Über die Durchführung Gruppen-sequentieller Tests für das Zweistichprobenproblem basierend auf robusten Lokations- und Skalenschätzern mit SAS

Dr. A. Christmann
Universität Dortmund, HRZ, D-44221 Dortmund

Für feste Stichprobenumfänge sind viele Eigenschaften robuster Schätzer wie die Influenzfunktion, der Bruchpunkt, die maximal bias curve oder die asymptotische Verteilung untersucht worden, vgl. z.B. Huber (1981), Hampel et al. (1986) und Davies (1990, 1993). In einigen Bereichen der angewandten Statistik, insbesondere bei biometrischen Fragestellungen, spielen gruppensequentielle Pläne eine große Rolle. Derartige Pläne können den erwarteten Stichprobenumfang im Vergleich zu Plänen mit festem Stichprobenumfang senken, was nicht nur aus zeitlichen oder finanziellen sondern oft auch aus ethischen Aspekten sinnvoll und notwendig ist, vgl. Pocock (1977). Allerdings wurden im Vergleich zu Plänen mit festem Stichprobenumfang für gruppensequentielle Pläne deutlich weniger Ergebnisse veröffentlicht über die Anwendung robuster Schätzer, vgl. Silvapulle & Sen (1993) und Mehta et al. (1994). In diesem Artikel wird das Verhalten eines gruppensequentiellen Zweistichprobentests untersucht, wenn man die klassischen nicht-robusten Schätzer in der t-Teststatistik durch moderne robuste Schätzer für die Lokations- und Skalenparameter ersetzt.

1. Modell

Gegeben seien unabhängige Zufallsvariablen $X_{j,i}$, i=1,..., N_j, j=1,2, wobei $X_{1,1},...,X_{1,N_1}$ unabhängig und identisch verteilt sind gemäß der Verteilungsfunktion $F((\cdot - \mu_1)/\sigma)$ und $X_{2,1},...,X_{2,N_2}$ unabhängig und identisch verteilt sind gemäß Verteilungsfunktion $G((\cdot - \mu_2)/\sigma)$. Hierbei sind die reellwertigen Lokationsparameter μ_1 und μ_2 sowie der Skalenparameter $\sigma > 0$ unbekannt. Der standardisierte Behandlungsunterschied sei mit $\Delta = (\mu_2 - \mu_1)/\sigma$ bezeichnet. Bei der klassischen Verteilungsannahme werden normalverteilte Zufallsvariablen vorausgesetzt, d.h. $F = G = \mathrm{N}(0,1)$.

2. Hypothesen

Das Ziel des Experiments sei die Durchführung eines zweiseitigen Tests auf Lokationsunterschied, d.h.

$$H_0 : \Delta = 0 \quad \text{vs.} \quad H_1 : \Delta \neq 0 \ .$$

Analog können auch einseitige Testprobleme untersucht werden.

3. Gruppensequentielle Tests

Die Idee gruppensequentieller Tests wird ausführlich z.B. in Pocock (1977, 1983) beschrieben. Die grundsätzliche Idee gruppensequentieller Tests besteht im Gegensatz zu Tests mit einem festen Stichprobenumfang darin, daß der gesamte Stichprobenumfang $N = N_1 + N_2$ in 2K Gruppen aufgeteilt wird und maximal K Tests durchgeführt werden können. Dies ermöglicht es, daß bei Gültigkeit von H_1 oft schon vorzeitig eine Entscheidung getroffen werden kann. Sei $N_j = n_{j,1} + ... + n_{j,K}$, j=1, 2, und

$$N_{1,k} = \sum_{i=1}^{k} n_{1,i} \ , \quad N_{2,k} = \sum_{i=1}^{k} n_{2,i} \ , \quad k = 1,...,K.$$

Sei $\hat{\mu}_{j,k}$ ein Schätzer für den Lokationsparameter und $\hat{\sigma}_{j,k}$ ein Schätzer für den Skalenparameter basierend auf $X_{j,1},...,X_{j,N_k}$. Sei T_k die Teststatistik auf der k-ten Stufe, wobei

$$T_k = \frac{\hat{\mu}_{2,k} - \hat{\mu}_{1,k}}{\sqrt{(\hat{\sigma}^2_{2,k} / N_{2,k}) + (\hat{\sigma}^2_{1,k} / N_{1,k})}} \ , \quad k = 1,...,K.$$

Der Test auf der k -ten Stufe, $1 \le k \le K$, sei gegeben durch:

$\lvert T_k \rvert > c(k)$:	STOP, Entscheidung für H_1
$\lvert T_k \rvert \le c(k)$ und $k < K$:	weiter mit Stufe $k+1$
$\lvert T_k \rvert \le c(k)$ und $k = K$:	STOP, Entscheidung für H_0 ,

wobei $c(k) = ck^{a-0.5}$, c und a fest. Als naiver Schätzer für den standardisierten Lokationsunterschied kann $\hat{\Lambda}$ verwendet werden, wobei

$$\hat{\Delta} = \sqrt{N_{1,k} + N_{2,k}} \, \frac{\hat{\mu}_{2,k} - \hat{\mu}_{1,k}}{\sqrt{N_{1,k}\hat{\sigma}^2_{1,k} + N_{2,k}\hat{\sigma}^2_{2,k}}} \ , \ k = 1,...,K.$$

Bei Annahme von $F = G = \mathrm{N}(0,1)$ sind die klassischen Schätzer für die Lokationsparameter das arithmetische Mittel $\overline{X} = \overline{X}_{j,k}$ und für den Skalenparameter die Standardabweichung $S = S_{j,k}$, j=1,2, k=1,...,K. In der Literatur wurden viele kompliziertere Schätzer für Δ vorgeschlagen, vgl. Kim (1988, 1989).

Es gibt mehrere Programme zur Planung und Durchführung gruppensequentieller Tests, z.B. SAS/IML mit den Makros SEQ, SEQSCALE, SEQSHIFT oder die Software EaSt von Cytel Software Corporation.

4. Robuste Schätzer

Es ist bekannt, daß die klassischen Schätzer $\overline{X}_{j,k}$ und $S_{j,k}$ nicht robust gegenüber der Verletzung der Verteilungsannahme $F = G = \mathrm{N}(0,1)$ sind, und daß Ausreißer einen starken Einfluß auf diese Schätzer haben können, vgl. Andrews et al. (1972), Huber (1981) und Hampel et al. (1986).

Im folgenden werden für den obigen gruppensequentiellen Test die folgenden Fragen untersucht:

- Welche Auswirkungen kann dies auf Niveau und Power des Tests haben ?
- Welche Auswirkungen kann dies auf den erwarteten Stichprobenumfang (ASN) des Tests haben ?
- Welche Auswirkungen kann dies auf den Bias des naiven Schätzers $\hat{\Delta}$ für den standardisierten Behandlungsunterschied haben ?

In der Literatur wurden viele robuste Schätzer für die Lage- und Skalenparameter vorgeschlagen, die unter schwachen Annahmen konsistent und asymptotisch normal verteilt sind. Bezeichne die Gaußklammerfunktion mit $[\cdot]$. Sei $h = h_{j,k} = [N_{j,k}/2]+1$, $L = L_{j,k} = h(h-1)/2$, $d_{j,N_{j,k}} = N_{j,k}/(N_{j,k}+1.4)$ für $N_{j,k}$ ungerade und $d_{j,N_{j,k}} = N_{j,k}/(N_{j,k}+3.8)$ für $N_{j,k}$ gerade. Mit $X_{(j,1:N_k)} \le \ldots \le X_{(j,N_k:N_k)}$ seien die Ordnungsstatistiken von $X_{j,1},\ldots, X_{j,N_k}$ bezeichnet. Als Skalenschätzer werden im folgenden die von Rousseeuw und Leroy (1988) vorgeschlagene Länge der kürzesten Hälfte

$$SH = SH(X_{j,1}, \ldots, X_{j,N_{j,k}}) =$$

$$0.7413 \cdot \min\{X_{(j,h+k-1:N_{j,k})} - X_{(j,k:N_{j,k})};\ k = 1, \ldots, [(N_{j,k}+1)/2]\},$$

und der von Rousseeuw und Croux (1993) vorgeschlagene Schätzer

$$Q = Q(X_{j,1}, \ldots, X_{j,N_{j,k}}) =$$

$$2.2219 \cdot d_{j,N_{j,k}} \cdot \{|X_{j,i} - X_{j,m}|;\ 1 \le i < m \le N_{j,k}\}_{L:(N_{j,k}(N_{j,k}-1)/2)},\ j = 1,2,$$

verwendet. Eigenschaften dieser Schätzer und weitere robuste Skalenschätzer sind außer in den beiden obigen Artikeln beschrieben in Croux und Rousseeuw (1992a,b) und Christmann, Gather und Scholz (1994). Als Lokationsschätzer wird der in der Princeton study (Andrews et al., 1972) untersuchte M-Schätzer 25A verwendet, der implizit definiert ist als Lösung von

$$\sum_{i=1}^{N_{j,k}} \psi\left(\frac{X_{j,i}-\mu_j}{\hat{\sigma}}\right)=0 \ ,$$

wobei

$$\begin{aligned}
\psi(r) = \ & r && \text{für} \quad 0\le|r|\le a \\
= \ & a\cdot\text{sign}(r) && a<|r|\le b \\
= \ & a\cdot\frac{c-|r|}{c-b}\text{sign}(r) && b<|r|\le c \\
= \ & 0 && |r|>c \ ,
\end{aligned}$$

und $a=1.645$, $b=3.0$, $c=6.5$ und $\hat{\sigma}$ ein robuster Skalenschätzer ist.

5. Studiendesign

Einige Eigenschaften der oben beschriebenen gruppensequentiellen Tests werden anhand einer Simulationsstudie für mittlere Stichprobenumfänge untersucht. Hierbei werden drei verschiedene Pläne betrachtet, vgl. Tabelle 1.

Tabelle 1: Parameter für die gruppensequentiellen Pläne;
α = Wahrscheinlichkeit für den Fehler I. Art

Plan	Beschreibung	α	K	$n_{j,1}=\ldots=n_{j,K}$	a
1	Pocock	0.05	3	40	0.5
2	Pocock	0.01	5	34	0.5
3	Wang und Tsiatis (1987)	0.01	5	34	0.4695

Die kritischen Werte für die untersuchten gruppensequentiellen Tests sind in Tabelle 2 angegeben. Sie wurden durch Simulation basierend auf 10000 Läufen erhalten und sind so gewählt, daß bei Gültigkeit der klassischen Verteilungsannahme $F=G=\text{N}(0,1)$ die Tests eine Power von approximativ 95% bei $\Delta=(\mu_2-\mu_1)/\sigma=0.5$ besitzen.

Tabelle 2: Kritische Werte für die gruppensequentiellen Tests

Plan	Beschreibung	$(\overline{X},S)$	$(25A,SH)$	$(25A,Q)$
1	Pocock	2.2934	2.6473	2.3355
2	Pocock	3.0442	3.5630	3.1241
3	Wang und Tsiatis (1987)	3.1308	3.6210	3.1902

Im einzelnen wurden die folgenden Kombinationen bei der Simulation betrachtet, wobei jeweils 10000 Simulationsläufe durchgeführt wurden.

- 3 gruppensequentielle Pläne gemäß Tabelle 1 und Tabelle 2
- 3 Schätzerpaare: $(\overline{X},S)$, $(25A,SH)$, $(25A,Q)$
- 3 Verteilungen: Standardnormalverteilung N(0,1), t-Verteilung mit 3 Freiheitsgraden t_3 , Mischungsverteilung MIXN=0.9 N(0,1) + 0.1 N(10,100)
- 7 Paare von Verteilungsfunktionen (F,G):
  (N(0,1), N(0,1)), (N(0,1), t_3), (N(0,1), MIXN), (t_3 , t_3),
  (t_3 , MIXN), (MIXN, MIXN), (MIXN, N(0,1))
- 5 Werte von Δ : 0, 0.25, 0.5, 0.75, 1.

6. Ergebnisse

Da die Ergebnisse sich für die drei untersuchten Pläne qualitativ stark ähneln, werden hier nur einige typische Ergebnisse zu Plan 1 in den Tabellen 3 bis 5 dargestellt.

Tabelle 3: Geschätzte Güte der Tests in %.

Δ	Verteilungen		gruppensequentieller Test basierend auf		
	F	G	$(\overline{X},S)$	$(25A,SH)$	$(25A,Q)$
0	N(0,1)	N(0,1)	5.0	5.0	5.0
		t3	4.9	6.2	5.5
		MIXN	59.2	4.6	3.6
	t3	t3	5.2	7.5	6.0
		MIXN	53.8	5.8	4.3
	MIXN	MIXN	4.1	3.9	2.8
		N(0,1)	59.3	4.5	3.8
0.25	N(0,1)	N(0,1)	42.3	37.9	40.7
		t3	26.5	35.1	34.3
		MIXN	88.7	35.7	36.5
	t3	t3	20.0	32.6	29.4
		MIXN	82.6	33.5	31.3
	MIXN	MIXN	6.4	28.3	24.9
		N(0,1)	26.5	29.7	27.0
0.5	N(0,1)	N(0,1)	95.2	93.1	94.5
		t3	76.1	87.9	87.9
		MIXN	98.7	91.3	92.0
	t3	t3	60.1	82.9	80.4
		MIXN	96.3	86.0	84.8
	MIXN	MIXN	13.7	85.7	83.0
		N(0,1)	10.8	88.7	87.2

Tabelle 4: Geschätzte Werte für den erwarteten Stichprobenumfang (ASN) pro Gruppe

Δ	Verteilungen		gruppensequentieller Test basierend auf		
	F	G	$(\bar{X}, S)$	$(25A, SH)$	$(25A, Q)$
0	N(0,1)	N(0,1)	117.5	117.3	117.6
		t3	117.6	116.6	117.3
		MIXN	104.1	117.5	118.3
	t3	t3	117.5	116.1	117.1
		MIXN	105.1	116.8	117.8
	MIXN	MIXN	118.2	117.9	118.7
		N(0,1)	103.9	117.4	118.0
0.25	N(0,1)	N(0,1)	103.9	104.6	104.7
		t3	109.7	105.3	107.1
		MIXN	86.2	106.0	107.1
	t3	t3	112.3	106.5	108.9
		MIXN	89.1	106.3	108.5
	MIXN	MIXN	117.1	108.7	111.3
		N(0,1)	114.0	108.0	110.4
0.5	N(0,1)	N(0,1)	67.8	69.7	69.6
		t3	85.2	74.6	77.2
		MIXN	66.6	72.5	74.1
	t3	t3	94.6	78.8	83.2
		MIXN	71.5	77.1	80.7
	MIXN	MIXN	114.2	78.5	83.6
		N(0,1)	116.1	75.7	79.7

Tabelle 5: Geschätzte Werte für den Median von $\hat{\Delta} - \Delta$

Δ	Verteilungen		gruppensequentieller Test basierend auf		
	F	G	$(\overline{X}, S)$	$(25A, SH)$	$(25A, Q)$
0	N(0,1)	N(0,1)	-0.001	0.001	0.000
		t3	0.001	0.003	0.003
		MIXN	0.320	0.012	0.017
	t3	t3	0.002	0.002	0.002
		MIXN	0.307	0.013	0.016
	MIXN	MIXN	0.000	-0.003	-0.002
		N(0,1)	-0.320	-0.014	-0.018
0.25	N(0,1)	N(0,1)	0.011	0.030	0.010
		t3	-0.057	0.015	-0.018
		MIXN	0.159	0.024	-0.003
	t3	t3	-0.091	-0.004	-0.042
		MIXN	0.147	0.009	-0.026
	MIXN	MIXN	-0.192	-0.009	-0.045
		N(0,1)	-0.482	-0.006	-0.041
0.5	N(0,1)	N(0,1)	0.047	0.123	0.049
		t3	-0.079	0.096	-0.004
		MIXN	0.019	0.101	0.017
	t3	t3	-0.145	0.046	-0.054
		MIXN	-0.008	0.060	-0.039
	MIXN	MIXN	-0.382	0.037	-0.061
		N(0,1)	-0.650	0.069	-0.029

7. Zusammenfassung

Für die in der Simulation untersuchten Bedingungen lassen sich die Ergebnisse verkürzt wie folgt zusammenfassen.

- Die Verletzung der üblichen Normalverteilungsannahme kann gruppensequentielle Tests basierend auf $(\overline{X}, S)$ extrem beeinflussen bzgl. Niveau, Power, erwarteter Stichprobenumfang und des Bias des naiven Schätzers $\hat{\Delta}$ für die standardisierte Behandlungsdifferenz Δ.
- Die Verwendung robuster Schätzer, z.B. $(25A, SH)$ und $(25A, Q)$ kann in diesen Fällen bzgl. der obigen 4 Kriterien zu deutlich stabileren Resultaten führen. Bei Gültigkeit der klassischen Annahme $F = G = \mathrm{N}(0,1)$ führen die Tests basierend auf den beiden Paaren von robusten Schätzern zu Ergebnissen, die sich nur wenig von denen des Tests basierend auf $(\overline{X}, S)$ unterscheiden. Bei Verletzung der klassischen Verteilungsannahme schneidet der gruppensequentielle Test basierend auf $(\overline{X}, S)$ oft deutlich schlechter ab als die beiden

hier untersuchten Tests basierend auf robusten Schätzern. Unter H_0 sind selbst Wahrscheinlichkeiten für den Fehler I. Art von 10α möglich, vgl. Tabelle 3. Als Nachteil dieser gruppensequentieller Tests basierend auf diesen beiden Paaren von robusten Schätzer ist ein etwas höherer numerischer Aufwand für die Bestimmung der Schätzwerte zu nennen. Diese Schätzer sind jedoch z.B. mit SAS/IML-Macros oder mit FORTRAN Programmen berechenbar.

Theoretische Ergebnisse über finite oder asymptotische Eigenschaften dieser gruppensequentiellen Tests basierend auf $(25A, SH)$ oder $(25A, Q)$ sind bisher nicht bekannt. Da derzeit die kritischen Werte der Tests nur über Simulationen erhalten werden können, erscheint für die praktische Anwendung solcher Tests die Erstellung einer Tabelle der kritischen Werte für übliche Pläne sinnvoll.

8. Literatur

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., & Tukey, J.W. (1972). *Robust Estimates of Location. Survey and Advances.* Princeton University Press, Princeton, N.J..

Christmann, A., Gather, U., Scholz, G. (1994). Some properties of the length of the shortest half. *Statistica Neerlandica* **48**, 209-213.

Croux, C. & Rousseeuw, P.J. (1992a). A Class of High-Breakdown Scale Estimators Based on Subranges, *Comm.Statist.Theor.Meth.* **21**, 1935-1951.

Croux, C. & Rousseeuw, P.J. (1992b). Time efficient algorithms for two highly robust estimators of scale. *Comput.*Statist. **1**, 411-428.

Davies, P.L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann.Statist.* **18**, 1651-1675.

Davies, P.L. (1993). Aspects of robust linear regression. *Ann.Statist.* **21**, 1843-1899.

EaSt (1993). *A Software Package of the Design and Interim Monitoring of Group Sequential Clinical Trials.* Cambridge, Massachusetts: Cytel Software Corporation.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986). *Robust statistics. The approach based on influence functions.* J. Wiley & Sons, New York.

Huber , P.J. (1981). *Robust statistics.* J. Wiley & Sons, New York.

Kim, K. (1988). Improved approximation for estimation following closed sequential tests. *Biometrika* **75**, 121-128.

Kim, K. (1989). Point estimation following group sequential tests. *Biometrics* **45**, 613-617.

Mehta, C.R., Patel, N., Senchaudhuri, P. & Tsiatis, A. (1994). Exact permutational tests for group sequential clinical trials. *Biometrics* **50**, 1042-1053.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

Pocock, S.J. (1983). *Clinical Trials. A practical approach.* Wiley, Chichester.

Rousseeuw, P.J. & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *J. Amer. Statist. Assoc.* **88**, 1273-1283.

Rousseeuw, P.J. & Leroy, A.M. (1988). A robust scale estimator based on the shortest half. *Statistica Neerlandica* **42**, 103-116.

SAS Institute Inc., *SAS/IML Software: Changes and Enhancements through Release 6.11*, Cary, NC., 1995.

Silvapulle, M.J. & Sen, P.K. (1993). Robust tests in group sequential analysis: one- and two-sided hypotheses in the linear model. *Ann. Inst. Statist. Math.* **45**, 159-171.

Wang, S.K. & Tsiatis, A.A. (1987). Approximately Optimal One-Parameter Boundaries for Group Sequential Trials. *Biometrics* **43**, 193-199.