß dieser Ansatz hierfür konzipiert ist und insbesondere das Niveau α *exakt* einhält. Der *worst case scenario*-Ansatz weist eine sich aus der Konstruktionsmethode ergebende Konservativität auf. Interessanterweise sind die Unterschiede der Verfahren in bezug auf $power$ und ASN jedoch nicht sehr gravierend und in den meisten Fällen sogar äußerst gering. In den von Wassmer (1999a) betrachteten Fällen, in denen ein fairer Vergleich zwischen den Verfahren gegeben ist, ergibt sich ein $power$-Unterschied von höchstens 2-3% (wachsend in K). Dies folgt aus der relativ stark ausgeprägten Unempfindlichkeit des klassischen Ansatzes auf ungleiche Sequenzgrößen.

Der *worst case scenario*-Ansatz besitzt in der Regel höhere $power$, falls in der letzten Auswertung deutlich mehr Beobachtungen als erwartet erhoben worden sind (d.h. $t_K \gg 1$). Will man definitiv nicht am maximalen Stichprobenumfang N festhalten und geschieht $random\ overrunning$ bereits in früheren als in der letzten Sequenz, so ist der $worst\ case$ *scenario*-Ansatz für die Anwendung besser geeignet. Hierfür ist der α-$spending$-Ansatz jedoch nicht konzipiert und sollte in solchen Fällen nicht angewandt werden. Der von Wassmer (1999a) vorgeschlagene Ansatz hat im Vergleich zum α-$spending$-Ansatz deutlich kleinere $power$, falls $t_K \ll 1$. Dies ist jedoch durch die am Ende des letzten Abschnitts beschriebene Modifizierung der α-$spending$-Funktion und dem damit verbundenen Wechsel der Designvariante verursacht. Ist z.B. $(t_1, t_2, t_3, t_4) = (0.25, 0.375, 0.5, 0.625)$ (d.h. Halbierung der Sequenzgrößen nach der ersten Stufe) und $\alpha = 0.05$, so ergeben sich bei Verwendung der Funktion $\tilde{\alpha}_1^*(t_k)$, die $random\ underrunning$ berücksichtigt (d.h. $\tilde{\alpha}_1^*(0.675) := 0.05$), die kritischen Werte $u_1 = 2.37$, $u_2 = 2.49$, $u_3 = 2.49$, $u_4 = 2.14$. Dies ist weit davon entfernt, *konstante* kritische Werte im Sinn von Pocock zu benutzen. Wassmers Ansatz kann diese Art von Designwechsel nicht berücksichtigen und besitzt hier dementsprechend geringe $power$. Darüber hinaus ist zu beachten, daß die so gegebene Modifizierung der α-$spending$-Funktion bereits vor Durchführung der letzten Sequenz festgelegt sein muß und – wie bereits bemerkt – einer ungenügenden Planung des gruppensequentiellen Designs zuzuschreiben ist.

Zusammenfassend läßt sich feststellen, daß sich die beiden Verfahren in bezug

auf die $power$ und den ASN nur wenig unterscheiden. Insbesondere ergibt sich durch die von Wassmer (1999a) vorgeschlagene, relativ „brutale“ Adjustierung kein gravierender $power$-Verlust, falls vergleichbare Situationen zugrundegelegt werden. Die auf der nächsten Seite folgende Übersicht faßt die wesentlichen Eigenschaften und Unterschiede der beiden Verfahren zusammen.

Prinzipiell sind die Verfahren in unterschiedlichen Situationen einsetzbar, die sich in den zu spezifizierenden Vorgaben für den gruppensequentiellen Plan unterscheiden. Der α-$spending$-Ansatz stellt ein sehr flexibles Verfahren dar, mit dem gruppensequentielle Studien mit nicht vorhersagbaren Sequenzgrößen geplant und durchgeführt werden können. Ein Nachteil des Ansatzes ist die Vorgabe des maximalen Stichprobenumfangs N in der Planungsphase der Studie. Obwohl dies bei sorgfältiger Planung der Studie feststehen sollte, ist eine Spezifizierung von N in der Praxis oft nur schwer möglich oder nicht praktikabel. Beispielsweise sollen auch externe Faktoren (z.B. neue wissenschaftliche Erkenntnisse) den weiteren Verlauf der Studie leiten und der geplante Umfang N ist nur als sehr grobe „Leitlinie“ gedacht. Diese Fälle werden in dem *worst case scenario*-Ansatz geeigneter berücksichtigt, der die Vorgabe der maximalen Anzahl K von Sequenzen verlangt. Beiden Verfahren gemein ist die Tatsache, daß die Planung des Stichprobenumfangs auf der Annahme von in der Regel gleichen oder fest gegebenen Sequenzgrößen zu geschehen hat.

Bei beiden Verfahren muß der weitere Verlauf einer Studie datenunabhängig geplant werden. Dies bedeutet insbesondere, daß die Sequenzgrößen nicht aufgrund der Ergebnisse der Zwischenauswertung bestimmt sein dürfen. In der Praxis ist ein solches Vorgehen aber durchaus erwünscht. Beispielsweise kann es sinnvoll sein, die Sequenzgröße der folgenden Stufe klein zu halten (im Sinne eines möglichst effizienten gruppensequentiellen Plans), falls eine Ablehnung der Nullhypothese bereits kurz bevor stand, und vice versa. Die Möglichkeit der Durchführung eines solchen Vorgehens bieten adaptive Pläne, die in Kapitel 3 beschrieben werden. Im nächsten Abschnitt wird ein Überblick über die Anwendungsmöglichkeiten der in diesem Kapitel besprochenen gruppensequentiellen Verfahren gegeben. Es zeigt sich nämlich, daß die Verfahren nicht nur für den Parallelgruppenvergleich, sondern für eine Vielzahl anderer Studientypen – bisweilen geeignet modifiziert – verwendet werden können.

Übersicht

Vergleich des $worst\ case\ scenario$- mit dem α-$spending$-Ansatz

	$worst\ case\ scenario$	α-$spending$
Maximale Anzahl K der Sequenzen	vorgegeben	flexibel
Maximale Größe N der Erhebung	flexibel, Beschränkung vorgesehen	in der Planung festzulegen, Modifizierung der α-$spending$-Funktion möglich
Kritische Werte	einmalige Berechnung, im Prüfprotokoll festgelegt	sukzessive Berechnung im Verlauf der Studie
Designvarianten	festgelegt durch Vorgabe der kritischen Werte	festgelegt durch die Wahl der α-$spending$-Funktion
Planungsaspekte	in der Regel durch Annahme gleicher Sequenzgrößen	
Planung der nachfolgenden Sequenzen	datenunabhängig	

2.4 Andere Studientypen

Die bisher beschriebenen Verfahren sind für den Parallelgruppenvergleich normalverteilter Daten bei bekannter Varianz σ^2 konzipiert. Die Annahme einer bekannten Varianz ist für die Praxis kaum geeignet. Ergebnisse von Simulationsuntersuchungen (Pocock, 1977) zeigen jedoch, daß durch die Schätzung von σ^2 mit der Stichprobenvarianz bei genügend großem Stichprobenumfang das tatsächliche Niveau des Testverfahrens kaum beeinflußt wird. Darüber hinaus ist auch die Normalverteilungsannahme von keiner maßgeblichen Bedeutung, da bei der Verwendung der dargestellten Entscheidungsbereiche für nicht normalverteilte Merkmale (z.B. exponentialverteilte Merkmale) bzw. für die hieraus resultierenden Statistiken das Niveau approximativ eingehalten wird (Pocock, 1977). Die üblichen Ein- oder Zwei-Stichprobentests für Proportionen sind in gruppensequentiellen Designs ebenso durchzuführen. Allgemein gilt, daß die Verfahren immer dann zu approximativ gültigen Verfahren führen, wenn die „Zuwächse“ pro Sequenz approximativ normalverteilt sind. Aus diesem Grund lassen sich gruppensequentielle Verfahren auch sofort für verteilungsfreie Situationen (z.B. für den Wilcoxon-Test, vgl. auch Slud und Wei, 1982) oder Überlebenszeitanalysen mit *logrank*-Tests angeben (DeMets und Gail, 1985; Gu, Follmann und Geller, 1999; Kim et al., 1995; Kim und Tsiatis, 1990; Köpcke, 1984; Lan und Lachin, 1990; Lan und Zucker, 1993; Li, 1999; Lin, Shen, Ying und Breslow, 1996; Olschewski und Schumacher, 1986; Tsiatis, 1982; Tsiatis, Rosner und Tritchler, 1985; Tsiatis, Boucher und Kim, 1995). Diese Tatsache ist auch der Grund dafür, daß entsprechende exakte Verfahren kaum entwickelt wurden. Es gibt einige wenige Arbeiten, in denen zum Teil umfangreiche Tabellen der kritischen Werte für exakte Verfahren angegeben und deren Berechnungsmethoden beschrieben sind (vgl. z.B. Lin, Wei und DeMets, 1991; Jennison und Turnbull, 1991a). Vom theoretischen Standpunkt ist die Verwendung von nur approximativ gültigen Entscheidungsbereichen als Manko anzusehen. Im nächsten Kapitel dieser Arbeit werden Verfahren beschrieben, bei denen das Niveau des Tests exakt eingehalten wird.

Die Anwendung gruppensequentieller Pläne wurde auch für andere, klinisch relevante Fragestellungen wie Analysen mit multiplen Endpunkten (z. B. Jennison und Turnbull, 1993a; Lachin, 1997; Lee, 1994; Lee, Kim und Tsiatis, 1996; Su

und Lachin, 1992; Tang, Gnecco und Geller, 1989), Dosis-Wirkungs-Studien oder mehrarmige Studien (z. B. Follmann, Proschan und Geller, 1994; Geller, Proschan und Follmann, 1995; Hughes, 1993; Liu, 1995; Lui, 1993, 1994; Proschan, Follmann und Geller, 1994), Bioäquivalenzstudien (z. B. Durrlemann und Simon, 1990; Jennison und Turnbull, 1993b; Müller und Schäfer, 1999; Kittelson und Emerson, 1999; Whitehead, 1996) und Crossover-Versuchsplänen (z. B. Cook, 1994, 1995, 1996; Lehmacher, 1997) betrachtet.

Diese Verfahren sind erweiterbar, beispielsweise auf multiple Fragestellungen in Dosis-Wirkungs-Studien oder Studien mit multiplen Endpunkten, allgemeinere Designs, etc. Erste Ansätze für diese und ähnliche Fragestellungen sind in Bauer (1989b, 1991), Bauer und Kieser (1999), Friede und Kieser (2000a), Friede, Miller, Bischoff und Kieser (2001), Hellmich (2001), Hommel (2001), Kieser, Bauer und Lehmacher (1999), Kieser und Lehmacher (1995), Kropf, Hommel, Schmidt, Brickwedel und Jepsen (2000), Lang, Auterith und Bauer (2000), Lehmacher, Kieser und Hothorn (2000), Proschan (1999a) und Tang und Geller (1999) zu finden. Die Arbeiten von Bauer und Kieser (1999), Hellmich (2001), Hommel (2001), Kieser et al. (1999), Kropf et al. (2000), Lang et al. (2000) und Lehmacher et al. (2000) beziehen sich auf die Anwendung gruppensequentieller Verfahren auf multiple Fragestellungen in adaptiven Designs. Dabei deutet sich an, daß gerade bei komplexeren Fragestellungen ein adaptives Design bedeutende Einsparungen bringen kann, z.B. durch Weglassen von deutlichen („signifikanten“) oder voraussichtlich unerheblichen Effekten und durch die Planung der nachfolgenden Stichprobenumfänge aufgrund des vorliegenden Datenmaterials. Die Grundlagen der adaptiven Testverfahren werden im nächsten Kapitel besprochen.

3 Adaptive Testverfahren

Die bisher besprochenen Verfahren sind dadurch charakterisiert, daß die bis zu einer Zwischenauswertung erhobenen Daten nicht zur weiteren Planung der Studie benutzt werden dürfen. So ist es beispielsweise nicht erlaubt, bei einem Ergebnis, das nahe der Signifikanz ist, die Stichprobengröße der nächsten Stufe entsprechend kleiner zu planen, da zu erwarten ist, daß bereits eine Folgebeobachtung mit weniger Patienten zum gewünschten Ergebnis führen kann. Diese Vorgehensweise ist jedoch erstrebenswert, da sie zu einer effizienteren Ausnutzung der Information in den Daten führen würde. Wie bereits mehrfach bemerkt wurde, ist dies bei den auf der Idee der wiederholten Signifikanztests beruhenden Verfahren nicht vorgesehen. Ausgehend von dieser theoretischen Konzeption ist es nicht zugelassen, eine derart *adaptive* Vorgehensweise anzuwenden.

Schon Stein (1945) beschrieb ein zweistufiges Testverfahren, das die Varianz der ersten Stufe für die Fallzahlberechnung der zweiten Stufe benutzt. Obwohl dieses Verfahren die erste Stufe des Experiments nur zur Varianzschätzung vorsieht, ist es als eines der ersten adaptiv konzipierten Vorgehensweisen zu verstehen. Hayre (1985) schlug vor, die jeweils nächsten Gruppengrößen adaptiv unter Benutzung der beobachteten Effektstärke zu bestimmen. Tendenziell kleine Gruppengrößen ergeben sich, falls ein Abbruch der Studie „fast“ erreicht wird, und größere Gruppengrößen, falls dies nicht der Fall ist. Als Abbruchgrenzen wurden dabei die vom *SPRT* (Wald, 1947) abgeleiteten Schranken verwendet. Hayre (1985) zeigte mit Hilfe von Simulationen und einigen asymptotischen Überlegungen, daß bei Verwendung dieses Verfahrens auch schon für kleine Gruppengrößen die tatsächliche Fehlerrate 1. Art vom vorgegebenen Niveau α nicht stark abweicht. Vom theoretischen Standpunkt ist dies jedoch unbefriedigend, da sich die Ergebnisse auf den *SPRT* beschränken und allgemeinere und exakte adaptive Methoden wünschenswert sind. In der Übersichtsarbeit von Jennison und Turnbull (1991b) wird vor der nicht korrekten adaptiven Verwendung gruppensequentieller Pläne gewarnt, da sie üblicherweise zu einer zwar kleinen, aber

bezeichnenden und in extremen Fällen auch zu einer gravierenden Erhöhung des tatsächlichen Niveaus führen können (vgl. auch Fleming et al., 1984). Insbesondere ist nicht klar, wie groß die Fehlerwahrscheinlichkeit 1. Art bei beliebiger datengesteuerter adaptiver Planung ausfallen kann.

Eine in den Arbeiten von Gould (Gould, 1992, 1995, 1997; Gould und Shih, 1992, 1998) beschriebene Möglichkeit des datengesteuerten Monitoring von klinischen Studien bezieht sich auf das Revidieren des Stichprobenumfangs aufgrund der Schätzung der Variabilität in der laufenden Studie. In Simulationen wurde gezeigt, daß das Niveau des Tests nicht stark davon beeinflußt wird. Insbesondere wurde ein auf dem EM-Algorithmus basierendes Verfahren vorgeschlagen, das die Schätzung der Varianz vor der Entblindung der Studie ermöglicht (vgl. insbesondere Gould und Shih, 1992). Obwohl es sich dabei um ein für der Praxis klinischer Studien sehr ansprechendes Vorgehen handelt, kann nur die Variabilität des interessierenden Merkmals und nicht die beobachtete Effektstärke für die weitere Planung herangezogen werden (vgl. auch Friede und Kieser, 2000b, 2001; Kieser und Friede, 2000a,b).

Eine in der Praxis immer häufiger verwendete Methode geht auf die Arbeit von Bauer (1989a) zurück und wurde in verschiedener Form von Bauer und Köhne (1994), Bauer und Röhmel (1995) und Bauer und Kieser (1999), sowie von Bauer et al. (1998) und Kieser et al. (1999) vorgeschlagen. Diese Verfahren beruhen auf der Kombination der p-Werte der einzelnen Sequenzen. Sie erlauben nicht nur eine Bestimmung des Stichprobenumfangs auf der Basis der gemachten Beobachtungen, sondern auch generell eine Adaption des Studiendesigns. Dies ist z.B. bei *many-to-one*-Vergleichen von Bedeutung, wenn die Anzahl der durchgeführten Vergleiche in den folgenden Stufen des gruppensequentiellen Plans aufgrund der Beobachtungen reduziert werden soll, da für bestimmte Vergleiche ein Effekt schon in der ersten Zwischenauswertung nachgewiesen werden kann; bei einem Testproblem mit multiplen Endpunkten kann die Anzahl der untersuchten Endpunkte reduziert werden, etc.

In diesem Kapitel wird wie in dem vorigen der Parallelgruppenvergleich univariat normalverteilter Daten behandelt, so daß sich die adaptive Planung des Experiments lediglich auf die datengesteuerte Berechnung des Stichprobenumfangs beziehen wird. Alternativ zu dem von Bauer konzipierten Ansatz entwickelten Proschan und Hunsberger (1995) eine speziell für diese Situation anzuwendende

adaptive Testprozedur, die auf der Ermittlung der maximal möglichen Fehlerwahrscheinlichkeit 1. Art und einer entsprechenden Adjustierung der kritischen Werte beruht. Diese Verfahren werden in den folgenden Abschnitten besprochen, Vor- und Nachteile ausgearbeitet und erweitert. Da die Verfahren im wesentlichen nur für den zweistufigen Fall vorgeschlagen wurden, wird dabei der Verallgemeinerung auf mehr als $K = 2$ Stufen besondere Beachtung geschenkt. Insbesondere können adaptive Verfahren in einer sehr allgemeinen Klasse von Verfahren formuliert werden, die auf speziellen Kombinationstestverfahren beruhen. Der zweistufige Fall wird zuerst behandelt.

3.1 Zweistufige Designs

Beim Mittelwertsvergleich zweier normalverteilter Populationen mit konstanter, aber unbekannter Varianz σ^2 wird das Testen von $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$ gegen eine einseitige Alternative betrachtet, die o.B.d.A. durch

$$\mathrm{H}_1 : \mu_1 - \mu_2 > 0$$

gegeben sei. Die Überprüfung von H_0 gegen H_1 wird in zwei Stufen durchgeführt. Für beliebige Stichprobenumfänge n_k pro Stufe und Population, $k = 1, 2$, kann formal ein t-Test durchgeführt werden, der den (exakten) p-Wert

$$p_k = 1 - G_{2(n_k-1)}(t_k)$$

liefert, wobei $G_{2(n_k-1)}(\cdot)$ die Verteilungsfunktion der t-Verteilung mit $2(n_k-1)$ Freiheitsgraden bezeichnet. t_k ist die Realisation der t-Statistik T_k aus Stufe k, die gegeben ist durch

$$T_k = \frac{\bar{X}_{1k} - \bar{X}_{2k}}{S_k} \sqrt{\frac{n_k}{2}} \,,$$

falls S_k^2 die aus Stufe k ermittelte gepoolte Varianzschätzung bezeichnet. Beide im folgenden beschriebenen Verfahren zeichnen sich dadurch aus, daß die Zwischenauswertung zum Abbruch der Studie führt, falls ein genügend großer oder auch zu kleiner oder sogar negativer Effekt beobachtet wird. Dazu wird eine Schranke α_0 für den p-Wert p_1 der ersten Stufe spezifiziert, und die Studie

mit der Nichtablehnung (Annahme) von H_0 beendet, falls $p_1 \geq \alpha_0$. Diese frühe Annahme von H_0 wurde bereits im Kontext von klassischen gruppensequentiellen Verfahren behandelt (vgl. auch Gould und Pecore, 1982) und die Verbindung dieser Pläne mit den adaptiven Verfahren wird in den folgenden Ausführungen diskutiert.

3.1.1 Der Ansatz von Bauer und Köhne

Bauer und Köhne (1994) schlugen ein Verfahren vor, das auf der Verknüpfung der beiden p-Werte p_1 und p_2 mit Fishers Kombinationstest beruht (Fisher, 1932). Dieser Test wird üblicherweise in Meta-Analysen, d.h. bei der Verknüpfung mehrerer unabhängiger Studien, verwendet (vgl. z.B. Hedges und Olkin, 1985; Sonnemann, 1991). Offensichtlich kann dieses Verfahren auch für gruppensequentielle Studien verwendet werden, indem die Ergebnisse der Stufen als unabhängige Realisationen des selben Experiments aufgefaßt werden und die Testentscheidung mit der bei Fishers Kombinationstest verwendeten „Produktregel“ durchgeführt wird (Bauer, 1989c). H_0 wird dabei abgelehnt, falls

$$p_1 \cdot p_2 \leq c_\alpha = \exp(-\frac{1}{2}\,\chi^2_{4,\alpha}), \tag{3.1}$$

wobei $\chi^2_{4,\alpha}$ das $(1-\alpha)$-Perzentil der χ^2-Verteilung mit 4 Freiheitsgraden bezeichnet. Diese (globale) Testentscheidung erfüllt die Niveaubedingung, da bei Gültigkeit von H_0 die p-Werte stochastisch unabhängig und auf dem Intervall $[0;\,1]$ stetig gleichverteilt sind. Deshalb sind $-2 \cdot \ln(p_k)$, $k = 1, 2$, stochastisch unabhängig exponentialverteilt mit Parameter $\lambda = 2$, und $-2 \cdot (\ln(p_1) + \ln(p_2))$ χ^2-verteilt mit 4 Freiheitsgraden (vgl. z.B. Pfanzagl, 1991, S.24f.). Durch (3.1) wird also ein Niveau-α-Test definiert. Gilt in der Zwischenauswertung bereits $p_1 \leq c_\alpha$, dann ist bei Verwendung dieses Testverfahrens die Studie mit der Ablehnung von H_0 abzubrechen, da die Durchführung des zweiten Teils der Studie keine Änderung bringen kann: Da p_2 in jedem Fall kleiner als 1 ist, wird für jedes Ergebnis der zweiten Stufe (3.1) erfüllt sein. Dies ist unter dem Begriff *nonstochastic curtailment* zu verstehen, da die Herleitung der Abbruchbedingung auf nicht-stochastischen Überlegungen beruht.

Entsprechend dieser Darstellung gilt aufgrund der Gleichverteilung der p-Werte

unter Gültigkeit von H_0:

$$\begin{aligned}\alpha \;=\;& P(\text{Fehler 1. Art in erster Stufe}) + P(\text{Fehler 1. Art in zweiter Stufe}) = \\ & \int_0^{c_\alpha} dp_1 + \int_{c_\alpha}^1 \int_0^{c_\alpha/p_1} dp_2\, dp_1 = \\ & c_\alpha + \int_{c_\alpha}^1 c_\alpha/p_1\, dp_1 = \\ & c_\alpha - c_\alpha \cdot \ln c_\alpha \;,\end{aligned}$$

was einer Berechnung der Fläche unter der durch $p_1 \cdot p_2 = c_\alpha$ definierten Funktion entspricht[1]. Eine auf dieser Vorgehensweise beruhende Modifikation des Kombinationstests, die von Bauer und Köhne (1994) vorgeschlagen wurde, beruht auf der Tatsache, daß ein Studienabbruch auch für $p_1 \geq \alpha_0$ geschehen soll. Der Entscheidungsalgorithmus dieser zweistufigen Testprozedur lautet:

- Ist $p_1 \geq \alpha_0$, so wird die Studie mit der Annahme von H_0 beendet.
- Ist $p_1 \leq \alpha_1$ $(\alpha_1 > c_\alpha)$, so wird die Studie mit der Ablehnung von H_0 beendet.
- Ist $\alpha_1 < p_1 < \alpha_0$, so wird die zweite Stufe der Studie durchgeführt.
- In der zweiten Stufe wird H_0 abgelehnt, falls $p_1 \cdot p_2 \leq c_\alpha$.

Für die Fehlerrate 1. Art gilt:

$$\begin{aligned}& P(\text{Fehler 1. Art}) = \\ & P(\text{Fehler 1. Art in erster Stufe}) + P(\text{Fehler 1. Art in zweiter Stufe}) = \\ & \int_0^{\alpha_1} dp_1 + \int_{\alpha_1}^{\alpha_0} \int_0^{c_\alpha/p_1} dp_2\, dp_1 = \\ & \alpha_1 + c_\alpha \cdot (\ln\, \alpha_0 - \ln\, \alpha_1) \;.\end{aligned}$$

Um einen Niveau-α-Test zu erhalten, muß α_1 daher die Gleichung

$$\alpha_1 + c_\alpha \cdot (\ln\, \alpha_0 - \ln\, \alpha_1) = \alpha$$

1. Man beachte, daß dies als Nebenprodukt eine alternative Berechnungsmöglichkeit der Perzentile der χ^2-Verteilung mit 4 Freiheitsgraden mit sich bringt. Dies folgt aus der Gültigkeit von $P(X \geq x) = \exp(-x/2)(1 + x/2)$, falls X χ^2-verteilt mit 4 Freiheitsgraden ist (Johnson und Kotz, 1970, S.173). Eine allgemeine Lösung für geradzahlige Freiheitsgrade wird sich aus den Ausführungen der folgenden Abschnitte ergeben.

erfüllen, was eine iterative Bestimmungsmethode für α_1 erfordert. Tabelle 3.1 enthält für $\alpha = 0.05$, 0.025, 0.01, 0.005, $\alpha_0 = 0.1$, 0.15, 0.2, 0.25, 0.3,. . .,1.0 die Werte α_1 und c_α, die für die Durchführung der Testprozedur erforderlich sind.

Tabelle 3.1: *Signifikanzniveau α_1 und Schranke c_α im zweistufigen Verfahren nach Bauer und Köhne (1994) mit Abbruch für H_0, falls $p_1 \geq \alpha_0$. Die in Klammern angegebenen kritischen Werte für die Teststatistik T_1 sind für große n_1, d.h. für den Fall berechnet, daß die t-Verteilung durch die Standardnormalverteilung approximiert werden kann.*

		$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
α_0	c_α	0.00870	0.00380	0.00131	0.00059
0.10		0.0426 (1.722)	0.0186 (2.084)	0.0064 (2.489)	0.0029 (2.759)
0.15		0.0381 (1.774)	0.0166 (2.129)	0.0057 (2.529)	0.0026 (2.795)
0.20		0.0348 (1.815)	0.0152 (2.165)	0.0052 (2.560)	0.0024 (2.824)
0.25		0.0321 (1.850)	0.0140 (2.196)	0.0048 (2.587)	0.0022 (2.849)
0.30		0.0299 (1.882)	0.0131 (2.224)	0.0045 (2.612)	0.0020 (2.872)
0.40		0.0263 (1.938)	0.0115 (2.274)	0.0040 (2.656)	0.0018 (2.913)
0.50		0.0233 (1.990)	0.0102 (2.319)	0.0035 (2.696)	0.0016 (2.950)
0.60		0.0207 (2.040)	0.0090 (2.364)	0.0031 (2.736)	0.0014 (2.987)
0.70		0.0183 (2.091)	0.0080 (2.410)	0.0027 (2.777)	0.0012 (3.025)
0.80		0.0159 (2.147)	0.0069 (2.460)	0.0024 (2.822)	0.0011 (3.067)
0.90		0.0133 (2.216)	0.0058 (2.522)	0.0020 (2.877)	0.0009 (3.119)
1.00		0.0087 (2.378)	0.0038 (2.669)	0.0013 (3.009)	0.0006 (3.242)

Wie man erkennt, kann für kleiner gewähltes α_0 bereits „früher“ (d.h. für größeres α_1) eine Ablehnung von H_0 in der ersten Stufe getroffen werden. Dies wird aus der in Abbildung 3.1 für $\alpha = 0.05$ und $\alpha_0 = 0.50$ illustrierten Ablehnregion in der p_1-p_2-Ebene deutlich. Der Wert α_1 ist so gewählt, daß die schraffierte Fläche gleich α ist. Die Fläche des durch die Bedingung ‘$p_1 \leq \alpha_1$’ definierten Rechtecks kann deshalb umso größer sein, je mehr durch die Bedingung ‘$p_1 \geq \alpha_0$’ ausgespart und bei der schraffierten Fläche darangesetzt werden kann. Man erkennt hier auch leicht das Verhalten für die Grenzfälle $\alpha_0 \to 1$ und $\alpha_0 \to \alpha$: Geht $\alpha_0 \to 1$, so ist $\alpha_1 = c_\alpha$, was einem Abbruch mit der Ablehnung von H_0 entspricht, falls die Testentscheidung nach Durchführung der ersten Stu-

fe bereits feststeht (*nonstochastic curtailment*); der Fall $\alpha_0 \to \alpha$ entspricht dem Design mit festem Stichprobenumfang n_1, und es ist $\alpha_1 = \alpha$.

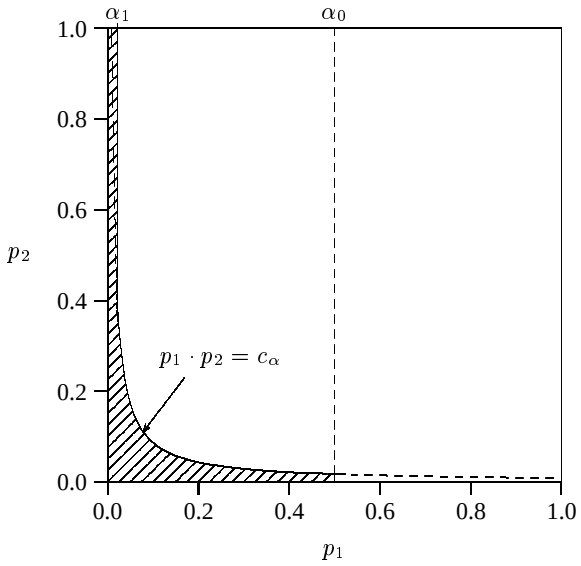

Abbildung 3.1: *Ablehnregion (schraffierte Fläche) für das zweistufige Verfahren nach Bauer und Köhne (1994);* $\alpha = 0.05$, $\alpha_0 = 0.50$, $\alpha_1 = 0.0233$.

Die beschriebene Modifikation des Kombinationstests beruht auf der Möglichkeit der „Verlagerung“ des Ablehnbereichs für die erste Stufe der Prozedur bei Berücksichtigung der frühen Annahme von H_0, falls $p_1 \geq \alpha_0$. Nach Durchführung der zweiten Stufe wird ein Kombinationstest nach Fisher zum vollen Niveau α durchgeführt. Eine allgemeinere Vorgehensweise verwendet den Kombinationstest zum Niveau α_2, d.h. die Ablehngrenze $c_{\alpha_2} = \exp(-1/2 \cdot \chi^2_{4,\alpha_2})$, $\alpha_2 < \alpha$, bei Durchführung der zweiten Stufe (Bauer und Röhmel, 1995). Der Entscheidungsalgorithmus lautet wie oben beschrieben, indem c_α durch c_{α_2} ersetzt wird, und der Wert $\alpha_1 \in [c_{\alpha_2};\, \alpha]$ durch die Bedingung

$$\alpha_1 + c_{\alpha_2} \cdot (\ln\, \alpha_0 - \ln\, \alpha_1) = \alpha$$

zu bestimmen ist. Die von Bauer und Köhne (1994) vorgeschlagene Prozedur

entspricht damit dem Fall $\alpha_2 = \alpha$. Eine andere Wahl von α_2 korrespondiert mit einer anderen Aufteilung des globalen Niveaus α auf die Zwischen- und die Endauswertung. In Tabelle 3.2 sind einige Werte α_1 und c_{α_2} für den Fall $\alpha_1 = \alpha_2$ angegeben. Diese Werte ergeben ein einseitiges gruppensequentielles Design mit konstanter Aufteilung des Niveaus auf die beiden Stufen der Studie. Dies entspricht einem Design nach Pocock, das im klassischen gruppensequentiellen Kontext in Abschnitt 2.2.3 besprochen wurde. Um eine Verbindung mit den in Tabelle 2.5 angegebenen Werte aufzuzeigen, wurde $\alpha_0 = 0.31$, 0.50, 0.84 und 1.0 gewählt. Dort wurde ein Abbruch mit der Annahme von H_0 mit $u^L = 0.5$, 0, –1.0 und $-\infty$ postuliert, was für den Fall bekannter Varianzen den angeführten α_0-Werten entspricht.

Tabelle 3.2: *Signifikanzniveau α_1 und Schranke c_{α_2} im zweistufigen Verfahren mit $\alpha_1 = \alpha_2$ und Abbruch für H_0, falls $p_1 \geq \alpha_0$. Die in Klammern angegebenen kritischen Werte für die Teststatistik T_1 sind für große n_1, d.h. für den Fall berechnet, daß die t-Verteilung durch die Standardnormalverteilung approximiert werden kann.*

α_0		$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
0.31	α_1	0.0371 (1.785)	0.0177 (2.103)	0.0068 (2.468)	0.0033 (2.715)
	c_{α_2}	0.00608	0.00254	0.00084	0.00037
0.50	α_1	0.0349 (1.813)	0.0169 (2.123)	0.0065 (2.483)	0.0032 (2.727)
	c_{α_2}	0.00566	0.00240	0.00080	0.00036
0.84	α_1	0.0329 (1.839)	0.0160 (2.143)	0.0063 (2.498)	0.0031 (2.739)
	c_{α_2}	0.00527	0.00226	0.00076	0.00034
1.00	α_1	0.0323 (1.848)	0.0158 (2.150)	0.0062 (2.502)	0.0030 (2.743)
	c_{α_2}	0.00515	0.00222	0.00075	0.00034

Der Vergleich mit den in Tabelle 2.5 angegebenen Werten zeigt, daß die für die erste Stufe benötigte Schranke für die Ablehnung von H_0 für das auf Fishers Kombinationstest beruhende Verfahren stets kleiner als für das einseitige gruppensequentielle Verfahren mit konstanten kritischen Schranken, der Möglichkeit der frühen Annahme von H_0 und der Voraussetzung $n_1 = n_2$ ist. Dies bedeutet, daß mit dem Kombinationstest die Ablehnung von H_0 in der Zwischenauswertung häufiger geschieht und demgemäß höhere $power$ auf der ersten Stufe

aufweist. Daraus folgt jedoch nicht, daß dieses Verfahren auch höhere globale $power$ besitzt. Die kritische Grenze für die für das klassische gruppensequentielle Verfahren anzuwendende Statistik $Z_2^* = (T_1 + T_2)/\sqrt{2}$ (vgl. Abschnitt 2.1) ergibt nämlich bei Verwendung der kritischen Schranken des dazugehörenden Kombinationstests den Wert $(\Phi^{-1}(1-\alpha_1) + \Phi^{-1}(1 - c_{\alpha_2}/\alpha_1))/\sqrt{2}$. Dieser Wert ist stets größer als der kritische Wert für die erste Stufe. Beispielsweise ist für $\alpha = 0.05$ und $\alpha_0 = 0.50$ der kritische Wert für Z_2^* durch $1.979 > 1.813$ gegeben. Somit liegen keine konstanten kritischen Schranken für die „übliche“ Verknüpfung der Testresultate vor, sondern lediglich eine konstante Aufteilung der Fehlerraten der beiden Stufen bei Verwendung von Fishers Kombinationstest. Dies verdeutlicht die prinzipiell unterschiedliche Vorgehensweise der beiden Ansätze. Insbesondere wird sich zeigen, daß der auf dem Kombinationstest beruhende Ansatz im Vergleich zum klassischen gruppensequentiellen Ansatz stets niedrigere $power$ besitzt.

Der entscheidende Vorteil des in den Arbeiten von Bauer vorgeschlagenen Verfahrens der Verknüpfung der separaten p-Werte besteht darin, daß sämtliche in der Zwischenauswertung zur Verfügung stehende Information zur Planung der nächsten Stufe verwendet werden darf. Trotz datengesteuerter Planung der zweiten Stufe ergibt sich ein exakter Niveau-α-Test. Im wesentlichen folgt dies aus der Tatsache, daß für jede Realisation der ersten Stufe der p-Wert p_2 der zweiten Stufe *bei Gültigkeit der Nullhypothese* auf dem Intervall $[0;\ 1]$ gleichverteilt ist. Somit besitzt p_2 eine von der ersten Stufe unabhängige Verteilung, wie auch immer die Daten und insbesondere der Stichprobenumfang der zweiten Sequenz zustande gekommen sind. Der formale Beweis, daß das so bestimmte Verfahren das globale Niveau α einhält, läßt sich dann folgendermaßen führen (vgl. Bauer, 1989a; Bauer und Kieser, 1999):

p_1 und p_2 seien die p-Werte der ersten bzw. zweiten Stufe der Studie. Die Stichprobe der ersten Stufe sei mit $\mathfrak{X}_1$ bezeichnet. Der Stichprobenumfang der zweiten Sequenz sei – beispielsweise durch den beobachteten Effekt in der ersten Stufe – adaptiv gewählt. Die Stichprobengröße der zweiten Sequenz ist damit eine Zufallsgröße N_2 (abhängig von den Ergebnissen $\mathfrak{X}_1$ der ersten Sequenz); deren Realisationen seien mit n_{2j}, $j \in J$, bezeichnet, wobei J eine endliche oder abzählbar unendliche Menge ist. Mit $p_0(n_{2j} \mid \mathfrak{X}_1)$, $j \in J$, werde die bedingte Wahrscheinlichkeit für die Wahl des Stichprobenumfangs n_{2j}, gegeben

der Beobachtung $\mathfrak{x}_1$, bezeichnet. Die konkrete Form von p_0 muß dabei nicht spezifiziert sein.

Die bedingte Dichte von p_2 unter H_0, gegeben der Wahl n_{2j} des Stichprobenumfangs und der Beobachtung $\mathfrak{x}_1$, sei mit $f_0(p_2 \mid n_{2j},\, \mathfrak{x}_1)$ bezeichnet. Diese bedingte Dichte exisiert, falls die erste Stufe nicht zum Abbruch der Studie geführt hat. Ist α_1 die Schranke für p_1 bei gegebenem Niveau α, dann ist f_0 definiert, falls $\alpha_1 < p_1 < \alpha_0$, und gegeben durch die Dichte der Gleichverteilung auf dem Intervall $[0;\, 1]$. Führt die zweite Stufe für beliebig und auch datenabhängig gewählten Stichprobenumfang der zweiten Sequenz zur Ablehnung von H_0, falls $p_1 \cdot p_2 \leq c_{\alpha_2}$, so gilt für die Fehlerwahrscheinlichkeit erster Art:

$P(\text{Fehler 1. Art}) =$

$$
\begin{aligned}
&P_{\mathrm{H}_0}(p_1 \leq \alpha_1) + \int_{\alpha_1}^{\alpha_0} \sum_{j \in J} P_{\mathrm{H}_0}\left(p_1 \cdot p_2 \leq c_{\alpha_2} \;\middle|\; n_{2j},\, \mathfrak{x}_1\right) p_0(n_{2j} \mid \mathfrak{x}_1)\, dp_1 = \\
&\int_0^{\alpha_1} dp_1 + \int_{\alpha_1}^{\alpha_0} \sum_{j \in J} \left(\int_0^{c_{\alpha_2}/p_1} f_0(p_2 \mid n_{2j},\, \mathfrak{x}_1)\, dp_2 \right) p_0(n_{2j} \mid \mathfrak{x}_1)\, dp_1 = \\
&\alpha_1 + \int_{\alpha_1}^{\alpha_0} \sum_{j \in J} \left(\int_0^{c_{\alpha_2}/p_1} dp_2 \right) p_0(n_{2j} \mid \mathfrak{x}_1)\, dp_1 = \\
&\alpha_1 + \int_{\alpha_1}^{\alpha_0} \underbrace{\sum_{j \in J} p_0(n_{2j} \mid \mathfrak{x}_1)}_{\leq\ 1} \left(\int_0^{c_{\alpha_2}/p_1} dp_2 \right) dp_1 \leq \\
&\alpha_1 + \int_{\alpha_1}^{\alpha_0} c_{\alpha_2}/p_1\, dp_1 = \alpha
\end{aligned}
$$

nach Bestimmungsmethode von α_1 und c_{α_2}.

□

Der Beweis der Aussage beruht also auf der Tatsache, daß unter Gültigkeit der Nullhypothese der p-Wert der zweiten Stufe durch die Dichte f_0 der stetigen Gleichverteilung gegeben ist (die Bedingung $\sum_{j \in J} p_0(n_{2j} \mid \mathfrak{x}_1) \leq 1$ ist trivialerweise erfüllt). Die für die Herleitung der kritischen Schranken von Fishers Kombinationstest benötigte stochastische Unabhängigkeit der p-Werte wird in

der Beweisführung *nicht* benötigt[2], sondern nur die *Bestimmungsmethode* der kritischen Schranken für dieses Testverfahren. Man beachte auch, daß die Grenzen des Integrals für f_0 unabhängig sind von n_{2j}. Wäre dies nicht der Fall, so wäre die letzte Abschätzung nicht gültig.

Die Berechnung des Stichprobenumfangs der zweiten Stufe auf der Basis der Daten der ersten Stufe kann auf verschiedene Arten geschehen. Beispielsweise können bedingte $power$-Berechnungen der zweiten Stufe aufgrund des beobachteten Effekts der ersten Stufe durchgeführt werden. Dies geschieht wie folgt. Für den Fall, daß die Varianz σ^2 als bekannt vorausgesetzt werden kann, ist der beobachtete standardisierte Effekt durch

$$\Phi^{-1}(1-p_1)\cdot\sqrt{\frac{2}{n_1}}$$

gegeben und bei gegebenem p_1 muß für eine Ablehnung von H_0 die Bedingung $p_2 \leq c_{\alpha_2}/p_1$ erfüllt sein. Für den Stichprobenumfang n_2 ergibt sich die Formel

$$n_2 = n_1 \cdot \frac{(\Phi^{-1}(1-c_{\alpha_2}/p_1)+\Phi^{-1}(1-\beta_2))^2}{(\Phi^{-1}(1-p_1))^2} \tag{3.2}$$

bei Vorgabe von $power = 1-\beta_2$ für den zweiten Teil der Studie. (3.2) kann dazu benutzt werden, um eine Neuberechnung von n_2 durchzuführen und über den weiteren Verlauf der Studie zu entscheiden. Wurde beispielsweise ein zweistufiges adaptives Design mit $\alpha_2 = \alpha = 0.05$, $\alpha_0 = 0.50$ und $n_1 = 30$ geplant und hat sich in der ersten Stufe ein p-Wert $p_1 = 0.027$ ergeben, so wird die Studie weitergeführt, da $\alpha_1 = 0.0233$ (vgl. Tabelle 3.1). Als Stichprobenumfang n_2 ergibt sich bei Vorgabe von 0.80 für die bedingte $power$ gemäß (3.2) der Wert $n_2 = 13.72$, womit die zweite Stufe pro Behandlungsgruppe 14 Beobachtungen erfordert, um bei gegebenem beobachteten Effekt der ersten Stufe mit Wahrscheinlichkeit 80 % zu einem signifikanten Resultat in der zweiten Stufe zu gelangen. Um mit Wahrscheinlichkeit 90 % zu einem signifikanten Resultat in

2. Sie ist auch streng genommen nicht erfüllt, da die Verteilung von p_2 nur existiert, falls $\alpha_1 < p_1 < \alpha_0$. Dieses Problem ließe sich umgehen, wenn für diesen Fall die Gleichverteilung festgelegt werden würde. Man beachte aber, daß der hier geführte Beweis auch für den allgemeineren Fall gültig ist, in dem für die Verteilung der p-Werte unter H_0 eine Verteilung gefordert wird, die stochastisch größer als die Gleichverteilung ist und evtl. vom Ergebnis der ersten Stufe abhängen kann. Dieser Fall tritt beispielsweise bei diskret verteilten Teststatistiken auf.

der zweiten Stufe zu gelangen, sind $n_2 = 25$ Beobachtungen erforderlich. Falls aufgrund derartiger Berechnungen entschieden wird, die Studie sogar mit gleich vielen Beobachtungen wie in der ersten Stufe durchzuführen (d.h. $n_2 = 30$), dann ergibt sich die bedingte $power$ durch die Formel

$$1 - \Phi(\Phi^{-1}(1 - c_{\alpha_2}/p_1) - \Phi^{-1}(1 - p_1)\sqrt{\frac{n_2}{n_1}})$$

und ist in dem Zahlenbeispiel durch 0.929 gegeben. Dieser Wert ist groß im Vergleich zu der $power$ 0.611 einer neu geplanten Studie mit dem in der Zwischenauswertung erhobenen Effekt und einem (festen) Stichprobenumfang $n = 30$. Dies verdeutlicht den in der ersten Stufe bereits beobachteten, relativ starken Effekt, der schon „fast“ zur Ablehnung der Nullhypothese führte.

Diese Vorgehensweise wird im folgenden Abschnitt noch weiter vertieft. Es soll an dieser Stelle jedoch schon darauf hingewiesen werden, daß sie ad absurdum geführt werden kann, falls sich ein weniger starker Effekt (d.h. ein p-Wert, der nicht nahe der Signifikanzgrenze α_1 ist) in der ersten Stufe ergibt. Beispielsweise führt für $p_1 = 0.10$ ein nach dieser Vorgehensweise berechneter Stichprobenumfang bei $1 - \beta_2 = 0.80$ ceteris paribus zu $n_2 = 89$, bei $p_1 = 0.30$ sogar zu $n_2 = 818$, was praktisch nicht mehr durchführbar erscheint. Es ist zu beachten, daß diese Eigenschaft nicht ohne weiteres dazu verleiten darf, den Wert α_0 für den frühen Abbruch für H_0 zu verkleinern. Dieser Wert ist vor Beginn der Studie unter Betrachtung des Ereignisses ‘$p_1 \geq \alpha_0$’ bei Gültigkeit einer vorab spezifizierten Alternativhypothese festzulegen. α_0 sollte generell nicht zu klein gewählt werden, da dieses Ereignis unter H_1 dann relativ hohe Wahrscheinlichkeit besitzen kann (vgl. Bauer und Köhne, 1994). In bisher mit diesem Plan durchgeführten Studien wurde meist der Wert $\alpha_0 = 0.50$ gewählt. Ein Abbruch der Studie mit negativem Studienresultat wird also (zwingend!) gefordert, falls der beobachtete Effekt nicht die gewünschte Richtung aufweist. n_1 darf nicht zu klein gewählt werden, da sonst die Wahrscheinlichkeit des Ereignisses ‘$p_1 \geq \alpha_0$’ unter Gültigkeit von H_1 recht groß ausfallen kann. Eine Wahl von $n_1 \geq 20$ sollte unter diesen Bedingungen als Faustregel gelten, was sich auch durch andere Betrachtungsweisen rechtfertigen läßt (vgl. z.B. Birkett und Day, 1994; Kieser und Wassmer, 1997).

Bei der Planung der Studie ist neben der Wahl von α_1, α_2 und α_0 die globale $power$ des zweistufigen Verfahrens bei fest vorgegebenen n_1 und n_2 zu be-

rücksichtigen. Da die Planung der zweiten Stufe datenabhängig ist, sind $power$-Betrachtungen bei vorgegebenen n_1 und n_2 zwar artifiziell und können per se den adaptiven Charakter der zweistufigen Prozedur nicht berücksichtigen. Dennoch erscheint es nicht nur der einfacheren theoretischen Handhabung wegen sinnvoll, diesen Fall zu betrachten, da sich hieraus Planungsaspekte für die Durchführung einer zweistufigen adaptiven Prozedur ergeben (vgl. Wassmer, 1997).

Um die $power$ des Kombinationstests berechnen zu können, muß die gemeinsame Verteilung der p-Werte p_k, $k = 1, 2$, unter der Alternativhypothese spezifiziert werden. Wegen der stochastischen Unabhängigkeit der p-Werte im Szenario mit fest vorgegebenen n_1 und n_2 ist die Verteilung gegeben durch das Produkt der Dichten von p_1 und p_2 unter H_1. Aus dem Transformationssatz für Dichten folgt die folgende Darstellung für die Dichte von p_k (vgl. Banik, Köhne und Bauer, 1996):

$$f_{\vartheta_k^*}(p_k) = \frac{g_{2\cdot n_k-2,\,\vartheta_k^*}(G^{-1}_{2\cdot n_k-2}(1-p_k))}{g_{2\cdot n_k-2,0}(G^{-1}_{2\cdot n_k-2}(1-p_k))}\ ,$$

wobei $g_{2\cdot n_k-2,\vartheta_k^*}(\cdot)$ die Dichte der nichtzentralen t-Verteilung mit $2 \cdot n_k - 2$ Freiheitsgraden und Nichtzentralitätsparameter $\vartheta_k^* = \frac{\mu_1-\mu_2}{\sigma}\sqrt{n_k/2}$ bezeichnet; $G^{-1}_{2\cdot n_k-2}(\cdot)$ ist die Umkehrfunktion der Verteilungsfunktion der t-Verteilung mit $2 \cdot n_k - 2$ Freiheitsgraden, $k = 1, 2$. Die gemeinsame Dichtefunktion

$$f_{\vartheta^*}(p_1, p_2) = f_{\vartheta_1^*}(p_1) \cdot f_{\vartheta_2^*}(p_2) \tag{3.3}$$

ist damit abhängig von $\vartheta^* = (\vartheta_1^*, \vartheta_2^*)$ und den Freiheitsgraden $2 \cdot n_k - 2$, $k = 1, 2$. Die $power$ ergibt sich durch zweidimensionale Integration der Ablehnbereiche und ist gegeben durch

$$1 - \int_{\alpha_0}^{1} f_{\vartheta_1^*}(p_1)\,dp_1 - \int_{\alpha_1}^{\alpha_0}\int_{c_{\alpha_2}/p_1}^{1} f_{\vartheta^*}(p_1, p_2)\,dp_2\,dp_1\ . \tag{3.4}$$

Die Dichte und die Berechnung des durch die zweidimensionale Integration sich ergebenden Volumens ist in Abbildung 3.2 illustriert.

Die Dichtefunktion (3.3) besitzt Singularitäten für $p_1 \to 0$ und $p_2 \to 0$. Das in Anhang A.7 beschriebene SAS-Programm benutzt das Integrationsmodul QUAD

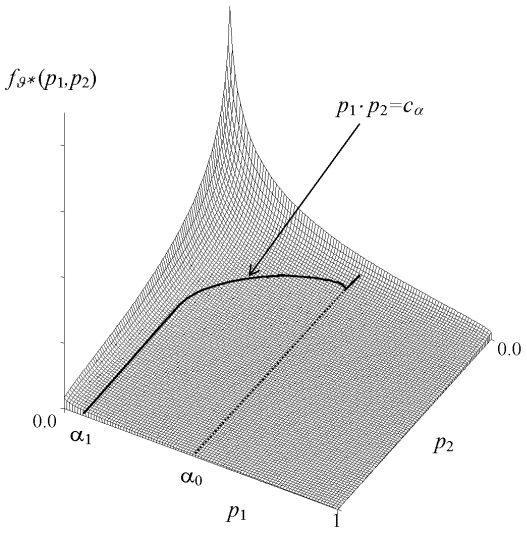

Abbildung 3.2: *Dichtefunktion $f_{\vartheta*}(p_1, p_2)$ unter der Alternativhypothese. Standardisierter Effekt=1, $\alpha_2 = \alpha$, $n_1 = n_2$ so groß, daß die t-Verteilung durch die Standardnormalverteilung approximiert werden kann.*

(verfügbar in SAS/IML ab Version 6.11, SAS Institute Inc., 1995), das die Berechnung von (3.4) trotz dieser numerischen Schwierigkeiten problemlos durchführt (Wassmer, 1997, vgl. auch Banik et al., 1996; Bauer und Köhne, 1994). Daneben wird die in SAS verfügbare Wahrscheinlichkeitsfunktion PDF verwendet, um die Dichtefunktion der nichtzentralen t-Verteilung zu berechnen [3]. In Wassmer (1997) wurde diese Möglichkeit genutzt, um einige Eigenschaften von zweistufigen Plänen im Parallelgruppenvergleich normalverteilter Daten zu untersuchen. Beispielsweise ergibt sich das interessante Resultat, daß der $power$-Verlust im Vergleich zu einem Plan ohne Zwischenauswertung mit fest vorgegebenem Stichprobenumfang $n_1 + n_2$ nicht besonders groß ist (vgl. auch Banik

3. Auf eine noch einfachere Berechnungsmöglichkeit der $power$ eines zweistufigen, auf Fishers Kombinationstest beruhenden Verfahrens wird am Ende von Abschnitt 3.1.4 eingegangen.

et al., 1996; Bauer und Köhne, 1994). Darüber hinaus konnte gezeigt werden, daß es bei im Vergleich zu n_2 großem Stichprobenumfang n_1 (d.h. n_1/n_2 groß) zu einem leichten $power$-Gewinn kommen kann, falls die frühe Annahme von H_0 in Betracht gezogen wird (vgl. Table 1 in Wassmer, 1997). Dies erscheint auf den ersten Blick widersprüchlich, da eine frühe Annahme von H_0 den Gesamtstichprobenumfang und demgemäß die globale $power$ kleiner macht. Bei dieser Überlegung bleibt die Tatsache jedoch unberücksichtigt, daß durch die Betrachtung des Ereignisses '$p_1 \geq \alpha_0$' die Schranke α_1 für die frühe Ablehnung von H_0 größer wird. Unter H_1 kann aber die Wahrscheinlichkeit des Ereignisses '$p_1 \leq \alpha_1$' größer sein als der „Verlust", der durch die Nichtberücksichtigung der Wahrscheinlichkeit des Ereignisses '$p_1 \geq \alpha_0$ und $p_1 \cdot p_2 \leq c_{\alpha_2}$' entsteht. Dies ist aber mehr von theoretischem denn praktischem Interesse, da der Gewinn an $power$ im Promille-Bereich liegt.

Die Berechnung der globalen $power$ kann dazu benutzt werden, den benötigten Stichprobenumfang beispielsweise bei Annahme gleicher oder fest vorgegebener Sequenzgrößen zu bestimmen und auf diese Art die Planung der Stichprobenumfänge eines zweistufigen Designs zu ermöglichen. Ein zweistufiges Design mit $\alpha_2 = \alpha = 0.05$ und $\alpha_0 = 0.50$ benötigt z.B. die Stichprobenumfänge $n_1 = n_2 = 28$, um zu einer globalen $power$ von mindestens 0.80 bei Vorgabe des standardisierten Effekts $\frac{\mu_1 - \mu_2}{\sigma} = 0.50$ zu gelangen. Zum Vergleich: Ein Plan ohne Zwischenauswertung benötigt ceteris paribus $n_1 + n_2 = 51$ Beobachtungen pro Gruppe, was den vergleichsweise geringen Verlust an $power$ verdeutlicht. Der durchschnittliche Stichprobenumfang ASN ist gegeben durch

$$ASN = n_1 + n_2 \cdot P_{H_1}(\alpha_1 < p_1 < \alpha_0) =$$
$$n_1 + n_2 \cdot \left(G_{2 \cdot n_1 - 2, \vartheta_1^*}(G^{-1}_{2 \cdot n_1 - 2}(1 - \alpha_0)) - G_{2 \cdot n_1 - 2, \vartheta_1^*}(G^{-1}_{2 \cdot n_1 - 2}(1 - \alpha_1))\right),$$

wobei $G_{2 \cdot n_1 - 2, \vartheta_1^*}(\cdot)$ die Verteilungsfunktion der nichtzentralen t-Verteilung mit $2 \cdot n_1 - 2$ Freiheitsgraden und Nichtzentralitätsparameter $\vartheta_1^* = \frac{\mu_1 - \mu_2}{\sigma}\sqrt{n_1/2}$ bezeichnet. Er ist für dieses Beispiel gegeben durch $ASN = 42.8$, was den Vorteil bei der Verwendung eines Verfahrens mit Zwischenauswertung widerspiegelt. In Tabelle 3.3 sind für $\alpha = 0.05$, 0.025, 0.01, $\alpha_0 = 0.30$, 0.50, 1.0, gegebene $power = 0.50$, 0.80, 0.90, 0.95, standardisierte Effektstärke $\frac{\mu_1 - \mu_2}{\sigma} = 0.20$, 0.50, 0.80 und das Design mit $\alpha_2 = \alpha$ die benötigten Stichprobenumfänge und der ASN angegeben. Zum Vergleich sind in dieser Tabelle auch die benötigten Stichprobenumfänge im Design ohne Zwischenauswertung enthalten.

Tabelle 3.3: *Benötigter maximaler Stichprobenumfang und ASN bei gegebenem standardisiertem Effekt $\frac{\mu_1-\mu_2}{\sigma}$, $power$ $1-\beta$ und $n_1 = n_2$ im zweistufigen Verfahren nach Bauer und Köhne (1994) ($\alpha_2 = \alpha$) mit Abbruch für H_0, falls $p_1 \geq \alpha_0$. In Klammern sind die Stichprobenumfänge im Design ohne Zwischenauswertung angegeben.*

			$\alpha = 0.05$		$\alpha = 0.025$		$\alpha = 0.01$	
$\frac{\mu_1-\mu_2}{\sigma}$	$1-\beta$	α_0	n_1+n_2	ASN	n_1+n_2	ASN	n_1+n_2	ASN
0.20	0.50	0.30	162 (136)	121.8	220 (194)	176.5	300 (272)	254.9
		0.50	152	126.8	212	184.0	294	264.0
		1.00	148	139.1	208	196.8	292	278.2
	0.80	0.30	362 (310)	255.1	440 (394)	328.8	546 (504)	430.3
		0.50	338	258.4	424	336.1	536	439.8
		1.00	330	283.0	418	363.0	534	470.1
	0.90	0.30	498 (429)	328.9	584 (527)	407.7	704 (653)	517.2
		0.50	466	330.7	564	414.9	692	527.5
		1.00	456	364.4	558	452.9	688	568.8
	0.95	0.30	630 (542)	391.4	722 (651)	472.2	850 (790)	584.2
		0.50	586	388.8	694	476.8	834	594.0
		1.00	574	429.1	686	522.2	832	646.0
0.50	0.50	0.30	28 (23)	21.2	38 (32)	30.7	50 (45)	42.8
		0.50	26	21.8	36	31.4	50	45.0
		1.00	26	24.4	36	34.1	50	47.7
	0.80	0.30	60 (51)	42.3	72 (64)	54.0	90 (82)	71.1
		0.50	56	42.8	70	55.6	88	72.4
		1.00	54	46.5	70	60.8	88	77.7
	0.90	0.30	82 (70)	54.1	96 (86)	67.0	114 (106)	84.2
		0.50	76	54.0	92	67.8	114	86.9
		1.00	74	59.4	92	74.7	112	93.1
	0.95	0.30	102 (88)	63.5	118 (105)	77.1	138 (128)	95.1
		0.50	96	63.6	114	78.2	136	97.0
		1.00	94	70.3	112	85.5	136	105.9
0.80	0.50	0.30	12 (10)	9.2	16 (14)	13.1	22 (19)	19.0
		0.50	12	10.1	16	14.0	22	19.9
		1.00	12	11.3	16	15.2	22	21.0
	0.80	0.30	24 (21)	17.1	30 (26)	22.5	36 (33)	28.8
		0.50	24	18.3	28	22.5	36	29.8
		1.00	22	19.1	28	24.6	36	32.0
	0.90	0.30	32 (28)	21.4	38 (34)	26.8	46 (43)	34.2
		0.50	32	22.6	38	28.0	46	35.3
		1.00	30	24.3	38	30.9	46	38.4
	0.95	0.30	42 (35)	26.0	48 (42)	31.3	56 (51)	38.6
		0.50	38	25.4	46	31.6	54	38.9
		1.00	38	28.6	46	35.1	54	42.6

Tabelle 3.3 illustriert den Einfluß von α_0 für den Fall gleicher Sequenzgrößen. Im Fall kleiner bis mittlerer Effektstärken ($\frac{\mu_1-\mu_2}{\sigma} = 0.20$ und 0.50) und $\alpha_0 = 0.30$ ist ein deutlicher Verlust an $power$ bzw. ein deutlich höher geforderter Stichprobenumfang im Vergleich zum „reinen“ Kombinationstest (d.h. $\alpha_0 = 1$) zu beobachten. Dagegen ist der $power$-Verlust für $\alpha_0 = 0.50$ vergleichsweise gering, da für mittlere bis starke Effektstärken ($\frac{\mu_1-\mu_2}{\sigma} = 0.50$ und 0.80) die benötigten Stichprobenumfänge um höchstens 2, d.h. die benötigten Stichprobenumfänge pro Stufe um höchstens 1, differieren. Der ASN ist für hohe geforderte $power$ i.d.R. am kleinsten, falls $\alpha_0 = 0.50$. Ist er am kleinsten für $\alpha_0 = 0.30$, so muß dies mit einem höheren Gesamtstichprobenumfang $n_1 + n_2$ „bezahlt“ werden.

Die Tabelle kann benutzt werden, um für gegebenes α, α_0, $1-\beta$ und standardisierte Effektstärke $\frac{\mu_1-\mu_2}{\sigma}$ den benötigten Stichprobenumfang und den ASN für ein zweistufiges Verfahren nach Bauer und Köhne (1994) abzulesen. Dies erscheint sinnvoll und das resultierende Verfahren ist in gewissem Sinne optimal, falls keine Adaption des Stichprobenumfangs in der Zwischenauswertung durchgeführt wird. Es sei jedoch noch einmal betont, daß derartige Betrachtungen die Einführung von adaptiven Verfahren *nicht* motivieren. Trotzdem können sie als Werkzeug oder „Krücke“ bei der anfänglichen Planung einer adaptiv konzipierten Studien dienen. Insbesondere korrespondieren die bei klassischen einseitigen zweistufigen Plänen (vgl. Abschnitt 2.2.3) sich ergebenden Werte für den maximalen Stichprobenumfang und den ASN mit den Werten in Tabelle 3.3. Man beachte, daß die Varianz σ^2 in dieser Betrachtung als nicht bekannt vorausgesetzt wird.

Ein Vergleich der $power$ einer spezifischen adaptiven Vorgehensweise mit der $power$ eines klassischen einseitigen zweistufigen Plans kann ebenso geführt werden. Die hierzu durchzuführenden Betrachtungen werden jedoch erst im nächsten Abschnitt besprochen. Dieser Abschnitt enthält die Beschreibung eines alternativen adaptiven Planes nach Proschan und Hunsberger (1995), der praktisch zeitgleich mit dem Ansatz von Bauer und Köhne (1994) vorgeschlagen wurde.

3.1.2 Der Ansatz von Proschan und Hunsberger

Proschan und Hunsberger (1995) schlugen ein zweistufiges Verfahren vor, das

auf der Betrachtung der maximalen Fehlerwahrscheinlichkeit 1. Art beruht, die bei einer datenabhängigen Wahl des Stichprobenumfangs der zweiten Stufe entsteht. Darauf aufbauend beschrieben sie ein allgemeines Verfahren zur Bestimmung einer adaptiven Testprozedur. Dieser Abschnitt beinhaltet die Beschreibung dieser Vorgehensweisen im Rahmen und in der Terminologie der bisher betrachteten Verfahren.

Proschan und Hunsberger entwickelten ihren Ansatz für den Fall des einseitigen Parallelgruppenvergleichs der Mittelwerte von normalverteilten Daten bei bekannter Varianz σ^2. In dem Testverfahren soll nach n_1 Beobachtungen aus dem beobachteten Wert z_1 von Z_1 der Stichprobenumfang $n_2 = n_2(z_1)$ bestimmt werden, wobei $n_2 = 0$ zugelassen ist. Für diesen Fall wird untersucht, welche Auswirkung die datengesteuerte Wahl von n_2 auf die Fehlerwahrscheinlichkeit 1. Art bei Vorgabe eines kritischen Wertes u hat, falls nach der zweiten Stufe die gepoolte Teststatistik Z_2^* (vgl. Abschnitt 2.1) verwendet wird. Ist beispielsweise u das $(1-\alpha)$-Quantil der Standardnormalverteilung, d.h. $u = \Phi^{-1}(1-\alpha)$, so entsteht die Frage, wie hoch das tatsächliche Niveau des sich ergebenden Testverfahrens ist. Die Beantwortung dieser Frage wird durch die Betrachtung der bedingten Fehlerrate 1. Art $P_{\mathrm{H}_0}(Z_2^* > u \mid z_1)$ ermöglicht. Für diese gilt:

$$P_{\mathrm{H}_0}(Z_2^* > u \mid z_1) = P_{\mathrm{H}_0}\left(\frac{\sqrt{n_1}Z_1 + \sqrt{n_2(Z_1)}Z_2}{\sqrt{n_1 + n_2(Z_1)}} > u \mid z_1\right) =$$

$$P_{\mathrm{H}_0}\left(\frac{\sqrt{n_1}z_1 + \sqrt{n_2(z_1)}Z_2}{\sqrt{n_1 + n_2(z_1)}} > u\right) , \tag{3.5}$$

da $E(\zeta(X,Y) \mid Y = y) = E(\zeta(X,y))$, falls X und Y unabhängig sind und ζ eine beliebige Funktion bezeichnet (vgl. z.B. Witting, 1985, S. 130). (3.5) ergibt

$$1 - \Phi\left(\frac{u\sqrt{1+\tau_2(z_1)} - z_1}{\sqrt{\tau_2(z_1)}}\right) , \tag{3.6}$$

falls $\tau_2(z_1) = n_2(z_1)/n_1$ und $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Die unbedingte Fehlerrate 1. Art ergibt sich dann durch Mittelung mit der Verteilung von Z_1; sie ist gegeben durch

$$\int_{-\infty}^{\infty} \left(1 - \Phi\left(\frac{u\sqrt{1+\tau_2(z_1)} - z_1}{\sqrt{\tau_2(z_1)}}\right)\right)\phi(z_1)\, dz_1 , \tag{3.7}$$

und die Frage ist, wie groß (3.7) bei Vorgabe des kritischen Wertes $u = \Phi^{-1}(1-\alpha)$ werden kann. Die maximale Fehlerwahrscheinlichkeit 1. Art kann mit Hilfe von Standardmethoden der Variationsrechnung (vgl. z.B. Bronstein und Semendjajew, 1987, S.379 ff.) durch die folgende Vorgehensweise ermittelt werden (vgl. das Appendix in Proschan und Hunsberger, 1995):

Ist $z_1 \geq u$, so ist (3.6) maximal für $\tau_2(z_1) \to 0$, d.h. für $n_2 \to 0$, da hierfür (3.6) identisch 1 ist.

Ist $z_1 < u$, so ist die Funktion

$$f(\tau_2) = \frac{u\sqrt{1+\tau_2} - z_1}{\sqrt{\tau_2}} \tag{3.8}$$

zu betrachten. (3.8) ist für $0 < z_1 < u$ minimal, falls $\tau_2 = \left(\frac{u}{z_1}\right)^2 - 1$, denn es gilt:

$$\frac{\partial f}{\partial \tau_2} = \frac{1}{\tau_2}\left(\frac{u}{2\sqrt{1+\tau_2}}\sqrt{\tau_2} - (u\sqrt{1+\tau_2} - z_1)\frac{1}{2\sqrt{\tau_2}}\right) =$$

$$\frac{u\tau_2 - u(1+\tau_2) + z_1\sqrt{1+\tau_2}}{2\tau_2\sqrt{\tau_2}\sqrt{1+\tau_2}} = \frac{-u + z_1\sqrt{1+\tau_2}}{2\tau_2\sqrt{\tau_2}\sqrt{1+\tau_2}} = 0$$

$$\Leftrightarrow \quad \tau_2 = \left(\frac{u}{z_1}\right)^2 - 1 \ .$$

Hierfür ist (3.6) maximal mit maximalem Wert

$$1 - \Phi\left(\frac{u\sqrt{1+\left(\frac{u}{z_1}\right)^2 - 1} - z_1}{\sqrt{\left(\frac{u}{z_1}\right)^2 - 1}}\right) = 1 - \Phi(\sqrt{u^2 - z_1^2})$$

für alle u und z_1 mit $0 < z_1 < u$.

(3.8) ist für $z_1 \leq 0$ minimal, falls $\tau_2 \to \infty$, da hierfür $\frac{\partial f}{\partial \tau_2} < 0$. Für $\tau_2 \to \infty$ geht (3.8) gegen u und (3.6) gegen den maximalen Wert $1 - \Phi(u)$.

Die maximale Fehlerwahrscheinlichkeit 1. Art ist also gemäß (3.7) durch

$$\int_{-\infty}^{0} (1-\Phi(u))\phi(z_1)\,dz_1 + \int_{0}^{u} \left(1-\Phi(\sqrt{u^2-z_1^2})\right)\phi(z_1)\,dz_1 + \int_{u}^{\infty} \phi(z_1)\,dz_1 =$$

$$\frac{3}{2}\,\alpha + \int_{0}^{u} \left(1-\Phi(\sqrt{u^2-z_1^2})\right)\phi(z_1)\,dz_1 \tag{3.9}$$

gegeben. Darüber hinaus kann gezeigt werden, daß

$$\int_{0}^{u} \left(1-\Phi(\sqrt{u^2-z_1^2})\right)\phi(z_1)\,dz_1 = \frac{1}{4}\,\exp(-\frac{u^2}{2}) - \frac{1}{2}\,\alpha$$

(vgl. das Appendix in Proschan und Hunsberger, 1995), womit (3.9) den Wert

$$\alpha + \frac{1}{4}\exp(-\frac{u^2}{2}) \tag{3.10}$$

ergibt.

Ist $\alpha = 0.01$ oder 0.05, so ist (3.10) gleich 0.0267 bzw. 0.1146. Eine datenabhängige Wahl von n_2 und die Verwendung von unadjustierten kritischen Werten u kann also zu einer mehr als Verdoppelung des vorgegebenen Niveaus α führen. Dies kann vermieden werden, wenn man vereinbart, die Studie nur dann fortzusetzen, falls der p-Wert der ersten Sequenz (mindestens) kleiner als 0.50 ist, und für die beiden Stufen eine geeignet größere Schranke $\tilde{c}_{\,\mathrm{PH}}(\alpha) > u$ zu verwenden.

Bei Vorgabe der kritischen Schranke $u_{\alpha_0} = \Phi^{-1}(1-\alpha_0)$, $\alpha_0 \leq 0.50$, für die Annahme von H_0 lautet der Entscheidungsalgorithmus des auf Proschan und Hunsberger (1995) zurückgehenden zweistufigen adaptiven Verfahrens:

- In der ersten Stufe wird H_0 angenommen und die Studie beendet, falls $z_1 \leq u_{\alpha_0}$.
- Ist $z_1 \geq \tilde{c}_{\mathrm{PH}}(\alpha)$, so wird die Studie mit der Ablehnung von H_0 beendet.
- Ist $u_{\alpha_0} < z_1 < \tilde{c}_{\mathrm{PH}}(\alpha)$, so wird die zweite Stufe der Studie durchgeführt, wobei beliebige Fallzahl-Neukalkulationen durchgeführt werden dürfen.
- In der zweiten Stufe wird H_0 abgelehnt, falls $z_2^* \geq \tilde{c}_{\mathrm{PH}}(\alpha)$.

(3.11)

α_0 wird vor Studienbeginn festgelegt. $\tilde{c}_{\mathrm{PH}}(\alpha)$ ist ein zu ermittelnder kritischer Wert, der die Einhaltung des Niveaus α bei Verwendung dieser Testprozedur sichert. Die obige Herleitung zeigt, daß dies bei der Bestimmung von $\tilde{c}_{\mathrm{PH}}(\alpha)$ durch die Bedingung

$$1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} \left(1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2})\right) \phi(z_1)\, dz_1 = \alpha \quad (3.12)$$

gewährleistet ist. Für ausgewählte Werte von α_0 und α sind $\tilde{c}_{\mathrm{PH}}(\alpha)$ Tabelle 3.4 zu entnehmen. Die Bestimmung von $\tilde{c}_{\mathrm{PH}}(\alpha)$ geschieht dabei iterativ und mit Hilfe univariater numerischer Integration. Ein SAS-Programm für die Berechnung der Werte ist in Anhang A.8 enthalten. Zum Vergleich sind in Tabelle 3.4 die kritischen Werte enthalten, die sich durch die von Wassmer (1999a) vorgeschlagene Vorgehensweise, d.h. die Betrachtung des *worst case scenarios*, ergeben (vgl. Abschnitt 2.3.2). Die Bestimmungsgleichung für den adjustierten Wert $\tilde{u}$ ist unter den dortigen Bedingungen durch

$$P_{\mathrm{H}_0}(Z_1 \geq \tilde{u}) + \sup_{0<t_1<1} P_{\mathrm{H}_0}(u_{\alpha_0} < Z_1 < \tilde{u}, Z_2^* \geq \tilde{u}) = \alpha$$

gegeben. $\tilde{u}$ kann – wie in Abschnitt 2.3.2 beschrieben – mit Hilfe von nichtlinearen Optimierungsmethoden bestimmt werden. Man beachte, daß diese Maximierungsaufgabe anders als die in diesem Abschnitt beschriebene über datenunabhängige $t_1 = n_1/(n_1 + n_2)$ geführt wird, während im Ansatz von Proschan und Hunsberger (1995) in einem von $\tau_2(z_1)$ erzeugten Funktionenraum maximiert wird. Dies erklärt die unterschiedlichen Ergebnisse. In Tabelle 3.4 sind darüber hinaus die einseitigen kritischen Werte nach DeMets und Ware (1980) für den Fall $n_1 = n_2$ und mit $u^L = \Phi^{-1}(1 - \alpha_0)$ angegeben (vgl. Abschnitt 2.2.3).

Klarerweise ist für alle α und α_0 der kritische Wert $\tilde{c}_{\mathrm{PH}}(\alpha)$ am größten, und der kritische Wert für $n_1 = n_2$ am kleinsten. Wie aus der Tabelle ersichtlich ist, ist der auf die Adjustierung wirkende Effekt einer beliebig datengesteuerten Wahl von n_2 größer als das Zulassen von beliebigen (datenunabhängig gewählten) Stichprobenumfängen. Für den einseitigen Fall mit Berücksichtigung eines frühen Abbruchs mit der Annahme von H_0 ist der Effekt der Zulassung von datenunabhängig gewählten n_2 sogar vernachlässigbar klein, während die adaptive Planung eine deutliche Erhöhung der zu fordernden kritischen Schranke mit sich bringt. Dieser Vergleich der kritischen Werte verdeutlicht auf sehr anschauliche

Tabelle 3.4: *Kritische Werte $\tilde{c}_{\mathrm{PH}}(\alpha)$ des zweistufigen Verfahrens nach Proschan und Hunsberger (1995) mit Abbruch für H_0, falls $p_1 \geq \alpha_0$. Der in Klammern stehende erste Wert bezeichnet den kritischen Wert des worst case scenario-Ansatzes (Wassmer, 1999a); der zweite in Klammern stehende Wert ist unter der Annahme $n_1 = n_2$ bestimmt.*

α_0	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
0.10	1.773 (1.76, 1.75)	2.134 (2.11, 2.11)	2.530 (2.50, 2.49)	2.792 (2.75, 2.75)
0.15	1.821 (1.80, 1.80)	2.167 (2.14, 2.14)	2.554 (2.51, 2.51)	2.811 (2.76, 2.76)
0.20	1.852 (1.83, 1.83)	2.189 (2.15, 2.15)	2.570 (2.52, 2.52)	2.825 (2.77, 2.77)
0.25	1.875 (1.84, 1.84)	2.207 (2.16, 2.16)	2.584 (2.53, 2.52)	2.836 (2.77, 2.77)
0.30	1.894 (1.86, 1.85)	2.222 (2.17, 2.17)	2.595 (2.53, 2.53)	2.846 (2.78, 2.77)
0.40	1.925 (1.88, 1.87)	2.246 (2.19, 2.17)	2.614 (2.54, 2.53)	2.862 (2.78, 2.77)
0.50	1.951 (1.89, 1.87)	2.267 (2.20, 2.18)	2.630 (2.55, 2.53)	2.876 (2.79, 2.77)

Weise den „Preis“, der für eine datenabhängige bzw. datenunabhängige Modifizierung des Stichprobenumfangs der Studie zu bezahlen ist.

Eine gleichmäßige Verbesserung des Verfahrens ergibt sich durch die Überlegung, daß der kritische Wert für die zweite Sequenz der Studie von z_1, n_1 und n_2 abhängen kann. Setzt man für den kritischen Wert $\tilde{u}_2$ der Prozedur nicht $\tilde{c}_{\mathrm{PH}}(\alpha)$, sondern

$$\tilde{u}_2 = \tilde{u}_2(n_1, n_2, z_1) = \frac{\sqrt{n_1}z_1 + \sqrt{n_2}\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2}}{\sqrt{n_1 + n_2}} \,,$$

so ergibt sich ein Niveau-α-Test, denn aus (3.12) folgt:

$$\begin{aligned}
\alpha \;=\;& 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} \left(1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2})\right) \phi(z_1)\, dz_1 = \\
& 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} P_{\mathrm{H}_0}(Z_2 > \sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2})\phi(z_1)\, dz_1 = \\
& 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) + \\
& \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} P_{\mathrm{H}_0}(\frac{\sqrt{n_1}z_1 + \sqrt{n_2}Z_2}{\sqrt{n_1 + n_2}} > \frac{\sqrt{n_1}z_1 + \sqrt{n_2}\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2}}{\sqrt{n_1 + n_2}})\phi(z_1)\, dz_1 =
\end{aligned}$$

$$1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} P_{\mathrm{H}_0}(Z_2^* > \tilde{u}_2 \mid z_1)\phi(z_1)\, dz_1 \,. \tag{3.13}$$

Die Prozedur, die in der zweiten Sequenz die Nullhypothese ablehnt, falls $Z_2^* > \tilde{u}_2$, erfüllt also die Niveaubedingung. Für die Funktion $\tilde{u}_2$ gilt

$$\tilde{u}_2(n_1, n_2, z_1) \leq \tilde{c}_{\mathrm{PH}}(\alpha) \;\text{ für alle } n_1, n_2 \text{ und } z_1 > 0, \tag{3.14}$$

was sich durch die Betrachtung von

$$\tilde{u}_2(n_1, n_2, z_1) = \frac{z_1 + \sqrt{\tau_2}\sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha) - z_1^2}}{\sqrt{1+\tau_2}}$$

($\tau_2 = n_2/n_1$) und die Berechnung von $\frac{\partial \tilde{u}_2}{\partial \tau_2}$ und $\frac{\partial \tilde{u}_2}{\partial z_1}$ ergibt:

$$\frac{\partial \tilde{u}_2}{\partial \tau_2} = \frac{\sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha) - z_1^2} - \sqrt{\tau_2} z_1}{2\sqrt{\tau_2}\sqrt{(1-\tau_2)^3}} = 0 \quad \Leftrightarrow \quad \tau_2 = \frac{\tilde{c}_{\mathrm{PH}}^2(\alpha)}{z_1^2} - 1 \,,$$

und $\tilde{u}_2$ ist maximal mit maximalen Wert

$$\frac{z_1 + (\tilde{c}_{\mathrm{PH}}^2(\alpha) - z_1^2)/z_1}{\sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha)/z_1^2}} = \tilde{c}_{\mathrm{PH}}(\alpha) \,.$$

Ebenso ist

$$\frac{\partial \tilde{u}_2}{\partial z_1} = \frac{1}{\sqrt{1+\tau_2}} \left(1 - \frac{\sqrt{\tau_2}\, z_1}{\sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha) - z_1^2}} \right) = 0 \quad \Leftrightarrow \quad z_1 = \frac{\tilde{c}_{\mathrm{PH}}(\alpha)}{\sqrt{1+\tau_2}}$$

und $\tilde{u}_2$ ist maximal mit maximalen Wert

$$\frac{\tilde{c}_{\mathrm{PH}}(\alpha)/\sqrt{1+\tau_2} + \sqrt{\tau_2}\sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha) - \tilde{c}_{\mathrm{PH}}^2(\alpha)/(1+\tau_2)}}{\sqrt{1+\tau_2}} = \tilde{c}_{\mathrm{PH}}(\alpha) \,.$$

Damit ist die Gültigkeit von (3.14) gezeigt.

Der kritische Wert des Testverfahrens kann auf der zweiten Stufe verkleinert, und damit die Prozedur gleichmäßig verbessert werden. Der Entscheidungsalgorithmus der auf diese Art modifizierten Prozedur lautet wie (3.11) mit der Modifikation, daß in der zweiten Stufe H_0 abgelehnt wird, falls

$$\begin{aligned} z_2^* &\geq \tilde{u}_2 \quad \text{bzw.} \\ z_2 &\geq \sqrt{\tilde{c}_{\mathrm{PH}}^2(\alpha) - z_1^2} \,. \end{aligned}$$

Es ergibt sich die gegenüber (3.11) gleichmäßig verbesserte Prozedur

- In der ersten Stufe wird H_0 angenommen und die Studie beendet, falls $Z_1 \leq u_{\alpha_0}$.
- Ist $z_1 \geq \tilde{c}_{\mathrm{PH}}(\alpha)$, so wird die Studie mit der Ablehnung von H_0 beendet.
- Ist $u_{\alpha_0} < z_1 < \tilde{c}_{\mathrm{PH}}(\alpha)$, so wird die zweite Stufe der Studie durchgeführt.
- In der zweiten Stufe wird H_0 abgelehnt, falls $z_2 \geq \sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2}$.

(3.15)

Bevor auf eine aus (3.13) sich ergebende viel allgemeinere Betrachtungsweise eingegangen wird, einige Bemerkungen zu den statistischen Eigenschaften des (verbesserten) Proschan und Hunsberger-Verfahrens (3.15):

Ein Vergleich mit dem auf Fishers Kombinationstest beruhenden Verfahren ist naheliegend. Wassmer (1998) zeigte, daß die beiden Verfahren sehr ähnliche bzw. praktisch ununterscheidbare Resultate bzgl. der resultierenden Entscheidungsbereiche und der $power$ liefern. Die $power$ wurde im „nicht-adaptiven Szenario“, d.h. unter der Annahme fest vorgegebener Stichprobenumfänge n_1 und n_2, ermittelt (vgl. Table 3 in Wassmer, 1998). Hier zeigten sich nur minimale Unterschiede, die sich durch die leicht unterschiedlichen ASN-Werte erklären lassen (vgl. ebd.).

Das Verfahren zur adaptiven Berechnung des für die zweite Stufe benötigten Stichprobenumfangs kann auf die in Abschnitt 3.1.1 beschriebene Art erfolgen. Dies geschieht durch Berechnung von

$$P(Z_2^* > \tilde{u}_2 \mid z_1) = P(Z_2 > \sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2}) =$$
$$1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} - \frac{\mu_1 - \mu_2}{\sigma} \cdot \sqrt{\frac{n_2}{2}}) \, .$$

Schätzt man $\mu_1 - \mu_2$ aus der ersten Stichprobe durch $\bar{x}_1 - \bar{x}_2 = z_1 \cdot \sqrt{2\sigma^2/n_1}$ und fordert man $P(Z_2^* > \tilde{u}_2 \mid z_1) = 1 - \beta_2$, so ergibt sich sofort

$$n_2 = n_1 \cdot \frac{\left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} + \Phi^{-1}(1-\beta_2)\right)^2}{z_1^2} \, . \qquad (3.16)$$

Die Formel (3.16) führt zu praktisch denselben Resultaten wie (3.2) und kann analog benutzt werden, um eine Neuberechnung von n_2 durchzuführen und über den weiteren Verlauf der Studie zu entscheiden.

Ein Nachteil des so beschriebenen Verfahrens ist die Tatsache, daß die Varianz σ^2 als bekannt vorausgesetzt werden muß. Proschan und Hunsberger nahmen an, '*that n_1 is fairly large, so that σ^2 is essentially known after n_1 observations per group.*' (Proschan und Hunsberger, 1995, S. 1316). Wassmer (1998) zeigte durch Berechnung der exakten Fehlerwahrscheinlichkeit 1. Art, daß diese Annahme gerechtfertigt ist, sobald $n_1 > 10$ und $n_2 > 10$ (bzw. $n_2 = 0$). Nichtsdestoweniger beruht dieses Verfahren auf der Annahme spezifischer Verteilungsannahmen, woraus sich prinzipiell approximative Entscheidungsbereiche für die praktische Anwendung ergeben.

Das in diesem Abschnitt besprochene Verfahren kann weiter modifiziert und in einen sehr allgemeinen Kontext gestellt werden (Proschan und Hunsberger, 1995). Die bisher dargestellten Überlegungen zeigen nämlich, daß eine beliebige, monoton steigende Funktion $\alpha(z_1)$, deren Wertebereich das Intervall $[0;\ 1]$ umfaßt, und für die

$$\int_{-\infty}^{\infty} \alpha(z_1)\phi(z_1)\,dz_1 \;=\; \alpha \tag{3.17}$$

erfüllt ist, zu einem adaptiven Test zum Niveau α führt, falls $\alpha(z_1)$ bei Studienbeginn fest vorgegeben wird. Die Funktion $\alpha(z_1)$ gibt die bedingte Fehlerrate nach Beendigung der Studie, gegeben $Z_1 = z_1$, an und wird als *conditional error function* bezeichnet[4]. Bedingung (3.17) garantiert, daß ungeachtet der datengesteuerten Bestimmung des Stichprobenumfangs $n_2 = n_2(z_1)$ der Test das Niveau α einhält.

Setzt man

$$\alpha(z_1) = \begin{cases} 0 & , \text{ falls } z_1 < \Phi^{-1}(1-\alpha) \\ 1 & , \text{ falls } z_1 \geq \Phi^{-1}(1-\alpha)\ , \end{cases}$$

so ergibt sich ein Test zum Niveau α mit fest vorgegebenem Stichprobenumfang n_1.

4. Diese Funktion ist nicht mit der in Abschnitt 2.3.3 behandelten α-$spending$-Funktion zu verwechseln!

Die *conditional error function*

$$\alpha(z_1) = \begin{cases} 0 & \text{, falls } z_1 \leq u_{\alpha_0} \\ 1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2}) & \text{, falls } u_{\alpha_0} < z_1 < \tilde{c}_{\mathrm{PH}}(\alpha) \\ 1 & \text{, falls } z_1 \geq \tilde{c}_{\mathrm{PH}}(\alpha) \end{cases}$$

ergibt das in (3.15) beschriebene Verfahren. Proschan und Hunsberger (1995) schlugen auch die Klasse von Funktionen

$$\alpha(z_1) = \begin{cases} 0 & \text{, falls } z_1 \leq u_{\alpha_0} \\ 1 - \Phi(a + b \cdot z_1) & \text{, falls } u_{\alpha_0} < z_1 < \tilde{c}_{\mathrm{PH}}(\alpha) \\ 1 & \text{, falls } z_1 \geq \tilde{c}_{\mathrm{PH}}(\alpha) \end{cases} \tag{3.18}$$

vor, die in Abschnitt 3.1.4 dieser Arbeit einer genaueren Betrachtung unterzogen und in einen noch allgemeineren Rahmen gestellt wird. Bevor auf diese Möglichkeiten eingegangen wird, werden im folgenden Abschnitt die Vorteile bei der Verwendung adaptiver Verfahren diskutiert.

3.1.3 *Power*-Gewinn bei adaptiver Planung

In den beiden letzten Abschnitten wurden zwei Verfahren für die Durchführung von zweistufigen Testverfahren besprochen, die die adaptive, d.h. die von den Ergebnissen der ersten Stufe abhängige Planung des Stichprobenumfangs der zweiten Sequenz ermöglicht. Diese Prozeduren erfüllen die Niveaubedingung unter den gemachten Annahmen exakt, wobei zu beachten ist, daß bei dem von Proschan und Hunsberger (1995) vorgeschlagenen Verfahren die Varianz als bekannt vorausgesetzt wird. Es ist intuitiv naheliegend, daß diese Verfahren vorteilhaft für die Planung einer Studie sind, da man die volle Information der Daten in der Zwischenauswertung benutzen und verwerten darf. Die Frage ist jedoch: Welcher Art sind die Vorteile? bzw.: Wie läßt sich die Überlegenheit der adaptiven Verfahren im Vergleich zu den herkömmlichen Verfahren messen? Prinzipiell sollte es einen Vorteil geben in der erreichten $power$ und im ASN im Vergleich zu einem herkömmlichen Test mit festem Stichprobenumfang, aber auch im Vergleich zu einem gruppensequentiellen Testverfahren, das die Möglichkeit der adaptiven Planung nicht in Betracht zieht. Die generelle Beantwortung der Frage ist zweifellos schwierig, da sie von sehr vielen Faktoren abhängt. Insbesondere sind *beliebige* Fallzahladaptionen möglich, was die

Bewertung des Vorgehens schwierig macht. Zudem kann die Berechnung der $power$ bereits recht schwierig sein, was in Abschnitt 3.1.1 für das Verfahren nach Bauer und Köhne (1994) illustriert wurde. Man ist bei der Bewertung des Vorteils der adaptiven Vorgehensweise in der Regel auf Monte-Carlo-Studien angewiesen, die bestimmte Algorithmen der adaptiven Planung simulieren, und diese so bewerten. Bauer et al. (1998) und Bauer und Röhmel (1995) stellen Simulations-Ergebnisse vor, die sich insbesondere durch die adaptive Auswahl der zu testenden Nullhypothesen ergeben. Dies steht nicht im Zentrum der vorliegenden Arbeit, gehört aber zu den Forschungsaktivitäten des Autors.

Die $power$ und der ASN des Verfahrens nach Proschan und Hunsberger (1995) läßt sich unter den in ihrer Arbeit gemachten Voraussetzungen numerisch berechnen. Dies kann z.B. unter der Annahme geschehen, daß die Berechnung des Stichprobenumfangs der zweiten Sequenz durch die Formel (3.16) geschieht. Es sei weiter angenommen, daß der Stichprobenumfang n_1 der ersten Sequenz so bestimmt wurde, daß ein Test zum Niveau α mit festem Stichprobenumfang n_1 die $power$ $1-\beta$ zum Testen der Hypothese H_0 gegen die Alternative $\frac{\mu_1-\mu_2}{\sigma} = \delta^*$ besitzt, wobei δ^* die angenommene standardisierte Differenz bezeichnet. Dies führt unter der Voraussetzung, daß die Varianz bekannt ist, zu

$$n_1 \;=\; 2 \cdot \Big(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta^*} \Big)^2 \quad \text{bzw.} \tag{3.19}$$

$$\delta^* \cdot \sqrt{\frac{n_1}{2}} \;=\; \Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) \;=\; \vartheta_1^* \;=\; E(Z_1)\;.$$

Ist in Wirklichkeit der Effekt nicht δ^*, sondern $\lambda \cdot \delta^*$, so ist

$$E(Z_1) = \lambda \cdot \vartheta_1^* = \lambda \cdot \big(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) \big)\;, \tag{3.20}$$

und die $power$ der ersten Stufe der in (3.15) beschriebenen Prozedur ist durch

$$1 - \Phi\Big(\tilde{c}_{\mathrm{PH}} - \lambda \cdot \big(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) \big) \Big)$$

gegeben. Unter der Alternative $\frac{\mu_1-\mu_2}{\sigma} = \lambda \cdot \delta^*$ ist

$$E(Z_2) \quad = \quad \lambda \cdot \vartheta_2^* \;=\; \lambda \cdot \delta^* \sqrt{\frac{n_2}{2}}\;, \text{ wobei} \tag{3.21}$$

$$n_2 = n_1 \cdot \frac{\left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} + \Phi^{-1}(1-\beta_2)\right)^2}{z_1^2} = \tag{3.22}$$

$$2 \cdot \left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta^*}\right)^2 \cdot \frac{\left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} + \Phi^{-1}(1-\beta_2)\right)^2}{z_1^2},$$

falls die erste Sequenz mit der $power\ 1-\beta$ und die zweite Sequenz mit der bedingten $power\ 1-\beta_2$ geplant wird (vgl. (3.16)). Der Nichtzentralitätsparameter (3.21) ist damit gegeben durch

$$\lambda \cdot \frac{\left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right) \cdot \left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} + \Phi^{-1}(1-\beta_2)\right)}{z_1}, \tag{3.23}$$

und die $power$ der Prozedur (3.15) ist durch

$$\Phi(\lambda \cdot \vartheta_1^* - \tilde{c}_{\mathrm{PH}}) + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} \left(1 - \Phi\left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} - \lambda \cdot \vartheta_2^*\right)\right) \cdot \phi(z_1 - \lambda \cdot \vartheta_1^*)\, dz_1 \tag{3.24}$$

berechenbar, wobei $\lambda \cdot \vartheta_1^*$ durch den Ausdruck (3.20) und $\lambda \cdot \vartheta_2^*$ durch den Ausdruck (3.23) gegeben ist (vgl. Proschan und Hunsberger, 1995, S. 1320). Völlig analog läßt sich auch der ASN der sich ergebenden Prozedur berechnen. Er ist gemäß (3.19) und (3.22) durch

$$n_1 + E\Big(n_2(Z_1)\Big) = 2 \cdot \left(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta^*}\right)^2 \times$$

$$\left(1 + \int_{u_{\alpha_0}}^{\tilde{c}_{\mathrm{PH}}(\alpha)} \frac{\left(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - z_1^2} + \Phi^{-1}(1-\beta_2)\right)^2}{z_1^2} \cdot \phi(z_1 - \lambda \cdot \vartheta_1^*)\, dz_1\right) \tag{3.25}$$

numerisch berechenbar, wobei $\lambda \cdot \vartheta_1^*$ wie oben durch den Ausdruck (3.20) bestimmt ist. Insbesondere gilt, daß die $power$ nicht von δ^* abhängig ist, was die Beschreibung der sich ergebenden Eigenschaften einfacher macht.

In Tabelle 3.5 sind für $\alpha = 0.05$, $\alpha_0 = 0.15, 0.30$ und 0.40 und $1 - \beta_2 = 0.50, 0.80$ einige Ergebnisse dieses Vorgehens angegeben. Dabei ist in der Spalte '$power_{\mathrm{Fix}}$' die $power$ des Tests mit festem Stichprobenumfang n_1 berechnet, falls in Wahrheit der Effekt $\lambda \cdot \delta^*$ vorliegt. Dies ergibt für $\lambda = 1$ den Wert $1 - \beta$, d.h. die vorgegebene $power$ der ersten Stufe. $power_{\mathrm{PH}}$ bezeichnet die durch (3.24) berechnete $power$ des zweistufigen Verfahrens, die natürlich größer als die $power$ des einstufigen Verfahrens ist. $n_{power_{\mathrm{PH}}}$ bezeichnet den Stichprobenumfang, der für ein einstufiges Verfahren nötig ist, um die $power$ $power_{\mathrm{PH}}$ des zweistufigen adaptiven Verfahrens zu besitzen. ASN_{PH} ist der durchschnittliche Stichprobenumfang nach (3.25). $n_{power_{\mathrm{PH}}}$ und ASN_{PH} sind unter der Annahme berechnet, daß in Wirklichkeit der Effekt $\lambda \cdot \delta^*$ vorliegt, wobei o.B.d.A. der Fall $\delta^* = 1$ betrachtet wird (vgl. auch Table 2 in Proschan und Hunsberger, 1995).

Die Tabelle zeigt, daß das zweistufige Verfahren in der Regel mit einem niedrigerem ASN die $power$ des einstufigen Verfahrens erreicht. Dies ist in den hier angegebenen Fällen für $\alpha_0 = 0.15$ immer sowie für $\alpha_0 = 0.30, 0.40$ und einer Unterschätzung des wahren Effekts in der Planungsphase ($\lambda = 1.5$) sowie für $\alpha_0 = 0.30$ und $\lambda = 1$ der Fall. Der durchschnittlich benötigte Stichprobenumfang des zweistufigen Verfahrens kann für $\alpha_0 = 0.30$ und $\alpha_0 = 0.40$ auch größer als der (feste) Stichprobenumfang eines entsprechend einstufigen Verfahrens ausfallen. Dies liegt aber darin begründet, daß für solche Fälle ein nach (3.16) berechneter Stichprobenumfang und damit (3.25) sehr groß sein kann (vgl. dazu die Bemerkung auf S. 83). Für diese Fälle sind andere Regeln für die Ermittlung des Stichprobenumfangs n_2 erforderlich. Proschan und Hunsberger (1995) schlugen in ihrer Arbeit die Vorgabe von $\alpha_0 = 0.15$ vor, was im Kontext ihrer Arbeit begründet liegt. Dieser bezieht sich weniger auf die Bereitstellung eines gruppensequentiellen Testverfahrens, sondern vielmehr auf die Möglichkeit der kontrollierten Fortführung einer Studie, die „fast“ die Signifikanz erreicht hat („*Designed Extension of Studies Based on Conditional Power*“).

Der Vergleich mit einem klassischen (zweistufigen) gruppensequentiellen Plan kann ebenso geführt werden. Tabelle 3.6 enthält die Ergebnisse der folgenden Vorgehensweise: Zuerst werden die Stichprobenumfänge $n_1 = n_2$ ermittelt, die zur $power$ $1 - \beta$ bei dem gruppensequentiellen Verfahren nach DeMets und Ware (1980) führen. Die $power$ $power_{\mathrm{DW}}$ dieses Verfahrens ist für verschiedene Werte λ berechnet, d.h. wie oben, falls in Wirklichkeit der Effekt $\lambda \cdot \delta^*$

Tabelle 3.5: *Vergleich von $power$ und ASN im adaptiven Design nach Proschan und Hunsberger (1995) mit Abbruch für H_0, falls $p_1 \geq \alpha_0$, $\alpha = 0.05$. (vgl. Text für die Erklärung von $power_{\mathrm{Fix}}$, $power_{\mathrm{PH}}$, $n_{power_{\mathrm{PH}}}$ und ASN_{PH}).*

		$1-\beta_2 = 0.50$			$1-\beta_2 = 0.80$		
λ	$power_{\mathrm{Fix}}$	$power_{\mathrm{PH}}$	ASN_{PH}	$n_{power_{\mathrm{PH}}}$	$power_{\mathrm{PH}}$	ASN_{PH}	$n_{power_{\mathrm{PH}}}$
$\alpha_0 = 0.15, 1-\beta = 0.80$							
0.25	0.153	0.179	14.7	16.8	0.210	19.0	22.5
0.50	0.344	0.424	15.2	16.9	0.507	20.6	22.1
0.75	0.587	0.697	14.8	16.6	0.771	19.4	20.3
1.00	0.800	0.882	13.8	16.0	0.921	16.5	18.7
1.50	0.981	0.993	12.5	14.9	0.996	12.8	16.7
$\alpha_0 = 0.15, 1-\beta = 0.90$							
0.25	0.181	0.214	20.6	23.3	0.256	26.9	31.4
0.50	0.428	0.525	21.0	23.3	0.612	28.3	29.8
0.75	0.709	0.810	19.7	22.6	0.865	24.8	26.8
1.00	0.900	0.950	18.2	21.6	0.969	20.3	24.6
1.50	0.997	0.999	17.2	20.1	1.000	17.2	22.2
$\alpha_0 = 0.30, 1-\beta = 0.80$							
0.25	0.153	0.243	29.1	28.8	0.327	50.4	45.8
0.50	0.344	0.559	28.0	25.7	0.674	48.5	35.1
0.75	0.587	0.796	22.6	21.7	0.880	36.5	28.3
1.00	0.800	0.925	17.1	19.0	0.968	23.8	24.4
1.50	0.981	0.995	12.7	15.8	0.999	13.3	20.4
$\alpha_0 = 0.30, 1-\beta = 0.90$							
0.25	0.181	0.298	40.7	39.8	0.396	70.7	61.1
0.50	0.428	0.656	36.5	33.5	0.764	62.3	44.7
0.75	0.709	0.876	26.9	27.9	0.938	40.6	36.0
1.00	0.900	0.968	20.2	24.4	0.990	24.7	31.3
1.50	0.997	0.999	17.2	20.6	1.000	17.3	26.5
$\alpha_0 = 0.40, 1-\beta = 0.80$							
0.25	0.153	0.327	64.5	45.7	0.424	123.9	67.5
0.50	0.344	0.628	53.0	31.1	0.745	100.3	42.4
0.75	0.587	0.828	34.6	23.9	0.915	61.5	32.3
1.00	0.800	0.934	21.1	19.9	0.979	32.1	27.1
1.50	0.981	0.995	12.9	15.9	1.000	13.6	22.0
$\alpha_0 = 0.40, 1-\beta = 0.90$							
0.25	0.181	0.386	88.0	58.8	0.492	169.0	84.4
0.50	0.428	0.711	64.4	38.8	0.821	120.2	52.6
0.75	0.709	0.894	36.7	29.7	0.958	60.8	40.4
1.00	0.900	0.970	22.2	24.9	0.994	29.0	34.3
1.50	0.997	0.999	17.2	20.5	1.000	17.4	28.3

vorliegt. $power_{\mathrm{PH}}$ ist die $power$ des zweistufigen adaptiven Verfahrens mit bedingter $power$ $1-\beta_2$ und Stichprobenumfang n_1 der ersten Sequenz. ASN_{PH} ist der durchschnittliche Stichprobenumfang dieses adaptiven Designs. Analog zu Tabelle 3.5 sind $power_{\mathrm{PH}}$ und ASN_{PH} unter der Annahme bestimmt, daß der Effekt $\lambda \cdot \delta^*$ vorliegt. Zum Vergleich wurde ein gruppensequentieller Plan mit ungleichen Sequenzgrößen und Stichprobenumfang der ersten Sequenz gleich n_1 berechnet (d.h. geeignet gewähltem $\boldsymbol{\tau} = (1, n_2/n_1)$, vgl. Abschnitt 2.3.1), der die $power$ $power_{\mathrm{PH}}$ besitzt. ASN_{DW} bezeichnet den durchschnittlichen Stichprobenumfang dieses gruppensequentiellen Plans, falls in Wirklichkeit der Effekt $\lambda \cdot \delta^*$ vorliegt. ASN_{DW} und ASN_{PH} sind wie oben o.B.d.A. für den Fall $\delta^* = 1$ berechnet und die Ergebnisse für $\alpha = 0.05$, $\alpha_0 = 0.15, 0.30$ und 0.40 und $1 - \beta_2 = 0.50, 0.80$ angegeben.

Tabelle 3.6 zeigt den Vorteil des adaptiven Verfahrens, der hauptsächlich in der Erhöhung der $power$ liegt. Der ASN-Verlust gegenüber dem klassischen Verfahren wird durch einen i.d.R. deutlichen $power$-Gewinn ausgeglichen. Ein $power$-Gewinn für $\alpha_0 = 0.15$ korrespondiert mit einem nur marginal höheren oder sogar gleich hohen ASN. Die $power$ des adaptiven Verfahrens ist größer als die $power$ des klassischen Verfahrens, falls die bedingte $power$ für die Durchführung der zweiten Sequenz groß genug (hier $1 - \beta_2 = 0.80$) bzw. α_0 nicht zu klein gewählt wird. Man beachte, daß der so geführte Vergleich „künstlich“ ist, da der Effekt in der konkreten Anwendung unbekannt ist, und daher die Bestimmung der Parameter des klassischen Verfahrens nur zu Vergleichszwecken dient.

Wie oben erwähnt, „kollabieren“ die Berechnungen relativ schnell für größer werdendes α_0, was in der Bestimmung von n_2 aus der bedingten $power$-Betrachtung begründet ist. Die so bestimmten Werte von n_2 werden für größere α_0 schnell sehr groß und damit ein Vergleich nicht fair, da ein stark „überhöhter“ ASN zugrundeliegen würde (man vergleiche hierzu auch Posch und Bauer, 2000; Friede und Kieser, 2001). Eine realistischere Vorgehensweise würde darin bestehen, eine Obergrenze für den Stichprobenumfang der zweiten Stufe anzugeben. Für die Bewertung dieser Vorgehensweise ist es jedoch kaum möglich, analytische oder numerische Resultate anzugeben. Man ist hier auf Simulationen angewiesen, deren Behandlung den Rahmen dieser Arbeit sprengen würde. Prinzipiell ist zu erwarten, daß die adaptive Planung einer Studie zu einer effizi-

Tabelle 3.6: *Vergleich von $power$ und ASN im adaptiven Design nach Proschan und Hunsberger (1995) mit Abbruch für H_0, falls $p_1 \geq \alpha_0$, $\alpha = 0.05$. (vgl. Text für die Erklärung von $power_{\mathrm{DW}}$, $power_{\mathrm{PH}}$, ASN_{PH} und ASN_{DW}).*

		$1-\beta_2 = 0.50$			$1-\beta_2 = 0.80$		
λ	$power_{\mathrm{DW}}$	$power_{\mathrm{PH}}$	ASN_{PH}	ASN_{DW}	$power_{\mathrm{PH}}$	ASN_{PH}	ASN_{DW}
$\alpha_0 = 0.15, 1-\beta = 0.80$							
0.25	0.151	0.147	9.9	9.6	0.169	12.5	12.5
0.50	0.344	0.327	10.3	10.0	0.395	13.7	13.6
0.75	0.590	0.557	10.2	10.0	0.643	13.8	13.2
1.00	0.800	0.762	9.8	9.5	0.827	12.6	11.7
1.50	0.977	0.963	8.8	8.6	0.978	9.6	9.2
$\alpha_0 = 0.15, 1-\beta = 0.90$							
0.25	0.180	0.174	14.0	13.6	0.204	18.0	17.9
0.50	0.433	0.410	14.5	14.1	0.491	19.5	19.1
0.75	0.716	0.678	14.1	13.7	0.755	18.6	17.5
1.00	0.900	0.869	13.2	12.8	0.911	16.0	14.8
1.50	0.995	0.991	11.9	11.8	0.995	12.3	12.1
$\alpha_0 = 0.30, 1-\beta = 0.80$							
0.25	0.148	0.178	16.8	14.7	0.236	28.8	26.0
0.50	0.336	0.414	17.2	14.8	0.528	30.0	23.4
0.75	0.583	0.642	15.7	12.6	0.751	26.8	18.3
1.00	0.800	0.808	13.0	10.3	0.889	20.9	13.9
1.50	0.981	0.963	8.7	7.9	0.988	10.9	8.8
$\alpha_0 = 0.30, 1-\beta = 0.90$							
0.25	0.175	0.215	23.5	20.6	0.288	40.5	36.1
0.50	0.420	0.500	23.3	19.5	0.616	40.5	29.8
0.75	0.706	0.738	19.8	15.7	0.834	32.9	21.9
1.00	0.900	0.885	15.4	12.6	0.944	23.0	15.9
1.50	0.997	0.988	10.7	10.3	0.997	11.9	10.8
$\alpha_0 = 0.40, 1-\beta = 0.80$							
0.25	0.148	0.242	37.1	28.1	0.323	71.1	47.7
0.50	0.335	0.493	34.5	20.3	0.608	66.1	30.6
0.75	0.581	0.692	27.5	14.6	0.804	51.5	21.2
1.00	0.800	0.832	19.6	11.0	0.917	34.6	15.0
1.50	0.983	0.964	9.7	7.8	0.992	13.1	8.8
$\alpha_0 = 0.40, 1-\beta = 0.90$							
0.25	0.174	0.287	50.9	37.0	0.378	97.8	60.2
0.50	0.417	0.568	44.6	24.8	0.686	84.9	36.9
0.75	0.703	0.772	32.0	17.2	0.872	58.6	24.3
1.00	0.900	0.896	20.6	12.9	0.959	34.1	16.6
1.50	0.997	0.987	10.9	10.0	0.998	12.6	10.6

enteren Vorgehensweise führt als die herkömmlichen gruppensequentiellen Testverfahren. Neben per se flexibleren Planungsmöglichkeiten wird sich dies aller Erwartung nach auch in verbesserten Gütekriterien bemerkbar machen. Ebenso ist zu erwarten, daß es adaptive Planungsstrategien gibt, die eine erwünschte $power$-Forderung trotz Unter- oder Überschätzung des Effekts in angemessenen Rahmen erfüllt.

3.1.4 Ein allgemeines Prinzip für die Konstruktion adaptiver Pläne

Eine Verallgemeinerung der Prozedur nach Proschan und Hunsberger (1995) ergibt sich durch die Betrachtung des Ablehnbereichs von (3.15) in der p_1-p_2-Ebene. Er ist durch die p-Werte p_1 und p_2 gegeben, für die

$$p_1 \leq 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) \quad \text{oder}$$
$$\left(p_2 \leq 1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - (\Phi^{-1}(1-p_1))^2}) \;\text{ und }\; p_1 \leq \alpha_0\right)$$

erfüllt ist. Dieser Ablehnbereich ist für $\alpha = 0.05$ und $\alpha_0 = 0.50$ in Abbildung 3.3 graphisch veranschaulicht und mit dem Ablehnbereich des Verfahrens nach Bauer und Köhne (1994) verglichen.

Abbildung 3.3 illustriert die minimalen Unterschiede in den Ablehnregionen der beiden Testverfahren. Die Entscheidungsbereiche sind für beide Verfahren so bestimmt, daß die Fläche des Ablehnbereichs gleich α ist. Viel wichtiger jedoch: Diese Art der Betrachtung führt zu einer *exakten* Lösung des Proschan und Hunsberger (1995)-Ansatzes auch für den Fall einer unbekannten Varianz σ^2, falls die Entscheidungsregel auf den p-Werten beruht. Der sich ergebende Entscheidungsalgorithmus lautet:

- In der ersten Stufe wird H_0 angenommen und die Studie beendet, falls $p_1 \geq \alpha_0$.
- Ist $p_1 \leq 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha))$, so wird die Studie mit der Ablehnung von H_0 beendet.
- Ist $1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) < p_1 < \alpha_0$, so wird die zweite Stufe der Studie durchgeführt.
- In der zweiten Stufe wird H_0 abgelehnt, falls $p_2 \leq 1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - (\Phi^{-1}(1-p_1))^2})$.

(3.26)

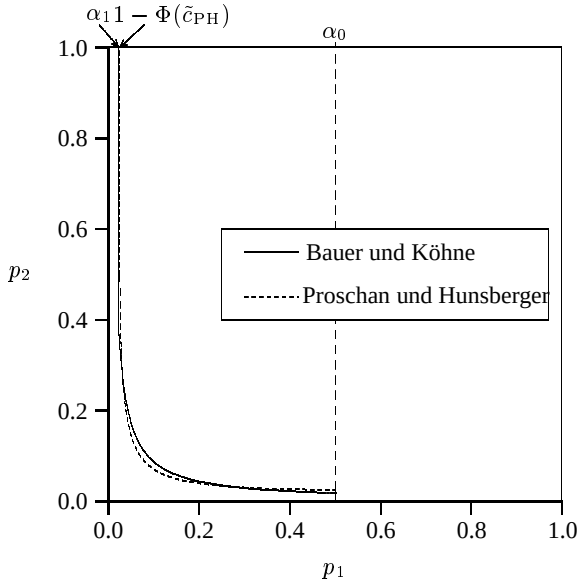

Abbildung 3.3: *Vergleich der Ablehnregionen der adaptiven zweistufigen Verfahren nach Bauer und Köhne (1994) bzw. Proschan und Hunsberger (1995);* $\alpha = 0.05$, $\alpha_0 = 0.50$, $\alpha_1 = 0.0233$, $\tilde{c}_{\mathrm{PH}} = 1.951$.

Das Verfahren (3.26) hält bei adaptiver Planung das Niveau α, da sich die Testentscheidung durch eine spezifische Verknüpfung von p-Werten p_1 und p_2 ergibt, die beide auf dem Intervall $[0;\ 1]$ gleichverteilt sind. Der in Abschnitt 3.1.1 geführte Beweis, daß das adaptive Verfahren das Niveau α einhält, läßt sich nämlich problemlos auf den Fall übertragen, daß nicht der Kombinationstest nach Fisher, sondern ein beliebig anderer Kombinationstest verwendet wird (Bauer, 1989a). Wie man leicht sieht, werden die p-Werte beim obigen Verfahren durch

$$\sqrt{(\Phi^{-1}(1-p_1))^2 + (\Phi^{-1}(1-p_2))^2} \tag{3.27}$$

kombiniert, woraus sich mit der zusätzlichen Angabe der Struktur des Fortsetzungsbereichs das angegebene Verfahren (3.26) ergibt.

Die am Ende des letzten Abschnitts beschriebene allgemeine Vorgehensweise, die sich durch die Vorgabe der *conditional error function* $\alpha(z_1)$ ergibt, kann modifiziert werden, indem (3.17) in Abhängigkeit von p_1 formuliert wird. Es gilt: Jedes zweistufige Testverfahren, das bei Vorgabe einer monoton fallenden Funktion $\alpha(p_1)$ mit Wertebereich $[0;\ 1]$ die Bedingung

$$\int_0^1 \alpha(p_1)\, dp_1 \ = \ \alpha \tag{3.28}$$

erfüllt, führt zu einem adaptiven Test zum Niveau α, falls $\alpha(p_1)$ bei Studienbeginn fest vorgegeben wird. Die Funktion $\alpha(p_1)$ korrespondiert mit einer Kombination $\varrho(p_1, p_2)$ der p-Werte p_1 und p_2, die beispielsweise durch Fishers Kombinationstest, durch (3.27) oder eine beliebig andere Funktion definiert ist. $\alpha(p_1)$ gibt die bedingte Fehlerrate nach Beendigung der Studie (unter Verwendung des durch $\varrho(p_1, p_2)$ definierten Testverfahrens), gegeben der beobachtete p-Wert p_1, an. Die Behauptung, daß der Test das Niveau α einhält, folgt wie oben aus der Betrachtung der bedingten Tests. $\alpha(p_1)$ ist in der Regel abhängig von weiteren Parametern, die die Struktur des Fortsetzungsbereichs spezifizieren.

Aus der Theorie der bedingten Tests ergibt sich so eine sehr allgemeine Vorgehensweise, die ein breites Spektrum von Anwendungen ermöglicht (vgl. auch Bauer, Brannath und Posch, 2001; Posch und Bauer, 1999; Wassmer, 2000). Insbesondere gilt:

1. Die von Proschan und Hunsberger (1995) vorgeschlagene Prozedur (3.15) kann für den Fall einer unbekannten Varianz σ^2 verallgemeinert werden, indem von der von Bauer (1989a) vorgeschlagenen Möglichkeit der Verknüpfung von p-Werten mittels Kombinationstestverfahren Gebrauch gemacht wird.
2. Die von Proschan und Hunsberger (1995) vorgeschlagene Spezifikation einer *conditional error function* $\alpha(z_1)$ kann verallgemeinert werden, so daß die von Bauer und Köhne (1994), Bauer und Röhmel (1995) und Bauer und Kieser (1999) vorgeschlagenen Verfahren als Spezialfälle resultieren.

Ein sehr interessanter Typ von Anwendungen ergibt sich durch die Vorgabe von

$$\alpha(p_1) = \begin{cases} 0 & , \text{ falls } p_1 \geq \alpha_0 \\ 1 - \Phi(\sqrt{2} \cdot \tilde{c}(\alpha) - \Phi^{-1}(1-p_1)) & , \text{ falls } 1 - \Phi(\tilde{c}(\alpha)) < p_1 < \alpha_0 \\ 1 & , \text{ falls } p_1 \leq 1 - \Phi(\tilde{c}(\alpha)) \ , \end{cases} \tag{3.29}$$

was mit der Kombinationsregel

$$\varrho(p_1, p_2) = \Phi^{-1}(1-p_1) + \Phi^{-1}(1-p_2)$$

korrespondiert, die gemeinhin als *inverse normal method* bekannt ist (Hedges und Olkin, 1985). Dies hat seinen Ursprung darin, daß $\Phi^{-1}(1-p_k)$, $k = 1, 2$, unter der Nullhypothese standardnormalverteilt ist, falls p_k auf $[0;\ 1]$ stetig gleichverteilt ist.

Wie man leicht sieht, führt Bedingung (3.28) bei Vorgabe von (3.29) zu

$$1 - \Phi(\tilde{c}(\alpha)) + \int_{1-\Phi(\tilde{c}(\alpha))}^{\alpha_0} (1 - \Phi(\sqrt{2} \cdot \tilde{c}(\alpha) - \Phi^{-1}(1-p_1))\, dp_1 =$$

$$P_{\mathrm{H}_0}(Z_1 \geq \tilde{c}(\alpha)) + \int_{u_{\alpha_0}}^{\tilde{c}(\alpha)} P_{\mathrm{H}_0}(\frac{Z_1 + Z_2}{\sqrt{2}} \geq \tilde{c}(\alpha) \mid z_1)\, \phi(z_1) dz_1 =$$

$$P_{\mathrm{H}_0}(Z_1 \geq \tilde{c}(\alpha)) + P_{\mathrm{H}_0}(u_{\alpha_0} < Z_1 < \tilde{c}(\alpha), Z_2^* \geq \tilde{c}(\alpha)) = \alpha \ . \tag{3.30}$$

(3.30) ist die Bestimmungsgleichung für die kritischen Werte eines gruppensequentiellen Plans nach DeMets und Ware (1980) (für $K = 2$), der in Abschnitt 2.2.3 besprochen wurde (vgl. auch Tabelle 3.4). Für dieses Verfahren sind also die im angegebenen Abschnitt hergeleiteten kritischen Werte $u = \tilde{c}(\alpha)$ zu verwenden, wobei beispielsweise auch kritische Werte innerhalb der Δ-Klasse zu einer gültigen Wahl der kritischen Schranken in einem adaptiven Design führen (vgl. ebd.). Die Vorgabe von (3.29) führt zu einem Verfahren, das rechentechnisch sehr vorteilhaft ist, da die kritischen Werte eines gruppensequentiellen Plans übernommen werden können. Insbesondere ist dieses Verfahren in enger Verwandschaft zu der von Proschan und Hunsberger (1995) vorgeschlagenen Funktion (3.18) mit $a = \sqrt{2} \cdot \tilde{c}(\alpha)$ und $b = -1$ zu sehen. Diese Vorgehensweise (d.h. die Verwendung der *inverse normal method* zur Verknüpfung der p-Werte im gruppensequentiellen Kontext) wurde auch schon von Bauer (1989a) und

Bauer und Köhne (1994) vorgeschlagen. Lehmacher und Wassmer (1999) stellten sie in einen allgemeineren Zusammenhang und beschrieben insbesondere deren Anwendung auf mehr als $K = 2$ Stufen. Das resultierende Verfahren sei aus diesem Grund das Verfahren nach Lehmacher und Wassmer (1999) genannt.

Auch die Ablehnregion dieses Verfahren ist in der p_1-p_2-Ebene darstellbar. Die Nullhypothese wird abgelehnt, falls

$$p_1 \leq 1 - \Phi(u) \text{ oder } \left(p_2 \leq 1 - \Phi(\sqrt{2} \cdot u - \Phi^{-1}(1 - p_1)) \text{ und } p_1 \leq \alpha_0\right) ,$$

was in Abbildung 3.4 im Vergleich zur Ablehnregion nach Bauer und Köhne (1994) graphisch illustriert ist. Man erkennt einen deutlicheren Unterschied zum Verfahren nach Bauer und Köhne (1994) als der in Abbildung 3.3 illustrierte. Insbesondere ist es bei einem p-Wert p_1 der ersten Sequenz von nahe 0.50 sehr unwahrscheinlich, in der zweiten Stufe zu einem signifikanten Resultat zu kommen: Für $p_1 = 0.50$ wird in dem in der Abbildung angegebenen Fall $p_2 = 0.0041$ benötigt, was praktisch ausgeschlossen ist bzw. einen sehr hohen Stichprobenumfang der zweiten Stufe erfordern würde. Hier ergibt sich also als „Indiz“ für eine vernünftige Weiterplanung der Studie die Vorgabe eines kleineren α_0. Andererseits ist die obige Eigenschaft als Vorteil des Verfahrens nach Lehmacher und Wassmer (1999) anzusehen, da es (z.B. vor einer Zulassungsbehörde) nur schwer vertretbar ist, daß ein in zweiten Stufe nur moderat kleiner p-Wert ($p_2 = 0.0173$, falls $p_1 = 0.50$ beim Verfahren nach Bauer und Köhne, 1994, in dem in der Abbildung angegebenen Fall) zur Ablehnung der Nullhypothese führen soll, obwohl der erste p-Wert keinen Effekt zeigt (ein angemessener Stichprobenumfang n_1 sei vorausgesetzt).

Die auf der nächsten Seite folgende Übersicht faßt die behandelten zweistufig adaptiven Testverfahren mit dem Werkzeug der durch (3.28) beschriebenen allgemeinen Vorgehensweise zusammen. Wie bereits bemerkt, läßt sich auch der Fall einer Studie mit festem Stichprobenumfang damit charakterisieren. Dieses Werkzeug läßt sich auch benutzen, um die Bereitstellung von adaptiven Verfahren für die zweiseitige Fragestellung zu ermöglichen. Die Beschreibung dieser Verfahren geschieht im nächsten Abschnitt.

Zuvor wird jedoch noch dargestellt, wie sich die in Abschnitt 3.1.3 dargestellten Betrachtungen zum $power$-Gewinn der adaptiven Planung für die verschiedenen Verfahren in dieser einheitlichen Formulierung durchführen lassen. Die dort

Übersicht

Funktionen $\varrho(p_1, p_2)$ und $\alpha(p_1)$ für die beschriebenen Testverfahren

Fester Stichprobenumfang:

$$\varrho(p_1, p_2) = p_1$$

$$\alpha(p_1) = \begin{cases} 0 & , \text{ falls } p_1 > \alpha \\ 1 & , \text{ falls } p_1 \leq \alpha \end{cases}$$

Bauer und Köhne (1994):

$$\varrho(p_1, p_2) = p_1 \cdot p_2$$

$$\alpha(p_1) = \begin{cases} 0 & , \text{ falls } p_1 \geq \alpha_0 \\ c_{\alpha_2}/p_1 & , \text{ falls } \alpha_1 < p_1 < \alpha_0 \\ 1 & , \text{ falls } p_1 \leq \alpha_1 \end{cases}$$

Proschan und Hunsberger (1995):

$$\varrho(p_1, p_2) = \sqrt{(\Phi^{-1}(1-p_1))^2 + (\Phi^{-1}(1-p_2))^2}$$

$$\alpha(p_1) = \begin{cases} 0 & , \text{ falls } p_1 \geq \alpha_0 \\ 1 - \Phi(\sqrt{\tilde{c}^2_{\mathrm{PH}}(\alpha) - (\Phi^{-1}(1-p_1))^2}), & \\ \qquad \text{falls } 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) < p_1 < \alpha_0 & \\ 1 & , \text{ falls } p_1 \leq 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha)) \end{cases}$$

Lehmacher und Wassmer (1999):

$$\varrho(p_1, p_2) = \Phi^{-1}(1-p_1) + \Phi^{-1}(1-p_2)$$

$$\alpha(p_1) = \begin{cases} 0 & , \text{ falls } p_1 \geq \alpha_0 \\ 1 - \Phi(\sqrt{2} \cdot \tilde{c}(\alpha) - \Phi^{-1}(1-p_1)), & \\ \qquad \text{falls } 1 - \Phi(\tilde{c}(\alpha)) < p_1 < \alpha_0 & \\ 1 & , \text{ falls } p_1 \leq 1 - \Phi(\tilde{c}(\alpha)) \ , \end{cases}$$

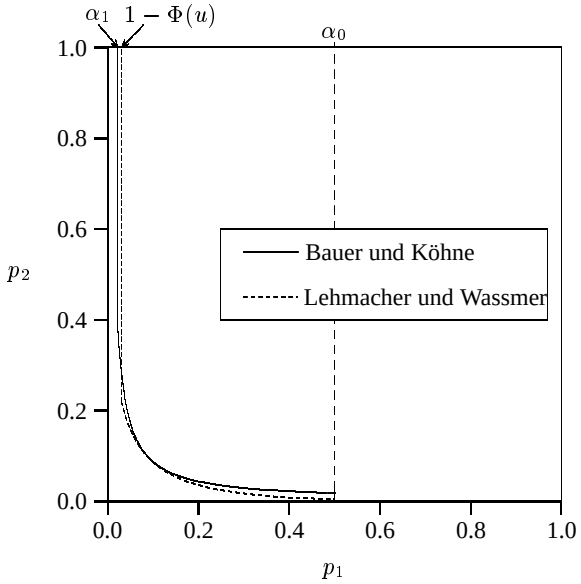

Abbildung 3.4: *Vergleich der Ablehnregionen der adaptiven zweistufigen Verfahren nach Bauer und Köhne (1994) bzw. Lehmacher und Wassmer (1999);* $\alpha = 0.05$, $\alpha_0 = 0.50$, $\alpha_1 = 0.0233$, $u = 1.871$.

abgeleiteten Formeln (3.24) und (3.25) für die $power$ bzw. den ASN des zweistufigen adaptiven Verfahrens ergeben sich bei Verwendung der Funktion $\alpha(p_1)$ zu

$$power = \Phi(\lambda \cdot \vartheta_1^* - \Phi^{-1}(1-\alpha_1)) + \int_{\alpha_1}^{\alpha_0} \Big(1 - \Phi\big(\Phi^{-1}(1-\alpha(p_1)) - \lambda \cdot \vartheta_2^*\big)\Big) \cdot \frac{\phi(\Phi^{-1}(1-p_1) - \lambda \cdot \vartheta_1^*)}{\phi(\Phi^{-1}(1-p_1))}\, dp_1$$

und

$$ASN = 2 \cdot \Big(\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{\delta^*}\Big)^2 \Big(1 + \int_{\alpha_1}^{\alpha_0} \frac{\big(\Phi^{-1}(1-\alpha(p_1)) + \Phi^{-1}(1-\beta_2)\big)^2}{(\Phi^{-1}(1-p_1))^2} \cdot \frac{\phi(\Phi^{-1}(1-p_1) - \lambda \cdot \vartheta_1^*)}{\phi(\Phi^{-1}(1-p_1))}\, dp_1\Big)\ ,$$

wobei $\lambda \cdot \vartheta_1^*$ durch (3.20) und $\lambda \cdot \vartheta_2^*$ durch

$$\lambda \cdot \frac{\left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right) \cdot \left(\Phi^{-1}(1-\alpha(p_1)) + \Phi^{-1}(1-\beta_2)\right)}{\Phi^{-1}(1-p_1)}$$

gegeben sind. n_1 wird also so gewählt, daß die erste Sequenz die $power$ $1-\beta$ besitzt. α_1 bezeichnet wie bei Verwendung der Bauer und Köhne-Prozedur den kritischen Wert für den p-Wert der ersten Stufe. Er ist gegeben durch $\alpha_1 = 1 - \Phi(\tilde{c}_{\mathrm{PH}}(\alpha))$ bei Verwendung der Proschan und Hunsberger-Prozedur bzw. $\alpha_1 = 1 - \Phi(\tilde{c}(\alpha))$ bei Verwendung des eben besprochenen Verfahrens nach Lehmacher und Wassmer. Die Ableitung der Ausdrücke für die $power$ und den ASN geschieht völlig analog zu den Ausführungen in Abschnitt 3.1.3, wobei konsequent der Entscheidungsalgorithmus in Abhängigkeit der p-Werte zu formulieren ist. Darüber hinaus ist zu beachten, daß die Dichte $f_{\vartheta_1^*}(p_1)$ durch

$$f_{\vartheta_1^*}(p_1) = \frac{\phi(\Phi^{-1}(1-p_1) - \vartheta_1^*)}{\phi(\Phi^{-1}(1-p_1))}$$

gegeben ist. In Ergänzung zu Tabelle 3.5 illustrieren Abbildung 3.5 und Abbildung 3.6 die $power$ bzw. den ASN der drei betrachteten adaptiven Verfahren nach Bauer und Köhne (1994), Lehmacher und Wassmer (1999) sowie Proschan und Hunsberger (1995)[5].

Die Unterschiede der verschiedenen Verfahren sind eher marginal. In der Tat unterscheiden sich die Verfahren von Bauer und Köhne (1994) und Proschan und Hunsberger (1995) kaum, wie aufgrund der in Wassmer (1998) dargestellten Vergleiche auch nicht anders zu erwarten war. Beim Vergleich dieser Verfahren mit dem Verfahren nach Lehmacher und Wassmer (1999) ist festzuhalten, daß sich je nach Wahl von α_0 unterschiedliche Ergebnisse ergeben: Bei kleinem α_0 (z.B. $\alpha_0 = 0.15$) besitzt das Bauer und Köhne-Verfahren die höchste $power$, während die höchste $power$ für größere Werte von α_0 (z.B. $\alpha_0 = 0.30$ oder 0.40) das Lehmacher und Wassmer-Verfahren liefert. Dabei ist ein $power$-„Gewinn“

5. Insbesondere ist die $power$ bzw. der ASN rechentechnisch einfacher zu handhaben als die in Abschnitt 3.1.1 beschriebene Möglichkeit der $power$-Berechnung beim modifizierten Kombinationstest nach Fisher, bei der eine zweidimensionale Integration durchzuführen war. Man beachte jedoch, daß „das innere“ Integral nicht verschwunden ist, sondern in dem mit der Verteilungsfunktion formulierten Ausdruck für die $power$ enthalten ist.

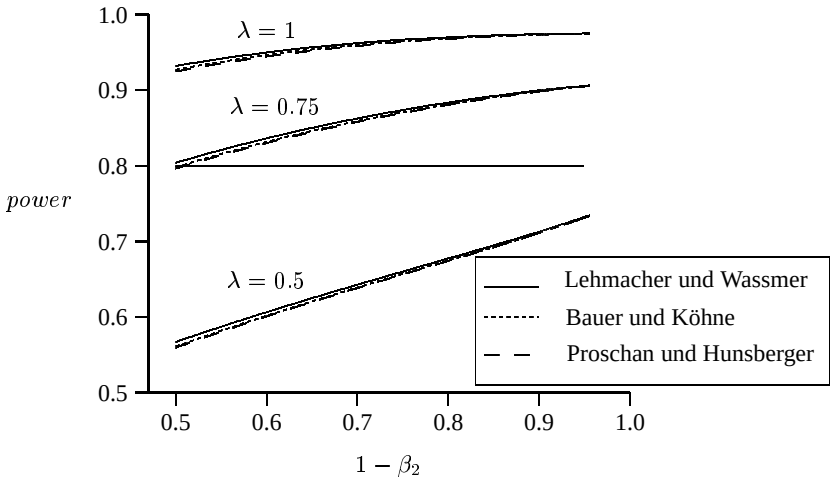

Abbildung 3.5: *Vergleich der globalen $power$ der adaptiv zweistufigen Verfahren nach Bauer und Köhne (1994), Lehmacher und Wassmer (1999) sowie Proschan und Hunsberger (1995);* $\alpha = 0.05, \alpha_0 = 0.30, 1 - \beta = 0.80$ *(vgl. Tabelle 3.5).*

jeweils mit einer Erhöhung des ASN des betrachteten Verfahrens verbunden (vgl. Abbildung 3.6). Dies macht einen direkten Vergleich der Verfahren schwieriger. Weitere Untersuchungen über die Bewertung dieser Verfahren hinsichtlich geeigneter Gütekriterien sowie günstige Vorgehensweisen in adaptiv geplanten Studien sind Gegenstand von aktuellen Forschungstätigkeiten (vgl. auch Posch und Bauer, 1999).

3.1.5 Zweiseitige Pläne

Die bisher behandelten adaptiven Verfahren wurden für die einseitige Fragestellung beschrieben. Die Bereitstellung von adaptiven Plänen für die zweiseitige

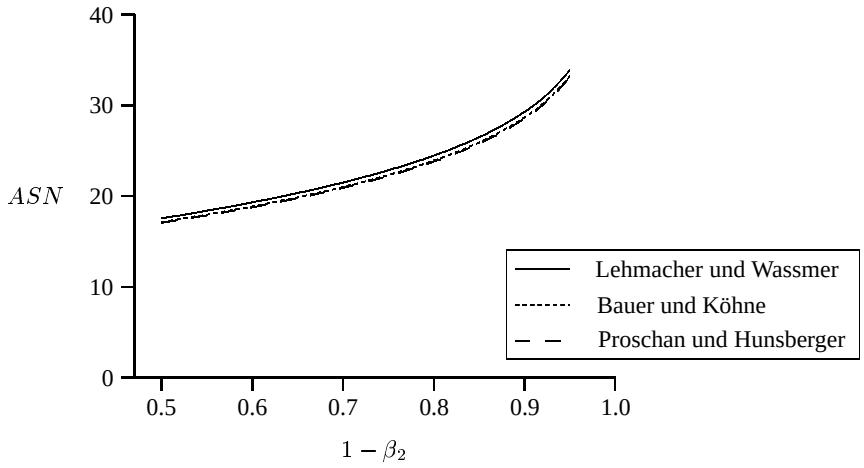

Abbildung 3.6: *Vergleich des ASN der adaptiv zweistufigen Verfahren nach Bauer und Köhne (1994), Lehmacher und Wassmer (1999) sowie Proschan und Hunsberger (1995);* $\alpha = 0.05$, $\alpha_0 = 0.30$, $1-\beta = 0.80$, $\delta^* = 1$, $\lambda = 1$ *(vgl. Tabelle 3.5).*

Fragestellung, d.h. das Testen von H_0 gegen die zweiseitige Alternative

$$\mathrm{H}_1 : \ \mu_1 - \mu_2 \neq 0 \ ,$$

ist jedoch für die praktische Anwendbarkeit dieser Verfahren sehr wichtig. Ohne auf die in der Literatur heftig diskutierte Kontroverse eingehen zu wollen, ob prinzipiell zweiseitige oder auch für bestimmte Fragestellungen einseitige Verfahren zulässig sind, sei betont, daß die aktuellen ICH-*guidelines* (ICH, 1998) die Verwendung von einseitigen Verfahren zum Niveau $\alpha/2$ als Richtlinie vorgeben. Dies sichere die Vergleichbarkeit und Konsistenz der Testergebnisse mit entsprechenden Konfidenzintervallen. Die bisher in der vorliegenden Arbeit beschriebenen kritischen Werte für die adaptiven Verfahren wurden auch für $\alpha = 0.025$ und $\alpha = 0.005$ tabelliert, da deren „ICH-gerechte“ Verwendung mit den üblichen Niveauvorgaben 0.05 bzw. 0.01 korrespondiert. Die in diesem Kapi-

tel bisher beschriebenen Verfahren kommen daher nicht in den Konflikt, die Richtlinien-Anforderung nicht erfüllen zu können.

Trotzdem erscheint es sinnvoll, auch „wirklich“ zweiseitige Pläne bereitzustellen, die die erwünschte Richtung der Testentscheidung offenhalten. Es ist klar, daß die Verwendung zweiseitiger p-Werte und entsprechender Kombinationstestverfahren zu direktionalen Konflikten führen kann. Denn ein zweiseitiger p-Wert kann klein sein, wenn der Effekt in die eine oder in die entgegengesetzte Richtung geht. Eine einfache Verwendung von zweiseitigen p-Werten in den auf Kombinationstests beruhenden adaptiven Verfahren ist also nicht oder nur eingeschränkt möglich. Ein solches Verfahren müßte zusätzlich fordern, daß der Effekt in der ersten und in der zweiten Stufe in die gleiche Richtung zeigt. Dieses Vorgehen besitzt den Nachteil, daß ein gegenläufiger Trend *nie* zur Ablehnung von H_0 führt. Es ist jedoch nicht einzusehen, warum dieser Fall theoretisch ausgeschlossen werden soll (man denke an einen schwachen Effekt mit sehr kleinem Stichprobenumfang n_1 und einen starken gegenläufigen Effekt mit sehr großem Stichprobenumfang n_2).

Eine Vorgehensweise, bei der dieses Problem nicht entsteht, beruht auf der Verwendung von einseitigen p-Werten auch für die zweiseitige Fragestellung. Die Nullhypothese wird abgelehnt, falls der p-Wert bzw. der geeignet kombinierte p-Wert klein (d.h. nahe an 0) oder groß (d.h. nahe an 1) ist. Dies läßt sich bei der Verwendung von Fishers Produktregel sehr einfach verdeutlichen. H_0 wird nach der zweiten Sequenz abgelehnt, falls

$$p_1 \cdot p_2 \leq c_{\alpha/2} \quad \text{oder} \quad (1-p_1) \cdot (1-p_2) \leq c_{\alpha/2} \ ,$$

wobei p_1 und p_2 die einseitigen p-Werte der ersten bzw. zweiten Sequenz bezeichnen. Die Studie wird mit der Annahme von H_0 in der ersten Sequenz beendet, falls

$$\alpha_0 \leq p_1 \leq 1-\alpha_0 \ ,$$

wobei $\alpha_0 \leq 0.50$ zu wählen ist. Entsprechend lassen sich Werte $\alpha_1 > c_{\alpha/2}$ bestimmen, die zur Ablehnung von H_0 in der ersten Stufe führen, falls

$$p_1 \leq \alpha_1 \quad \text{oder} \quad 1-p_1 \leq \alpha_1 \ .$$

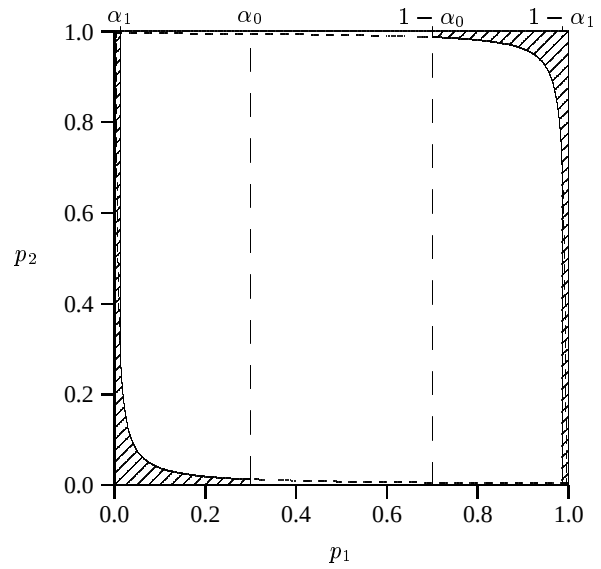

Abbildung 3.7: *Ablehnregion (schraffierte Fläche) für das zweiseitige zweistufige Verfahren mit Fishers Produktregel;* $\alpha = 0.05$, $\alpha_0 = 0.30$, $\alpha_1 = 0.0131$.

In Abbildung 3.7 sind die Ablehnregionen dieses zweiseitigen Verfahrens für $\alpha = 0.05$ und $\alpha_0 = 0.30$ graphisch illustriert.

Man sieht sofort, daß bei der Berücksichtigung der frühen Annahme von H_0 die oben erwähnte Möglichkeit der Ablehnung von H_0 nach der zweiten Sequenz trotz gegenläufigem Effekt nicht mehr gegeben ist. Insbesondere ist aufgrund der Symmetrie der Ablehnbereiche auch leicht einzusehen, daß diese Vorgehensweise der am Anfang dieses Abschnitts besprochenen entspricht: Führt man zwei einseitige Tests zum Niveau $\alpha/2$ durch, und zwar den Test für H_0 gegen $\mathrm{H}_1 : \mu_1 - \mu_2 > 0$ sowie den Test für H_0 gegen $\mathrm{H}_1 : \mu_1 - \mu_2 < 0$, so ist das Verfahren ein valider Test für die zweiseitige Fragestellung. Man beachte, daß bei dieser in Abbildung 3.7 illustrierten Möglichkeit die Richtung des Effekts nicht vorgegeben zu sein braucht. Bei Nichtberücksichtigung der frühen Annahme von H_0 ist das resultierende Testverfahren nicht mit dem einseitigen

Test zum Niveau $\alpha/2$ identisch. Allerdings ergeben sich numerisch kaum unterscheidbare Ergebnisse, da lediglich die Wahrscheinlichkeit für die Ereignisse '$p_1 < c_{\alpha/2}$ und $p_2 > 1 - c_{\alpha/2}$' und '$p_1 > 1 - c_{\alpha/2}$ und $p_2 < c_{\alpha/2}$' unter Gültigkeit von H_0 mit berücksichtigt werden muß. Diese ist durch $c^2_{\alpha/2}$ gegeben und bei üblicher Wahl von α (bzw. α_2) verschwindend klein. Eine analoge Eigenschaft wurde schon in Abschnitt 2.2.3 bei der Behandlung klassischer einseitiger gruppensequentieller Pläne bemerkt.

Die im letzten Abschnitt beschriebene allgemeine Vorgehensweise läßt sich auf die zweiseitige Fragestellung übertragen. Wird eine Funktion $\alpha(p_1)$ vorgegeben, die als durch p_1 bedingte Fehlerrate 1. Art definiert ist, so führt ein Testverfahren mit

$$\int_0^1 \alpha(p_1)\, dp_1 = \alpha$$

zu einem Niveau-α adaptiven Testverfahren. Die Funktion $\alpha(p_1)$ ist im zweiseitigen Testproblem monoton fallend für $p_1 \leq 0.50$ und monoton steigend für $p_1 \geq 0.50$. Ein Verfahren ohne Zwischenauswertung führt zu

$$\alpha(p_1) = \begin{cases} 1 & , \text{ falls } p_1 \leq \alpha/2 \\ 0 & , \text{ falls } \alpha/2 < p_1 < 1 - \alpha/2 \\ 1 & , \text{ falls } p_1 \geq 1 - \alpha/2\ , \end{cases}$$

was dem üblichen zweiseitigen Test entspricht. Im eben beschriebenen Verfahren ist $\alpha(p_1)$ bei Vorgabe von $\varrho(p_1, p_2) = p_1 \cdot p_2$ durch

$$\alpha(p_1) = \begin{cases} 1 & , \text{ falls } p_1 \leq \alpha_1 \\ c_{\alpha_2}/p_1 & , \text{ falls } \alpha_1 < p_1 < \alpha_0 \\ 0 & , \text{ falls } \alpha_0 \leq p_1 \leq 1 - \alpha_0 \\ c_{\alpha_2}/(1-p_1) & , \text{ falls } 1 - \alpha_0 < p_1 < 1 - \alpha_1 \\ 1 & , \text{ falls } p_1 \geq 1 - \alpha_1 \end{cases}$$

gegeben, wobei $\alpha_0 \leq 0.50$ vorausgesetzt werden muß, um dieses Prinzip (d.h. die Vorgabe der Funktion $\alpha(p_1)$ und die hierdurch sich ergebenden Entscheidungsbereiche) anwenden zu können.

Auch die beiden anderen beschriebenen Testverfahren lassen sich so beschreiben bzw. deren zweiseitige Version auf ein solides theoretisches Fundament stellen.

Beim Verfahren nach Proschan und Hunsberger (1995) ergibt sich aufgrund der Symmetrie der Ablehnregionen in der p_1-p_2-Ebene die Verwendung zweier einseitiger Tests zum Niveau $\alpha/2$. Damit können die in Tabelle 3.4 angegebenen kritischen Werte verwendet werden. Eine zweiseitige Version der von Lehmacher und Wassmer (1999) vorgeschlagenen Vorgehensweise führt durch die Vorgabe einer *conditional error function* zu entsprechenden zweiseitigen Varianten, die die kritischen Werte der klassischen gruppensequentiellen Pläne übernehmen können. In der Planungsphase können somit die mit den beschriebenen Vor- und Nachteilen behafteten gruppensequentiellen Testverfahren angegeben werden, um den adaptiven Plan zu definieren. So können beispielsweise die Eigenschaften der Verfahren nach Pocock (1977) und O'Brien und Fleming (1979) diskutiert und evtl. ein optimales Design nach Wang und Tsiatis (1987) verwendet werden. Hierzu sind die Schranken zu benutzen, die den Abbruch mit der Annahme von H_0 berücksichtigen und deren Bestimmung mit den in dieser Arbeit beschriebenen Methoden zu geschehen hat. Man beachte, daß aus der Verwendung des durch $\varrho(p_1, p_2) = \Phi^{-1}(1-p_1) + \Phi^{-1}(1-p_2)$ definierten Kombinationstests bei Nichtberücksichtigung eines Abbruchs der Studie mit der Annahme von H_0 auch zweiseitige adaptive Verfahren angegeben werden können, die auf der direkten Verwendung der zweiseitigen kritischen Werte beruhen. Die resultierenden Verfahren werden im nächsten Abschnitt ausführlicher diskutiert und einige weitere Eigenschaften dieser schon prima facie sehr ansprechenden Pläne angegeben. Dies geschieht bei der sehr naheliegenden Verallgemeinerung der besprochenen adaptiven Verfahren auf mehr als $K = 2$ Stufen.

3.2 Mehrstufige Designs

Dieser Abschnitt behandelt adaptive Pläne für mehr als $K = 2$ Sequenzen. Die Möglichkeit der Bereitstellung dieser Pläne beruht auf der Tatsache, daß ein gruppensequentielles Verfahren, das auf einer spezifischen, vorab festgelegten Kombination ϱ von p-Werten p_k, $k = 1, 2, \ldots, K$, der einzelnen Sequenzen beruht, zu einem globalen Test zum Niveau α führen kann. Die Eigenschaft, daß auf Informationen der bereits durchgeführten Stufen zurückgegriffen werden kann, falls die Entscheidungsregel auf der Kombination der K separaten p-Werte beruht, wurde in Bauer (1989a) beschrieben. Es wurde dort formal be-

wiesen, daß das so bestimmte Testverfahren bei geeigneter Wahl der Entscheidungsbereiche die Niveaubedingung erfüllt. Der Beweis ergibt sich analog dem in Abschnitt 3.1.1 durchgeführten Beweis für den mit Fishers Kombinationstest formulierten zweistufigen Fall.

In den beiden folgenden Abschnitten wird gezeigt, wie für $K > 2$ auf der Basis dieser allgemeinen Vorgehensweise die Entscheidungsbereiche für das resultierende adaptiv gruppensequentielle Verfahren zu bestimmen sind. Dabei wird Fishers und das auf der *inverse normal method* beruhende Kombinationstestverfahren behandelt.

3.2.1 Mehrstufige Designs auf der Basis von Fishers Kombinationstest

Wie in Abschnitt 3.1 wird zunächst das einseitige Testproblem im Parallelgruppenvergleich behandelt, bei dem die Varianz nicht als bekannt vorausgesetzt werden muß. Die Studie bestehe aus maximal K Sequenzen, wobei (maximal) K t-Tests durchgeführt und K (einseitige) p-Werte $p_1, p_2, \ldots, p_K$ ermittelt werden. Die p-Werte werden separat bestimmt, d.h. daß jeweils nur die Daten der k-ten Sequenz, $k = 1, 2, \ldots, K$, zugrundeliegen. Um diese p-Werte geeignet zu kombinieren, werde Fishers Produktregel mit folgendem Entscheidungsalgorithmus verwendet: Die Studie wird nach Durchführung der Sequenz k mit der Annahme von H_0 beendet, falls $p_k \geq \alpha_0^{(k)}$, $k \in \{1, 2, \ldots, K-1\}$. Ist $\varrho(p_1, p_2, \ldots, p_k) = p_1 \cdot p_2 \cdot \ldots \cdot p_k \leq c_{\alpha_k} := \exp(-1/2 \cdot \chi^2_{2k,\alpha_k})$, so wird auf Stufe k mit der Ablehnung von H_0 abgebrochen. In Sequenz k wird also formal ein mit Fishers Produktregel durchgeführter Kombinationstest zum lokalen Niveau α_k durchgeführt, wobei der frühzeitige Abbruch mit der Annahme von H_0 zu berücksichtigen ist.

Die Fehlerwahrscheinlichkeit 1. Art π_k in Sequenz k ist gegeben durch

$$\pi_k = \int_{\alpha_1}^{\alpha_0^{(1)}} \int_{c_{\alpha_2}/p_1}^{\alpha_0^{(2)}} \int_{c_{\alpha_3}/(p_1 p_2)}^{\alpha_0^{(3)}} \ldots \int_{c_{\alpha_{k-1}}/(p_1 p_2 \cdots p_{k-2})}^{\alpha_0^{(k-1)}} \int_0^{c_{\alpha_k}/(p_1 p_2 \cdots p_{k-1})} dp_k \, dp_{k-1} \cdots dp_1 \tag{3.31}$$

($\alpha_1 = c_{\alpha_1}$), $k = 1, 2, \ldots, K$, und das globale Niveau des K-stufigen Verfahrens

ist bei Verwendung der kritischen Werte c_{α_k} durch

$$\pi_1 + \pi_2 + \ldots + \pi_K \tag{3.32}$$

bestimmbar. Die für die Durchführung der Prozedur benötigten Werte c_{α_k} bzw. α_k werden ermittelt, indem (3.32) identisch α gesetzt wird. Ein geschlossener, rekursiv definierter Ausdruck für (3.31) wurde von Wassmer (1999b) angegeben:

$$\begin{aligned}\pi_k = c_{\alpha_k} \sum_{\tilde{k}=1}^{k} \left(\prod_{i=1}^{\tilde{k}-1} \ln(\alpha_0^{(k-i)}) \right) \times \\ \left(\frac{1}{(k-\tilde{k})!} \ln^{k-\tilde{k}} \left(\frac{\prod_{i=1}^{k-\tilde{k}-1} \alpha_0^{(i)}}{c_{\alpha_{k-\tilde{k}}}} \right) - \right. \\ \left. \sum_{i=1}^{k-\tilde{k}-1} \frac{1}{(k-\tilde{k}+1-i)!} \ln^{k-\tilde{k}+1-i} \left(\frac{c_{\alpha_i} \prod_{j=i+1}^{k-\tilde{k}-1} \alpha_0^{(j)}}{c_{\alpha_{k-\tilde{k}}}} \right) \frac{\pi_i}{c_{\alpha_i}} \right),\end{aligned} \tag{3.33}$$

$k = 1, 2, \ldots, K$, wobei für $j' > j$ $\prod_{i=j'}^{j} x_i = 1$ und $\sum_{i=j'}^{j} x_i = 0$ zu setzen ist; ferner ist $\ln^k(x) = \big(\ln(x)\big)^k$, $0^0 = 1$ und für c_{α_0} ein beliebiger Wert zu setzen.

Die Gültigkeit von (3.33) kann mit Hilfe eines Induktionsbeweises gezeigt werden (vgl. Appendix A in Wassmer, 1999b). Die Berechnung des K-dimensionalen Integrals (3.31) mittels der Formel (3.33) ist eine Alternative und Verallgemeinerung der in Bauer (1989c) für den Fall $\alpha_0^{(k)} = 1$, $k = 1, 2, \ldots, K-1$, angegebenen Berechnungsmethode. (3.33) kann nicht oder nur unwesentlich vereinfacht werden. Man beachte jedoch, daß für spezielle Wahl der $\alpha_0^{(k)}$ und α_k diese Möglichkeit besteht[6]. In Bauer und Köhne (1994), S.1034, und Bauer und

6. Für $\alpha_0^{(k)} = 1$ und $c_{\alpha_k} = \exp(-\frac{1}{2}\chi^2_{2K,\alpha}) =: c_\alpha$, $k = 1, 2, \ldots, K$, ergibt sich der einfache Spezialfall $\pi_k = c_\alpha/(k-1)!\ln^{k-1}(1/c_\alpha)$, was eine alternative Berechnungsmöglichkeit der Perzentile der χ^2-Verteilung mit geradzahligen Freiheitsgraden mit sich bringt. Dies folgt aus der Gültigkeit von $P(X \geq x) = \exp(-x/2)\sum_{j=0}^{\nu-1}(\frac{1}{2}x)^j/j!$, falls X χ^2-verteilt mit 2ν Freiheitsgraden ist (Johnson und Kotz, 1970, S.173) (vgl. Fußnote 1 auf S.75).

Röhmel (1995), S.1600, wurden beispielsweise für $K = 3$ Ausdrücke angegeben, die aus der hier angegebenen Formel als Spezialfälle resultieren.

Die Ausdrücke für $\pi_1, \pi_2, \pi_3, \pi_4$ lauten wie folgt:

$$
\begin{aligned}
\pi_1 &= \alpha_1 \ , \\
\pi_2 &= c_{\alpha_2} \Big(\ln(\alpha_0^{(1)}) - \ln(\alpha_1) \Big) \ , \\
\pi_3 &= c_{\alpha_3} \left(\frac{1}{2} \ln^2(\frac{\alpha_0^{(1)}}{c_{\alpha_2}}) - \frac{1}{2} \ln^2(\frac{\alpha_1}{c_{\alpha_2}}) + \right. \\
&\qquad \left. \ln(\alpha_0^{(2)}) \Big(\ln(\alpha_0^{(1)}) - \ln(\alpha_1) \Big) \right) \ , \\
\pi_4 &= c_{\alpha_4} \left(\frac{1}{6} \ln^3(\frac{\alpha_0^{(1)}\, \alpha_0^{(2)}}{c_{\alpha_3}}) - \frac{1}{6} \ln^3(\frac{\alpha_1\, \alpha_0^{(2)}}{c_{\alpha_3}}) + \right. \\
&\qquad \frac{1}{2} \ln^2(\frac{c_{\alpha_2}}{c_{\alpha_3}}) \Big(\ln(\alpha_1) - \ln(\alpha_0^{(1)}) \Big) + \\
&\qquad \frac{1}{2} \ln(\alpha_0^{(3)}) \Big(\ln^2(\frac{\alpha_0^{(1)}}{c_{\alpha_2}}) - \ln^2(\frac{\alpha_1}{c_{\alpha_2}}) \Big) + \\
&\qquad \left. \ln(\alpha_0^{(2)}) \ln(\alpha_0^{(3)}) \Big(\ln(\alpha_0^{(1)}) - \ln(\alpha_1) \Big) \right) \ ,
\end{aligned}
$$

und werden beliebig komplexer für größer werdendes K. In Anhang A.9 ist ein SAS-Programm beschrieben, das die Berechnung von (3.32) für beliebiges (bzw. die Rechnerkapazität nicht überschreitendes) K ermöglicht.

Für die konkrete Berechnung der kritischen Schranken müssen noch Bedingungen an die Entscheidungsbereiche formuliert werden. Wassmer (1999b) schlug einige Möglichkeiten vor, die hier kurz vorgestellt und erläutert werden. Bei allen diesen Verfahren wird vorausgesetzt, daß die Werte $\alpha_0^{(k)}$ konstant gleich α_0 gewählt werden (d.h. $\alpha_0^{(1)} = \alpha_0^{(2)} = \ldots = \alpha_0^{(K-1)} =: \alpha_0$). Die erste Möglichkeit beruht auf konstanten lokalen Signifikanzniveaus.

Konstante lokale Signifikanzniveaus

Benutzt man konstante lokale Signifikanzniveaus α^* auf jeder Stufe k, d.h. setzt man $\alpha_1 = \alpha_2 = \ldots = \alpha_K =: \alpha^*$ und entsprechend $c_{\alpha_k} = \exp(-1/2 \cdot \chi^2_{2k,\alpha^*})$,

$k = 1, 2, \ldots, K$, in (3.31), so läßt sich α^* bestimmen, indem (3.32) identisch α gesetzt wird. In Tabelle 3.7 sind für $K = 2, 3, 4, 5$, $\alpha = 0.05, 0.025, 0.01, 0.005$ und $\alpha_0 = 1.0$, 0.50, 0.30 die kritischen Werte c_{α_k} für das Produkt der p-Werte auf Stufe k, $k = 1, 2, \ldots, K$, angegeben. Tabellen von α^* für einige andere Werte von α_0 finden sich in Wassmer (1999b). Man beachte auch, daß der Fall $\alpha_0 = 1$ bereits in Bauer (1989c) vorgeschlagen und kritische Werte für bis zu $K = 10$ Stufen angegeben wurden. Das SAS-Programm in Anhang A.9 reproduziert diese und die in Tabelle 3.7 angegebenen allgemeineren Entscheidungsschranken für beliebige Werte von α und α_0.

Mit kleiner werdendem α_0 werden die lokalen Signifikanzniveaus α^* größer, was dem bereits für $K = 2$ beobachteten Gewinn entspricht, falls die frühe Annahme von H_0 berücksichtigt wird. Es sei an dieser Stelle jedoch noch einmal betont, daß bei der Wahl einer derartigen Prozedur die Studie nach Durchführung einer Stufe k abgebrochen werden *muß*, falls tatsächlich $p_k \geq \alpha_0$ beobachtet wird. Man beachte auch, daß die Wahl konstanter Signifikanzniveaus α^* nicht bedeutet, daß die Ablehnwahrscheinlichkeiten π_k, $k = 1, 2, \ldots, K$, konstant sind. Dies ist in völliger Analogie zu den in Kapitel 2 besprochenen klassischen gruppensequentiellen Plänen zu sehen, bei denen eine konstante Aufteilung des Niveaus nach Pocock (1977) auch nicht mit einer linearen α-*spending*-Funktion korrespondierte (vgl. Abschnitt 2.3.3).

Eine bei der Anwendung von Fishers Kombinationstest auftretende Problematik wurde von Bauer und Köhne (1994) als sog. '*qualitative treatment-stage interaction*' beschrieben: Es kann passieren, daß ein p-Wert p_k, $k \in \{1, 2, \ldots, K\}$, formal zu einer Ablehnung von H_0 auf Stufe k führt, obwohl $p_k \geq \alpha_0$. In diesem Fall ist also nicht klar, ob die Studie mit der Annahme oder der Ablehnung von H_0 beendet werden soll. Ein Zahlenbeispiel: Ist $\alpha = 0.05$, $K = 3$ und $\alpha_0 = 0.20$, so ergeben sich die kritischen Werte $\alpha_1 = \alpha^* = 0.0375$, $c_{\alpha_2} = 0.006163$ und $c_{\alpha_3} = 0.00125$. Ist $p_1 = 0.0376$ $(> \alpha_1)$ und $p_2 = 0.1642$ $(p_1 \cdot p_2 > c_{\alpha_2})$, dann führt $p_3 = 0.202 > \alpha_0$ zur Ablehnung von H_0, da $p_1 \cdot p_2 \cdot p_3 < c_{\alpha_3}$. Die „Interaktion zwischen Effekten und Stufen" ist hier also in Stufe K aufgetreten, was nach Meinung des Autors jedoch keine prinzipielle Problematik darstellt. Denn auf Stufe K genügt es zu fordern, daß der p-Wert in „die richtige Richtung" bzw. sogar „nicht zu sehr in die falsche Richtung" zeigt. Die Berechnungen zeigen überdies, daß diese Art der Interaktion (d.h. für

Tabelle 3.7: *Kritische Werte c_{α_k} für das Produkt der k p-Werte, $k = 1, 2, \ldots, K$, basierend auf konstanten lokalen Signifikanzniveaus $\alpha_1 = \alpha_2 = \ldots = \alpha_K = \alpha^* = c_{\alpha_1}$. α_0 ist die untere Schranke für p_k, die zur frühen Annahme von H_0 führt.*

			α			
α_0	K		0.05	0.025	0.01	0.005
1.00	2	c_{α_1}	0.032308	0.015788	0.006168	0.00304
		c_{α_2}	0.005154	0.002221	0.000753	0.000338
	3	c_{α_1}	0.025514	0.012309	0.004744	0.002319
		c_{α_2}	0.003897	0.001664	0.000559	0.000249
		c_{α_3}	0.000748	0.000291	0.000088	0.000037
	4	c_{α_1}	0.021758	0.010405	0.003973	0.001931
		c_{α_2}	0.003231	0.00137	0.000457	0.000203
		c_{α_3}	0.000607	0.000235	0.000071	0.000029
		c_{α_4}	0.000128	0.000046	0.000013	4.923E-6
	5	c_{α_1}	0.019318	0.009177	0.00348	0.001683
		c_{α_2}	0.00281	0.001186	0.000394	0.000174
		c_{α_3}	0.00052	0.000201	0.00006	0.000025
		c_{α_4}	0.000108	0.000039	0.000011	4.127E-6
		c_{α_5}	0.000024	8.058E-6	2.045E-6	7.564E-7
0.50	2	c_{α_1}	0.034942	0.01687	0.006519	0.003194
		c_{α_2}	0.005659	0.002399	0.000802	0.000357
	3	c_{α_1}	0.030025	0.014195	0.005368	0.002595
		c_{α_2}	0.004725	0.001962	0.000643	0.000283
		c_{α_3}	0.000928	0.000349	0.000102	0.000042
	4	c_{α_1}	0.027885	0.012999	0.004841	0.002318
		c_{α_2}	0.004328	0.001772	0.000572	0.000249
		c_{α_3}	0.000841	0.000312	0.00009	0.000037
		c_{α_4}	0.000182	0.000062	0.000016	6.235E-6
	5	c_{α_1}	0.026863	0.012405	0.00457	0.002173
		c_{α_2}	0.004141	0.001679	0.000536	0.000232
		c_{α_3}	0.000801	0.000294	0.000084	0.000034
		c_{α_4}	0.000173	0.000058	0.000015	5.734E-6
		c_{α_5}	0.00004	0.000012	2.99E-6	1.07E-6
0.30	2	c_{α_1}	0.037256	0.01779	0.006809	0.003319
		c_{α_2}	0.00611	0.002552	0.000843	0.000373
	3	c_{α_1}	0.033996	0.015818	0.005892	0.002824
		c_{α_2}	0.005477	0.002226	0.000715	0.000311
		c_{α_3}	0.001095	0.000401	0.000115	0.000047
	4	c_{α_1}	0.033056	0.015173	0.005562	0.002637
		c_{α_2}	0.005297	0.00212	0.000669	0.000288
		c_{α_3}	0.001054	0.000381	0.000107	0.000043
		c_{α_4}	0.000233	0.000077	0.00002	7.373E-6
	5	c_{α_1}	0.032802	0.014967	0.005441	0.002564
		c_{α_2}	0.005248	0.002087	0.000653	0.000279
		c_{α_3}	0.001044	0.000374	0.000104	0.000041
		c_{α_4}	0.00023	0.000076	0.000019	7.107E-6
		c_{α_5}	0.000054	0.000016	3.82E-6	1.342E-6

$k = K$) für $0.30 \leq \alpha_0 \leq 1$ in den hier betrachteten Fällen nicht auftritt. Tritt bei einer Wahl von kritischen Schranken das Problem hingegen für $k < K$ auf, so ist die Anwendung der Prozedur höchst fraglich.

Volle Ausschöpfung des Niveaus auf der letzten Stufe und konstante lokale Niveaus

Eine alternative Möglichkeit ergibt sich, wenn auf der letzten Stufe K formal ein Kombinationstest zum Niveau α durchgeführt wird, d.h. man setzt $c_{\alpha_K} = \exp(-1/2 \cdot \chi^2_{2K,\alpha})$. Berücksichtigt man den frühen Abbruch der Studie mit der Annahme von H_0 nicht, so kann die Studie in einer Stufe $k < K$ mit der Ablehnung von H_0 abgebrochen werden, falls das Produkt der p-Werte kleiner als c_{α_K} ist. Diese Tatsache wurde in früheren Abschnitten mit dem Begriff *nonstochastic curtailment* beschrieben. Berücksichtigt man die frühe Annahme von H_0, so lassen sich analog zu der in Bauer und Köhne (1994) für $K = 2$ angegebenen Vorgehensweise die kritischen Schranken entsprechend vergrößern. Wählt man konstante lokale Niveaus $\alpha_1 = \alpha_2 = \ldots = \alpha_{K-1} =: \alpha^\dagger$ für die Stufen $k < K$, so ergeben sich die in Tabelle 3.8 für α = 0.05, 0.025, 0.01, $\alpha_0 = 0.50, 0.30$ und $K = 2, 3, 4, 5$ angegebenen Werte.

Interessanterweise ist der Einfluß von K auf die lokalen Signifikanzniveaus $\alpha^\dagger$ nicht besonders groß. Für $\alpha = 0.05$ und $\alpha_0 = 0.50$ kann z.B. ein Kombinationstest zu ungefähr der Hälfte des globalen Niveaus unabhängig von der Anzahl der durchzuführenden Stufen angewendet werden. Bei der Anwendung dieser Methode können für $K > 2$ die im vorigen Abschnitt beschriebenen *qualitative treatment-stage interactions* auftreten. Auch hier zeigen jedoch die durchgeführten Berechnungen, daß Interaktionen nur in Stufe K auftreten können und daher keinen wesentlichen Einwand für die Verwendung dieses Verfahrens darstellen.

Eine schwerwiegendere Problematik ergibt sich für $\alpha_0 > 0.60$ und $K > 2$. In diesem Fall ist eine Verletzung der Bedingung

$$c_{\alpha_1} \geq c_{\alpha_2} \geq \ldots \geq c_{\alpha_K} \tag{3.34}$$

möglich. Dies bedeutet, daß die Entscheidungsregel für einen gewissen p-Wert nicht zur Ablehnung von H_0 auf Stufe k führt, obwohl für diesen p-Wert die Nullhypothese in Stufe $k + 1$ in jedem Fall abgelehnt werden würde, da $p_{k+1} \leq$

Tabelle 3.8: *Kritische Werte c_{α_k} für das Produkt der k p-Werte, $k = 1, 2, \ldots, K$, und lokale Signifikanzniveaus α_k mit voller Ausschöpfung des Niveaus α auf Stufe K und konstante Signifikanzniveaus $\alpha_1 = \alpha_2 = \ldots = \alpha_{K-1} = \alpha^\dagger$. α_0 ist die untere Schranke für p_k, die zur frühen Annahme von H_0 führt.*

			α = 0.05		α = 0.025		α = 0.01	
α_0	K	k	α_k	c_{α_k}	α_k	c_{α_k}	α_k	c_{α_k}
0.50	2	1	0.0233	0.023315	0.0102	0.010189	0.0035	0.003506
		2	0.0500	0.008705	0.0250	0.003804	0.0100	0.001309
	3	1	0.0238	0.023791	0.0105	0.010467	0.0036	0.003625
		2	0.0238	0.003588	0.0105	0.00138	0.0036	0.000412
		3	0.0500	0.001844	0.0250	0.000728	0.0100	0.000224
	4	1	0.0245	0.024453	0.0109	0.01085	0.0038	0.003794
		2	0.0245	0.003706	0.0109	0.001438	0.0038	0.000434
		3	0.0245	0.000708	0.0109	0.000248	0.0038	0.000067
		4	0.0500	0.000429	0.0250	0.000156	0.0100	0.000043
	5	1	0.0250	0.024971	0.0112	0.011161	0.0039	0.003936
		2	0.0250	0.003799	0.0112	0.001486	0.0039	0.000452
		3	0.0250	0.000727	0.0112	0.000257	0.0039	0.00007
		4	0.0250	0.000155	0.0112	0.00005	0.0039	0.000012
		5	0.0500	0.000106	0.0250	0.000036	0.0100	9.124E-6
0.30	2	1	0.0299	0.029938	0.0131	0.013084	0.0045	0.004503
		2	0.0500	0.008705	0.0250	0.003804	0.0100	0.001309
	3	1	0.0314	0.031406	0.0139	0.013935	0.0049	0.004872
		2	0.0314	0.004984	0.0139	0.001921	0.0049	0.000576
		3	0.0500	0.001844	0.0250	0.000728	0.0100	0.000224
	4	1	0.0322	0.032212	0.0145	0.01446	0.0051	0.005124
		2	0.0322	0.005136	0.0145	0.002005	0.0051	0.00061
		3	0.0322	0.001019	0.0145	0.000358	0.0051	0.000097
		4	0.0500	0.000429	0.0250	0.000156	0.0100	0.000043
	5	1	0.0326	0.032553	0.0147	0.014717	0.0053	0.005264
		2	0.0326	0.005201	0.0147	0.002046	0.0053	0.000629
		3	0.0326	0.001033	0.0147	0.000366	0.0053	0.0001
		4	0.0326	0.000228	0.0147	0.000074	0.0053	0.000018
		5	0.0500	0.000106	0.0250	0.000036	0.0100	9.124E-6

1. Die Anwendung eines derartigen Plans ist deshalb nicht zu empfehlen. Eine Modifizierung ergibt sich dadurch, indem man $\alpha^{\dagger}$ entsprechend verändert, so daß (3.34) erfüllt ist. Die durchgeführten Berechnungen zeigen, daß das Problem der Verletzung von (3.34) für $\alpha_0 \leq 0.60$ nicht auftritt. Insbesondere erweist sich aber die Einführung der frühen Annahme von H_0 trotz intuitiv klarer Heuristik als relativ starker Eingriff in das logische Gefüge der Ablehnregionen.

Volle Ausschöpfung des Niveaus auf der letzten Stufe und Vermeidung von Interaktionen

Wie bei der eben beschriebenen Methode sei auf Stufe K ein Kombinationstest zum Niveau α durchzuführen. Setzt man $c_{\alpha_k} = c_{\alpha_{k+1}}/\alpha_0$, $k = 2, 3, \ldots, K-1$, so ist leicht einzusehen, daß eine *qualitative treatment-stage interaction* nach dieser Wahl der kritischen Schranken für das Produkt der p-Werte auf den Stufen $k > 1$ nicht auftreten kann. Numerische Berechnungen zeigen überdies, daß für $\alpha_0 \geq 0.30$ angenommen werden kann, daß $\alpha_1 > c_{\alpha_2}/\alpha_0$ (vgl. Bauer und Köhne, 1994). Um einen Niveau-α-Test zu definieren, muß bei dieser Wahl der kritischen Schranken für $k > 1$ also α_1 derart gewählt werden, daß (3.32) gleich α ist. In Tabelle 3.9 sind die resultierenden Werte α_1 für α = 0.05, 0.025, 0.01, α_0 = 0.50, 0.30 und $K = 2, 3, 4, 5$ angegeben.

Die ermittelten lokalen Signifikanzniveaus erfüllen die Monotoniebedingung

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_K \tag{3.35}$$

in den hier betrachteten Fällen für $K > 3$ nicht. Dies kann kritisiert werden, denn das Verfahren entbehrt somit – abgesehen von der Vermeidung von Interaktionen – einer heuristischen Begründung der Wahl der lokalen Signifikanzniveaus. Andererseits sind Signifikanzniveaus α_k, die (3.35) erfüllen, „im Geist“ der klassischen gruppensequentiellen Pläne und werden auch bei der Verwendung derer adaptiver Varianten entsprechende Vor- und Nachteile aufweisen.

In Wassmer (1999b) ist darüber hinaus die Wahl der kritischen Schranken für Fishers Produktregel gemäß einer vorab definierten α-$spending$-Funktion beschrieben und diskutiert. Diese Vorgehensweise wurde für $K = 3$ und $\alpha_0 = 1$ bereits in Bauer und Röhmel (1995) vorgeschlagen. Es zeigt sich, daß auch hier die Wahl von α_0 entscheidenden Einfluß auf die Eigenschaften der resultierenden Entscheidungsbereiche besitzt. Es kann zu Inkonsistenzen wie bei den eben

Tabelle 3.9: *Kritische Werte c_{α_k} für das Produkt der k p-Werte, $k = 1, 2, \ldots, K$, und lokale Signifikanzniveaus α_k mit voller Ausschöpfung des Niveaus α auf Stufe K. Um Interaktionen zwischen Effekten und Stufen zu vermeiden, sind die kritischen Werte c_{α_k} für $k = 2, 3, \ldots, K-1$ durch $c_{\alpha_k} = c_{\alpha_{k+1}}/\alpha_0$ gegeben. α_0 ist die untere Schranke für p_k, die zur frühen Annahme von H_0 führt.*

			α = 0.05		α = 0.025		α = 0.01	
α_0	K	k	α_k	c_{α_k}	α_k	c_{α_k}	α_k	c_{α_k}
0.50	2	1	0.0233	0.023315	0.0102	0.010189	0.0035	0.003506
		2	0.0500	0.008705	0.0250	0.003804	0.0100	0.001309
	3	1	0.0236	0.023593	0.0103	0.010279	0.0035	0.003521
		2	0.0244	0.003688	0.0110	0.001457	0.0039	0.000447
		3	0.0500	0.001844	0.0250	0.000728	0.0100	0.000224
	4	1	0.0281	0.028137	0.0127	0.012685	0.0045	0.004513
		2	0.0126	0.001717	0.0052	0.000623	0.0017	0.000174
		3	0.0283	0.000858	0.0130	0.000311	0.0047	0.000087
		4	0.0500	0.000429	0.0250	0.000156	0.0100	0.000043
	5	1	0.0329	0.032861	0.0153	0.015268	0.0056	0.005614
		2	0.0068	0.000847	0.0026	0.000285	0.0008	0.000073
		3	0.0165	0.000423	0.0070	0.000143	0.0023	0.000036
		4	0.0309	0.000212	0.0144	0.000071	0.0053	0.000018
		5	0.0500	0.000106	0.0250	0.000036	0.0100	9.124E-6
0.30	2	1	0.0299	0.029938	0.0131	0.013084	0.0045	0.004503
		2	0.0500	0.008705	0.0250	0.003804	0.0100	0.001309
	3	1	0.0291	0.029141	0.0126	0.012627	0.0043	0.004302
		2	0.0374	0.006147	0.0170	0.002428	0.0061	0.000745
		3	0.0500	0.001844	0.0250	0.000728	0.0100	0.000224
	4	1	0.0317	0.031651	0.0143	0.014314	0.0051	0.005136
		2	0.0303	0.004769	0.0127	0.001731	0.0042	0.000482
		3	0.0415	0.001431	0.0193	0.000519	0.0071	0.000145
		4	0.0500	0.000429	0.0250	0.000156	0.0100	0.000043
	5	1	0.0341	0.034056	0.0160	0.016014	0.0060	0.006014
		2	0.0256	0.00392	0.0101	0.001321	0.0030	0.000338
		3	0.0359	0.001176	0.0157	0.000396	0.0053	0.000101
		4	0.0438	0.000353	0.0207	0.000119	0.0077	0.00003
		5	0.0500	0.000106	0.0250	0.000036	0.0100	9.124E-6

erwähnten Methoden kommen, die die Wahl einer spezifischen Aufteilung des Gesamtniveaus auf die Stufen der Studie erschwert. Insbesondere ist der Einfluß von α_0 größer als die bei den gruppensequentiellen Plänen diskutierte frühe Annahme von H_0. Der maßgebliche Grund für diesen Sachverhalt ist darin zu sehen, daß die Wahrscheinlichkeit für das Ereignis '$p_k > \alpha_0$' bei Verwendung eines gruppensequentiellen Plans generell kleiner als bei Verwendung von Fishers Kombinationstest ist. In Abbildung 3.4 ist dies für den Fall $K = 2$ illustriert. Ein weiterer Grund hierfür liegt darin, daß die beschriebenen, auf der Produktregel beruhenden Methoden zur frühen Annahme kommen, falls ein *separater* p-Wert den Wert α_0 überschreitet. Für klassische gruppensequentielle Pläne ist die Bedingung für den frühen Abbruch stärker, da die *kumulierte* Teststatistik den kritischen Wert unterschreiten muß. Dies erklärt neben der unterschiedlichen Gestalt der Ablehnregionen, warum der Einfluß des frühen Abbruchs mit der Annahme von H_0 auf die kritischen Schranken bei der Verwendung von Fishers Kombinationstest größer als bei der Verwendung entsprechender gruppensequentieller Pläne ist (vgl. z.B. Tabelle 2.5 mit den in diesem Abschnitt angegebenen kritischen Schranken).

Eine weitere Variante der Kombinationstestverfahren ergibt sich, wenn gemäß dieser Überlegung die Bedingung für die Annahme von H_0 stärker gewählt wird (z.B. in Abhängigkeit zu allen in den vorigen Stufen beobachteten p-Werten). Dies ist prinzipiell möglich, wird jedoch hier nicht weiter verfolgt. Eine optimale Wahl der Methode der Bestimmung der Entscheidungsbereiche ist auch deshalb schwierig, weil die $power$ und der ASN rechentechnisch aufwendig ist und daher Simulationsuntersuchungen notwendig sind, um Eigenschaften dieser Verfahren im adaptiven Design zu untersuchen. Im nächsten Abschnitt wird ein Verfahren vorgestellt, das auf altbekannte Ergebnisse zurückgreifen kann und entsprechend die Wahl der Entscheidungsbereiche erleichtern wird.

Abschließend sei bemerkt, daß die in diesem Abschnitt dargestellten Vorgehensweisen für das Testen von einseitigen Hypothesen formuliert wurden. Gemäß der ICH-Richtlinien (ICH, 1998) wird – wie bereits erwähnt – die Verwendung dieser Tests zum Niveau $\alpha/2$ empfohlen. Entsprechende kritische Werte sind den Tabellen zu entnehmen bzw. mit dem in Anhang A.9 angegebenen SAS-Programm zu berechnen.

3.2.2 Mehrstufige Designs auf der Basis der *inverse normal method*

In Abschnitt 3.1 wurde für das zweistufige Design beschrieben, wie unter Benutzung der *inverse normal method* für die Verknüpfung der p-Werte die Entscheidungsbereiche für ein adaptiv gruppensequentielles Design hergeleitet werden können. Es stellte sich heraus, daß die für das klassische (nicht-adaptive) gruppensequentielle Design abgeleiteten Schranken verwendet werden können. Hieraus ergeben sich zweiseitige und einseitige Varianten, wobei die Ergebnisse über die Eigenschaften der klassischen Verfahren in der Planungsphase einer Studie berücksichtigt werden können. Diese Vorgehensweise ist wie die im letzten Abschnitt besprochene auf den Fall mehrerer Sequenzen mühelos verallgemeinerbar (Lehmacher und Wassmer, 1999).

Werden höchstens K Sequenzen durchgeführt, so seien in Stufe k die einseitigen p-Werte gemäß

$$\varrho(p_1, p_2, \ldots, p_k) = \sum_{\bar{k}=1}^{k} \Phi^{-1}(1 - p_{\bar{k}}) \tag{3.36}$$

verknüpft, $k = 1, 2, \ldots, K$. Offensichtlich sind die Summanden standardnormalverteilt, falls p_k auf $[0;\, 1]$ stetig gleichverteilt ist, was für den hier wie im letzten Abschnitt angenommenen Parallelgruppenvergleich der Fall ist. Damit ist (3.36) eine Summe aus k standardnormalverteilten Variablen und mit der in Abschnitt 2.2 definierten Größe S_k (vgl. (2.20)) identisch. Geeignete Entscheidungsbereiche im Rahmen eines gruppensequentielles Verfahrens sind bereits bestimmt. Denn sämtliche in Abschnitt 2.2 angegebenen Entscheidungsbereiche sind verwendbar, solange sich daraus valide Niveau-α-Tests ergeben. Diese Bereiche sind für (3.36) selbst oder für

$$\frac{1}{\sqrt{k}} \cdot \varrho(p_1, p_2, \ldots, p_k) = \frac{1}{\sqrt{k}} \sum_{\bar{k}=1}^{k} \Phi^{-1}(1 - p_{\bar{k}}) = Z_k^* \tag{3.37}$$

angebbar. So lassen sich beispielsweise die Schranken nach Wang und Tsiatis (1987) verwenden, die die klassischen Pläne nach Pocock (1977) oder O'Brien und Fleming (1979) als Spezialfälle enthalten. Ebenso sind die von DeMets und

Ware (1980, 1982) vorgeschlagenen einseitigen Varianten mit Abbruch der Studie zugunsten der Annahme von H_0 verwendbar. Diese Verfahren halten im adaptiven Szenario das Niveau α ein, da die Entscheidungsregel auf einer Verknüpfung von k separaten p-Werten beruht (Bauer, 1989a).

Die beschriebene Vorgehensweise beruht im Fall einer bekannten Varianz σ^2 auf der Verwendung der Teststatistik $Z_k^* = \sum_{\tilde{k}=1}^{k} Z_{\tilde{k}} / \sqrt{k}$, $k = 1, 2, \ldots, K$, für beliebige, auch datenabhängig gewählte Stichprobenumfänge $n_1, n_2, \ldots, n_K$. Die wesentliche Eigenschaft des Verfahrens besteht in der Fixierung der Gewichtung der Studienabschnitte vor Durchführung der Untersuchung (vgl.auch Cui, Hung und Wang, 1999, die eine auf dem α-spending-Ansatz beruhende Vorgehensweise vorschlagen, die der hier dargestellten sehr ähnlich ist). Die von (3.36) bzw. (3.37) implizierte Gleichgewichtung korrespondiert mit einer vermuteten Aufteilung der Stufen in Sequenzen mit gleichen Stichprobenumfängen. In diesem Fall entspricht die Gleichgewichtung der im klassischen Sinne optimalen Gewichtung. Wichtiger – und wesentlicher – Punkt ist, daß bei einer Zwischenauswertung die transformierten p-Werte der Studienabschnitte *immer* gemäß (3.36) bzw. (3.37) zu bestimmen sind. Dies bedeutet, daß eine Gleichgewichtung beibehalten werden muß, obwohl eine Adaption des Stichprobenumfangs stattgefunden hat, und die Sequenzen ungleich groß geplant werden. Andererseits bedeutet dies, daß bei keiner durchzuführenden Adaption der Stichprobenumfänge das resultierende Verfahren *identisch* mit dem korrespondierenden gruppensequentiellen Verfahren ist. Genauer: Im Fall einer bei den klassischen Verfahren vorausgesetzten bekannten Varianz σ^2 sind die Verfahren in diesem Falle völlig identisch; im realistischeren Fall (σ^2 unbekannt) beruht das vorgeschlagene Verfahren auf einer gleichgewichteten Verknüpfung der *separaten* Teststatistiken der einzelnen Stufen, die mit der t-Teststatistik auf Stufe k nur approximativ übereinstimmt. Nichtsdestoweniger sei bemerkt, daß diese Vorgehensweise quasi als Nebenprodukt eine exakte Lösung des Problems der unbekannten Varianz bei gruppensequentiellen Verfahren mit sich bringt. Für die klassische gruppensequentielle Vorgehensweise sind hier nur die auf Jennison und Turnbull (1991a) zurückgehenden exakten Tabellen für andere Verteilungssituationen bekannt.

Eine Gleichgewichtung der gemäß der *inverse normal method* transformierten p-Werte trotz ungleicher Sequenzgrößen scheint prima facie dazu zu neigen, mit einem $power$-Verlust zu reagieren, da die im herkömmlichen Sinne „be-

ste“ – nämlich gewichtete – Teststatistik (2.4) bzw. (2.11) nicht verwendet wird. Der tatsächliche $power$-Verlust ist jedoch erstaunlich gering. Dieses Ergebnis erhält man dadurch, daß man das vorgeschlagene Verfahren dem „optimalen“ Verfahren (d.h. dem Verfahren mit der üblichen Gewichtung der Teststatistiken) für verschiedene Aufteilungen der Stichprobenumfänge auf die einzelnen Sequenzen gegenüberstellt. Einige Resultate dieses Vergleichs sind in Tabelle 3.10 für den Fall $K = 5$ und $\alpha = 0.05$ enthalten (vgl. auch Table 1 in Lehmacher und Wassmer, 1999). Als Effektgröße wurde die Größe zugrundegelegt, die bei einem Design mit konstanten Abbruchschranken (nach Pocock, 1977) bei $\boldsymbol{V} = (n_1, n_1+n_2, \ldots, n_1+n_2+\ldots+n_K)$ = (20,40,60,80,100) die $power$ $1-\beta$ ergibt, $1 - \beta = 0.80, 0.90$. Das einseitige Verfahren berücksichtigt den frühen Abbruch mit der Annahme von H_0, falls $Z_k^* \leq u^L = 0$. Verschiedene Aufteilungen der Stichprobenumfänge wurden betrachtet, die einer Erhöhung oder Erniedrigung der Stichprobenumfänge der Sequenzen entsprechen. Gemäß der in Abschnitt 2.3.1 beschriebenen Methode können die kritischen Werte, die $power$ und der ASN des Verfahrens für eine spezifische Sequenzaufteilung berechnet und mit der $power$ und dem ASN des vorgeschlagenen Verfahrens verglichen werden. Dieser Vergleich beruht auf der Annahme einer bekannten Varianz σ^2.

Tabelle 3.10 zeigt, daß ein maßgeblicher $power$-Verlust lediglich im Fall einer vorzeitigen Zwischenauswertung ($n_1 = 10$) auftritt. Dieses Verhalten ist für das zweiseitige Verfahren (ohne Abbruch mit der Annahme von H_0) deutlicher ausgeprägt als für das einseitige Verfahren. Der Grund hierfür liegt darin, daß für diesen Fall das einseitige Verfahren bereits durch die Berücksichtigung der frühen Annahme von H_0 einen beträchtlichen $power$-Verlust aufzuweisen hat, da ein früher Abbruch durch die geringe $power$ in der ersten Stufe häufig geschehen kann. Eine einseitige Variante ohne frühen Abbruch mit der Annahme von H_0 verhält sich analog zur zweiseitigen Variante, und vice versa. Der $power$-Verlust kann dabei noch größer werden: Im zweiseitigen Fall ist ceteris paribus bei $\boldsymbol{V}$ = (5,10,15,90,100) und $1 - \beta = 0.90$ die $power$ des optimalen Verfahrens gleich 0.886, während die $power$ des vorgeschlagenen Verfahrens nur 0.743 beträgt. Generell wird der $power$-Verlust mit wachsender Anzahl früher Zwischenauswertungen größer. Dies sind jedoch Fälle, die in der Praxis kaum auftreten bzw. zu vermeiden sind. In allen anderen Fällen ist der Unterschied in der $power$ und dem ASN verschwindend klein. Überdies treten sogar Fälle mit einem leichten $power$-Gewinn beim vorgeschlagenen Verfahren im Vergleich zum „optimalen“

Tabelle 3.10: *Vergleich von $power$ und ASN im adaptiven Verfahren nach Lehmacher und Wassmer (1999) mit entsprechendem einseitigen und zweiseitigen gruppensequentiellen Verfahren bei fest vorgegebener Aufteilung $\boldsymbol{V} = (n_1, n_1 + n_2, \ldots, n_1 + n_2 + \ldots + n_K)$ der Stichprobenumfänge. Einseitiger Test mit Abbruch mit der Annahme von H_0, falls $Z_k^* \leq u^L = 0$; Design nach Pocock (1977) mit konstanten Schranken; $K = 5$, $\alpha = 0.05$.*

		Einseitiger Test				Zweiseitiger Test			
		Adaptiv		$\boldsymbol{V}$ vorgegeben		Adaptiv		$\boldsymbol{V}$ vorgegeben	
$1-\beta$	$\boldsymbol{V}$	$power$	ASN	$power$	ASN	$power$	ASN	$power$	ASN
0.80	(10,20,30,60,100)	0.683	44.3	0.701	43.9	0.722	69.0	0.775	68.5
	(10,20,30,100,160)	0.754	59.5	0.767	57.6	0.878	98.3	0.938	93.7
	(10,20,100,140,200)	0.786	68.8	0.789	66.2	0.955	109.3	0.978	102.0
	(20,40,60,80,100)	0.800	51.3	0.800	51.3	0.800	65.0	0.800	65.0
	(20,40,80,120,160)	0.873	62.2	0.875	61.9	0.937	83.1	0.944	82.6
	(20,80,120,160,200)	0.900	74.4	0.899	72.8	0.978	95.4	0.981	93.2
	(40,60,80,90,100)	0.861	61.0	0.861	60.4	0.823	68.0	0.824	66.9
	(40,100,140,150,160)	0.948	80.2	0.947	78.1	0.957	91.5	0.957	89.0
	(40,80,100,150,200)	0.958	74.9	0.960	75.2	0.978	87.4	0.981	87.6
0.90	(10,20,30,60,100)	0.798	39.3	0.809	39.0	0.844	60.5	0.884	60.0
	(10,20,30,100,160)	0.835	51.2	0.839	50.3	0.953	83.5	0.982	80.1
	(10,20,100,140,200)	0.847	61.3	0.848	59.8	0.989	93.7	0.996	89.0
	(20,40,60,80,100)	0.900	44.9	0.900	44.9	0.900	56.8	0.900	56.8
	(20,40,80,120,160)	0.934	51.8	0.934	51.8	0.982	68.8	0.984	68.7
	(20,80,120,160,200)	0.942	64.4	0.942	63.3	0.996	81.3	0.997	79.7
	(40,60,80,90,100)	0.946	54.4	0.946	53.7	0.914	60.9	0.915	59.8
	(40,100,140,150,160)	0.984	68.5	0.983	66.8	0.988	79.3	0.989	77.2
	(40,80,100,150,200)	0.986	62.4	0.986	62.6	0.996	72.8	0.997	73.0

Verfahren auf.

Es ist zu berücksichtigen, daß die Testverfahren mit optimalen Gewichten die Möglichkeit der Adaption der Stichprobenumfänge in keiner Weise berücksichtigen. Der hier dargestellte Vergleich ist also „künstlich“, da Äpfel mit Birnen verglichen werden. Er ist lediglich als ein „quasi“- oder „als ob“-Vergleich anzusehen, der voraussetzt, daß die Stichprobenumfänge bekannt sind, die aus dem adaptiven Vorgehen zu ermitteln sind. Auf der anderen Seite ist das Ergebnis dieser artifiziellen Vorgehensweise aus diesem Grund umso interessanter. Denn wenn die Unterschiede in den betrachteten Parametern so klein ausfallen, dann

wird eine Evaluation des adaptiven Verfahrens in einem adaptiven Design einen umso deutlicheren Gewinn im Vergleich zu den herkömmlichen gruppensequentiellen Verfahren erbringen. Eine Untersuchung der $power$ bei adaptiver Planung läßt sich wie in dem in Abschnitt 3.1.3 für den zweistufigen Fall durchgeführten Szenario durchführen. Allerdings ist man dabei auf Simulationsuntersuchungen angewiesen, da aus den in Abschnitt 3.1.3 erwähnten Gründen kaum analytische bzw. numerische Resultate zu erwarten sind. Die Untersuchung der Eigenschaften dieser Verfahren im adaptiven Design ist der Gegenstand weiterer Untersuchungen.

Eine Modifikation des Verfahrens ergibt sich, falls a priori von der Gleichgewichtung der Teststatistiken abgesehen wird und eine spezifische Gewichtung $\boldsymbol{\tau} = (1, \tau_2, \ldots, \tau_K)$ gewählt wird (vgl. die Abschnitte 2.1 und 2.3.1). Beispielsweise kann sich in der Planungsphase der Studie ergeben, daß die $power$ in frühen Studienabschnitten recht hoch sein soll; dies entspricht einer höheren Gewichtung für kleine k. Administrative Gründe können zu einer möglichst frühen Testentscheidung zwingen; dies führt zu einer höheren Gewichtung in den späteren Stufen. Die dem adaptiven Verfahren zugrundeliegenden Teststatistiken sind gegeben durch

$$\sum_{\bar{k}=1}^{k} \sqrt{\tau_{\bar{k}}} \cdot \Phi^{-1}(1 - p_{\bar{k}}) \tag{3.38}$$

(vgl. (2.11)) bzw.

$$\frac{1}{\sqrt{\sum_{\bar{k}=1}^{k} \tau_{\bar{k}}}} \sum_{\bar{k}=1}^{k} \sqrt{\tau_{\bar{k}}} \cdot \Phi^{-1}(1 - p_{\bar{k}}), \tag{3.39}$$

$k = 1, 2, \ldots, K$, und die kritischen Werte können gemäß der in Abschnitt 2.3.1 beschriebenen Methode berechnet werden. Die in (3.38) definierte Kombination der p-Werte wird für den gesamten Verlauf der Studie beibehalten, d.h. daß bei einer Zwischenauswertung die transformierten p-Werte der Studienabschnitte *immer* gemäß (3.38) bzw. (3.39) zu bestimmen sind. Ein Abweichen von dieser Vorgehensweise führt zu einer Verletzung der Niveaubedingung.

Das folgende Zahlenbeispiel illustriere die eben dargestellte Methode. Man plane ein zweiseitiges adaptiv gruppensequentielles Verfahren mit konstanten kri-

tischen Schranken nach Pocock (1977) zum Niveau $\alpha = 0.05$. Es sei eine vierstufiges Design geplant, in der eine höhere $power$ im ersten Studienabschnitt vorliegen soll. Dies führe zur Vorgabe von $\boldsymbol{\tau} = (1, 0.5, 0.5, 0.5)$. Die konstanten kritischen Schranken für die Teststatistik (3.39) lauten $u = 2.319$ (vgl. Tabelle 2.7). Im ersten Studienabschnitt habe sich ein p-Wert $p_1 = 0.0339$ ergeben. Dies führt zur Fortsetzung der Studie, da $\Phi^{-1}(1 - p_1) = 1.826 < 2.319$; eine Beibehaltung der Sequenzplanung werde beschlossen, d.h. die nächste Stufe wird mit der Hälfte des Stichprobenumfangs der ersten Sequenz durchgeführt. Die zweite Sequenz ergebe einen p-Wert $p_2 = 0.1281$. Auch dies führt zu keiner Ablehnung von H_0, da (3.39) durch $(\Phi^{-1}(1-p_1)+\sqrt{0.5}\Phi^{-1}(1-p_2))/\sqrt{1.5} = 2.147 < 2.319$ gegeben ist. Abweichend von der anfänglich geplanten Vorgehensweise werde jetzt adaptiv eine Erhöhung des Stichprobenumfangs der dritten Sequenz beschlossen (beispielsweise eine Verdoppelung, d.h. ebenso viele Beobachtungen wie in der ersten Sequenz). Die dritte Sequenz ergebe einen p-Wert $p_3 = 0.1577$, was wegen $(\Phi^{-1}(1-p_1)+\sqrt{0.5}\Phi^{-1}(1-p_2)+\sqrt{0.5}\Phi^{-1}(1-p_3))/\sqrt{2} = 2.361 > 2.319$ zum Abbruch der Studie mit der Ablehnung von H_0 führt.

3.2.3 Adaptive Verfahren ohne Vorgabe der Stufenanzahl

Aktuell schlagen Müller und Schäfer (2001) eine Methode vor, die die Vorteile der klassischen gruppensequentiellen und der adaptiven Verfahren auf eine alternative Art und Weise miteinander verbinden. Dies führt zu einem Ansatz, in dem im Verlauf einer Studie von den Studienergebnissen abhängig sogar entschieden werden kann, ob und wieviel weitere Zwischenauswertungen durchgeführt werden sollen. Damit kann auch die Wahl der Entscheidungsbereiche neu gewählt werden. Dieser Vorschlag wird in Müller und Schäfer (2000) im Rahmen des *conditional error function* Ansatzes weiter verallgemeinert. Die rekursive Anwendung eines spezifischen zweistufigen Kombinationstests wird in Brannath, Posch und Bauer (1999) vorgeschlagen. Auch in diesem Ansatz ist die Anzahl der Zwischenauswertungen flexibel und die Wahl der Entscheidungsbereiche kann datenabhängig adaptiv geschehen.

4 Zusammenfassung

In dieser Arbeit werden statistische Testverfahren für die gruppensequentielle und adaptive Durchführung von klinischen Studien vorgestellt. Diese Verfahren stellen ein attraktives Instrument auch für die Planungsphase einer Untersuchung dar, mit der in den meisten Fällen eine Verringerung des benötigten Stichprobenumfangs und eine möglichst effiziente Nutzung der in einer Stichprobe enthaltenen Information erreicht werden kann. Dies hat sehr weitreichende ökonomische, ethische und auch organisatorische Konsequenzen, und eine Verwendung solcher sequentieller Pläne erscheint in vielen Fällen als vorteilhaft. Die Arbeit enthält neben der Beschreibung der eigenen, in jüngster Zeit publizierten Arbeiten des Autors und verschiedenen, bisher nicht veröffentlichten neuen Entwicklungen auch eine kritische Übersicht über die in der Literatur vorgeschlagenen Verfahren und eine Einbettung der neuen Verfahren in einen allgemeinen Kontext. Dabei wird versucht, die wesentlichen Charakteristika der vorgeschlagenen Pläne herauszuarbeiten. Durch die Bereitstellung von SAS-Programmen für die Berechnung der Entscheidungsbereiche und wesentlicher Kenngrößen der Tests (ASN und $power$) wird die Anwendung der Verfahren mit einem im universitären Bereich weit verbreiteten Statistik-Programmpaket ermöglicht.

Die klassischen gruppensequentiellen Pläne beruhen auf der Annahme, daß nach einer bestimmten, jeweils gleich großen Anzahl von Beobachtungen eine Zwischenauswertung durchgeführt wird. Es existieren relativ umfassende Kenntnisse über die Wahl der entsprechenden Entscheidungsbereiche. Die verschiedenen Verfahren werden vorgestellt. Für modernere Ansätze, die die Möglichkeit des Abbruchs einer Studie mit der Annahme der Nullhypothese explizit berücksichtigen, werden optimale Lösungen ermittelt und diskutiert. Nicht nur aus organisatorischen Gründen ist es häufig zweckmäßiger, Verfahren zu verwenden, die die Voraussetzung konstanter Sequenzgrößen nicht benötigen. Als prominentesten Vertreter dieser Klasse von Verfahren ist der *use function*-Ansatz zu sehen. Durch die Vorgabe einer *use function* ist man sogar frei in der Anzahl der durch-

zuführenden Zwischenauswertungen. Allerdings muß für die Planung einer Studie mit diesem Verfahren der maximale Stichprobenumfang festgelegt werden. Obwohl diese Forderung für eine korrekte Planung erfüllt sein sollte (und auch wünschenswert ist), ist de facto ein solches Vorgehen für die klinische Praxis eher restriktiv. Eine Möglichkeit, sich von dieser Forderung zu befreien, besteht in der Verwendung des vom Autor dieser Arbeit vorgeschlagenen *worst case scenario*-Ansatzes. In der Arbeit wird dargestellt, daß dieser Ansatz bzgl. der etablierten Gütekriterien (ASN und $power$) sehr vergleichbare Eigenschaften besitzt.

Ein Nachteil dieser Verfahren besteht in der Tatsache, daß der Stichprobenumfang der folgenden Sequenz des gruppensequentiellen Plans nicht datenabhängig ermittelt werden darf. Diese Möglichkeit bieten adaptive Pläne, die im zweiten Teil der Arbeit behandelt werden. Hier werden neue Verfahren entwickelt und untersucht, die für die datengesteuerte Planung einer Studie geeignet sind. Adaptive Pläne sind als eine Erweiterung von gruppensequentiellen Plänen zu betrachten, in denen die zu einem spezifischen Zeitpunkt verfügbare Information der Stichprobe in die weitere Planung der Studie mit einbezogen werden kann. War man früher auf die Durchführung von Pilotstudien zur Schätzung des nachweisbaren Effekts angewiesen, so können heute in einer adaptiv gruppensequentiell geplanten Studie die beobachteten Effekte effizient für die weitere Planung verwendet werden. Die ursprünglich im wesentlichen für den Fall von nur einer einzigen Zwischenauswertung konzipierten Pläne werden vorgestellt und einige neue Eigenschaften beschrieben. Es zeigt sich, daß sich adaptive Pläne ergeben

- in einer Klasse von durch p-Wert-Kombinationstests definierten Verfahren, oder
- in einer Klasse von Funktionen, die die bedingte Fehlerrate definiert.

Dieses Ergebnis ermöglicht die Angabe einer Vielzahl von Verfahren, die die adaptive Planung und Durchführung von klinischen Studien ermöglichen. Die Erweiterung auf mehrstufige Designs ist offensichtlich. Die Lösungen werden für zwei Typen von Kombinationstests (bzw. Funktionsklassen) beschrieben und für die praktische Anwendung zur Verfügung gestellt.

Literaturverzeichnis

ARMITAGE, P. (1975). *Sequential Medical Trials*. Blackwell, Oxford, 2. Auflage.

ARMITAGE, P. (1991). Interim analysis in clinical trials. *Statistics in Medicine* **10**: 925–937.

ARMITAGE, P., MCPHERSON, C. K., ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Stat. Soc. A* **132**: 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 J.* **38**: 25–37.

BAUER, M., BAUER, P., BUDDE, M. (1998). A simulation program for adaptive two-stage designs. *Computational Statistics & Data Analysis* **26**: 351–371.

BAUER, P. (1989a). Multistage testing with adaptive designs. *Biom. und Inform. in Med. und Biol.* **20**: 130–148.

BAUER, P. (1989b). A sequential elimination procedure for choosing the best population(s) based on multiple testing. *J. Stat. Plan. Inf.* **21**: 245–252.

BAUER, P. (1989c). Sequential tests of hypotheses in consecutive trials. *Biometrical J.* **31**: 663–676.

BAUER, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine* **10**: 871–890.

BAUER, P., BRANNATH, W., POSCH, M. (2001). Flexible two-stage designs. *Methods of Information in Medicine* to appear.

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

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

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

BAUER, P., SCHEIBER, V., WOHLZOGEN, F. X. (1986). *Sequentielle statistische Verfahren*. Gustav Fischer Verlag, Stuttgart.

BIRKETT, M. A., DAY, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**: 2455–2463.

BRANNATH, W., POSCH, M., BAUER, P. (1999). Recursive combination tests. Submitted.

BRITTAIN, E. H., BAILEY, K. R. (1993). Optimization of multistage testing times and critical values in clinical trials. *Biometrics* **49**: 763–772.

BRONSTEIN, I. N., SEMENDJAJEW, K. A. (1987). *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun und Frankfurt/Main, 23. Auflage.

CHANG, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**: 247–254.

CHANG, M. N., O'BRIEN, P. C. (1986). Confidence intervals following group sequential tests. *Contr. Clin. Trials* **7**: 18–26.

CHUANG, C. S., LAI, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**: 317–332.

COAD, D. S., WOODROOFE, M. B. (1996). Corrected confidence intervals after sequential testing with applications to survival analysis. *Biometrika* **83**: 763–777.

COBURGER, S., WASSMER, G. (2000). Conditional bias adjusted point estimates in adaptive group sequential test designs. *Biometrical J.* submitted.

COHEN, A., SACKROWITZ, H. B. (1996). Lower confidence bounds using pilot samples with an application to auditing. *J. Amer. Stat. Ass.* **91**: 338–342.

COOK, R. J. (1994). Interim monitoring of bivariate responses using repeated confidence intervals. *Contr. Clin. Trials* **15**: 187–200.

COOK, R. J. (1995). Interim analysis in 2 x 2 crossover trials. *Biometrics* **51**: 932–945.

COOK, R. J. (1996). Coupled error spending functions for parallel bivariate sequential tests. *Biometrics* **52**: 442–450.

CUI, L., HUNG, H. M. J., WANG, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**: 853–857.

DEMETS, D. L., GAIL, M. H. (1985). Use of logrank tests and group sequential methods at fixed calender times. *Biometrics* **41**: 1039–1044.

DEMETS, D. L., LAN, K. K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in Medicine* **13**: 1341–1352.

DEMETS, D. L., WARE, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* **67**: 651–660.

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

DENNE, J. S., JENNISON, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* **18**: 1575–1585.

DUFFY, D. E., SANTNER, T. J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **43**: 81–93.

DURRLEMANN, S., SIMON, R. (1990). Planning and monitoring of equivalence studies. *Biometrics* **46**: 329–336.

EMERSON, S. S. (1993). Computation of the uniform minimum variance unbiased estimator of the normal mean following a group sequential trial. *Computers and Biomedical Research* **26**: 68–73.

EMERSON, S. S. (1996). Statistical packages for group sequential methods. *The American Statistician* **50**: 183–192.

EMERSON, S. S., FLEMING, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45**: 905–923.

EMERSON, S. S., FLEMING, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**: 875–892.

EMERSON, S. S., KITTELSON, J. M. (1997). A computationally simpler algorithm for the UMVUE of a normal mean following a sequential trial. *Biometrics* **53**: 365–369.

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

FISHER, R. A. (1932). *Statistical Methods for Research Workers*. Oliver & Boyd, London, 4. Auflage.

FLEMING, T. R., HARRINGTON, D. P., O'BRIEN, P. C. (1984). Designs for group sequential trials. *Contr. Clin. Trials* **5**: 348–361.

FOLLMANN, D. A., PROSCHAN, M. A., GELLER, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**: 325–336.

FRICK, H. (2000). On confidence bounds for the Bauer-Köhne two-stage test. Submitted.

FRIEDE, T., KIESER, M. (2000a). A loss function based approach for dose selection in two-stage dose-response trials. *J. Epidemiology and Biostatistics, submitted* .

FRIEDE, T., KIESER, M. (2000b). On the inappropriatenes of an EM algorithm based procedure for blinded sample size reestimation. Submitted.

FRIEDE, T., KIESER, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine* **20**: to appear.

FRIEDE, T., MILLER, F., BISCHOFF, W., KIESER, M. (2001). A note on change point estimation in dose-response trials. *Computational Statistics & Data Analysis* **29**: in press.

GELLER, N. L. (1994). Discussion of 'Interim analysis: the alpha spending approach'. *Statistics in Medicine* **13**: 1353–1356.

GELLER, N. L., PROSCHAN, M. A., FOLLMANN, D. A. (1995). Group sequential monitoring of multi-armed clinical trials. *Drug Inf. J.* **29**: 705–713.

GHOSH, B. K. (1970). *Sequential Tests of Statistical Hypotheses*. Addison-Wesley, Reading, Massachusetts.

GOULD, A. L. (1992). Interim analysis for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine* **11**: 53–66.

GOULD, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine* **14**: 1039–1051.

GOULD, A. L. (1997). Issues in blinded sample size re-estimation. *Comm. Stat. - Sim. Comp.* **26**: 1229–1239.

GOULD, A. L., PECORE, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69**: 75–80.

GOULD, A. L., SHIH, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Comm. Stat. - Theory Meth.* **21**: 2833–2853.

GOULD, A. L., SHIH, W. J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* **17**: 89–100.

GU, M., FOLLMANN, D., GELLER, N. (1999). Monitoring a general class of two-sample survival statistics with applications. *Biometrika* **86**: 45–57.

HAYRE, L. S. (1985). Group sequential sampling with variable group sizes. *J. R. Statist. Soc. B.* **47**: 90–97.

HEDGES, L. V., OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Akademic Press.

HELLMICH, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics* to appear.

HOMMEL, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical J.* **43**: to appear.

HUGHES, M. (1993). Stopping guidelines for clinical trials with multiple treatments. *Statistics in Medicine* **12**: 901–915.

HWANG, I. K., SHIH, W. J., DECANI, J. S. (1990). Group sequential designs using a family of Type I error probability spending functions. *Statistics in Medicine* **9**: 1439–1445.

ICH (1998). *Note for Guidance on Statistical Principles for Clinical Trials*. (ICH Topic E9, Step 4, CPMP/ICH/363/96). ICH - Technical Coordination, European Agency for the Evaluation of Medicinal Products, London.

JENNISON, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74**: 155–165.

JENNISON, C., TURNBULL, B. W. (1989). Interim analysis: the repeated confidence interval approach. *J. R. Statist. Soc. B* **51**: 305–361.

JENNISON, C., TURNBULL, B. W. (1991a). Exact calculations for sequential t, chi-square and F tests. *Biometrika* **78**: 133–141.

JENNISON, C., TURNBULL, B. W. (1991b). Group sequential tests and repeated confidence intervals. In B. K. Ghosh, P. K. Sen, Hg., *Handbook of Sequential Analysis*, S. 283–311. Marcel Dekker, New York.

JENNISON, C., TURNBULL, B. W. (1993a). Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**: 741–752.

JENNISON, C., TURNBULL, B. W. (1993b). Sequential equivalence testing and repeated confidence intervals, with application to normal and binary responses. *Biometrics* **49**: 31–43.

JENNISON, C., TURNBULL, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, Boca Raton, London, New York, Washington, D.C.

JOHNSON, N. L., KOTZ, S. (1970). *Continous Univariate Distributions - 1*. Wiley, New York.

KIESER, M., BAUER, P., LEHMACHER, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical J.* **41**: 261–277.

KIESER, M., FRIEDE, T. (2000a). Blinded sample size reestimation in multiarmed clinical trials. *Drug. Inf. J.* **34**: 455–460.

KIESER, M., FRIEDE, T. (2000b). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**: 901–911.

KIESER, M., LEHMACHER, W. (1995). Multiples Testen bei klinischen Prüfungen mit Zwischenauswertungen und a-priori geordneten Hypothesen. *Proceedings der 40. Jahrestagung der GMDS, München MMV Medizin Verlag* S. 162–265.

KIESER, M., WASSMER, G. (1997). On the use of the upper confidence limit on the variance from a pilot sample for sample size determination. *Biometrical J.* **38**: 941–949.

KIM, K. (1989). Point estimation following group sequential tests. *Biometrics* **45**: 613–617.

KIM, K., BOUCHER, H., TSIATIS, A. A. (1995). Design and analysis of group sequential logrank tests in maximum duration versus information trials. *Biometrics* **51**: 988–1000.

KIM, K., DEMETS, D. L. (1987a). Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43**: 857–864.

KIM, K., DEMETS, D. L. (1987b). Design and analysis of group sequential tests based on the Type I error spending rate function. *Biometrika* **74**: 149–154.

KIM, K., TSIATIS, A. A. (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics* **46**: 81–92.

KITTELSON, J. M., EMERSON, S. S. (1999). A unifying family of group sequential test

designs. *Biometrics* **55**: 874–882.

KÖPCKE, W. (1984). *Zwischenauswertungen und vorzeitiger Abbruch von Therapiestudien.* Springer, Berlin, Heidelberg, New York, Tokyo.

KÖPCKE, W. (1989). Analyses of group sequential clinical trials. *Contr. Clin. Trials* **10**: 222–230S.

KROPF, S., HOMMEL, G., SCHMIDT, U., BRICKWEDEL, J., JEPSEN, M. S. (2000). Multiple comparison of treatments with stable multivariate tests in a two-stage adaptive design, including a test for non-inferiority. *Biometrical J.* **42**: 951–965.

LACHIN, J. M. (1997). Group sequential monitoring of distribution-free analyses of repeated measures. *Statistics in Medicine* **16**: 653–668.

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

LAN, K. K. G., DEMETS, D. L. (1989). Group sequential procedures: Calender versus information time. *Statistics in Medicine* **8**: 1191–1198.

LAN, K. K. G., LACHIN, J. M. (1990). Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* **46**: 759–770.

LAN, K. K. G., REBOUSSIN, D. M., DEMETS, D. L. (1994). Information and information fractions for design and sequential monitoring of clinical trials. *Comm. Stat. - Theory Meth.* **23**: 403–420.

LAN, K. K. G., ZUCKER, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine* **12**: 753–765.

LANG, T., AUTERITH, A., BAUER, P. (2000). Trendtests with adaptive scoring. *Biometrical J.* **42**: 1007–1020.

LEE, J. W. (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Statistics in Medicine* **13**: 101–113.

LEE, S. J., KIM, K., TSIATIS, A. A. (1996). Repeated significance testing in longitudinal clinical trials. *Biometrika* **83**: 779–789.

LEHMACHER, W. (1997). Residual effects in crossover trials. In J. Vollmar, Hg., *Biometrics in the Pharmaceutical Industry*, Band 7, S. 41–65. Fischer, Stuttgart.

LEHMACHER, W., KIESER, M., HOTHORN, L. (2000). Sequential and multiple testing for dose-response analysis. *Drug Inf. J.* **34**: 591–597.

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

LI, Z. (1999). A group sequential test for survival trials: An alternative to rank-based procedures. *Biometrics* **55**: 277–283.

LI, Z., GELLER, N. L. (1991). On the choice of times for data analysis in group sequential trials. *Biometrics* **47**: 745–750.

LIN, D. Y., SHEN, L., YING, Z., BRESLOW, N. E. (1996). Group sequential designs for monitoring survival probabilities. *Biometrics* **52**: 1033–1041.

LIN, D. Y., WEI, L. J., DEMETS, D. L. (1991). Exact statistical inference for group sequential trials. *Biometrics* **47**: 1399–1408.

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

LIU, W. (1995). A group sequential procedure for all-pairwise comparisons of k treatments based on range statistics. *Biometrics* **51**: 946–955.

LUI, K. J. (1993). A simple generalization of the O'Brien and Fleming group sequential test procedure to more than two treatment groups. *Biometrics* **49**: 1216–1219.

LUI, K.-J. (1994). A group sequential method for one standard control and more than one experimental treatment. *Biometrical J.* **36**: 515–529.

MCPHERSON, C. K., ARMITAGE, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J. Roy. Stat. Soc. A* **134**: 15–25.

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

MCPHERSON, K. (1990). Sequential stopping rules in clinical trials. *Statistics in Medicine* **9**: 595–600.

MÜLLER, H.-H., SCHÄFER, H. (1999). Optimization of testing times and critical values in sequential equivalence testing. *Statistics in Medicine* **18**: 1769–1788.

MÜLLER, H.-H., SCHÄFER, H. (2000). Monitoring clinical trials using a general statistical principle for design changes. Submitted.

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**: in press.

NEYMAN, J., PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A**: 263–295.

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

OLSCHEWSKI, M., SCHUMACHER, M. (1986). Sequential analysis of survival times in clinical trials. *Biometrical J.* **28**: 273–293.

PAMPALLONA, S., TSIATIS, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *J. Stat. Plan. Inf.* **42**: 19–35.

PFANZAGL, J. (1991). *Elementare Wahrscheinlichkeitsrechnung*. de Gruyter, Berlin, New York.

PINHEIRO, J. C., DEMETS, D. L. (1997). Estimating and reducing bias in group sequen-

tial designs with Gaussian independent increment structure. *Biometrika* **84**: 831–845.

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

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

POCOCK, S. J. (1996). The role of external evidence in data monitoring of a clinical trial. *Statistics in Medicine* **15**: 1285–1293.

POSCH, M., BAUER, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical J.* **41**: 689–696.

POSCH, M., BAUER, P. (2000). Interim analysis and sample size assessment. *Biometrics* **56**: 1170–1176.

PROSCHAN, M. A. (1999a). A multiple comparison procedure for three- and four-armed controlled clinical trials. *Statistics in Medicine* **18**: 787–798.

PROSCHAN, M. A. (1999b). Properties of spending function boundaries. *Biometrika* **86**: 466–473.

PROSCHAN, M. A., FOLLMANN, D. A., GELLER, N. L. (1994). Monitoring multi-armed trials. *Statistics in Medicine* **13**: 1441–1452.

PROSCHAN, M. A., FOLLMANN, D. A., WACLAWIW, M. A. (1992). Effects on assumption violations on type I error rate in group sequential monitoring. *Biometrics* **48**: 1131–1143.

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

PROSCHAN, M. A., WITTES, J. (2000). An improved double sampling procedure based on the variance. *Biometrics* **56**: 1183–1187.

REBOUSSIN, D. M., DEMETS, D. L., KIM, K., LAN, K. K. G. (1995). Programs for computing group sequential bounds using the Lan-DeMets method. *Manuskript* .

ROSNER, G. L., TSIATIS, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* **75**: 723–729.

SAS INSTITUTE INC. (1989). *SAS/IML Software: Usage and Reference, Version 6, First Edition*. Cary, NC: SAS Institute Inc.

SAS INSTITUTE INC. (1995). *SAS/IML Software: Changes and Enhancements through Release 6. 11*. Cary, NC: SAS Institute Inc.

SCHARFSTEIN, D. O., TSIATIS, A. A. (1998). The use of simulation and bootstrap in information-based group sequential studies. *Statistics in Medicine* **17**: 75–87.

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

ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal

distributions. *J. Amer. Stat. Ass.* **62**: 626–633.

SIEGMUND, D. (1985). *Sequential Analysis*. Springer, New York, Berlin, Heidelberg, Tokyo.

SKOVLUND, E., WALLØE, L. (1989). Estimation of treatment difference following a sequential clinical trial. *J. Amer. Stat Ass.* **84**: 823–828.

SLEPIAN, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell. System Technical Journal* **41**: 463–501.

SLUD, E., WEI, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Stat. Ass.* **77**: 862–868.

SONNEMANN, E. (1991). Kombination unabhängiger Tests. *In:Vollmar,J. (ed.): Biometrie in der chemisch-pharmazeutischen Industrie* **4**: 91–112.

STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16**: 243–258.

SU, J. Q., LACHIN, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* **48**: 1033–1042.

TANG, D.-I., GELLER, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**: 1188–1192.

TANG, D.-I., GNECCO, C., GELLER, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *J. Amer. Stat. Ass.* **84**: 776–779.

TODD, S., WHITEHEAD, J. (1997). Confidence interval calculation for a sequential clinical trial of binary responses. *Biometrika* **84**: 737–743.

TODD, S., WHITEHEAD, J., FACEY, K. M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika* **83**: 453–461.

TSIATIS, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Amer. Stat. Ass.* **77**: 855–861.

TSIATIS, A. A., BOUCHER, H., KIM, K. (1995). Sequential methods for parametric survival models. *Biometrika* **82**: 165–173.

TSIATIS, A. A., ROSNER, G. L., MEHTA, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**: 797–803.

TSIATIS, A. A., ROSNER, G. L., TRITCHLER, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* S. 365–373.

TURNBULL, B. W. (1997). Group sequential tests. *Encyclopedia of Statistical Sciences (Kotz, Read, Banks eds)* .

WALD, A. (1947). *Sequential Analysis*. Wiley, New York.

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

WASSMER, G. (1997). A technical note on the power determination for Fisher's combi-

nation test. *Biometrical J.* **39**: 831–838.

WASSMER, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **54**: 696–705.

WASSMER, G. (1999a). Group sequential monitoring with arbitrary inspection times. *Biometrical J.* **41**: 197–216.

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

WASSMER, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers* **41**: 253–279.

WASSMER, G., BILLER, C. (1998). Einführung in SAS/IML. In: SAS Anwenderhandbuch im Netz. http://www.urz.uni-heidelberg.de/statistik/sas-ah/.

WASSMER, G., BOCK, W. (1999). Tables of Δ-class boundaries for group sequential trials. *Inf., Biom. und Epid. in Med. und Biol.* **30**: 190–194.

WASSMER, G., EISEBITT, R. (2001). ADDPLAN 2001: Adaptive Designs - Plans and Analyses.

WASSMER, G., EISEBITT, R., COBURGER, S. (2001). Flexible interim analyses in clinical trials using multistage adaptive test designs. *Drug Inf. J.* **35**: to appear.

WASSMER, G., LEHMACHER, W. (1997). On the determination of one-sided confidence limits in adaptive interim analysis. *Proceedings der 42. Jahrestagung der GMDS, München MMV Medizin* S. 340–344.

WETHERILL, G. B. (1975). *Sequential Methods in Statistics.* Chapman and Hall, London.

WHITEHEAD, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**: 573–581.

WHITEHEAD, J. (1996). Sequential designs for equivalence studies. *Statistics in Medicine* **15**: 2703–2715.

WHITEHEAD, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley, New York, revised 2^{nd} Auflage.

WITTES, J. T., SCHABENBERGER, O., ZUCKER, D. M., BRITTAIN, E., PROSCHAN, M. (1999). Internal pilot studies I: Type I error rate of the naive t-test. *Statistics in Medicine* **18**: 3481–3491.

WITTING, H. (1985). *Mathematische Statistik*. Teubner, Stuttgart.

ZUCKER, D. M., WITTES, J. T., SCHABENBERGER, O., BRITTAIN, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine* **18**: 3493–3509.

A SAS/IML-Implementierung gruppensequentieller Pläne

Die Subroutine `CALL SEQ(prob,domain<,options>)` in SAS/IML berechnet die Matrix `prob` der Fortsetzungswahrscheinlichkeiten eines gruppensequentiellen Tests mit den in `domain` spezifizierten Fortsetzungsbereichen $\breve{C}_k$ für S_k (vgl. (2.14) in Abschnitt 2.1). Dabei kann auch `.M` und `.P` gesetzt werden, was $-\infty$ bzw. ∞ repräsentiert. Durch die Option `tscale` (bzw. durch Definition des IML $keywords$ `TSCALE`) wird der Vektor $(\tau_2, \tau_3, \ldots, \tau_K)$ spezifiziert, der standardmäßig auf $(1, 1, \ldots, 1)$ gesetzt ist (vgl. Abschnitt 2.1).

Beispiel
Die in Armitage et al. (1969) berechneten Wahrscheinlichkeiten für den Fehler erster Art bei Verwendung der $(1 - \alpha/2)$-Fraktile der Standardnormalverteilung in einem K-stufigen zweiseitigen gruppensequentiellen Design mit gleichen Gruppengrößen lassen sich mit dem SEQ-Modul wie folgt bestimmen:

```
PROC IML;
K=10;
alpha=0.05;
crit=PROBIT(1-alpha/2);
mm=-crit*SQRT(1:K)//crit*SQRT(1:K);
CALL SEQ(prob,mm);
size=1-(prob[2,]-prob[1,])[K];
PRINT K size;
QUIT;
```

Dies ergibt

```
      K        SIZE
      10       0.1933566
```

Die Wahrscheinlichkeit in einem 10-stufigen Plan, die Nullhypothese mindestens einmal abzulehnen, ist damit fast 20%, wenn eine Ablehnung nach jeder Stufe erfolgen kann. Dabei ist `size` das durch (2.17) angegebene Niveau des Testverfahrens bei Verwendung der durch `mm` spezifizierten Fortsetzungsbereiche. Entsprechend adjustierte kritische Werte sind durch einen geeigneten Suchalgorithmus zu finden, aber auch bei Angabe von `level` $= 1 - \alpha$ mit Hilfe des Moduls `SEQSCALE(prob,gscale,domain,level)`. Der Nichtzentralitätsparameter ϑ^*= `theta` (vgl. (2.22)) läßt sich durch

```
CALL SEQSHIFT(prob,theta,domain,beta);
```

bei gegebener $power$ `1-beta` bestimmen.

Die folgenden Beispiele dokumentieren die Programme, die zu einem Teil der in dieser Arbeit angegebenen Ergebnissen führen. In den in A.2 und A.5 beschriebenen Programmen sind zudem nichtlineare Optimierungsroutinen implementiert, die in SAS Institute Inc. (1995) ausführlich dokumentiert sind.

A.1 Kritische Werte nach Wang und Tsiatis

Das folgende Programm berechnet die Abbruchschranken und die dazugehörenden Grenzen für die p-Werte in der Δ-Klasse von Fortsetzungsbereichen nach Wang und Tsiatis (1987) im zweiseitigen Testproblem. `crit`=$c(K, \alpha, \Delta)$ ist dabei der in Tabelle 2.2 angegebene kritische Wert des Testverfahrens.

```
PROC IML;
delta=0;    ***    O'Brien und Fleming (1979)  ***;
delta=0.5;  ***    Pocock (1977)               ***;
alpha=0.05;
K=4;
mm =-(1:K)##delta//(1:K)##delta;
CALL SEQSCALE(prob,crit,mm,1-alpha);
bounds=crit*(1:K)##(delta-0.5);
alphas=2*(1-PROBNORM(bounds));
PRINT crit[FORMAT=6.4],bounds[FORMAT=6.4],alphas[FORMAT=7.5];
QUIT;
```

A.2 Schranken mit optimalem *ASN*

Bei gegebener $power\ 1-\beta$ berechnet das folgende Programm den Plan, der den ASN unter H_1 minimiert. Dabei sind beliebige Grenzen (ASN2) bzw. Grenzen innerhalb der Δ-Klasse (ASN1) zugelassen. Der benötigte Stichprobenumfang n (n1) für den approximativ optimalen Plan innerhalb der Δ-Klasse ist für $\frac{\mu_1-\mu_2}{\sigma}=1$ angegeben. Für $\frac{\mu_1-\mu_2}{\sigma}\neq 1$ muß n mit $\left(\frac{\sigma}{\mu_1-\mu_2}\right)^2$ multipliziert werden (vgl. 2.23)) Das Programm reproduziert die Ergebnisse von Tabelle 2.3 und gibt die Ablehnschranken und die Grenzen für die p-Werte an.

```
PROC IML;
***  ASN unter H1 in der Delta-Klasse   ***;
START ASN1(delta) GLOBAL(K,alpha,beta,scale,asn1,mm,n1);
  mm=(-(1:K)##delta)//((1:K)##delta);
  level=1-alpha;
  CALL SEQSCALE(prob,scale,mm,level);
  mm=scale*mm;
  CALL SEQSHIFT(prob,theta,mm,beta);
  p=prob[3,]-prob[2,]+prob[1,];
  n1=2*theta**2;
  asn1=n1*(K-p*T(K-1:0));
  RETURN(asn1);
FINISH ASN1;
***  ASN unter H1 fuer beliebige Bereiche   ***;
START ASN2(geom) GLOBAL(K,alpha,beta,scale,asn2,mm,n2);
  mm=(-abs(geom)#(1:K)##0.5 // abs(geom)#(1:K)##0.5);
  level=1-alpha;
  CALL SEQSCALE(prob,scale,mm,level);
  mm=scale*mm;
  CALL SEQSHIFT(prob,theta,mm,beta);
  p=prob[3,]-prob[2,]+prob[1,];
  n2=2*theta**2;
  asn2=n2*(K-p*T(K-1:0));
  RETURN(asn2);
FINISH ASN2;
alpha=0.05;
beta=0.2;
K=3;
```

```
***  Approximativ optimaler ASN nach Wang und Tsiatis (1987) ***;
start=0;
opt={0 0};
CALL NLPDD(calls,deltamin,"ASN1",start) OPT=opt;
bounds=scale*(1:K)##(deltamin-0.5);
alphas=2*(1-PROBNORM(bounds));
***     Optimaler ASN nach Pocock (1982)  ***;
start=mm[2,];
CALL NLPDD(calls,mm_min,"ASN2",start) OPT=opt;
rel=asn2/asn1*100;
PRINT alpha beta K deltamin[FORMAT=5.3] n1[FORMAT=5.2],
      asn1[FORMAT=7.3] asn2[FORMAT=7.3] rel[FORMAT=6.2],
      bounds[FORMAT=6.4],
      alphas[FORMAT=7.5];
QUIT;
```

A.3 Der Ansatz nach Pampallona und Tsiatis

Die beiden folgenden Programme reproduzieren die Ergebnisse von Pampallona und Tsiatis (1994) und geben die Abbruchschranken sowie den ASN an. Insbesondere beim zweiseitigen Fall ist die Rechenzeit relativ hoch, was durch die geforderte Genauigkeit (`eps=`) gesteuert werden kann. Ein PENTIUM 133MHz benötigt bei `eps=`10^{-7} für die Berechnung der kritischen Werte für den zweiseitigen Fall bei K=3 ca. eine Minute, bei K=4 ca. 2:30 Minuten, bei K=5 ca. 3:40 Minuten und bei K=10 ca. 15 Minuten. Dies sollte für praktische Anwendungen kein Problem darstellen.

Zweiseitiger Fall

```
PROC IML;
eps=1E—7; ***  Genauigkeit der kritischen Werte  ***;
alpha=0.01;
beta=0.20;
delta=0;
K=4;
null=REPEAT(0,1,K);
prec2=1;
ko2=6;
ku2=1;
DO UNTIL (prec2<eps);
  c2=(ku2+ko2)/2;
  prec1=1;
  ko1=10;
  ku1=PROBIT(1—alpha/2);
  DO UNTIL (prec1<eps);
      km=(ku1+ko1)/2;
      FREE b0 b1 crit;
      n=2*(km+c2)**2*K**(2*(delta—1));
      delst=SQRT(n/2);
      b0=(delst*(1:K)—c2*(1:K)##delta//null)[<>,];
      b1=km*(1:K)##delta+eps;
      crit=—b1//—b0//b0//b1;
      CALL SEQ(prob,crit);
      size=0;
```

```
      DO i=1 TO K;
         size=size+(prob[5,i]-prob[4,i]+prob[1,i]);
      END;
     IF size<alpha THEN ko1=km;
     ELSE ku1=km;
     prec1=ko1-ku1;
  END;
  c1=km;
  crit1=crit-(delst*(1:K)//delst*(1:K)//delst*(1:K)//
  delst*(1:K));
  CALL SEQ(prob,crit1);
  power=0;
  DO i=1 TO K;
     power=power+(prob[5,i]-prob[4,i]+prob[1,i]);
  END;
  IF power>1-beta THEN ko2=c2;ELSE ku2=c2;
  prec2=ko2-ku2;
END;
***  ASN unter H0   ***;
  CALL SEQ(prob,crit);
  rej=0;
  DO i=1 TO K;
    rej=rej+(prob[5,i]-prob[4,i]+prob[3,i]-prob[2,i]+
    prob[1,i])*(K-i);
  END;
  asn0=2*n*(K-rej);
***  ASN unter H01   ***;
  crit01=crit-(delst/2*(1:K)//delst/2*(1:K)//delst/2*(1:K)//
  delst/2*(1:K));
  CALL SEQ(prob,crit01);
  rej=0;
  DO i=1 TO K;
    rej=rej+(prob[5,i]-prob[4,i]+prob[3,i]-prob[2,i]+
    prob[1,i])*(K-i);
  END;
  asn01=2*n*(K-rej);
****  ASN unter H1   ***;
  crit1=crit-(delst*(1:K)//delst*(1:K)//delst*(1:K)//
  delst*(1:K));
  CALL SEQ(prob,crit1);
```

```
  rej=0;
  DO i=1 TO K;
    rej=rej+(prob[5,i]-prob[4,i]+prob[3,i]-prob[2,i]+
    prob[1,i])*(K-i);
  END;
  asn1=2*n*(K-rej);
DO j=1 TO K;
  crit[1,j]=((j*delst-c2*j**delta // 0)[<>])/SQRT(j);
  crit[2,j]=(c1*j**delta+eps)/SQRT(j);
END;
bounds=-crit[2,]//-crit[1,]//crit[1,]//crit[2,];
PRINT K delta alpha beta c1[FORMAT=6.4] c2[FORMAT=6.4],
      n[FORMAT=5.2] bounds[FORMAT=7.3], asn0[FORMAT=5.2]
      asn01[FORMAT=5.2] asn1[FORMAT=5.2];
QUIT;
```

Einseitiger Fall

```
PROC IML;
eps=1E—7;   ***  Genauigkeit der kritischen Werte  ***;
alpha=0.05;
beta=0.20;
delta=0;
K=4;
prec2=1;
ko2=6;
ku2=1;
DO UNTIL (prec2<eps);
    c2=(ku2+ko2)/2;
    prec1=1;
    ko1=10;
    ku1=PROBIT(1—alpha);
    DO UNTIL (prec1<eps);
        km=(ku1+ko1)/2;
        FREE b0 b1 crit;
        n=2*(km+c2)**2*K**(2*(delta—1));
        delst=SQRT(n/2);
        b0=delst*(1:K)—c2*(1:K)##delta;
        b1=km*(1:K)##delta;
        b0=(b0//b1)[><,]—eps;
        crit=b0//b1;
        CALL SEQ(prob,crit);
        size=0;
        DO i=1 TO K;
           size=size+(prob[3,i]—prob[2,i]);
        END;
       IF size<alpha THEN ko1=km;
       ELSE ku1=km;
       prec1=ko1—ku1;
    END;
    c1=km;
    crit1=crit—(delst*(1:K)//delst*(1:K));
    CALL SEQ(prob,crit1);
    power=0;
    DO i=1 TO K;
       power=power+(prob[3,i]—prob[2,i]);
```

```
    END;
    IF power>1-beta THEN ko2=c2;ELSE ku2=c2;
    prec2=ko2-ku2;
END;
***  ASN unter H0   ***;
     CALL SEQ(prob,crit);
     rej=0;
     DO i=1 TO K;
       rej=rej+(prob[1,i]+prob[3,i]-prob[2,i])*(K-i);
     END;
     asn0=2*n*(K-rej);
***  ASN unter H01   ***;
     crit01=crit-(delst/2*(1:K)//delst/2*(1:K));
     CALL SEQ(prob,crit01);
     rej=0;
     DO i=1 TO K;
       rej=rej+(prob[1,i]+prob[3,i]-prob[2,i])*(K-i);
     END;
     asn01=2*n*(K-rej);
*** ASN unter H1   ***;
     crit1=crit-(delst*(1:K)//delst*(1:K));
     CALL SEQ(prob,crit1);
     rej=0;
     DO i=1 TO K;
       rej=rej+(prob[1,i]+prob[3,i]-prob[2,i])*(K-i);
     END;
     asn1=2*n*(K-rej);
DO j=1 TO K;
  crit[1,j]=(j*delst-c2*j**delta);
  crit[2,j]=(c1*j**delta+eps);
END;
bounds=crit[1,]//crit[2,];
PRINT K delta alpha beta c1[FORMAT=6.4] c2[FORMAT=6.4],
      n[FORMAT=5.2] bounds[FORMAT=7.3], asn0[FORMAT=5.2]
      asn01[FORMAT=5.2] asn1[FORMAT=5.2];
QUIT;
```

A.4 Ungleiche Sequenzgrößen

Die Verwendung des Arguments TSCALE wird in dem folgenden Programmcode illustriert. Das angegebene Modul VKNOWN berechnet bei gegebenem α und Vektor der Analysezeitpunkte $t_1, t_2, \ldots, t_K$ (V=) die Abbruchschranken in der Δ-Klasse (vgl. Tabelle 2.7 in Abschnitt 2.3.1). Darüber hinaus wird bei Vorgabe des standardisierten Effekts und des maximalen Stichprobenumfangs N die $power$ und der ASN des resultierenden Testverfahrens berechnet.

```
PROC IML;
alpha=0.05;
delta=0.5;
N=100;
theta=0.4018;
START VKNOWN(theta,V) GLOBAL(alpha,delta,scale,power,N,asn);
  Vnew=V/V[1];
  K=NCOL(V);
  tscale=REPEAT(1,1,K-1);
  DO i=1 TO K-1;
    tscale[i]=(Vnew[i+1]-Vnew[i]);
  END;
    m=((1:K)##(-0.5+delta))#(Vnew##0.5);
    mm=-m//m;
    CALL SEQSCALE(prob,scale,mm,1-alpha) TSCALE=tscale;
    mm=scale*mm;
  nvec=N*V;
  nz=theta*(nvec/2#Vnew)##0.5;
  mmnz=mm-(nz//nz);
  CALL SEQ(prob,mmnz) TSCALE=tscale;
  power=1-(prob[2,]-prob[1,])[K];
  p=prob[3,]-prob[2,]+prob[1,];
  asn=nvec[K]-(nvec[K]-nvec)*T(p);
  PRINT tscale,
       theta K scale[FORMAT=6.3] power[FORMAT=6.3]
       asn[FORMAT=5.2];
FINISH VKNOWN;
```

```
V=0.2 || 0.4 || 0.6 || 0.8 || 1;
RUN VKNOWN(theta,V);
V=0.4 || 0.6 || 0.8 || 1;
RUN VKNOWN(theta,V);
V=0.8 || 1;
RUN VKNOWN(theta,V);
QUIT;
```

A.5 Implementierung des *worst case scenario*-Ansatzes

Für die Berechnung der kritischen Werte im *worst case scenario*-Ansatz nach Wassmer (1999a) mit oberer bzw. unterer Schranken für die Sequenzgrößen wird das SEQ-Modul zusammen mit der $conjugate\ gradient$ Optimierungsroutine NLPCG benutzt. Als untere Schranke für das O'Brien und Fleming-Verfahren wird 0.10 benutzt. Dies ergibt die in Abschnitt 2.3.2 angegebenen Ergebnisse (vgl. Tabelle 2.9 und 2.10). Darüber hinaus wird bei Vorgabe des Vektors der Analysezeitpunkte $t_1, t_2, \ldots, t_K$ (`V=`), des standardisierten Effekts (`theta=`) und des maximalen Stichprobenumfangs N die $power$ und der ASN des resultierenden Testverfahrens berechnet.

```
PROC IML;
***   Niveau eines Testverfahrens bei spezifierten tau  ***;
START SIZE(tau) GLOBAL(delta,z_alpha);
  K=NCOL(tau)+1;
  Vnew=REPEAT(1,1,K);
  DO i=2 TO K;
    Vnew[i]=Vnew[i-1]+tau[i-1];
  END;
  m=z_alpha*((1:K)##(-0.5+delta))#(Vnew##0.5);
  mm=-m//m;
  tscale=tau;
  CALL SEQ(prob,mm) TSCALE=tscale;
  size=1-(prob[2,]-prob[1,])[K];
  RETURN(size);
FINISH SIZE;
*** Suche des kritischen Wertes durch lineare Optimierung ***;
START CRITVAL(K) GLOBAL(alpha,delta,z_alpha,scale,restrict);
  mm=(-(1:K)##delta)//((1:K)##delta);
  CALL SEQSCALE(prob,scale,mm,1-alpha);* scale ist Startwert *;
  prec=1;
  ku=scale;
  ko=scale+1;
  DO UNTIL (prec<1E-4);
     z_alpha=(ku+ko)/2;
     start=REPEAT(1,1,K-1);
```

```
    optn={1 0};
    constr=REPEAT(0.1,1,K-1)//REPEAT(restrict,1,K-1);
    CALL NLPCG(CALLs,rmax,"SIZE",start,optn,constr);
    size=SIZE(rmax);
    IF size<alpha THEN ko=z_alpha;
    ELSE ku=z_alpha;
    prec=ko-ku;
  END;
FINISH CRITVAL;
*** Power bei Alternative theta und Analysezeitpunkte V ***;
START POWER(theta,V) GLOBAL(alpha,delta,z_alpha,power,N,asn);
  Vnew=V/V[1];
  K=NCOL(V);
  tscale=REPEAT(1,1,K-1);
  DO i=1 TO K-1;
    tscale[i]=(Vnew[i+1]-Vnew[i]);
  END;
  m=z_alpha*((1:K)##(-0.5+delta))#(Vnew##0.5);
  mm=-m//m;
  nvec=N*V/2;
  nz=theta*(nvec#Vnew)##0.5;
  mmnz=mm-(nz//nz);
  CALL SEQ(prob,mmnz) TSCALE=tscale;
  power=1-(prob[2,]-prob[1,])[K];
  p=prob[3,]-prob[2,]+prob[1,];
  asn=2*(nvec[K]-(nvec[K]-nvec)*T(p));
FINISH POWER;
***    Beispiel      ***;
alpha=0.05;
delta=0.5;
restrict=4;
N=100;
V=0.5 || 0.7 || 1;
theta=0.50;
K=NCOL(V);
RUN CRITVAL(K);
RUN POWER(0,V);
size=power;
```

```
run POWER(theta,V);
PRINT delta restrict z_alpha[FORMAT=4.2] scale[FORMAT=5.3],
      theta N V size[FORMAT=5.3] power[FORMAT=5.3]
      asn[FORMAT=5.1];
QUIT;
```

A.6 Implementierung des α-*spending*-Ansatzes

Der folgende Programmcode imlementiert den α-*spending*-Ansatz für den zweiseitigen Parallelgruppenvergleich mit den α-*spending*-Funktionen $\tilde{\alpha}_1^*$ (delta=0.5) und $\tilde{\alpha}_2^*$ (delta=0). Die kritischen Werte können bei spezifiziertem α und Vektor der Analysezeitpunkte $t_1, t_2, \ldots, t_K$ (V=) berechnet werden. Zusätzlich werden bei gegebener Effektgröße und maximalen Stichprobenumfang N die $power$ und der ASN berechnet. Bei der laufenden Studie werden bei Stufe k die kritischen Werte sukzessive bestimmt, indem $(t_1, t_2, \ldots, t_k)$ anstelle des kompletten Vektors $(t_1, t_2, \ldots, t_K)$ gesetzt wird.

```
PROC IML;
***     Definition der alpha spending Funktion      ***;
START ALPHA(t) GLOBAL(alpha,delta);
  IF t<=1 THEN DO;
    IF delta=0.5 THEN x=alpha*LOG(1+(EXP(1)-1)*t);
    IF delta=0 THEN x=4-4*PROBNORM(PROBIT(1-alpha/4)/SQRT(t));
    RETURN(x);
  END;
  ELSE RETURN(alpha);
FINISH ALPHA;
START FIND(y) GLOBAL(alphavec,K,Vnew,tscale);
  m=REPEAT(.,1,K);
  DO i=1 TO K-1;
    m[i]=PROBIT(1-alphavec[i]/2)*Vnew[i]**0.5;
  END;
  m[K]=PROBIT(1-y/2)*vnew[K]**0.5;
  mm=-m//m;
  CALL SEQ(prob,mm) TSCALE=tscale[1:K-1];
  size=(prob[3,]-prob[2,]+prob[1,])[K];
  RETURN(size);
FINISH FIND;
START DEMETS(theta,V)
  GLOBAL(alpha,alphavec,K,Vnew,tscale,power,N,asn,bounds);
  Vnew=V/V[1];
  Kmax=ncol(V);
  tscale=REPEAT(.,1,Kmax-1);
  DO i=1 TO Kmax-1;
```

```
    tscale[i]=(Vnew[i+1]-Vnew[i]);
  END;
  alphavec=ALPHA(V[1]);
  DO K=2 TO Kmax;
    prec=1;
    ko=alpha;
    ku=1E-7;
      DO UNTIL (prec<1E-7);
        alphaj=(ku+ko)/2;
        sizej=FIND(alphaj);
        IF sizej<(ALPHA(V[K])-ALPHA(V[K-1])) THEN ku=alphaj;
        ELSE ko=alphaj;
        prec=ko-ku;
      END;
    alphavec=alphavec || alphaj;
  END;
  bounds=PROBIT(1-alphavec/2);
***   Power-Berechnung bei spezifiziertem theta   ***;
  m=bounds#(vnew##0.5);
  mm=-m//m;
  nvec=N/2*V;
  nz=theta*(nvec#Vnew)##0.5;
  mmnz=mm-(nz//nz);
  CALL SEQ(prob,mmnz) TSCALE=tscale;
  power=1-(prob[2,]-prob[1,])[Kmax];
  p=prob[3,]-prob[2,]+prob[1,];
  asn=2*(nvec[Kmax]-(nvec[Kmax]-nvec)*T(p));
FINISH DEMETS;
***   Beispiel   ***;
alpha=0.05;
delta=0.5;
N=100;
V=0.1||0.3||0.7||0.9||1;
theta=0.50;
RUN DEMETS(theta,V);
PRINT theta N V,
      bounds[FORMAT=5.3] power[FORMAT=5.3] asn[FORMAT=5.2];
QUIT;
```

A.7 Bestimmung der *power* bei Fishers Produktregel

Mit dem unten beschriebenen Programm läßt sich die $power$ von Fishers Kombinationstest im zweistufigen Design bestimmen. Dies geschieht für das Parallelgruppendesign mit unbekannter Varianz bei einer fest vorgegebenen Alternative (`delta=`). `n1` und `n2` bezeichnen die Stichprobenumfänge der ersten bzw. zweiten Sequenz; `alpha0` spezifiziert die Grenze α_0, und mit `alpha1=, alpha2=` werden die kritischen Werte für den ersten p-Wert bzw. das Produkt der p-Werte definiert.

```
PROC IML;
***   Definition der gemeinsamen Dichtefunktion              ***;
START PDENS(p2) global(p1,nz1,nz2,df1,df2);
 f=pdf("T",tinv(1-p1,df1),df1,nz1)/pdf("T",tinv(1-p1,df1),df1,0)*
  pdf("T",tinv(1-p2,df2),df2,nz2)/pdf("T",tinv(1-p2,df2),df2,0);
 return(f);
FINISH;
***   Definition der Dichte von p1                           ***;
START PDENS1(p1) global(nz1,df1);
 f=pdf("T",tinv(1-p1,df1),df1,nz1)/pdf("T",tinv(1-p1,df1),df1,0);
 return(f);
FINISH;
***   Das erste Integral                                     ***;
START POWER1(crit1) global(eps);
   interval= crit1 || 1;
   CALL QUAD(pow1,"PDENS1",interval) eps=eps;
   return(pow1);
FINISH;
***   Zweidimensionale Integration   ***;
START PDENS2(y2) GLOBAL(p1,c_alpha2,eps);
  p1=y2;
  interval=c_alpha2/p1 || 1;
  CALL QUAD(p,"PDENS",interval) eps=eps;
  return(p);
FINISH;
START POWER2(crit2) global(alpha0,eps);
   interval= crit2 || alpha0;
```

```
   CALL QUAD(pow2,"PDENS2",interval) eps=eps;
   return(pow2);
FINISH;
***          Beispiel                                        ***;
eps=1E-7;  ** Genauigkeit der Integration                    ***;
alpha0=1;  ** kein frueher Abbruch mit der Annahme von H0***;
alpha2=0.05;                          ***   alpha            ***;
alpha1=EXP(-0.5*CINV(1-alpha2,4)); ***   spending          ***;
n1=8;
n2=8;
delta=1.30;
df1=n1-2;
df2=n2-2;
nz1=SQRT(n1)/2*delta;
nz2=SQRT(n2)/2*delta;
c_alpha2=exp(-0.5*CINV(1-alpha2,4));
power=1-power1(alpha0)-power2(alpha1);
PRINT alpha1 alpha2 alpha0 delta n1 n2 power;
QUIT;
```

A.8 Berechnung der kritischen Werte der Prozedur nach Proschan und Hunsberger

Mit Hilfe des Moduls `QUAD` lassen sich die zur Durchführung der Prozedur nach Proschan und Hunsberger (1995) benötigten kritischen Schranken einfach berechnen. Dies geschieht für gegebenes Niveau α (`alpha=`) und Annahmeschranke $\alpha_0 \leq 0.50$ (`alpha0=`).

```
PROC IML;
***    Definition der zu integrierenden Funktion    ***;
START f(x) GLOBAL(k);
  y=(1-PROBNORM(SQRT(k**2-x**2)))*PDF("NORMAL",x);
  RETURN(y);
FINISH f;
eps=1E-09;
alpha0=0.30;
alpha=0.05;
u_alpha0=PROBIT(1-alpha0);
prec=1;
ko=6;
ku=PROBIT(1-alpha);
DO UNTIL (prec<eps);
  k=(ku+ko)/2;
  interval=u_alpha0 || k;
  CALL QUAD(p,"f",interval);
  p=p+1-PROBNORM(k);
  IF p<alpha THEN ko=k;
  ELSE ku=k;
  prec=ko-ku;
END;
PRINT alpha alpha0 k[FORMAT=5.3];
QUIT;
```

A.9 Berechnung der kritischen Werte der mehrstufigen adaptiven Prozedur, basierend auf Fishers Produktregel

Mit dem folgenden Programm werden die lokalen Signifikanzniveaus α_k (`alphavec`) und die kritischen Werte c_{α_k} (`cvec`), $k = 1, 2, \ldots, K$, berechnet, die die Durchführung einer mehrstufigen adaptiven Prozedur, basierend auf Fishers Produktregel, zum Niveau α ermöglichen. Dabei sind die drei in Abschnitt 3.2.1 beschriebenen Methoden implementiert: `method=1` für konstante lokale Signifikanzniveaus α^*, `method=2` für volle Ausschöpfung des Niveaus auf der letzten Stufe und konstante lokale Niveaus $\alpha^\dagger$ sonst, sowie `method=3` für volle Ausschöpfung des Niveaus auf der letzten Stufe und Vermeidung von Interaktionen. Das SAS macro `%functs` erzeugt die SAS/IML-Funktionen `pi1,pi2,...,piK`, wobei die maximale Anzahl K der Stufen durch die macro-Variable `&kk` gegeben ist. Das macro `%size` bestimmt das Niveau des Verfahren bei Benutzung der Vektoren `cvec` und `a_0vec`.

```
OPTION NOMACROGEN NOSYMBOLGEN;
PROC IML SYMSIZE=1000;
      ***   Macro zur Erzeugung der Funktionen fuer pi_k    ***;
%MACRO functs;
%DO kmax=1 %TO &kk;
  START pi&kmax(cvec,a_0vec);** Definition der IML Funktionen**;
    %MACRO integral;
      kmax=&kmax;
      %DO k=1 %TO &kmax;
        alph_0&k=a_0vec[&k];
        c_alph&k=cvec[&k];
      %END;
      pi_K=0;
      %DO k=1 %TO &kmax;
        k=&k;
        %LET km1=%EVAL(&k-1);
        prod1=1;
        %DO i=1 %TO &km1;
          %LET Kmi= %EVAL(&kmax-&i);
          prod1=prod1*LOG(alph_0&Kmi);
```

```
        %END;
        %LET kmkm1=%EVAL(&kmax-&k-1);
        prod2=1;
        %DO i=1 %TO &kmkm1;
          prod2=prod2*alph_0&i;
        %END;
        %LET kmk=%EVAL(&kmax-&k);
        IF kmax>k THEN prod2=1/GAMMA(kmax-k+1)*
        (LOG(prod2/c_alph&kmk))**(kmax-k);
        sum=0;
        %DO i=1 %TO &kmkm1;
          prod3=1;
          %LET jp1=%EVAL(&i+1);
          %DO j=&jp1 %TO &kmkm1;
          prod3=prod3*alph_0&j;
          %END;
          i=&i;
          sum=sum+1/GAMMA(kmax-k+2-i)*(LOG(c_alph&i*prod3/
          c_alph&kmk))**(kmax-k+1-i)*pi&i(cvec,a_0vec)/c_alph&i;
        %END;
        pi_K=pi_K+prod1*(prod2-sum);
      %END;
    %MEND integral;
    %integral;
    pi_k=c_alph&kmax*pi_k;
    RETURN(pi_k);
  FINISH pi&kmax;
%END;
%MEND functs;
      *** Berechnung des Niveaus der Prozedur    ***;
%MACRO size;
  %DO i=1 %TO &kk;
    size=size+pi&i(cvec,a_0vec);
  %END;
%MEND;
***                Beispiel                       ***;
%LET kk=4;            ***   Maximale Anzahl der Stufen    ***;
%FUNCTS;              ***   Generierung der IML-Funktionen **;
method=2;
kmax=&kk;
```

```
alpha=0.05;
alpha_0=0.4;
a_0vec=REPEAT(alpha_0,kmax,1);
eps=1E-10;       *** Finden der lokalen Signifikanzniveaus  **;
prec=1;          *** mit Genauigkeit 1E-10                  **;
ku=0;
ko=alpha;
DO UNTIL (prec<eps);
  alpha_1=(ku+ko)/2;
  cvec=REPEAT(.,kmax,1);
  IF method=1 THEN DO;
    DO i=1 TO kmax;
      cvec[i]=EXP(-0.5*CINV(1-alpha_1,2*i));
    END;
  END;
  IF method=2 THEN DO;
    DO i=1 to kmax-1;
      cvec[i]=EXP(-0.5*CINV(1-alpha_1,2*i));
    END;
    cvec[kmax]=EXP(-0.5*CINV(1-alpha,2*kmax));
  END;
  IF method=3 THEN DO;
    DO i=2 TO kmax;
      cvec[i]=EXP(-0.5*CINV(1-alpha,2*kmax))/alpha_0**(kmax-i);
      cvec[1]=alpha_1;
    END;
  END;
  size=0;
  %size;               ***  Berechnung des Niveaus ***;
  IF size<alpha THEN ku=alpha_1;
  ELSE ko=alpha_1;
  prec=ko-ku;
END;
alphavec=REPEAT(.,kmax);
DO i=1 TO kmax;
    alphavec[i]=1-PROBCHI(-2*LOG(cvec[i]),2*i);
END;
PRINT method alpha_0, alphavec cvec;
QUIT;
```