

# **Powervergabe bei Klinischen Studien mit flexiblen Designs**

Susanne Menzler Tierärztliche Hochschule Inst. für Biometrie, Epidemiologie u. Informationsverarbeitung 30559 Hannover susanne.menzler@tiho-hannover.de	Guido Knapp Universität Dortmund Fachbereich Statistik 44221 Dortmund guido.knapp@uni-dortmund.de
--	---

## **Zusammenfassung**

Als Erweiterung gruppensequentieller Verfahren ermöglicht das Self-Designing-Prinzip in Klinischen Studien bei einer a priori nicht festgelegten Anzahl von Zwischenauswertungen die Anpassung des Stichprobenumfangs unter Einbeziehung der Information aus vorangegangenen Stufen. Das Studienende wird durch die ebenfalls adaptive Vergabe von Gewichten reguliert, die die Varianz einer finalen Teststatistik aufteilen.

Nach Konstruktion wird der vorgegebene Fehler erster Art eingehalten, während der Fehler zweiter Art durch die Wahl der Gewichte und der Stichprobenumfänge auf den einzelnen Stufen bestimmt wird. Anhand des Verfahrens von Hartung und Knapp (2003) werden Strategien zur adaptiven Wahl dieser Designparameter vorgestellt, die in SAS/IML umgesetzt werden.

Bei der Orientierung am Powerverlauf klassischer gruppensequentieller Designs können die Subroutinen SEQ und SEQSCALE verwendet werden.

**Keywords:** Gruppensequentielle Designs, Self-Designing, Power.

## **1 Einleitung**

Gruppensequentielle Verfahren wurden entwickelt, um zum einen neue wirksamere Therapien schneller allen Patienten zugänglich zu machen, zum anderen um Therapierisiken frühzeitig zu erkennen und aussichtslose Studien abzubrechen. Die bekanntesten Designs gehen auf Pocock (1977) und O'Brien und Fleming (1979) zurück, die jeweils Signifikanztests nach der

Beobachtung einer konstanten Anzahl von Personen durchführen, wobei insgesamt eine vorgegebene Irrtumswahrscheinlichkeit eingehalten wird.

Die maximale Anzahl von Stufen (Studienabschnitte, Interimsanalysen, Zwischenbewertungen) und damit der maximale Stichprobenumfang werden dabei im Voraus festgelegt. Diese ersten gruppensequentiellen Pläne legen die Normalverteilung der Response-Variablen mit bekannter Varianz zugrunde.

Neben der Möglichkeit frühe Entscheidungen zu treffen, erlauben adaptive Designs die Anpassung des Studienverlaufs an die bis zur jeweiligen Zwischenbewertung erhaltene Information. Denkbar sind hier unter anderem der Wechsel von einer parametrischen zu einer nicht-parametrischen Teststatistik oder vor allem die Neubestimmung des Stichprobenumfangs aufgrund veränderter Rahmenbedingungen wie größerer Varianz oder kleinerem zu erwartenden Effekt.

Bei diesen Verfahren, die ebenfalls die Anzahl möglicher Interimsanalysen vorgeben, werden meist die P-Werte der Teststatistiken aus den Studienteilen zu einer neuen Teststatistik kombiniert und diese mit entsprechend gewählten kritischen Werten verglichen. Der Kombination von P-Werten z.B. mittels der Inversen Normal Methode von Lehman und Wassmer (1999) oder mit der Fisherschen Kombinationsregel bei Bauer und Köhne (1994) erlaubt die Betrachtung nahezu beliebig verteilter Teststatistiken.

Eine Erweiterung der adaptiven klinischen Versuche bildet der von Fisher (1998) vorgeschlagene Ansatz des „Self-Designing Klinischer Studien“. In dieser Klasse von Studien ist die Gesamtanzahl der durchzuführenden Studienabschnitte a priori nicht festgelegt, sondern nach jedem Studienabschnitt während des Verlaufs der Studie wird adaptiv entschieden, ob nur noch ein weiterer oder mindestens zwei weitere Studienabschnitte durchgeführt werden. Zur adaptiven Entscheidung werden dabei die nicht verblindeten Ergebnisse aller bisher durchgeführten Zwischenbewertungen herangezogen. Die Steuerung des „Self-Designing“ erfolgt durch die adaptive Vergabe von Gewichten zu den einzelnen Studienabschnitten, wobei die Gewichte anteilig an der Varianz einer a priori festgelegten Teststatistik gewählt werden. Neben der adaptiven Gewichtsvergabe können auch die Stichprobenumfänge für die einzelnen Studienabschnitte adaptiv gewählt werden. Wenn die Varianz der finalen Teststatistik durch die Gewichtsvergabe aufgebraucht ist, endet die „Self-Designing“ Klinische Studie nach Fisher mit dem Signifikanztest auf Überlegenheit der neuen Behandlung. Für den Mittelwertvergleich zweier normalverteilter Zufallsvariablen mit bekannter Varianz geben

Shen und Fisher (1999) eine explizite Wahl der sequentiellen Gewichtung an, jedoch verwenden sie eine a priori festgelegte Sequenz von Stichprobenumfängen in ihrem Algorithmus.

Durch Verwendung von P-Werten der sequentiellen Teststatistiken zeigt Hartung (2001), dass die Restriktionen bezüglich der Verteilungsannahme der Zielvariablen aufgehoben werden können. Hartung leitet Algorithmen zum „Self-Designing“ Klinischer Studien her, indem die P-Werte der sequentiellen Teststatistiken mittels gewichteter Inverser Normal Methode kombiniert werden. Eine Testentscheidung ist in diesem Ansatz wie bei Fisher nur nach dem Verbrauch aller Gewichte möglich. Hartung und Knapp (2003) verwenden eine verallgemeinerte gewichtete Inverse Chi-Quadrat Methode zur Kombination der P-Werte der sequentiellen Teststatistiken und konstruieren damit Algorithmen zum „Self-Designing“ Klinischer Studien. Mit diesem Ansatz ist es möglich, die Studie frühzeitig zu Gunsten der Alternativhypothese unter Einhaltung der vorgegebenen Irrtumswahrscheinlichkeit abzubrechen.

In der vorliegenden Arbeit wird der „Self-Designing“-Ansatz von Hartung und Knapp benutzt und verfeinerte Algorithmen zum „Self-Designing“ Klinischer Studien werden vorgestellt.

## **2 Die verallgemeinerte gewichtete Inverse Chi-Quadrat Methode zur Kombination von Ergebnissen**

### **2.1 Grundprinzip**

Ausgangspunkt ist ein einseitiges Testproblem bezüglich eines reellwertigen Parameters  $\theta$ :

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta > 0.$$

Es wird angenommen, dass sich die Studie formal in  $K$  disjunkte Studienteile zerlegen lässt, die im Folgenden mit  $\text{stp}(k)$ ,  $k = 1, \dots, K$ , bezeichnet werden.

Aufgrund der Möglichkeit eines vorzeitigen Abbruchs werden jedoch nicht immer alle  $K$  Studienteile durchgeführt.

Auf jeder Stufe  $\text{stp}(k)$  wird eine geeignete Teststatistik  $T_k$  für das obige Testproblem  $H_0$  vs.  $H_1$  gewählt, deren Wahrscheinlichkeitsverteilung  $F_{k,0}$  unter der Nullhypothese stetig ist. Führen große Werte von  $T_k$  zu einer

Ablehnung der Nullhypothese, so wird in jedem  $\text{stp}(k)$  der P-Wert wie folgt definiert:

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

Aufgrund der stetigen Verteilung der Teststatistiken sind die P-Werte unter der Nullhypothese gleichverteilt auf  $(0, 1)$ . Transformiert man die P-Werte mit der inversen Verteilungsfunktion  $F_{\chi^2(v_k)}^{-1}$  einer  $\chi^2$ -Verteilung mit Freiheitsgraden  $v_k$

$$q_k(v_k) = F_{\chi^2(v_k)}^{-1}(1 - p_k), \quad k = 1, \dots, K,$$

so ist  $q_k(v_k)$  unter der Nullhypothese  $\chi^2$ -verteilt mit  $v_k$  Freiheitsgraden.

Nimmt man die Unabhängigkeit der disjunkten Studienteile an, erhält man bei einer vorgegebenen Folge von  $v_j, j=1, \dots, K$ , für die kombinierte Teststatistik bis zu Studienabschnitt  $\text{stp}(k)$  folgende Verteilungseigenschaft:

$$S_k = \sum_{j=1}^k q_j(v_j) \stackrel{H_0}{\sim} \chi^2(v_\Sigma(k)), \quad v_\Sigma(k) = \sum_{j=1}^k v_j, \quad k = 1, \dots, K.$$

Daher kann in Verbindung mit der Tatsache, dass  $q_j(v_j) \geq 0$  ist, die Nullhypothese zum Niveau  $\alpha$  verworfen werden, sobald gilt:

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

Dabei bezeichnet  $\chi^2(v_\Sigma(K))_{1-\alpha}$  das  $1-\alpha$ -Quantil der  $\chi^2$ -Verteilung mit  $v_\Sigma(K)$  Freiheitsgraden.

## 2.2 Adaptives Vorgehen

Um die Verteilungseigenschaft der finalen Teststatistik und damit das Signifikanzniveau zu erhalten, müssen bei adaptiver Wahl der Gewichte  $v_k$  zu Beginn der Studie die Gesamtfreiheitsgrade  $v_\Sigma(K)$  des gesamten sequentiellen Versuchs festgelegt werden. Dabei ist  $K$  die maximale Anzahl von Studienabschnitten. In der Regel wird eine relativ große Anzahl  $K$  vorgegeben. Basierend auf den Entscheidungsregeln in den Algorithmen wird in der Praxis jedoch nur ein kleiner Teil aller hypothetischen Studienabschnitte durchgeführt. Durch die Wahl der Gesamtfreiheitsgrade  $v_\Sigma(K)$  ergibt sich der globale kritische Wert als  $cv_\alpha = \chi^2(v_\Sigma(K))_{1-\alpha}$ .

Weiterhin empfiehlt es sich, ein minimales Gewicht  $v_{\min}$  für zu realisierende Studienabschnitte festzulegen. Zur Vereinfachung gelte im Folgenden  $v_{\min} = 1$  und  $v_{\Sigma}(K) = K$ . Damit ergibt sich ein relatives minimales Stufengewicht an den Gesamtfreiheitsgraden zu  $1/K$ .

Für die erste Stufe werden die  $K$  Freiheitsgrade in  $1 \leq v_1 \leq K$  als Gewicht der ersten Stufe und  $v_2^* = K - v_1 \geq 1$  aufgeteilt. So gilt, dass  $q_1(v_1) + q_2(v_2^*)$  unter der Nullhypothese  $\chi^2$ -verteilt ist mit  $K$  Freiheitsgraden. Nach der Realisierung der ersten Stufe wird die Studie mit der Ablehnung der Nullhypothese beendet, falls  $S_1 = q_1(v_1) \geq \chi^2(K)_{1-\alpha}$ . Andernfalls wird die Studie fortgesetzt, indem die Freiheitsgrade  $v_2^*$  auf die zweite Stufe und den verbleibenden Studienteil aufgeteilt oder vollständig für den zweiten Studienabschnitt verwendet werden. Generell ist bei der Aufteilung der Freiheitsgrade darauf zu achten, dass für den folgenden Studienteil mindestens  $v_{\min}$  Freiheitsgrade übrig bleiben. Stehen schließlich für einen Studienabschnitt  $k$  weniger als  $2v_{\min}$  Freiheitsgrade zur Verfügung, werden diese aufgebraucht und die Studie endet nach  $stp(k)$ .

Nach der Durchführung von  $stp(k-1)$  ergibt sich unter der Nullhypothese bei Verbrauch aller Freiheitsgrade folgende Verteilung für die kombinierte Teststatistik:

$$q_1(v_1) + \left\{ q_2(v_2) + \left[ q_3(v_3) + \left( \dots + \left[ q_{k-1}(v_{k-1}) + q_k \left( K - \sum_{j=1}^{k-1} v_j \right) \right] \right) \right] \right\}^{H_0} \sim \chi^2(K)$$

mit  $v_j \geq 1, i = 1, \dots, k-1$ , und  $K - v_{\Sigma}(k-1) \geq 1$ . Somit gilt weiterhin die Entscheidungsregel aus Abschnitt 2.1, dass  $H_0$  nach  $stp(k^*)$  verworfen und die Studie beendet wird, falls  $S_{k^*} \geq \chi^2(K)_{1-\alpha}$ .

Die Freiheitsgrade werden basierend auf dem Wissen aus den Studienabschnitten  $stp(1)$  bis  $stp(k-1)$  gewählt. Nach Durchführung von  $stp(k-1)$  stehen Schätzer für den interessierenden Parameter  $\hat{\theta}_{k-1}$  sowie für andere Parameter wie Varianzen zur Verfügung. Diese können zur Berechnung des Stichprobenumfangs in einer geeigneten Stichprobenfunktion  $f$  verwendet werden. Nach  $stp(k-1)$  lässt sich der Stichprobenumfang  $M_k$  berechnen, der benötigt wird, wenn in  $stp(k)$  die Studie nach Vergabe der verbleibenden Freiheitsgrade  $K - v_{\Sigma}(k-1)$  endet und der neu festgesetzte Wert des Parameters mit einer bedingten Power von  $1-\beta$  nachgewiesen werden soll:

$$M_k = f(1 - F_{\chi^2_{\{K - v_{\Sigma}(k-1)\}}} (cv_{\alpha} - S_{k-1}), \beta, \hat{\theta}_{k-1}).$$

Dabei gibt  $1 - F_{\chi^2_{\{K - v_{\Sigma}(k-1)\}}} (cv_{\alpha} - S_{k-1})$  den bedingten Fehler erster Art für den noch folgenden Studienteil an. Sollen jedoch noch mindestens zwei Studienabschnitte durchgeführt werden, ergibt sich mit einer geringeren Power  $1 - \beta_{g,k} < 1 - \beta$  für  $\text{stp}(k)$  der folgende Stichprobenumfang:

$$m_k = f(1 - F_{\chi^2_{\{K - v_{\Sigma}(k-1)\}}} (cv_{\alpha} - S_{k-1}), \beta_{g,k}, \hat{\theta}_{k-1}).$$

Es bietet sich an, die Freiheitsgrade  $v_k$  proportional an den verbleibenden Freiheitsgraden gemäß des Verhältnisses  $m_k/M_k$  zu wählen. Ersetzt man die Freiheitsgrade  $K - v_{\Sigma}(k-1)$  in  $m_k$  durch  $v_k = m_k/M_k \cdot (K - v_{\Sigma}(k-1))$ , lässt sich die adaptive Wahl von Umfang und Gewicht über einen Parameter steuern und der Stichprobenumfang führt exakt zur Power  $1 - \beta_{g,k}$  unter der Alternative  $\hat{\theta}_{k-1}$ .

Hartung und Knapp (2003) geben eine konstante bedingte Power  $1 - \beta_{g,k} = 1 - \beta_g$  bei der adaptiven Bestimmung der Stichprobenumfänge vor. Im Folgenden sollen weitere Strategien zur Wahl der  $\beta_{g,k}$  vorgestellt werden.

### 3 Strategien der Powervergabe

Bei Hartung und Knapp (2003) wird für die Berechnung des Stichprobenumfangs  $M_k$  bei Vergabe aller verbleibenden Freiheitsgrade in der letzten Stufe  $\text{stp}(k)$  stets die volle zu erreichende Gesamtpower  $1 - \beta$  vergeben. Ohne obere Schranken für die Stichprobenumfänge  $M_k$  und  $m_k$ , wie sie bei Hartung und Knapp eingeführt werden, führt dies zu einer sehr hohen Gesamtpower und entsprechend großen Stichprobenumfängen. Eine alternative Möglichkeit zur Kontrolle der Gesamtpower bietet die Schätzung des bereits vorliegenden Fehlers zweiter Art unter der neu bestimmten Alternative und darauf basierend die Berechnung des bedingten Fehler zweiter Art  $\beta_{k;K}$  bzw. der bedingten Power  $1 - \beta_{k;K}$  für den verbleibenden Studienteil. Ausgangspunkt dabei ist, dass sich der Gesamtfehler zweiter Art als Produkt der bedingten Fehlerwahrscheinlichkeiten zweiter Art darstellen lässt:

$$\begin{aligned}
 \beta &= P_{\theta} \left( \bigcap_{i=1}^K \{S_i \leq cv_{\alpha}\} \right) = P_{\theta} \left( \left( \bigcap_{j=1}^{k-1} \{S_j \leq cv_{\alpha}\} \right) \cap \left( \bigcap_{i=k}^K \{S_i \leq cv_{\alpha}\} \right) \right) \\
 &= P_{\theta} \left( \bigcap_{j=1}^{k-1} \{S_j \leq cv_{\alpha}\} \right) \cdot P_{\theta} \left( \bigcap_{i=k}^K \{S_i \leq cv_{\alpha}\} \mid \bigcap_{j=1}^{k-1} \{S_j \leq cv_{\alpha}\} \right) \\
 &:= \beta_{k-1} \cdot \beta_{k:K} \quad \Rightarrow \quad \beta_{k:K} = \frac{\beta}{\beta_{k-1}}.
 \end{aligned}$$

Dabei stellt  $\beta_{k-1}$  die Wahrscheinlichkeit dar, dass in den ersten  $k-1$  Stufen die Nullhypothese nicht abgelehnt und die Studie fortgesetzt wird.

Mit den in  $\text{stp}(1)$  bis  $\text{stp}(k-1)$  vergebenen Freiheitsgraden und Stichprobenumfängen kann  $\beta_{k-1}$  unter der neu berechneten Alternative geschätzt und somit die bedingte Fehlerwahrscheinlichkeit zweiter Art  $\hat{\beta}_{k:K}$  für den folgenden Teil der Studie bestimmt werden. Die für die adaptive Vergabe verwendeten Stichprobenfunktionen ergeben sich dann wie folgt:

$$\begin{aligned}
 M_k &= f(1 - F_{\chi^2_{\{K - v_{\Sigma}(k-1)\}}} (cv_{\alpha} - S_{k-1}), \hat{\beta}_{k:K}, \hat{\theta}_{k-1}), \\
 m_k &= f(1 - F_{\chi^2_{\{\frac{m_k}{M_k}(K - v_{\Sigma}(k-1))\}}} (cv_{\alpha} - S_{k-1}), \beta_{g,k}, \hat{\theta}_{k-1}).
 \end{aligned}$$

Für die Wahl der bedingten Power  $1 - \beta_{g,k}$  auf Stufe  $k$  stehen je nach Wünschen der Studienplaner verschiedene Strategien zur Verfügung. Die Aufteilung der Gesamtpower auf  $K$  Stufen orientiert sich ebenfalls an der Darstellung des Fehlers zweiter Art als Produkt der bedingten Fehler zweiter Art auf den einzelnen Stufen:

$$\begin{aligned}
 \beta &= P_{\theta} \left( \bigcap_{i=1}^K \{S_i \leq cv_{\alpha}\} \right) = P_{\theta} (S_1 \leq cv_{\alpha}) \cdot P_{\theta} (S_2 \leq cv_{\alpha} \mid S_1 \leq cv_{\alpha}) \\
 &\quad \cdot P_{\theta} \left( S_3 \leq cv_{\alpha} \mid \bigcap_{j=1}^K \{S_j \leq cv_{\alpha}\} \right) \cdot \dots \cdot P_{\theta} \left( S_K \leq cv_{\alpha} \mid \bigcap_{j=1}^{K-1} \{S_j \leq cv_{\alpha}\} \right) \\
 &:= \beta_1 \cdot \prod_{j=2}^K \beta_{g,j}.
 \end{aligned}$$

Im Folgenden werden drei Varianten zur Planung eines maximal fünfstufigen Designs mit einer Power von 90% vorgestellt, wobei die Beschränkung auf höchstens fünf Studienabschnitte durch die Vergabe aller verbleibenden Freiheitsgrade in  $\text{stp}(5)$  erreicht wird. In  $\text{stp}(5)$  wird die geschätzte Power  $1 - \hat{\beta}_{5,5}$  verwendet.

### 3.1 Konstante Power oder konstanter Powerzuwachs

Aus obiger Formel lässt sich für einen fünfstufigen Plan direkt ableiten, dass unter einer gegebenen Alternativen eine Power von 90% erreicht wird, wenn auf jeder Stufe eine konstante bedingte Power  $1 - \beta_{g,k}$  bzw.  $1 - \beta_1$  auf der ersten Stufe von  $37\% = \sqrt[5]{0.1}$  angesetzt wird. Dies bedeutet, dass bereits im ersten Schritt eine Power von 37% unter der vermuteten Alternativen erzielt wird. In den verbleibenden vier Stufen wird insgesamt nur noch eine absolute Power von 53% vergeben.

Alternativ, wenn ein gleichmäßiger Anstieg der Gesamtpower gewünscht wird, hier von  $90\%/5 = 18\%$ , wird folgende bedingte Fehlerwahrscheinlichkeit  $\beta_{g,k}$  verwendet:

$$\beta_{g,k} = \frac{\hat{\beta}_{k-1} - 0.18}{\hat{\beta}_{k-1}}.$$

Während bei einer konstanten bedingten Power von 37% der Zuwachs der Gesamtpower pro Stufe mit  $k$  sinkt, wächst die bedingte Power  $1 - \beta_{g,k}$  bei der Vorgabe eines konstanten Zuwachs pro Studienabschnitt.

### 3.2 Orientierung an gruppensequentiellem Design nach O'Brien und Fleming

Eine weitere Möglichkeit, die bedingte Power der einzelnen Studienabschnitte zu wählen, ist, den Powerverlauf in einem gruppensequentiellen Design nachzuahmen. Powerverläufe gruppensequentieller Designs nach Wang und Tsatis (1987) lassen sich mit Hilfe der Subroutinen SEQSCALE und SEQ aus SAS/IML bestimmen (siehe dazu auch Wassmer, 1999, und SAS, 1999). Diese Pläne, die die Normalverteilung der Zielvariablen voraussetzen und die maximale Anzahl an Interimsanalysen im Voraus festsetzen, ermöglichen eine frühzeitige Ablehnung der Nullhypothese, jedoch keine Anpassung des Studiendesigns an die erhaltenen Informationen.



Mit Hilfe der Routine SEQSCALE können kritische Schranken für ein vorgegebenes globales Niveau bestimmt werden. Nach Vorgabe der nicht adaptiven Stichprobenumfänge pro Stufe sowie einem Wert aus der Alternativen berechnet man mit der Subroutine SEQ die Power, die jeweils bis zu einer Stufe erreicht wird. Darauf basierend können gemäß den Rechenregeln für bedingte Wahrscheinlichkeiten die entsprechenden bedingten Fehlerraten zweiter Art  $\beta_{g,k}$  bestimmt werden.

Ein Spezialfall der Pläne nach Wang und Tsiatis geht auf O'Brien und Fleming (1979) zurück und ist charakterisiert durch eine geringe Power auf den ersten Stufen, was vorsichtige Planung im Sinne früher Kontrollen bedeutet.

Das folgende Programm berechnet die kritischen Schranken zum 2.5% Niveau für einen fünfstufigen Studienplan zum Vergleich zweier normalverteilter Stichproben. Zur Berechnung der Power wird neben der Varianz, die gleich 1 gesetzt wird, ein Unterschied in den Erwartungswerten der zwei Stichproben von 0.5 angenommen, der nach O'Brien und Fleming mit 90% Power nachgewiesen werden kann, wenn pro Stufe 34 Personen beobachtet werden.

```

/* Berechnung kritischer Werte und Powerverlauf im 5
stufigen gruppensequentiellen Design nach O'Brien und
Fleming*/
PROC IML;
alpha = 0.025;                               /* Niveau */
K = 5;                                       /* Stufenanzahl */
N = 34;                                     /* Umfang pro Stufe */
nmax = K*N;
V=(1:K)/K;
nvec=nmax/2*V;                               /* Vektor der Umfänge pro Stufe */
Theta = 0.5;                                /* Effekt, bei dem Power 90% */
alphazw = alpha *2; /* Niveau * 2 für einseitigen Test */

/* Struktur der kritischen Schranken */
mm = -J(1,K,1)// J(1,K,1);
call SEQSCALE(prob,critval,mm,1-alphazw);

/* kritische Schranken für kumulative Teststatistik*/
m=critval * J(1,K);

/* Powerberechnung */
mm = J(1,K,.M)//m;                          /* Integrationsgrenzen */

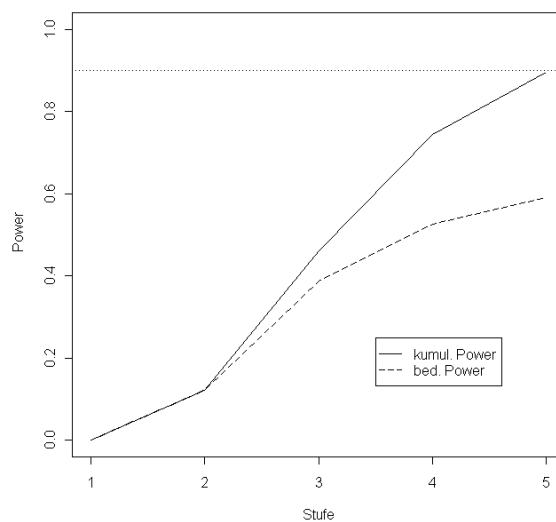
```

S. Menzler, G. Knapp

```
/* Nichtzentralitätsparameter */
nz= theta#(1:K)*sqrt(nmax/(K*4));

/* Verschobene Grenzen */
mmnz = J(1,K,.M)/(m-nz);
call seq(prob,mmnz) eps=1E-7 ;
power = 1- prob[2,];          /* Power bis Stufe k */
beta_gk = J(1,K,0);          /* bedingter Fehler 2. Art */
beta_gk[1] = 1- power[1];
DO j=2 TO K;
beta_gk[j] = 1 - (power[j]- power[j-1])/
(1-power[j-1]);
END;

PRINT critval power [format = 5.3] beta_gk
[format = 5.3];
QUIT;
```



**Abbildung 1:** Powerverlauf bei gruppensequentiellem Plan nach O'Brien und Fleming

In Abbildung 1 sind die Ergebnisse des obigen Programms grafisch dargestellt. Der langsame Anstieg der Power (power) von 0.1% im ersten Studienabschnitt über 46% im dritten bis zu 89.6% im fünften Studienteil wird deutlich. Der Verlauf der bedingten Power ( $1-\beta_{gk}$ ) ist durch die gestri-

chelte Linie gekennzeichnet und geht von 0.1% zu Beginn der Studie über 38.48% in der dritten bis zu 59.14% in der letzten Stufe.

## 4 Simulationsergebnisse

Die drei vorgestellten Strategien der Powervergabe wurden mittels Simulation verglichen, wobei jeweils 10000 Studien pro Szenario simuliert wurden. Als Teststatistik auf den einzelnen Stufen dient der einseitige Gaußtest. Insgesamt wird ein Niveau von 2.5% vorgegeben, während die angestrebte Power für Parameterwerte größer als 0.3 mindestens 90% betragen soll. Für geschätzte Parameterwerte kleiner als 0.3 wird für die Powerberechnungen der Wert 0.3 verwendet. Die Gesamtzahl der Freiheitsgrade ist 10, der minimale Wert liegt bei 1.

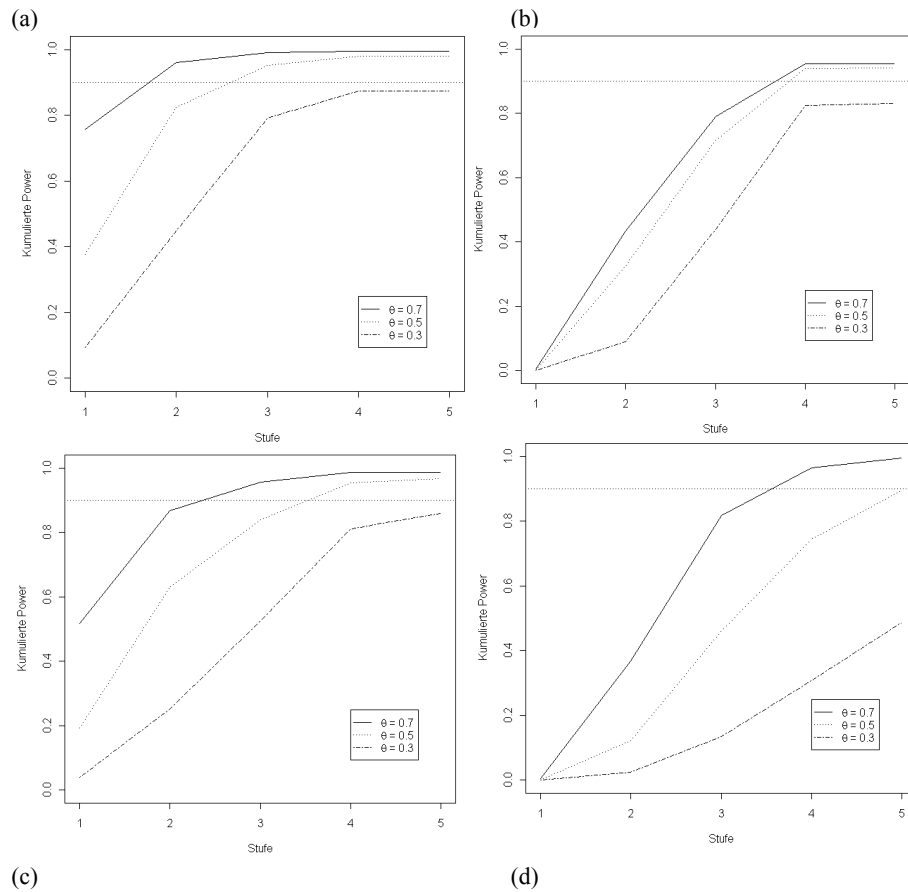
Ausgehend von einem Startwert  $\theta_1 = 0.5$  werden die Stichprobenumfänge und Freiheitsgrade adaptiv gewählt, wobei ab der zweiten Stufe die bedingte Power und der Stichprobenumfang basierend auf dem Kombinationsschätzer aus den Schätzern pro Stufe  $\theta_{stp(i)}$  ermittelt werden.

$$\hat{\theta}_{k-1} = \frac{\sum_{i=1}^{k-1} n_i \cdot \theta_{stp(i)}}{\sum_{i=1}^{k-1} n_i},$$

wobei  $n_i$  den Stichprobenumfang auf  $stp(i)$  darstellt. Als wahre Werte des Parameters  $\theta$  wurden 0.3, 0.5 und 0.7 gewählt.

In Abbildung 2 ist jeweils der Powerverlauf für die drei wahren Parameterwerte für eine der drei Strategien der Powervergabe dargestellt. Abbildungen 2(a)–2(c) stellen die Verfahren gemäß des Self-Designing-Prinzips dar. In Abbildung 2(d) ist zum Vergleich der Verlauf im nicht-adaptiven gruppensequentiellen Design nach O'Brien und Fleming dargestellt.

Die drei Self-Designing Strategien unterscheiden sich deutlich hinsichtlich der Power auf der ersten Stufe, die vor allem für den unterschiedlich schnellen Anstieg der kumulierten Gesamtpower verantwortlich ist. Bei konstanter bedingter Power von 37% wird zum Beispiel bereits in der zweiten Stufe, bei konstantem Powerzuwachs in der dritten und bei O'Brien und Fleming erst in der vierten Stufe bei einem wahren Wert von 0.7 eine Gesamtpower von 90% erreicht bzw. überschritten. Im Vergleich zum gruppensequentiellen Plan mit einer Power von 48.63% bei  $\theta = 0.3$  ist festzustellen, dass die Anpassung des Stichprobenumfangs bei allen Self-Designing Verfahren zu einer deutlichen Erhöhung der Gesamtpower führt (vgl. Tabelle 1).



**Abbildung 2:** Powerverlauf unter drei Alternativen für drei Strategien im Self-Designing (Konstante bedingte Power von 37% (a), konstanter Powerzuwachs von 18% (b), Powervergabe nach O'Brien und Fleming (c)) und für den gruppensequentieller Plan nach O'Brien und Fleming (d)

Die unterschiedlichen Strategien führen zu leichten Unterschieden hinsichtlich der mittleren Gesamtstichprobenumfänge und der Gesamtpower, die zusammen mit dem Median der simulierten Stichprobenverteilung in Tabelle 1 aufgeführt sind. Tendenziell ist zu bemerken, dass frühe Kontrollen, das heißt Interimsanalysen nach einer relativ kleinen Anzahl an Beobachtungen,

wie sie bei O'Brien und Fleming vorliegen, zu einem leichten Powerverlust bei gleichem oder sogar höherem Gesamtumfang führen.

Zum Vergleich sei hier noch der benötigte Stichprobenumfang von 468 in einem einstufigen Verfahren angeführt, um mit einer Power von 90% einen Wert von 0.3 nachzuweisen. Die Erhöhung des Stichprobenumfangs im Fall von  $\theta = 0.3$  ist also nicht unverhältnismäßig.

**Tabelle 1:** Simulierter mittlerer Stichprobenumfang (ASN), Median (kursiv) des Stichprobenumfangs und erzielte Power in %

Wahrer Wert $\theta$	0.3	0.5	0.7	0.3	0.5	0.7
	ASN ( <i>Median</i> )			Erzielte Power in %		
Self-Designing						
Konstante	342	168	114	87.48	98.00	99.58
bedingte Power	<i>328</i>	<i>112</i>	<i>102</i>			
Konstante	349	158	103	85.93	96.73	98.74
Powerzugabe	<i>314</i>	<i>116</i>	<i>82</i>			
nach O'Brien	347	184	114	83.18	94.13	95.48
und Fleming	<i>332</i>	<i>190</i>	<i>84</i>			
O'Brien und	154	123	97			
Fleming	<i>170</i>	<i>136</i>	<i>102</i>	48.63	89.57	99.49

## 5 Zusammenfassung

Mit Hilfe der Subroutinen SEQ und SEQSCALE können Powerverläufe bestimmter gruppensequentieller Designs berechnet werden und zur Planung von Klinischen Studien gemäß des Self-Designing-Prinzips nach Hartung und Knapp (2003) eingesetzt werden.

Die hier vorgestellten Schemata können jederzeit durchbrochen werden, indem zum Beispiel in einer folgenden Stufe alle verbleibenden Freiheitsgrade vergeben werden, eine höhere Power für den gesamten verbleibenden Teil oder die nächste Stufe angesetzt wird oder die Anzahl der Stufen nicht auf fünf beschränkt wird. Das Self-Designing-Prinzip bietet somit vielfältige

Planungsmöglichkeiten, je nach Vorwissen und Wünschen der Studienplaner.

## **Literatur**

- [1] Bauer, P. und Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50, 1029-1041.
- [2] Coburger, S., Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal*, 43, 821-833.
- [3] Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine*, 17, 1551-1562.
- [4] Hartung, J. (2001). A self-designing rule for clinical trials with arbitrary response variables. *Controlled Clinical Trials*, 22, 111-116.
- [5] Hartung, J., Knapp, G. (2003). A new class of completely self-designing clinical trials. *Biometrical Journal*, 45, 3-19.
- [6] Lehmacher, W. und Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55, 1286-1290.
- [7] O'Brien, P.C. und Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35, 549-556.
- [8] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199.
- [9] SAS Institute Inc. (1999). *SAS/IML user's guide, version 8*. Cary, NC: SAS Institute Inc.
- [10] Shen, Y. und Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics*, 55, 190-197.
- [11] Wang, S. und Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43, 193-199.
- [12] Wassmer, G. (1999). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien. Theoretische Konzepte und deren praktische Umsetzung mit SAS*. Verlag Alexander Mönch, Köln.