Fallzahladjustierung im Rahmen von Klinischen Studien

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
an der Universität Dortmund
Fachbereich Statistik

vorgelegt von

Michael Hennig

2002

Prüfungskommission:	Prof. Dr. K. Ickstadt	(Vorsitzende)
	Prof. Dr. A. Neiß	(Gutachter)
	Prof. Dr. S. Schach	(Gutachter)
	Prof. Dr. W. Urfer	(Gutachter)
	Dr. S. Selinski	(wissenschaftliche Mitarbeiterin)

Tag der mündlichen Prüfung: 9. Juli 2002

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Medizinische Statistik und Epidemiologie, Klinikum rechts der Isar. Ganz besonders möchte ich mich daher bei Herrn Prof. Dr. A. Neiß, Direktor des Institutes, bedanken für die großzügige und freundliche Förderung meiner Arbeit.

Herrn Prof. Dr. S. Schach vom Fachbereich Statistik der Universität Dortmund danke ich ganz herzlich für die Unterstützung und für seine Bereitschaft zur Begutachtung.

Mein Dank gilt auch Herrn Prof. Dr. W. Urfer vom Fachbereich Statistik der Universität Dortmund, der sich freundlicherweise als weiterer Gutachter zur Verfügung gestellt hat.

Beim International Steering Committee der ELSA Studie unter dem Vorsitz von Prof. Dr. A. Zanchetti, Universität Mailand, bedanke ich mich für die Erlaubnis, auf die Daten der ELSA-Studie zurückgreifen zu dürfen.

Schließlich möchte ich meiner Frau Susanne für ihre aktive Unterstützung und ihr großes Verständnis danken. Auch meine Kinder Nina, Simona und Tim hatten einen wesentlichen Anteil am Zustandekommen dieser Arbeit – Danke!

INHALTSVERZEICHNIS

1 Einleitung

Murphy's Law 13: „There is never enough time to do a job right the first time, but there is always time to do it again."

Die Wirksamkeit und Sicherheit eines neuen therapeutischen oder diagnostischen Verfahrens in der Medizin kann nur durch eine qualitativ hochwertige kontrollierte klinische Studie überprüft werden. Die randomisierte, doppelblinde Placebo-kontrollierte klinische Studie ist hier das Maß aller Dinge. Dazu stellt die Medizinische Statistik qualitätssichernde Verfahren zur Planung, Durchführung und Auswertung zur Verfügung.
Im Rahmen dieser Arbeit soll auf einen sehr wesentlichen Planungsaspekt eingegangen werden: die möglichst genaue Abschätzung der benötigten Fallzahl. Schließlich ist sowohl eine zu kleine als auch eine zu große klinische Studie aus ethischen, wissenschaftlichen und ökonomischen Gesichtspunkten problematisch.
Zur Fallzahlplanung werden Vorinformationen über die Zielvariable, wie z.B. deren Streuung benötigt. Diese Vorinformationen stammen in der Regel aus Vorgänger-Studien und sind oft mit einer großen Unsicherheit behaftet. Dies kann zum einen daran liegen, dass die Vor-Studien mit kleinen Fallzahlen durchgeführt wurden; zum anderen sind die Studiendesigns der Vor-Studie(n) und der zu planenden Haupt-Studie selten identisch. Auch gibt es oftmals überhaupt keine vergleichbaren Vor-Studien, so dass eine klassische Fallzahlplanung erst gar nicht durchgeführt werden kann. Aber auch eine Fallzahlplanung, die mit fehlerhaften Vorinformationen arbeitet, ist kritisch - vor allen Dingen bei großen Diskrepanzen zwischen den bei der Fallzahlplanung benutzten Parametern und den tatsächlichen Größen.
Die Bedeutung von Verfahren zur Adjustierung der Fallzahl wird auch in den für klinische Studien immer wichtiger werdenden Guidelines, wie z.B. die ICH-Guideline „Statistical Principles for Clinical Trials" (ICH-E9 1998), hervorgehoben.

Aus den oben beschriebenen Gründen wurden zahlreiche Methoden zur Adjustierung der Fallzahl entwickelt. Die Grundidee aller Adjustierungsverfahren besteht darin, die Vorinformationen mit den im Laufe der Studie beobachteten Daten abzugleichen und somit die Fallzahl entsprechend zu adjustieren. Die einzelnen Verfahren lassen sich grob in zwei Kategorien einteilen: Verfahren, die eine Entblindung der Studiendaten benötigen vs. Verfahren, die ohne Entblindung arbeiten.

In dieser Arbeit werden Verfahren zur Fallzahladjustierung im Rahmen von klinischen Studien zum Nachweis der Überlegenheit einer Therapie gegenüber gestellt, weiter entwickelt und schließlich bezüglich ihrer Eigenschaften verglichen.

Methoden zur Fallzahladjustierung können auch als Spezialfall von adaptiven Designs aufgefasst werden, s. dazu insbesondere (Bauer und Köhne 1994). Hier sind auch andere Veränderungen am Studiendesign denkbar, wie z.B. an der Anzahl der untersuchten Behandlungsgruppen oder an den Ein- und Ausschlusskriterien. Die im Rahmen dieser Arbeit diskutierten Methoden haben zum Ziel, die bei der Fallzahlplanung getroffenen Annahmen im Laufe der Studie zu überprüfen. Es geht also nicht um Zwischenauswertungen, bei denen die im Laufe der Studie gesammelten Daten zur sequentiellen Überprüfung einer oder mehrerer Hypothesen benutzt werden.

Im Kapitel 2 werden zunächst die Prinzipien der Fallzahlplanung beschrieben – und zwar für die im Rahmen von klinischen Studien am häufigsten anzutreffenden Situationen: Vergleich von Mittelwerten, Vergleich von Ereignisraten, Repeated Measurements.
Anschließend beschäftigt sich das Kapitel 3 mit den allgemeinen Methoden der Fallzahladjustierung. Hierzu gibt es eine ganze Reihe von Verfahren, die in diesem Kapitel hinsichtlich ihrer gemeinsamen Prinzipien zusammengefasst werden. Diese Zusammenfassung bildet die Grundlage einer Einteilung der Methoden in diverse Kategorien. Ein weiterer wesentlicher Aspekt ist hier auch die Erarbeitung von Kriterien zur Beurteilung von Adjustierungsverfahren. Neben den klassischen Kriterien wie Signifikanzniveau und Power werden mit der Verteilung der Fallzahl und der Breite des Konfidenzintervalls zwei zusätzliche wichtige Indikatoren vorgestellt.
Der für alle Verfahren relevanten Frage nach dem optimalen Zeitpunkt einer Fallzahladjustierung ist das Kapitel 4 gewidmet. Dazu werden die gängigen, teils kontroversen Ansätze vorgestellt und verglichen. Bei dem Vergleich werden sowohl methodische als auch praktische Gesichtspunkte berücksichtigt. Die in diesem Kapitel hergeleiteten Empfehlungen haben Gültigkeit für alle im Folgenden vorgestellten Verfahren.
Die Kapitel 5, 6 und 7 behandeln die unterschiedlichen Testsituationen, bei denen Adjustierungsverfahren angewandt werden können. Für alle vorgestellten Situationen wurden umfangreiche Simulationsstudien durchgeführt – mit dem Ziel, die jeweiligen Methoden hinsichtlich ihrer Eigenschaften zu vergleichen. Dazu wurden eine Reihe von relevanten

Szenarien bezüglich der Planungsparameter und der realen Größen betrachtet. Außerdem wurden die üblichen Beurteilungskriterien auch im Rahmen dieser Simulationen um die Punkte Verteilung der Fallzahl und Breite des Konfidenzintervalls (s.o.) erweitert, so dass ein differenzierter Vergleich zwischen den Methoden angestellt werden konnte. Die besondere Herausforderung bei den zugrunde liegenden Simulationsprogrammen lag dabei zum einem in der Komplexität der Verfahren und der Beurteilungskriterien. Zum anderem mussten durch die große Anzahl von Simulationen (100.000) sehr umfangreiche Dateien effektiv verarbeitet werden (s. dazu auch weitere Details im Anhang).

Im Kapitel 5 geht es um den Vergleich von Mittelwerten. Die Beurteilung wird hier zusätzlich mittels theoretischen Eigenschaften der Verfahren durchgeführt.
Methoden zur Fallzahladjustierung beim Vergleich von Ereignisraten werden im Kapitel 6 diskutiert. Hier zeigt sich für ein etabliertes Verfahren ein wesentliches Manko – eine entsprechende Korrektur wurde von mir entwickelt und anschließend beurteilt. Außerdem wird ein neues Verfahren vorgestellt, welches auch in den Vergleich mittels Simulationsstudien aufgenommen wird.
Kapitel 7 beschäftigt sich mit Repeated-Measurements-Designs – ein in der Praxis klinischer Studien sehr häufig vorzufindender Studientyp. Hierzu gibt es in der Literatur erst einen Vorschlag für Adjustierungsverfahren. Motiviert von einer konkreten Phase-III-Studie wird hierzu eine Methode unter Berücksichtigung von wesentlichen logistischen Studienaspekten entwickelt und ebenfalls mittels Simulationsstudien beurteilt.
Kapitel 8 fasst schließlich die Ergebnisse zusammen und gibt einen Ausblick auf die Strategien, die Gegenstand weiterer Forschungsaktivitäten sein können.

2 Fallzahlplanung

Ein sehr wesentlicher Aspekt bei der Planung eines Versuches stellt die Abschätzung der benötigten Fallzahl dar. Die große Bedeutung der Fallzahlplanung lässt sich u.a. dadurch dokumentieren, das laut CIS (Current Index of Statistics) in den Jahren von 1975 bis 2000 insgesamt 1347 Arbeiten zum Thema „sample size“ veröffentlicht wurden. In den Jahren 1990 – 2000 ist mit 54% ein Großteil dieser Arbeiten erschienen, wodurch die Aktualität dieses Themenkomplexes belegt wird.

Nun gibt es eine Fülle von Situationen für die spezielle Fallzahlplanungsmethoden existieren, s. dazu auch folgende Übersichtsarbeiten: (Kraemer und Thiemann 1987), (Desu und Raghovarao 1990), (Bock 1998), (Denenberg 1987). Zu den wichtigsten Vertretern zählen dabei folgende Versuchstypen: Studien zur Schätzung eines Parameters z.B. eines Normbereiches in der Medizin, epidemiologische Studien (Fall-Kontroll-Studien, Kohortenstudien), Studien zur Beurteilung der Güte eines diagnostischen Verfahrens, Phase-II-Studien im Zyklus der Arzneimittelentwicklung, Phase-III-Studien. Im Folgenden werde ich mich auf die Fallzahlplanung im Rahmen von kontrollierten klinischen Studien konzentrieren, s. dazu auch: (Machin, Campbell et al. 1997), (Lachin 1981), (Donner 1984), (Bristol 1989).

Bei der Planung einer klinischen Studie ist eine ganze Reihe von Designparametern zu spezifizieren. So müssen u.a. folgende Fragen beantwortet werden: Wie lange soll die Studie dauern? Welche Art des Vergleiches (Unterschied oder Äquivalenz oder Dosis-Wirkung) soll gewählt werden? Soll es sich um ein Parallel- oder Cross-over- oder faktorielles Design handeln? Wie soll verblindet werden? Wie soll randomisiert werden? Welches ist die primäre Zielvariable?

Darüber hinaus ist die Abschätzung der Anzahl der benötigten Patienten ein sehr wesentlicher Faktor, da eine falsch dimensionierte Studie sowohl ethisch als auch ökonomisch nicht vertretbar ist. Dies trifft für eine zu kleine Studie zu, mit der man einen Wirksamkeitsnachweis eventuell überhaupt nicht erbringen kann. Somit ist es unnütz und unwirtschaftlich Patienten mit der neuen Therapie zu behandeln. Bei einer zu groß angelegten Studie besteht das Problem zum einen darin, dass bei einer vorliegenden Wirksamkeit der neuen Therapie, die Etablierung bzw. Zulassung dieser Therapie unnötigerweise verzögert wird. Auf der anderen Seite werden bei einer nicht wirksamen Therapie Ressourcen nutzlos

vergeudet, bzw. bei einer schädlichen neuen Therapie Patienten fälschlicherweise mit dieser Therapie behandelt.

Somit kommt der möglichst genauen Abschätzung der benötigten Fallzahl eine sehr wesentliche Bedeutung bei der Planung einer klinischen Studie zu; s. dazu auch die für klinische Studien relevanten Guidelines (ICH-E9 1998), (Kolman, Meng et al. 1998), (Gardner und Altman 1989), (CPMP 1990), (ICH-E6 1996).

Wesentliches Ziel der Fallzahlplanung ist die Absicherung der Power, also der Wahrscheinlichkeit für das Auffinden eines Therapieunterschiedes (bei Studien zum Nachweis eines Unterschiedes) – so er denn vorliegt. Dabei besteht die grundlegende Idee aller Methoden darin, die Verteilung der Teststatistik unter der Alternative an einer vorgegebenen Stelle δ zu untersuchen. Die Verteilung unter der Alternative ist oft schwierig zu bestimmen, so dass in vielen Fällen mit der Normalapproximation gearbeitet wird.

Im Folgenden werde ich die Fallzahlplanungsmethoden für die wichtigsten Situationen bei klinischen Studien vorstellen.

2.1 Vergleich von Mittelwerten

Der Ausgangspunkt für die Herleitung der Fallzahl beim Vergleich von Mittelwerten ist die Verteilung der Zielvariablen X in der Behandlungs- bzw. Kontrollgruppe. Dabei geht man von folgendem Modell aus:

$$\mathbf{X_B \sim N(\mu_B, \sigma^2)} \qquad \mathbf{X_K \sim N(\mu_K, \sigma^2)} \qquad \text{mit unbekanntem } \sigma$$

Im Rahmen von klinischen Studien wird dann üblicherweise ein zweiseitiger Test durchgeführt, mit:

$$\mathbf{H_0 : \mu_B = \mu_K} \qquad \mathbf{vs.} \qquad \mathbf{H_1 : \mu_B \neq \mu_K}$$

Die Teststatistik des zugehörigen t-Tests lautet:

$$T = \frac{\overline{X_B} - \overline{X_K}}{s_p} \sqrt{\frac{n_B n_K}{n_B + n_K}} \tag{2.1}$$

mit :

$$s_p = \sqrt{\frac{(n_B - 1)s_B^2 + (n_K - 1)s_K^2}{n_B + n_k - 2}} \qquad (2.2)$$

"gepoolte" Standardabweichung

Im Weiteren werde ich mich auf den – im Rahmen von klinischen Studien häufigsten – Fall $n_B = n_K$ konzentrieren. Für den allgemeinen Fall $n_B \neq n_K$ ändert sich an der grundsätzlichen Vorgehensweise nichts.

Für $n_B = n_K$ vereinfacht sich die Teststatistik zu:

$$T = \frac{(\overline{X_B} - \overline{X_K})\sqrt{n}}{2\, s_p} \qquad (2.3)$$

mit: $n = n_B + n_K$.

Die Teststatistik T ist unter H_0 zentral t-verteilt mit n-2 Freiheitsgraden. Somit wird bei einem vorgegebenen Signifikanzniveau α die Nullhypothese abgelehnt, falls $|t| > t_{crit}$ – mit t_{crit}=(1-α/2)-Quantil der t-Verteilung mit (n-2) Freiheitsgraden.

Exakte Herleitung der Fallzahl

Unter H_1 ist T nicht-zentral t-verteilt mit (n-2) Freiheitsgraden und Nichtzentralitätsparameter

$$nzp = \frac{(\mu_B - \mu_K)\sqrt{n}}{2\sigma} = \frac{\delta\sqrt{n}}{2\sigma} \qquad (2.4)$$

Ziel der Fallzahlplanung ist es nun, das n so zu bestimmen, dass eine vorgegebene Wahrscheinlichkeit für den Fehler 2. Art (ß) eingehalten wird. Dabei wird die Gütefunktion des Tests an einer wiederum vorgegebenen Stelle $\delta = |\mu_B - \mu_K|$ (medizinisch relevanter Unterschied) zur Berechnung des ß's betrachtet *(im Folgenden werde ich diesen Sachverhalt durch die Formulierung „H_1 an der Stelle δ" beschreiben).* Im Rahmen von klinischen Studien sind Werte von 20% oder 10% für die Wahrscheinlichkeit des Fehlers 2. Art üblich – entsprechend einer Power (1-ß) von 80% bzw. 90%.

Somit muss n derart bestimmt werden, dass der Wert der Verteilungsfunktion unter der Alternative an der Stelle t_{crit} kleiner gleich ß ist:

$$F_{t,nzp,n-2}(t_{crit}) - F_{t,nzp,n-2}(-t_{crit}) \leq ß$$

mit:

$$F_{t,nzp,n-2}(t_{crit}) = \frac{1}{2^{\frac{n}{2}-2}\Gamma\left(\frac{n}{2}-1\right)} \int_0^{\infty} x^{n-3} e^{-\frac{1}{2}x^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{t_{crit}\,x}{\sqrt{n-2}}} e^{-\frac{1}{2}(u-nzp)^2} du\, dx \qquad (2.5)$$

Verteilungsfunktion der nicht-zentralen t-Verteilung mit Nichtzentralitätsparameter nzp und (n-2) Freiheitsgraden

Die Herleitung der Fallzahl kann nur iterativ erfolgen, da die (unbekannte) Fallzahl sowohl in t_{crit} als auch in die Verteilungsfunktion - über den Nichtzentralitätsparameter – eingeht.

Der gesamte Prozess der Fallzahlplanung besteht somit aus folgenden Schritten:

1. Vorgabe von α, β, δ, σ_0
2. n *(eventuell nach Vorgabe eines geeigneten Startwertes)*
3. $t_{crit}=f(n,\alpha)$
4. $nzp=g(n,\delta,\sigma_0)$
5. $F_{t,nzp,n-2}(t_{crit})$

Solange, bis minimales n erreicht, für das: $F_{t,nzp,n-2}(t_{crit}) - F_{t,nzp,n-2}(-t_{crit}) \leq ß$

Für das obengenannte Beispiel ergibt sich auf diesem Wege eine Fallzahl von n=128, da für n=127 gilt:

t_{crit} = 1.97912, nzp = 2.81736 und $F_{t,nzp,n-2}(t_{crit}) - F_{t,nzp,n-2}(-t_{crit})$ = 0.20167.

Für n=128 gilt:

t_{crit} = 1.97897, nzp = 2.82843 und $F_{t,nzp,n-2}(t_{crit}) - F_{t,nzp,n-2}(-t_{crit})$ = 0.19854.

Dabei wurden sowohl die kritischen Werte der t-Verteilung (t_{crit}) als auch die Werte der Verteilungsfunktion der nicht-zentralen t-Verteilung mittels SAS® berechnet (Funktionen: TINV, CDF).

Approximative Herleitung der Fallzahl

Als Alternative zu der exakten iterativen Vorgehensweise bietet sich folgende Vereinfachung an:

Ein wesentlicher Bestandteil der Teststatistik T ist die beobachtete Differenz der Mittelwerte

$$D = \bar{X}_B - \bar{X}_K .$$

Nun gilt, unter der zusätzlichen Annahme, dass σ bekannt ist, für die Verteilung von D unter H_0:

$$D \sim N\left(0, \frac{4\sigma^2}{n}\right) \tag{2.6}$$

mit: $n_B = n_K = n/2$.

Unter H_1 – an der Stelle δ – gilt für die Verteilung von D:

$$D \sim N\left(\delta, \frac{4\sigma^2}{n}\right) \tag{2.7}$$

Um das zweiseitige Signifikanzniveau α einzuhalten, wird H_0 abgelehnt, falls:

$$\frac{|d|}{\text{Standardfehler}(d)} = \frac{|d|\sqrt{n}}{2\sigma} > z_{1-\alpha/2} \tag{2.8}$$

bzw. falls:

$$|d| > \frac{2 z_{1-\alpha/2}\,\sigma}{\sqrt{n}} =: c_\alpha \text{ (kritischer Wert bzgl. } \alpha) \tag{2.9}$$

Somit gilt als Forderung für den Fehler 2. Art an der Stelle δ:

$$\begin{aligned} &P(H_0 \text{ nicht ablehnen} \mid H_1 \text{ an der Stelle } \delta) = \\ &P(-c_\alpha < D < c_\alpha \mid H_1 \text{ an der Stelle } \delta) = ß \end{aligned} \tag{2.10}$$

Da unter der Alternative (an der Stelle δ) gilt:

$$P(D < -c_\alpha \mid H_1 \text{ an der Stelle } \delta) \approx 0 \tag{2.11}$$

wird diese Forderung durch:

$$\begin{aligned} &P(H_0 \text{ ablehnen} \mid H_1 \text{ an der Stelle } \delta) = \\ &P(D > c_\alpha \mid H_1 \text{ an der Stelle } \delta) = 1 - ß \end{aligned} \tag{2.12}$$

ersetzt.

Zur Veranschaulichung:

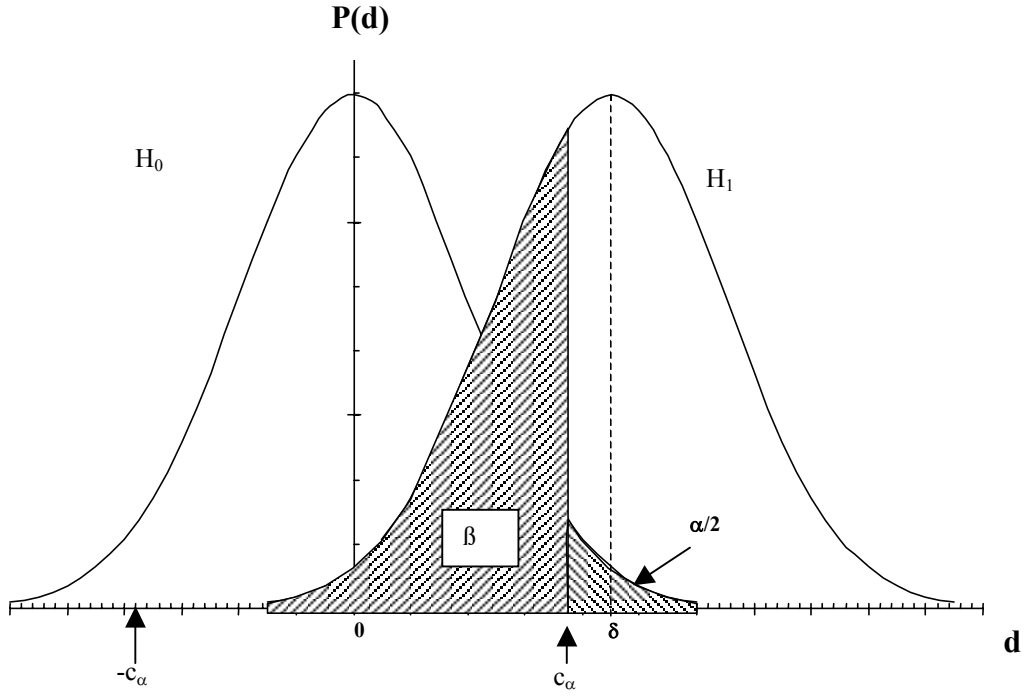


Abbildung 2.1: *Verteilung der Zwischen-Gruppen-Differenz D unter H_0 bzw. unter H_1*

Die Fallzahl wird dann folgendermaßen bestimmt:

$$\begin{aligned} &P(D > c_\alpha \mid H_1) = 1 - ß \\ \Leftrightarrow\ &P\left(\frac{(D - \delta)\sqrt{n}}{2\sigma} > \frac{(c_\alpha - \delta)\sqrt{n}}{2\sigma} \middle| H_1 \right) = 1 - ß \\ \Leftrightarrow\ &\frac{(c_\alpha - \delta)\sqrt{n}}{2\sigma} = z_ß \end{aligned} \tag{2.13}$$

da unter H_1: $\dfrac{(D-\delta)\sqrt{n}}{2\sigma} \sim N(0,1)$.

Löst man diese Gleichung nach n auf, so ergibt sich:

$$n = 4\left(\frac{\sigma}{\delta}\left(z_{1-\alpha/2} + z_{1-ß}\right) \right)^2 \tag{2.14}$$

Die so erhaltene “theoretische” Fallzahl wird dann für praktische Zwecke zur nächsten ganzen geraden Zahl aufgerundet, so dass in der Regel die tatsächliche Power leicht erhöht wird.

Zusammenfassung

Der Unterschied zwischen der exakten und der approximativ hergeleiteten Fallzahl ist marginal, wie folgende Tabelle des Zusammenhanges von $Q = \frac{\delta}{\sigma}$ und n deutlich macht:

Q	α=5%, β=20%		Unterschied n-exakt n-approximativ
	n – exakt mittels t-Verteilung	n – approximativ mittels Normalverteilung	
0.1	3142	3140	2
0.2	788	786	2
0.3	352	350	2
0.4	200	198	2
0.5	128	126	2
0.6	90	88	2
0.7	68	66	2
0.8	52	50	2
0.9	42	40	2
1.0	34	32	2
1.1	30	26	4
1.2	24	22	2
1.3	22	20	2
1.4	20	18	2
1.5	18	14	4

Tabelle 2.1: *Vergleich der exakten und approximativen Fallzahlen in Abhängigkeit von* $Q=\delta/\sigma$

Die Berechnungen wurden mittels SAS® durchgeführt.
Die approximative Vorgehensweise unterschätzt also die exakte Fallzahl um 2-4 Versuchseinheiten.

In der Praxis wird oftmals die so erhaltene Fallzahl in folgender Form nach oben korrigiert: Ausgehend von einer erwarteten drop-out rate (DOR) wird die Fallzahl mit dem Korrekturfaktor 1/(1-DOR) multipliziert.

Folgende Schlüsse lassen sich aus dieser fundamentalen Fallzahlformel (2.14) ziehen:

1. Die Standardabweichung der Zielgröße muss bei der Studienplanung bekannt sein. Dies ist in der Praxis klinischer Studien sehr selten der Fall. Wenn überhaupt Angaben über die Standardabweichung vorliegen, entstammen diese in der Regel aus kleinen

Pilotstudien, was zu einer großen Unsicherheit bei der Schätzung führt. Auch die Schätzung der Standardabweichung aus Angaben der Literatur kann sehr problematisch sein, da es sich bei publizierten Daten oft um andere Fragestellungen oder ein anderes Patientenkollektiv handelt. Eine weitere in der Praxis angewandte Methode ist die Abschätzung der Standardabweichung mittels der aus Vorstudien oftmals leichter zu bestimmenden Spannweite der Zielvariablen. Dabei wird als Standardabweichung 1/6 der Spannweite geschätzt, da bei vorliegender Normalverteilung fast der gesamte Wertebereich der Zielvariablen (99.87%) durch den Bereich $\mu \pm 3\sigma$ abgedeckt wird.

2. Die Standardabweichung der Zielgröße muss in beiden Gruppen gleich sein.
3. In die Fallzahl geht der Quotient (σ/δ) ein. Somit ist es bei der Fallzahlplanung ausreichend, δ in der Form: $\delta = k \bullet \sigma_0$ (z.B. k=1/2, 1, 2 oder 3) anzugeben.
4. Je größer das δ desto kleiner die Fallzahl - $1/\delta$ geht quadratisch in die Fallzahl ein.
5. Je größer die Power, desto größer die Fallzahl.
6. Je größer die Streuung, desto größer die Fallzahl - σ geht quadratisch in die Fallzahl ein, wie in folgender Grafik veranschaulicht wird:

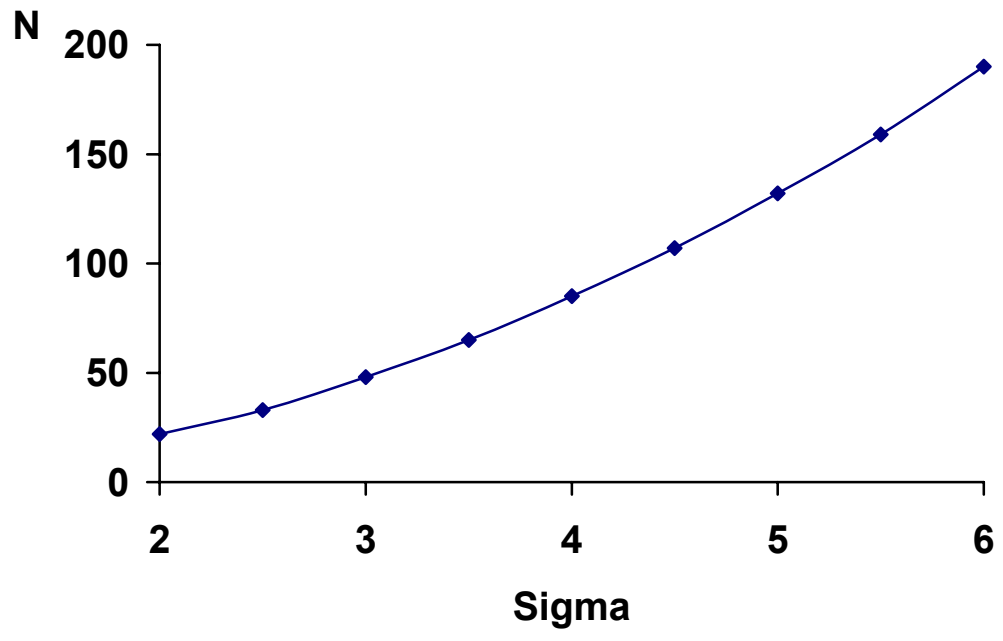


Abbildung 2.2: *Zusammenhang zwischen dem Fallzahlplanungsparameter σ und der resultierenden Fallzahl n (bei α = 5%, 1-β = 90%, δ = 2)*

Die o.g. Fallzahlformel kann auf vielfältige Weisen erweitert werden: Für den Fall ungleicher Gruppengrößen finden sich bei (Bock 1998) und (Schouten 1999a) weitere Details. Dies ist

dann von Bedeutung, wenn beispielsweise in der Behandlungsgruppe doppelt so viele Patienten wie in der Kontrollgruppe behandelt werden sollen.
Für den Vergleich von mehr als zwei Gruppen finden sich Strategien z.B. bei (Saville 1990) oder (Phillips 1998).

2.2 Vergleich von Ereignisraten

Handelt es sich bei der Zielvariablen um eine dichotome Größe, wie z.B. die Hospitalmortalität innerhalb von 28 Tagen, so wird in der Regel der Chi-Quadrat-Test benutzt, um auf Unterschied zwischen den Gruppen zu testen. Hier gilt für die Zufallsvariablen X_B, X_K:

$$\mathbf{X_B \sim B(n_B, p_B)} \qquad \mathbf{X_K \sim B(n_K, p_K)}$$

Zum Testen der Hypothese

$$\mathbf{H_0: p_B = p_K} \qquad \textbf{vs.} \qquad \mathbf{H_1: p_B \neq p_K}$$

lautet die Chi-Quadrat-Teststatistik:

$$X^2 = \frac{\left(\hat{p}_B - \hat{p}_K\right)^2}{\hat{p}(1-\hat{p})\dfrac{n_B + n_K}{n_B n_K}} \tag{2.15}$$

mit:

Gruppe	Ereignis		
	nein	ja	Σ
Behandlung	$n_{B,0}$	$n_{B,1}$	n_B
Kontrolle	$n_{K,0}$	$n_{K,1}$	n_K
Σ	n_0	n_1	n

$$\hat{p}_B = \frac{n_{B,1}}{n_B}$$
$$\hat{p}_K = \frac{n_{K,1}}{n_K} \tag{2.16}$$
$$\hat{p} = \frac{n_1}{n}$$

Nun gilt für $T = \sqrt{X^2}$:

- unter H_0 ist T approximativ N(0,1) verteilt
- unter H_1 ist T approximativ $N(\gamma,\varepsilon^2)$ verteilt, mit

$$\gamma = \sqrt{\frac{n_B n_K}{n_B + n_K}} \bullet \frac{p_B - p_K}{\sqrt{\frac{p_B n_B + p_K n_K}{n_B + n_K}\left(1 - \frac{p_B n_B + p_K n_K}{n_B + n_K}\right)}} \tag{2.17}$$

$$\varepsilon^2 = \frac{n_K p_B (1-p_B) + n_B p_K (1-p_K)}{(p_B n_B + p_K n_K)(n_B + n_K - p_B n_B - p_K n_K)} \bullet (n_B + n_K) \tag{2.18}$$

Aus der Forderung:

$$\Phi\left(\frac{z_{1-\alpha/2} - \gamma}{\varepsilon}\right) = \beta \tag{2.19}$$

(Einhaltung der Power an der Stelle $\delta = p_B - p_K$ der Alternative)

resultiert dann:

$$n_B = n_K = \left[\frac{\left(z_{1-\alpha/2} \sqrt{2 \frac{p_B + p_K}{2}(1 - \frac{p_B + p_K}{2})} + z_{1-\beta} \sqrt{p_B(1-p_B) + p_K(1-p_K)} \right)^2}{(p_B - p_K)^2} \right] + 1 \tag{2.20}$$

Bei der Anwendung dieser Fallzahlformel innerhalb klinischer Studien ist folgendes zu berücksichtigen:

1. Die Rate unter der Kontrollbehandlung muss bekannt sein, z.B. aus der Literatur.
2. Auch hier muss ein relevanter Unterschied zwischen Behandlungs- und Kontrollgruppe vorgegeben werden. Dafür gibt es folgende Möglichkeiten:

   2.1 Vorgabe eines relevanten additiven Unterschiedes: $\delta = |p_B - p_K|$.

   2.2 Vorgabe eines relevanten Faktors: $r = p_B/p_K$; z.B. $p_K = 0.8$, $r = 0.5$ (Halbierung der Ereignisrate unter Behandlung) → $p_B = 0.4$.

   2.3 Vorgabe eines relevanten Odds ratios (ROR): $ROR = \frac{p_B \bullet (1 - p_K)}{p_K \bullet (1 - p_B)}$ – daraus kann dann p_B hergeleitet werden: $p_B = \frac{ROR \bullet p_K}{(1 - p_K) + (ROR \bullet p_K)}$, z.B. p_K=0.6, ROR=0.5 → $p_B = 0.43$.

Bei allen Möglichkeiten kann p_B aus p_K und dem relevanten Unterschied berechnet werden, und anschließend Formel (2.20) angewandt werden.

Der Zusammenhang zwischen p_k und n (bei konstantem r) wird in folgender Grafik veranschaulicht:

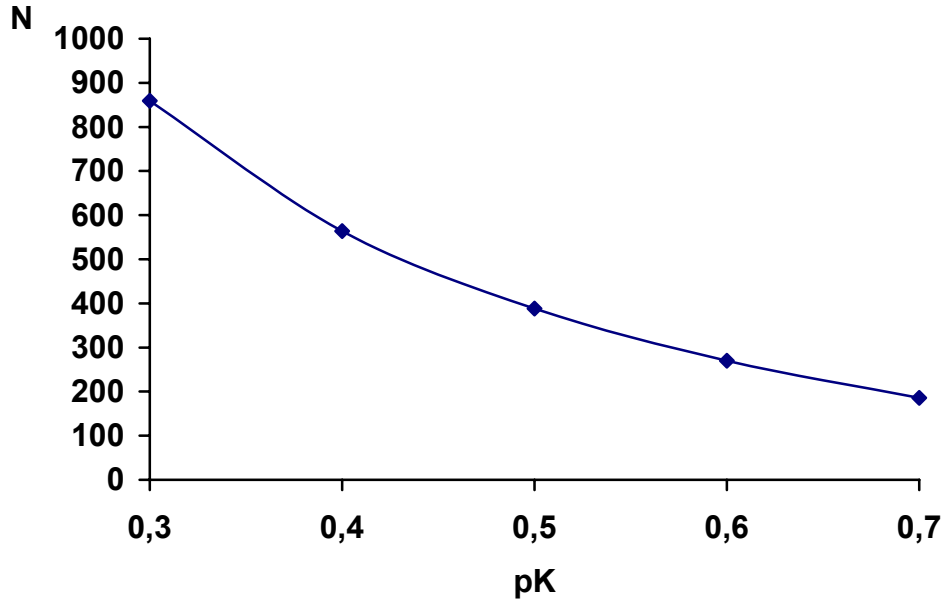


Abbildung 2.3: *Zusammenhang zwischen der Ereignisrate in der Kontrollgruppe (p_K) und der resultierenden Fallzahl n – bei konstantem $r = p_B / p_K = 0.8$ ($\alpha = 5\%$, $1\text{-}\beta = 80\%$)*

Eine andere Approximation liefert die Arcus-Sinus-Formel (Hartung 1985 S.420), die zu ähnlichen Ergebnissen führt. Eine Verbesserung der Approximation wird durch die Formel nach (Casagrande, Pike et al. 1978) erzielt. Ausführliche Tafeln für die exakten Werte finden sich ebenfalls in dieser Arbeit.
Erweiterungen zu dieser Fallzahlberechnung betreffen z.B. ungleiche Gruppengrößen (Donner 1984), (Machin, Campbell et al. 1997) oder relative Risiken (Donner 1984).

Der Vergleich von Ereignisraten kann auch als Spezialfall eines Mittelwertvergleiches angesehen werden. Schließlich lässt sich die hier zugrunde liegende Binomialverteilung durch die Normalverteilung approximieren. Somit können die Ergebnisse des vorigen Kapitels prinzipiell auf diese Fragestellung übertragen werden. Wegen der großen Bedeutung des Vergleichs von Ereignisraten für klinische Studien erfolgt jedoch hier und im weiteren Verlauf dieser Arbeit eine gesonderte Diskussion dieses Vergleichstyps.

2.3 Repeated Measurements

Wird eine Zielvariable im Laufe einer klinischen Studie mehrfach gemessen, so spricht man von repeated measurements. Dieses Konzept wird vor allen Dingen bei Studien angewandt, die zum Ziel haben, den Verlauf eines quantitativen klinischen Parameters über die Zeit zu beschreiben bzw. zwei (oder mehr) Behandlungsgruppen hinsichtlich dieses Verlaufs zu vergleichen. Dazu gibt es eine ganze Reihe von möglichen Auswertungsstrategien, s. dazu insbesondere (Crowder und Hand 1990).

Die einfachste Strategie besteht darin, die Mehrfachmessungen eines jeden Patienten zu einer „summary-statistic“ zusammenzufassen und anschließend univariat auszuwerten. So ist z.B. die Kalkulation einer Differenz „Messwert am Ende der Behandlung – Messwert zu Beginn der Behandlung“ die naheliegendste und auch eine medizinisch problemlos zu interpretierende „summary-statistic“. Weitere Vertreter dieses Ansatzes sind z.B.:

- Zeit bis zum Auftreten des maximalen Wertes
- Maximaler Wert
- Maximale Veränderung zum Ausgangswert
- Steigung (slope) der Messwerte
- Area under the curve (AUC)

Die zugehörige Fallzahlplanung kann über das im Abschnitt 2.1 (Vergleich von Mittelwerten) beschriebene Verfahren durchgeführt werden. Ausgangspunkt ist dabei die bei jedem Patienten aus den repeated measurements abgeleitete summary-statistic x_i. Hier müssen also Vorinformationen über die erwartete Streuung von X ebenso vorliegen wie ein a priori festgelegter relevanter Unterschied zwischen den Gruppen.

Bei (Schouten 1999b) werden auf der Basis von (2.14) Fallzahlformeln für folgende Szenarien entwickelt:

1) Der erwartete Unterschied zwischen den Gruppen ist ab einem bestimmten Zeitpunkt konstant. Ab diesem Zeitpunkt liegen mehrere Messungen der Zielvariablen vor.

2) Der erwartete Unterschied zwischen den Gruppen wächst im Laufe der Zeit linear an.

Für das 1. Szenario ist eine Kovarianzanalyse mit dem Mittelwert der Nach-Behandlungs-Messungen als abhängige Variable und mit dem Mittelwert der Vor-Behandlungs-Messungen als Kovariable die adäquate Auswertungsstrategie. Die für die Fallzahlplanung gemäß (2.14) benötigte Varianz ergibt sich bei diesem Ansatz zu:

$$\text{var} = \sigma^2 \bullet \left(\rho_{\text{nach}} + \frac{1-\rho_{\text{nach}}}{\text{r}} - \frac{\rho_{\text{mix}}^2}{\rho_{\text{vor}} + \frac{1-\rho_{\text{vor}}}{\text{p}}} \right) \tag{2.21}$$

mit: σ^2 : konstante Varianz der Zielvariablen pro Messzeitpunkt
p : Anzahl der Vorher-Messungen
r : Anzahl der Nachher-Messungen
ρ_{vor} : Korrelation zwischen Vorher-Messungen
ρ_{nach} : Korrelation zwischen Nachher-Messungen
ρ_{mix} : Korrelation zwischen Vorher- und Nachher-Messungen

In die Formel (2.14) ist dann anstelle von σ^2 *„var“* einzusetzen. Da

$$\left(\rho_{\text{nach}} + \frac{1-\rho_{\text{nach}}}{\text{r}} - \frac{\rho_{\text{mix}}^2}{\rho_{\text{vor}} + \frac{1-\rho_{\text{vor}}}{\text{p}}} \right) < 1$$

verringert sich die Fallzahl im Vergleich zu einem Design mit nur einer Messung.

Beim 2. Szenario werden die beiden Fälle „compound symmetry“ und „Autoregression 1. Art“ unterschieden. Beim ersten Fall liegt unabhängig vom zeitlichen Abstand zwischen den Messungen eine konstante Korrelation vor. Die resultierende optimale Auswertung ist eine Kovarianzanalyse mit der Steigung als abhängiger und der Vorher-Messung als unabhängiger Variable (d.h. bei jedem Patienten wird zunächst die Steigung in der Zielvariablen mittels KQ-Methode geschätzt). Die resultierende Formel für die Varianz, die anstelle von σ^2 in (2.14) einzusetzen ist, lautet hier:

$$\text{var} = \frac{\left[\sigma^2 \bullet (1-\rho) \bullet \sum_{j=0}^{k} \left(t_j - \bar{t} \right)^2 \right] - \left[\sigma^2 \bullet (1-\rho)^2 \bullet \left(t_0 - \bar{t} \right)^2 \right]}{\left[\sum_{j=0}^{k} \left(t_j - \bar{t} \right)^2 \right]^2} \tag{2.22}$$

mit: σ^2 : konstante Varianz der Zielvariablen pro Messzeitpunkt
$t_0, ..., t_k$: Messzeitpunkte (t_0: vor Randomisierung; t_i: nach Randomisierung $1 <= i <= k$; $t_k - t_0 = 1$)
ρ : konstante Korrelation zwischen den Messungen

Beim Fall der Autoregression 1. Art nimmt die Korrelation mit dem zeitlichen Abstand zwischen den Messungen exponentiell ab. Hier wird eine Kovarianzanalyse mit der letzten Messung als abhängige Variable und Vorher-Messung als unabhängige Variable empfohlen.
Um die Fallzahlplanung gemäß (Schouten 1999b) durchführen zu können, müssen auf der Grundlage von Vor-Studien also zunächst Annahmen über die Struktur des zu erwartenden Unterschiedes zwischen den Behandlungsgruppen getroffen werden. Anschließend müssen die in (2.21) bzw. (2.22) benötigten Parameter aus den Vor-Studien geschätzt werden.

Eine erweiterte Auswertungsstrategie von Repeated Measurements stellen multivariate Analyseverfahren dar, bei denen die Mehrfachmessungen (x_{i1}, ... , x_{ip}) eines jeden Patienten als multivariater Beobachtungsvektor x_i (allerdings ohne Berücksichtigung der zeitlichen Abfolge der Messungen) aufgefasst werden. Das multivariate Pendant zum univariaten t-Test ist Hotelling's T^2. Bei (Rochon 1991) finden sich für diese Auswertungsstrategie Tabellen für die Fallzahlplanung – für die beiden Kovarianzstrukturen compound symmetry und Autoregression, s.o.

Für komplexere Auswertungsstrategien, wie z.B. dem etablierten „2-stage-random-effects-model" nach (Laird und Ware 1982) liegen noch keine Fallzahlformeln vor.

3 Methoden der Fallzahladjustierung

„An interim check conducted on the blinded data may reveal that overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report. The steps taken to preserve blindness and the consequences, if any, for the type I error and the width of confidence intervals should be explained.“
aus: (ICH-E9 1998) §4.4 „Sample Size adjustment“

Die im vorigen Kapitel beschriebenen Methoden beruhen auf Annahmen, die zum Zeitpunkt der Studienplanung getroffen werden. Die Idee der Fallzahladjustierung besteht nun darin, diese Annahmen im Laufe der Studie zu überprüfen. Man spricht dann von internen Pilotstudien, im Gegensatz zu externen Pilotstudien, bei denen nach der Fallzahladjustierung eine neue, unabhängige Studie mit der aus der Pilotstudie resultierenden Fallzahl durchgeführt wird. Es handelt sich bei externen Pilotstudien somit um zwei getrennte Studien: eine Pilotstudie zur Ermittlung der benötigten Fallzahl und im Anschluss daran die eigentliche Studie. Diese sehr ressourcenaufwendige Vorgehensweise wird z.B. vorgeschlagen von (Browne 1995) und (Kieser und Wassmer 1996) – für den „klassischen“ Fall des Vergleichs zweier Gruppen hinsichtlich einer normalverteilten Zielgröße. Bei (Gould 1993) wird eine externe Pilotstudie benutzt, um in einer Äquivalenzstudie die Ereignisraten abzuschätzen. Auf diese Verfahren wird im Weiteren nicht näher eingegangen, da sie in der Praxis von sehr geringer Bedeutung sind.

Stellt sich z.B. heraus, dass man bei einem Zwei-Gruppen-Mittelwertsvergleich die Standardabweichung der Zielvariablen bei der Fallzahlplanung als zu niedrig angenommen hat, so hat dies massive Konsequenzen auf die benötigte Anzahl von Patienten. Nimmt man in diesem Fall keine Adjustierung der Fallzahl vor, so sinkt die Power der Studie, wie in der folgenden Grafik veranschaulicht wird:

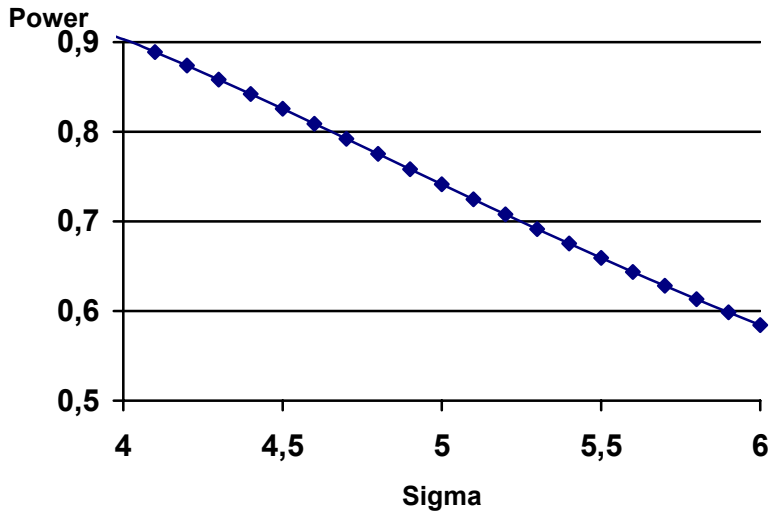


Abbildung 3.1: *Zusammenhang zwischen σ und Power bei einer Fallzahlplanung mit: α=5%, δ=2, n=85*

Hier wurde bei der Fallzahlplanung die Standardabweichung mit s_0=4 geschätzt und die Fallzahl für eine Power von 90% berechnet. Ist nun die tatsächliche Standardabweichung σ=6, so verringert sich die Power auf ca. 60% - sofern man an der ursprünglichen Fallzahl festhält. Somit ist das wesentliche Ziel aller Fallzahladjustierungsmethoden die Absicherung der Power der Studie

Alle Verfahren zur Fallzahladjustierung lassen sich in folgende Schritte aufteilen:

1. **Fallzahlplanung zu Beginn der Studie unter Verwendung aller verfügbaren Informationen → n_p**
2. **Rekrutierung eines „gewissen" Anteils der n_p Patienten: n_1**
3. **Überprüfung der bei der Fallzahlplanung getroffenen Annahmen – mittels der ersten n_1 Patientendaten**
4. **Berechnung einer neuen, adaptierten Fallzahl n_a**
5. **Weiterrekrutierung von Patienten, bis n_a erreicht ist**
6. **Abschließender Test**

Für den Fall des Mittelwertvergleiches zweier Gruppen mittels t-Test bedeutet dies, dass zunächst bei der Fallzahlplanung mit einer Standardabweichung s_0 gearbeitet wird. Zusammen mit den Annahmen bezüglich von α, β, und δ resultiert dann die Fallzahl n_p (s. 2.14). Nach $n_1 = q \bullet n_p$ Patienten erfolgt eine neue Schätzung für die Standardabweichung: s_1 . Damit wird die neue, adaptierte Fallzahl:

$$n_a = \left[n_p \left(\frac{s_1}{s_0} \right)^2 \right] + 1 \qquad (3.1)$$

Anschließend werden noch $n_2 = \max \left\{ 0 \; ; \; n_a - n_p \right\}$ Patienten rekrutiert – bis schließlich die Endauswertung mit allen n_a Patienten durchgeführt wird.

Bei (Stein 1945) *(im Folgenden wird diese Basis-Arbeit mit „Stein“ bezeichnet)* geht in den abschließenden Test (Schritt 6) dann nur die Standardabweichung der „ersten“ n_1 Patienten ein - und nicht die Standardabweichung aller Patienten. Beim Verfahren von (Wittes und Brittain 1990) *(wird im Folgenden mit „WB“ bezeichnet)* wird dagegen die Standardabweichung aller n_a Patienten für den Test benutzt. Ferner unterscheiden sich diese beiden klassischen Verfahren darin, dass WB nur eine Adjustierung nach oben erlauben ($n_a \geq n_p$) – während bei Stein auch eine Adjustierung nach unten möglich ist.

Der Ansatz von Stein

Im Detail besteht die Vorgehensweise nach Stein für den Zwei-Gruppen-Vergleich aus folgenden Schritten:

1. Fallzahlplanung auf Grundlage einer Schätzung von σ: s_0. Daraus resultiert n_p - geplante Gesamtzahl an Patienten, mit $n_{p,K}$ Fällen für die Kontrollgruppe und $n_{p,B}$ Fällen für die Behandlungsgruppe, also: $n_{p,K} = n_{p,B} = n_p/2$

2. Sammle n_1 Beobachtungen ($n_1/2$ pro Gruppe). *Dabei wird in der Arbeit von Stein kein Hinweis auf die Wahl von n_1 gegeben.*

3. Berechne die empirische Varianz s_1^2 (gepoolte Standardabweichung auf der Basis von $n_1/2$ Patienten pro Gruppe):

$$s_1 = \sqrt{\frac{s_{1,B}^2 + s_{1,K}^2}{2}}$$

4. Berechne:

$$n^* = 2\left[2\left(\frac{s_1}{\delta}\left(t_{1-\alpha/2,n_1-2} + t_{1-\beta,n_1-2}\right)\right)^2\right] \tag{3.2}$$

Die adjustierte Fallzahl ist nun:

$$n_a = \mathrm{Max}\left(n^*, n_1\right) \tag{3.3}$$

Somit ist hier auch eine Adjustierung nach unten möglich. Im Extremfall wird nach der ersten Phase die Rekrutierung beendet. Hier wird also die Fallzahlberechnung mit der t-Verteilung durchgeführt – der Unterschied zur Fallzahlplanung mittels Normalverteilung, analog zu (2.14) ist für $n_1 > 20$ marginal.

5. Sammle weitere $n_2 = n_a - n_1$ Beobachtungen.

6.1 Berechne die Teststatistik

$$t = \frac{\left|\bar{X}_B - \bar{X}_K\right| \sqrt{n_a}}{2\, s_1}$$

mit: $\bar{X}_B, \bar{X}_K$: Mittelwert in der Behandlungs- bzw. Kontrollgruppe nach Abschluss der gesamten Studie (n_a Beobachtungen)

s_1 : gepoolte Standardabweichung auf der Basis von n_1 Beobachtungen

Anders als in der klassischen Situation geht hier nur die beobachtete Standardabweichung aus der 1. Phase der Studie ein.

6.2 Lehne H_0 ab, falls $t > t_{n1-2,\alpha/2}$.

Anders als in der klassischen Situation wird hier die Teststatistik mit der t-Verteilung mit n_1-2 DF (entspricht dem Umfang der ersten Phase) verglichen.

Der Ansatz von Wittes&Brittain

Die entsprechenden Details bei der Methode von WB lauten:

1. Fallzahlplanung auf Grundlage einer Schätzung von σ: s_0. Daraus resultiert n_p - geplante Gesamtzahl an Patienten mit $n_{p,K}$ Fällen für die Kontrollgruppe und $n_{p,B}$ Fällen für die Behandlungsgruppe, also: $n_{p,K} = n_{p,B} = n_p/2$.

2. Sammle $n_1 = q \bullet n_p$ Beobachtungen pro Gruppe ($0 < q < 1$). *Dabei wird in der Arbeit von WB kein Hinweis auf die Wahl von q gegeben.*

3. Berechne die empirische Varianz s_1^2 (gepoolte Standardabweichung auf der Basis von n_1 Patienten pro Gruppen):

$$s_1 = \sqrt{\frac{s_{1,B}^2 + s_{1,K}^2}{2}}$$

4. Berechne

$$n_a = \text{Max}\left(\left[n_p\left(\frac{s_1}{s_0}\right)^2\right] + 1, n_p\right) \qquad (3.4)$$

   Somit ist hier nur eine Adjustierung nach oben möglich. Es werden mindestens so viele Patienten rekrutiert, wie ursprünglich geplant.

5. Sammle weitere $n_2 = n_a - n_1$ Beobachtungen.

6.1 Berechne die Teststatistik:

$$t = \frac{\left|\bar{X}_B - \bar{X}_K\right|\sqrt{n_a}}{2s_a}$$

mit:

$$s_a = \sqrt{\frac{s_{a,B}^2 + s_{a,K}^2}{2}}$$

(gepoolte Standardabweichung auf der Basis von n_a Patienten pro Gruppen)

Anders als bei Stein geht hier die beobachtete Standardabweichung aus der gesamten Studie ein.

6.2 Lehne H_0 ab, falls $t > t_{na-2,\alpha/2}$.

Ausgehend von diesen beiden klassischen Verfahren existiert eine Vielzahl von Vorschlägen zur Fallzahladjustierung. Die Unterschiede zwischen den Verfahren liegen dabei im Wesentlichen in den Adjustierungsregeln, in den Zeitpunkten der Adjustierung, in der „Blindheit" der Adjustierung und in den zugrundeliegenden Testsituationen. In (Gould 2001) findet sich eine aktuelle Übersicht über existierende Verfahren der Fallzahladjustierung – unter besonderer Berücksichtigung der praktischen Anwendbarkeit.

Darüber hinaus existieren zahlreiche Verfahren, bei denen die Fallzahladjustierung mit einer Zwischenauswertung kombiniert wird, z.B.: (Bauer und Köhne 1994), (Case, Morgan et al. 1987), (Guggerli, Maurer et al. 1993), (Proschan und Hunsberger 1995), (Fisher 1998), (Gould und Shih 1998), (Betensky und Tierney 1997) - auf diese Methoden wird im Rahmen dieser Arbeit nicht näher eingegangen.

3.1 Beurteilungskriterien

Wie ist die Güte einer Fallzahladjustierungsmethode zu beurteilen?
Die Fallzahl setzt sich zusammen aus einer vorgegebenen Stichprobengröße (= n_1) – für den ersten Teil der Studie - und einer zufälligen Komponente. Sie wird somit zu einer Zufallsvariablen, die von der zum Zeitpunkt der Fallzahladjustierung beobachteten Kenngröße (z.B. der Standardabweichung) abhängt. Die Verteilung dieser Zufallsvariablen stellt ein wesentliches Beurteilungskriterium der Güte einer Fallzahladjustierung dar. Für den einfachen Fall des Mittelwertvergleiches zweier Gruppen mittels t-Test lässt sich die Verteilung der adjustierten Fallzahl N_a wie folgt herleiten:

$$\begin{aligned} P\left(N_a \le n\right) &= P\left(\frac{S_1^2}{s_0^2} n_p \le n\right) \\ &= P\left(S_1^2 \le \frac{n s_0^2}{n_p}\right) \\ &= P\left(\frac{S_1^2 \left(n_1 - 2\right)}{\sigma^2} \le z\right), \text{mit } z = \frac{n s_0^2 \left(n_1 - 2\right)}{n_p \sigma^2} \\ &= F_{\chi^2_{n_1-2}}(z) \end{aligned} \tag{3.5}$$

Dabei hängt z natürlich insbesondere von der unbekannten wahren Standardabweichung σ ab, bzw. von dem Verhältnis s_0/σ.

Da es sich bei N_a um eine diskrete Zufallsvariable handelt, gilt:

$$P(N_a = n) = P(N_a \leq n) - P(N_a \leq n-1) \tag{3.6}$$

Außerdem muss berücksichtigt werden, dass für $n < n_1$ gilt:

$$P(N_a=n) = 0$$

- da die adjustierte Stichprobe aus mindestens n_1 Patienten besteht.

Somit können die relevanten Kenngrößen für die Verteilung von N_a , wie $E(N_a)$ oder $MSE(N_a)$ aus dieser Verteilung hergeleitet werden.

In den folgenden Abbildungen sind die Verteilungen von N_a für die 3 Szenarien $s_0 = \sigma$, $s_0 < \sigma$ und $s_0 > \sigma$ dargestellt. Dabei wurde von folgenden Fallzahlplanungsparametern ausgegangen:

- α=0.05
- 1-β=0.9
- δ=1

Die nach (2.14) resultierende Fallzahl ist somit n_p= 2• 86=172. Die Fallzahladjustierung wird nach der Hälfte der geplanten Patienten durchgeführt, also: n_1=86.

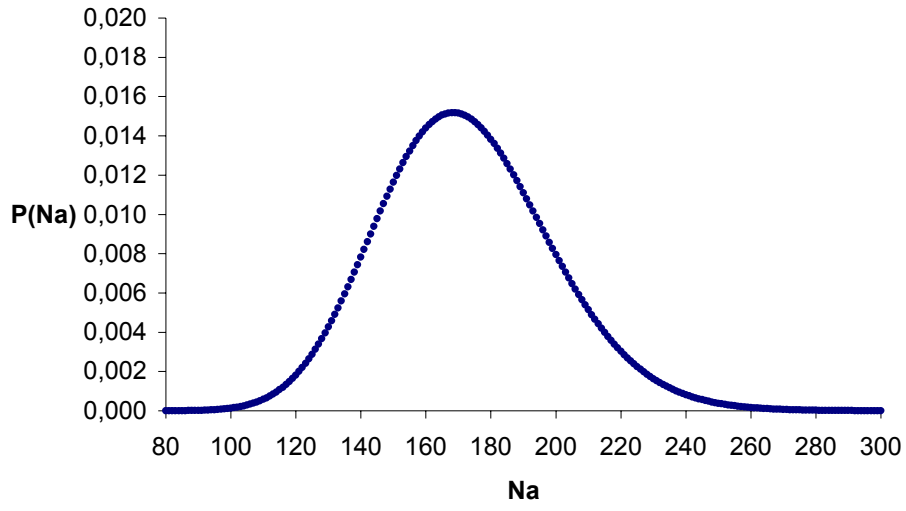


Abbildung 3.2: *Verteilung von N_a bei: $s_0 = \sigma = 2$ - $\alpha = 0.05$, $\beta=0.1$, $\delta=1$, $n_p=172$, $n_1=86$*

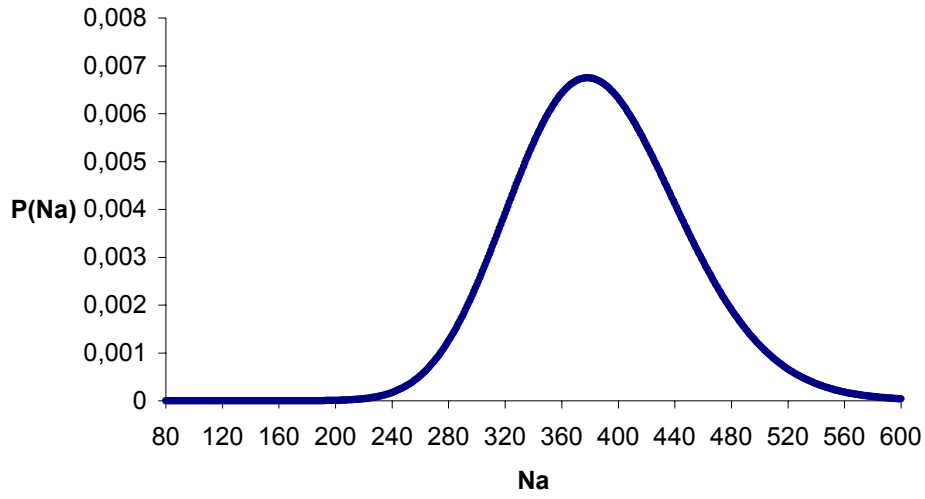


Abbildung 3.3: *Verteilung von N_a bei: s_0=2 < σ=3 - α = 0.05, β=0.1, δ=1, n_p=172, n_1=86*

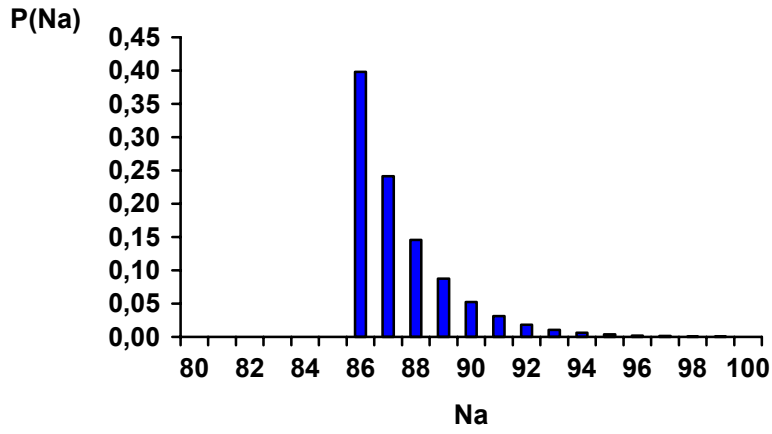


Abbildung 3.4: *Verteilung von N_a bei s_0 =2 > σ=1 - α = 0.05, β=0.1, δ=1, n_p=172, n_1=86*

Die Fallzahladjustierung hat Konsequenzen für die Verteilung der abschließenden Teststatistik, da die Verteilung der Fallzahl berücksichtigt werden muss. Daher liegt es nahe, die Auswirkungen der Fallzahladjustierung auf den Fehler erster Art zu untersuchen. Im Fall der oben erwähnten Methode von WB bedeutet dies konkret, dass bei der Verteilung der Teststatistik T auch die Verteilung der Standardabweichung S_1 berücksichtigt werden muss. Dabei geht die Dichte von T aus dem Produkt einer Normalverteilung mit Erwartungswert 0 (für δ), einer χ^2-Verteilung mit n_1-1 DF (für S_1) und einer χ^2-Verteilung mit n_2-1 DF (für S_2) hervor. S_2 ist die in der 2. Studienhälfte – nach der Adjustierung – beobachtete gepoolte Standardabweichung. Insbesondere hängt hier n_2 von s_1 ab. Exakte Formeln zur Herleitung der Dichte und somit zur Berechnung des tatsächlichen alphas finden sich bei (Gould und Shih 1992). Mit Hilfe von Simulationen zeigten WB, dass der Fehler 1. Art bei ihrem

Verfahren nur unwesentlich vergrößert wird – insbesondere nimmt diese Inflation mit größer werdendem n_1 ab. (Kieser und Friede 2000) führen eine Niveauadjustierung durch, so dass beim abschließenden Test das Niveau α nicht überschritten wird – so muss man beim t-Test und einer Fallzahladjustierung nach n_1=100 Patienten mit einem alpha von 0.0475 arbeiten um das gewünschte Niveau von 0.05 genau einzuhalten. Weitere Details hierzu finden sich auch in Kapitel 5.1.1.

Da das Ziel einer Fallzahladjustierung die Absicherung der Power ist, stellt natürlich die durch die Fallzahladjustierung tatsächlich erreichte Power ein weiteres wesentliches Beurteilungskriterium dar. Dabei ist wiederum zu berücksichtigen, dass die Power von den zum Zeitpunkt der Adjustierung vorliegenden Daten abhängt und somit auch zu einer Zufallsvariablen wird.

Neben dem Testen, ob eine neue Therapie wirksam ist, ist auch das Schätzen des Therapieeffektes im Rahmen von klinischen Studien von großer Bedeutung. Daher sind die Auswirkungen einer Fallzahladjustierung auf die Breite der zugehörigen Konfidenzintervalle ebenfalls von Interesse.

Da eine Fallzahladjustierung immer einen Zwischenauswertungscharakter hat, ist die vertrauliche Behandlung der Ergebnisse auch hier von großer Bedeutung. Schließlich werden Daten der ersten Patienten ausgewertet. Somit sind Verfahren, die keine Entblindung der Daten notwendig machen, von großem Vorteil - so dass als weiteres Beurteilungskriterium die „Blindheit“ der Methode berücksichtigt wird.

Die hier beschriebenen Beurteilungskriterien werden im folgenden Kapitel für die jeweiligen Verfahren angewandt.

3.2 Einteilung der unterschiedlichen Methoden

Die Vielzahl der existierenden Fallzahladjustierungsmethoden lassen sich in folgende Kategorien einteilen:
Methoden, die die **Entblindung** bezüglich der Behandlungsgruppe benötigen vs. Methoden, die ohne Entblindung auskommen.

Darüber hinaus unterscheiden sich die Verfahren hinsichtlich der zugrundeliegenden **Testsituation**, d.h. im Wesentlichen gibt es die beiden Kategorien: Test bei normalverteilten Daten / Test bei dichotomen Daten.

Alle Verfahren gehen von der Schätzung eines für die Fallzahlplanung benötigten Parameters, wie z.B. der Varianz im Fall von normalverteilten Daten, aus. Dabei gibt es Unterschiede hinsichtlich der **Art der Schätzung**: entweder werden Punktschätzer oder aber Intervallschätzer (Grenzen eines Konfidenzintervalls für den Punktschätzer) benutzt.

Schließlich gibt es eine ganze Reihe von unterschiedlichen Strategien bezüglich des **Zeitpunktes der Fallzahladjustierung**. Die „Standard-Strategie“ sieht eine Fallzahladjustierung nach der Hälfte der ursprünglich vorgesehenen Fallzahl n_P vor.

4 Zeitpunkt der Fallzahladjustierung

Welches ist der optimale Zeitpunkt zur Durchführung einer Fallzahladjustierung?
Bei der Beantwortung dieser Frage müssen folgende Punkte berücksichtigt werden:

Der Zusammenhang zwischen geplanter Fallzahl n_P und der zur Fallzahladjustierung benutzten Fälle n_1 sei:

$$n_1 = q \bullet n_P, \text{ mit: } 0 < q \leq 1 \tag{4.1}$$

Die der Fallzahladjustierung zugrunde liegende Schätzung eines Fallzahlplanungsparameters, wie z.B. der Standardabweichung einer normalverteilten Zielvariablen, sollte möglichst „stabil“ sein. Stellt nun diese „Stabilität“ das einzige Optimalitätskriterium dar, so läuft es auf einen möglichst späten Zeitpunkt hinaus, d.h. q=1. Dies bedeutet insbesondere bei der Strategie von WB, bei der eine Adjustierung nur nach oben vorgesehen ist, dass man die Fallzahladjustierung möglichst spät durchführen sollte.
(Sandvik, Erikssen et al. 1996) geben folgendes Optimalitätskriterium für den Fall normalverteilter Daten an: Eine vorgegebene Wahrscheinlichkeit W, dass mehr Patienten in der internen Pilotstudie sind als in der gesamten Studie überhaupt notwendig wären, darf nicht überschritten werden. Zusammen mit der Anzahl der Beobachtungen n_0, die für die Schätzung der Standardabweichung bei der Fallzahlplanung benutzt wurden, kann dann das entsprechend optimale q bestimmt werden. So ist z.B. für W=5% und n_0=50 das optimale q=0.74. (Singer 1999) zeigt, wie dieses Verfahren noch verbessert werden kann, wenn man auch die Rekrutierungszeit und –geschwindigkeit berücksichtigt.
Auch bei (Denne und Jennison 1999) wird die (Un)-Sicherheit der Schätzung s_0 zur Herleitung eines optimalen Zeitpunktes benutzt. Lässt sich diese Zuverlässigkeit im Vergleich zur wahren, unbekannten Standardabweichung σ in der Form

$$\sigma^2/\lambda < s_0^2 < \lambda\sigma^2$$

ausdrücken, so lautet die Empfehlung für n_1:

$$n_1 = 4\left(\frac{s_0/\lambda}{\delta}\left(z_{1-\alpha/2} + z_{1-\beta}\right)\right)^2 \tag{4.2}$$

Somit hängt die Größe der ersten Studienphase von der unteren Grenze der erwarteten Standardabweichung ab. Besteht also eine große Unsicherheit bezüglich der Schätzung der Standardabweichung ($\lambda >> 1$), sollte die Fallzahladjustierung sehr frühzeitig durchgeführt werden.

(Birkett und Day 1994) benutzen kein Optimalitätskriterium, sondern ermitteln aufgrund von Simulationen die Regel: $n_1 > 20$ – also unabhängig von n_p. Bei dieser Strategie kann sogar auf die Berechnung von n_p verzichtet werden, d.h. die Fallzahlplanung der Studie wird erst nach den ersten n_1 Patienten durchgeführt. Zu einem ähnlichen Ergebnis kommen (Jennison und Turnbull 2000) in Kapitel 14.
(Zucker, Wittes et al. 1999) empfehlen dagegen, auf der Grundlage eines Vergleiches mehrerer Adjustierungsstrategien: $n_1 > 40$ - während (Wittes, Schabenberger et al. 1999) für den Fall „Adjustierung nur nach oben“ zu dem Schluss kommen: $n_1 > 30$. Für den Fall „Adjustierung nach oben oder unten“ sollte $q > 0.4$ sein. Bei beiden Strategien führt ein kleines q von 0.1 zu einer geringen Power - diese steigt mit größer werdendem q. Als Daumenregel für q wird hier abschließend $0.25 < q < 0.75$ angegeben.
(Coffey und Muller 1999) untersuchen eine ganze Reihe unterschiedlicher Szenarien bezüglich der Adjustierungsregeln im Hinblick auf die Einhaltung des Fehlers 1. Art – mit dem Ergebnis, dass die optimale Wahl des Zeitpunktes von der konkreten Situation abhängt. So ist die pauschale Aussage von (Birkett und Day 1994) „$n_1 > 20$“ nicht richtig. Insbesondere bei der Strategie, dass auch eine Adjustierung unterhalb der geplanten Fallzahl möglich ist, kann diese Regel zu einer Inflation des alphas führen.

Eine Fallzahladjustierung kann erst dann durchgeführt werden, wenn die Daten aller n_1 Patienten zur Auswertung vorliegen. Dies bedeutet in der Praxis klinischer Studien, dass folgende Schritte durchgeführt werden müssen:

- Rekrutierung / Randomisierung von n_1 Patienten
- Erhebung der Zielvariablen bei den n_1 Patienten
- Dateneingabe für n_1 Patienten
- Daten-Cleaning für n_1 Patienten

In der Zeitspanne t_A zwischen der Rekrutierung des n_1-ten Patienten und der Beendigung der Fallzahladjustierung läuft die Rekrutierung unterdessen weiter, wie im folgenden Bild veranschaulicht wird:

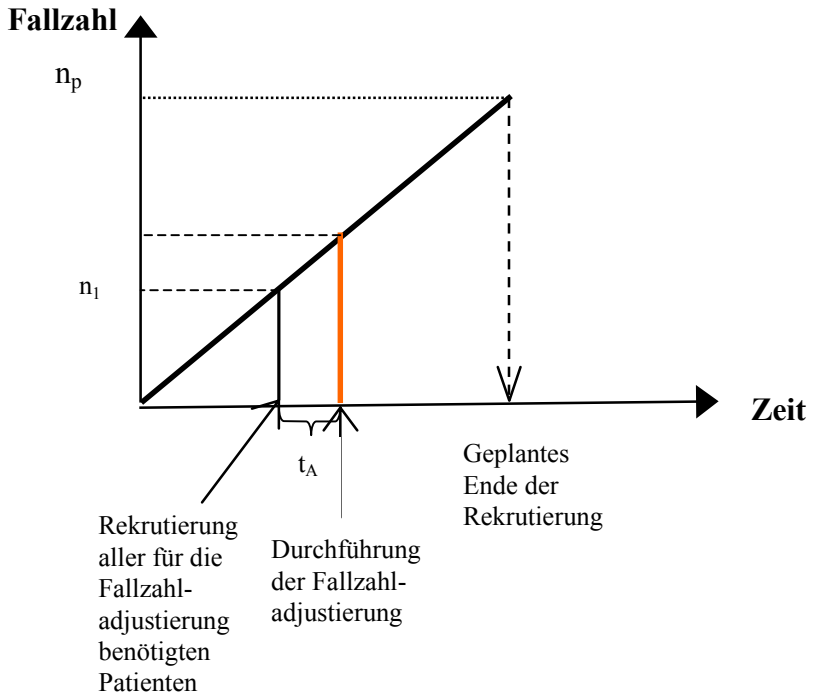


Abbildung 4.1: *Zeitlicher Ablauf der Rekrutierung und Fallzahladjustierung*

Unter logistischen Gesichtspunkten sollte die Fallzahladjustierung noch während der geplanten Rekrutierungszeit abgeschlossen sein, so dass für q gilt: $q < 1$. Dabei ist die Obergrenze von q von der studienspezifischen Zeit t_A und der geplanten Rekrutierungszeit abhängig. (Bolland, Sooriyarachchi et al. 1998) berichten von einer Studie, bei der eine geringe Rekrutierungsgeschwindigkeit zusammen mit einer langen Beobachtungszeit bis zum Erreichen der Zielvariable dazu führt, die Fallzahladjustierung spätestens nach 25% der geplanten Patienten durchführen zu müssen. Eine Erhöhung des q hätte hier die Beendigung der Rekrutierungszeit bedeutet.

In der Praxis klinischer Studien werden Fallzahladjustierungen üblicherweise nach der Hälfte der geplanten Patienten durchgeführt, d.h. $q = 0.5$. Diese Strategie erscheint sowohl unter theoretischen als auch unter logistischen Aspekten als sinnvoll für die meisten Studientypen.

5 Fallzahladjustierung beim Vergleich von Mittelwerten

Die in diesem Kapitel vorgestellten Verfahren gehen von folgender, im Rahmen von klinischen Studien sehr häufig vorzufindenden, Situation aus: Verglichen werden zwei Gruppen (Behandlungsgruppe vs. Kontrollgruppe) hinsichtlich einer normalverteilten Zielvariablen X:

$\mathbf{X_B \sim N(\mu_B,\sigma^2)}$ $\mathbf{X_K \sim N(\mu_K,\sigma^2)}$ **-** mit unbekanntem σ

Getestet wird:

$$H_0 : \mu_B = \mu_K \qquad \text{vs.} \qquad H_1 : \mu_B \neq \mu_K \tag{5.1}$$

Die Teststatistik des zugehörigen t-Tests lautet:

$$T = \frac{\overline{X_B} - \overline{X_K}}{s_p} \bullet \sqrt{\frac{n_B n_K}{n_B + n_K}} \tag{5.2}$$

mit :

$$s_p = \sqrt{\frac{(n_B - 1)s_B^2 + (n_K - 1)s_K^2}{n_B + n_k - 2}} \tag{5.3}$$

„gepoolte“ Standardabweichung

Für $n_B = n_K$ vereinfacht sich die Teststatistik zu:

$$T = \frac{\left(\bar{X}_B - \bar{X}_K\right)\sqrt{n}}{2\, s_p} \tag{5.4}$$

mit: $n = n_B + n_K$.

Die Teststatistik T ist unter H_0 zentral t-verteilt mit n-2 Freiheitsgraden. Somit wird bei einem vorgegebenen Signifikanzniveau α die Nullhypothese abgelehnt, falls $|t| > t_{crit}$ – mit t_{crit}= (1-α/2)-Quantil der t-Verteilung mit n-2 Freiheitsgraden.

Alle Fallzahladjustierungsverfahren haben zum Ziel, die Standardabweichung im Laufe der Studie zu schätzen, um dann mit dieser Schätzung die Fallzahl zu korrigieren. Dabei gibt es hinsichtlich der Adjustierung folgende Strategien:

Adjustierung nur nach oben

Bei dieser Strategie (s. z.B. WB) ist die ursprünglich geplante Fallzahl die Untergrenze für die adjustierte Fallzahl, d.h.

$$n_{a,o} = \max\{n_p \ , n(s_1)\} \tag{5.5}$$

mit: n_p : geplante Fallzahl zu Beginn der Studie

$n(s_1)$: Fallzahl auf der Basis der im Laufe der Studie neu geschätzten Standardabweichung s_1

$n_{a,o}$: nach oben adjustierte Fallzahl

Adjustierung nach oben oder unten

Bei dieser Strategie (s. z.B. Stein, (Birkett und Day 1994)) ist die bei der Fallzahladjustierung vorliegende geplante Fallzahl n_1 die natürliche Untergrenze für die adjustierte Fallzahl, d.h.

$$n_{a,o/u} = \max\{n_1 \ , n(s_1)\} \tag{5.6}$$

Adjustierung unter Berücksichtigung einer Obergrenze

Ausgangspunkt für diese Strategie(s. z.B. (Coffey und Muller 1999)) ist eine vorgegebene Obergrenze n_{max} (z.B. $n_{max} = 2 \bullet n_p$) für die adjustierte Fallzahl, d.h.

$$n_{a,max} = \min\{n_{max} \ , n(s_1)\} \tag{5.7}$$

Diese Strategie lässt sich natürlich mit den beiden anderen Strategien kombinieren.

Adjustierung in Abhängigkeit von Ober/Untergrenzen für s_1

Hier besteht die Idee darin, erst dann zu adjustieren, wenn die beobachtete Standardabweichung s_1 „wesentlich“ von der zur Fallzahlplanung benutzten Standardabweichung s_0 abweicht:

$$n_{a,s} = \begin{cases} \max\{n_1, n(s_1)\} & - \text{falls } s_1 < a \bullet s_0 \\ n_p & - \text{falls } a \bullet s_0 \le s_1 \le b \bullet s_0 \\ n(s_1) & - \text{falls } s_1 > b \bullet s_0 \end{cases} \qquad (5.8)$$

mit geeignet zu wählenden $a < 1$ und $b > 1$.
Auch hier lassen sich die im Vorfeld vorgestellten Strategien entsprechend integrieren.

5.1 Verfahren für entblindete Daten

Die im Folgenden diskutierten Verfahren setzen voraus, dass zum Zeitpunkt der Fallzahladjustierung die Information über die Gruppenzugehörigkeit für jeden Patienten vorliegt. Dabei reicht es aus, dass diese Information in der Form „Gruppe 1“ bzw. „Gruppe 2“ besteht – ohne Kenntnis darüber, welche der beiden Gruppen (1 / 2) nun die Behandlungsgruppe und welche die Kontrollgruppe ist. Die Fallzahladjustierung beruht somit auf partiell entblindeten Daten.
Die im Kapitel 3 bereits angesprochenen Arbeiten von Stein und WB bilden dabei die Grundlage für eine ganze Reihe von Weiterentwicklungen:
(Birkett und Day 1994) nehmen wie Stein eine Adjustierung nach oben oder unten vor; der abschließende Test beruht jedoch bezüglich der Standardabweichung auf allen n_a Beobachtungen. Bei (Browne 1995) und (Kieser und Wassmer 1996) besteht die Idee darin, nicht den Punktschätzer für die Standardabweichung sondern die obere Grenze eines einseitigen $100 \bullet (1-\gamma)$ Konfidenzintervalls für die Schätzung zu benutzen. Dabei garantiert diese Wahl, dass die geplante Power der Studie mit Wahrscheinlichkeit $(1-\gamma)$ eingehalten wird. Diese Arbeiten gehen von einer einseitigen Fragestellung und einer externen Pilotstudie aus. Bei (Kieser und Friede 2000) wird dieses Verfahren dann auf interne Pilotstudien übertragen.

(Denne und Jennison 1999) versuchen folgende Defizite bei Stein und WB auszugleichen: Bei Stein wird die in der zweiten Studienhälfte gesammelte Information bezüglich der Standardabweichung nicht für den abschließenden Test benutzt; während bei WB das Verwenden dieser Information dazu führt, dass insbesondere das vorgegebene

Signifikanzniveau nicht genau eingehalten wird. Dazu wird der abschließende Test mit der gepoolten Standardabweichung auf der Basis aller n_a Patienten durchgeführt; jedoch wird eine Adjustierung der Freiheitsgrade vorgenommen. Diese Adjustierung lautet:

$$2 \bullet (n_1 + \varepsilon\,(n_a - n_1) - 1)$$

Der Parameter ε steuert somit den Abstand zwischen der Strategie von Stein (ε=0) und der Strategie von WB (ε=1). Auf der Grundlage von Simulationen empfehlen die Autoren: ε=1/8.

(Coffey und Muller 1999) und (Coffey und Muller 2001) führen eine Weiterentwicklung von WB auf den allgemeinen Fall von linearen Modellen durch. Das zugrundeliegende Modell lautet dann:

$$Y = X\,ß + \varepsilon$$

mit: $Y \in \mathbb{R}^n$ – Zielvariable

$X \in \mathbb{R}^{n \times p}$ – Designmatrix

$ß \in \mathbb{R}^p$ – zu schätzende Parameter

$\varepsilon \in \mathbb{R}^n$ – Fehler: $\varepsilon \sim N(0,\sigma^2)$

Für dieses Modell wird eine Fallzahladjustierung entwickelt, die sich an WB orientiert, d.h. die Schätzung der Fehlerquadratsumme wird zur Korrektur der Fallzahl benutzt. Dabei werden folgende Szenarien untersucht: „Adjustierung nur nach oben“ / „Adjustierung nach oben oder unten“ / „Adjustierung nur nach oben – mit vorgegebener Obergrenze“.

5.1.1 Auswirkungen auf alpha

Wie bereits im Kapitel 3 erwähnt, haben WB aufgrund von Simulationen festgestellt, dass ihr Verfahren zu einer leichten Inflation des Signifikanzniveaus führt. Dies liegt daran, dass bei der Verteilung der Teststatistik T auch die Verteilung der Standardabweichung S_1 berücksichtigt werden muss, s. dazu auch (Gould und Shih 1992) und (Kieser und Friede 2000).

Theoretische Betrachtungen

Im Folgenden werde ich mich mit der einseitigen Testsituation beschäftigen, d.h.

$$H_0 : \mu_B = \mu_K \text{ vs. } H_1 : \mu_B > \mu_K$$

Die resultierenden Ergebnisse können entsprechend auf die zweiseitige Testsituation übertragen werden.

Die Teststatistik T basiert auf n_a Beobachtungen. Für den Fall gleicher Gruppengrößen $n_B = n_K = n_a/2$ gilt somit:

$$T = \frac{\left(\overline{X_B} - \overline{X_K}\right) \bullet \sqrt{n_a}}{2\, s_p} \tag{5.9}$$

mit:

$$s_p = \sqrt{\frac{s_B^2 + s_K^2}{2}} \tag{5.10}$$

Sei

$$d := \overline{x_B} - \overline{x_K}$$

die mittlere Differenz auf der Basis von allen n_a Beobachtungen, dann gilt:

$$d = \frac{n_1 \bullet d_1 + n_2 \bullet d_2}{n_a} \tag{5.11}$$

mit: d_i : mittlere Differenz für Studienphase $i \in \{1, 2\}$

$n_a = n_1 + n_2$ – d.h. nach der Adjustierung auf der Basis von n_1 Beobachtungen werden weitere n_2 Patienten rekrutiert

Betrachtet man nun anstelle von T die Teststatistik

$$T^* = \frac{d \bullet \sqrt{n_a}}{2 \bullet s_p^*} \tag{5.12}$$

mit:

$$s_p^* = \sqrt{\frac{(n_1-2)\bullet s_1^2 + (n_2-2)\bullet s_2^2}{n_a-4}}$$
$$s_1 = \sqrt{\frac{\left(s_{B,1}^2 + s_{K,1}^2\right)}{2}},\ s_2 = \sqrt{\frac{\left(s_{B,2}^2 + s_{K,2}^2\right)}{2}} \qquad (5.13)$$

(gepoolte Standardabweichung aus der Studienphase i=1, 2)

so lässt sich T* - entsprechend der beiden Studienphasen – schreiben als:

$$T^* = \frac{\frac{n_1\bullet d_1 + n_2\bullet d_2}{n_a}\ \sqrt{n_a}}{2\ \bullet\sqrt{\frac{(n_1-2)\bullet s_1^2 + (n_2-2)\bullet s_2^2}{n_a-4}}} \qquad (5.14)$$

Nun ist zu beachten, dass n_2 von s_1 abhängt, und zwar in der Form:

$$n_2 = \left(\max\left\{n_p, \left[n_p \bullet \left(\frac{s_1}{s_0}\right)^2\right] + 1\right\}\right) - n_1 \qquad (5.15)$$

Weiter gilt:

S_1^2 ist χ^2 – verteilt mit n_1-2 Freiheitsgraden.

Bedingt unter S_1^2 ist S_2^2 χ^2 – verteilt mit n_2-2 Freiheitsgraden.

Bedingt unter S_1^2 ist D normalverteilt mit Erwartungswert $(\mu_B-\mu_K)$ und Varianz:

$4\ \sigma^2 / (n_1+n_2)$

Dann gilt für die gemeinsame Verteilung von D, S_1^2 und S_2^2:

$$f(d, s_1^2, s_2^2) = g_{N\left(\mu_B-\mu_K, \left(4\sigma^2/(n_1+n_2)\right)\right)}(d) \bullet g_{\chi^2_{n_1-2}}\left(\frac{(n_1-2)\bullet s_1^2}{\sigma^2}\right) \bullet g_{\chi^2_{n_2-2}}\left(\frac{(n_2-2)\bullet s_2^2}{\sigma^2}\right) \qquad (5.16)$$

mit: $g_{N(\mu,\sigma 2)}$: Dichte der Normalverteilung mit Erwartungswert μ und Varianz σ^2

$g_{\chi^2_n}$: Dichte einer χ^2 – Verteilung mit n Freiheitsgraden

Durch die Vereinfachungen:

$$u = \sqrt{\frac{n_1 + n_2}{4}} \bullet \frac{d}{s_p^*}$$

$$\upsilon_i = \frac{(n_i - 2) \bullet s_i^2}{\sigma^2}, i = 1,2$$

lässt sich für die Dichte unter H_0 schreiben:

$$f(u, \upsilon_1, \upsilon_2 \mid H_0) = g_{N(0,(n_1+n_2-4)/(\upsilon_1+\upsilon_2))}(u) \bullet g_{\chi^2_{n_1-2}}(\upsilon_1) \bullet g_{\chi^2_{n_2-2}}(\upsilon_2) \tag{5.17}$$

Für den tatsächlichen Fehler 1. Art gilt dann:

$$\alpha = \int_0^\infty \int_0^\infty \int_{t_{n_1+n_2-4,1-\alpha}}^\infty f(u, \upsilon_1, \upsilon_2 \mid H_0)\, du\, d\upsilon_2\, d\upsilon_1 \tag{5.18}$$

Die konkrete Berechnung dieses Terms wurde bei (Kieser und Friede 2000) für den Fall „Adjustierung nach oben oder unten“ mittels numerischer Integration durchgeführt. Dabei ist zu beachten, dass α sowohl von dem unbekannten σ^2 als auch von dem studien-spezifischen n_1 abhängt. Für jedes n_1 gibt es jedoch eine eindeutige Obergrenze für α - diese liegt insbesondere für kleine n_1 erheblich über dem geplanten α von 0.05. Die Autoren empfehlen nun, das α für den abschließenden Test so zu wählen, dass das tatsächliche α genau eingehalten wird. So ist z.B. bei n_1=10 mit einem adjustierten α von 0.0387 zu arbeiten um ein tatsächliches α von 0.05 zu garantieren. Mit größer werdenden n_1 nähern sich das adjustierte α und das tatsächliche α mehr und mehr an.

Diese Ergebnisse basieren auf dem „t*-Test“, der anstelle von der üblichen über die beiden Behandlungsgruppen gepoolten Standardabweichung S_p mit der über die beiden Studienphasen gepoolten Standardabweichung S_p^* arbeitet. Der t*-Test ist nicht äquivalent zum t-Test, da $s_p^* \neq s_p$. Jedoch kann man mit Hilfe der Ergebnisse für den t*-Test Obergrenzen auch für den t-Test angeben – diese resultieren aus dem Zusammenhang zwischen S_p^* und S_p .

Zu ähnlichen Ergebnissen kommen (Wittes, Schabenberger et al. 1999) und (Zucker, Wittes et al. 1999): Hier werden die Strategien „Adjustierung nur nach oben“ und „Adjustierung nach oben oder unten“ verglichen – mit der abschließenden Empfehlung, dass für die meisten Situationen ein adjustiertes α von 0.047 ausreicht, um ein tatsächliches α von 0.05 zu erreichen.

(Coffey und Muller 1999) kommen für diese beiden Strategien zu dem Ergebnis, dass bei der „Adjustierung nach oben oder unten“ bei einem kleinen n_1 (n_1=10) eine starke Inflation des alphas resultiert (bis zu α=0.065) – unabhängig von der tatsächlichen Standardabweichung σ. Bei der Adjustierung nur nach oben findet diese Inflation auf niedrigerem Niveau statt und ist für den Fall einer starken Unterschätzung von σ ($s_0 << \sigma$) am größten (α=0.056).

Simulationen

Ausgehend von einer Fallzahlplanung mit folgenden Parametern:

- α=5% (zweiseitig)
- 1-β=90%
- μ_1=0, μ_2=1 $\Rightarrow$ δ=1
- s_0=$\sqrt{2}$
- $\Rightarrow$ n_0=2 • 43=86

habe ich mittels SAS® Simulationen bezüglich der Adjustierungsmethoden:

- Stein (**ST**) – „Adjustierung nach oben oder unten“
- WB (**WB**) – „Adjustierung nur nach oben“
- (Birkett und Day 1994) (**BD**) – „Adjustierung nach oben oder unten“

im Vergleich zur Strategie

- „keine Adjustierung“ (**KA**)

durchgeführt.

Dabei habe ich folgende Szenarien bzgl. σ untersucht:

- σ=1 < s_0
- σ=$\sqrt{2}$ = s_0
- σ=2 > s_0

Ferner habe ich bezüglich des Zeitpunktes der Adjustierung die beiden Situationen:

- p=0.5$\Rightarrow$ n_1=2 • 22=44
- p=0.2 $\Rightarrow$ n_1=2 • 9=18

betrachtet.

Insgesamt wurden pro Szenario 100.000 Simulationen durchgeführt – als Motivation für diese Anzahl diente folgende Überlegung: Die Angabe eines 95% Konfidenzintervall für α bzw.

1-β sollte mit folgender Genauigkeit (maximale Länge des zweiseitigen Konfidenzintervalls) erfolgen: 0.001 für α / 0.002 für 1-β. Das Design der Simulationen beruht dabei im Wesentlichen auf der in der Arbeit von WB beschriebenen Vorgehensweise. Das verwendete Programm zur Durchführung dieser Simulationsstudie findet sich im Anhang.

Es wurden somit für folgende 12 Szenarien Simulationen durchgeführt:

Szenario	**q**	**H_0 / H_1**	**σ**
1	0.5	H_0	1
2			√2
3			2
4		H_1	1
5			√2
6			2
7	0.2	H_0	1
8			√2
9			2
10		H_1	1
11			√2
12			2

Tabelle 5.1: *Übersicht Simulationsszenarien*

Die folgende Grafik zeigt den Anteil abgelehnter Nullhypothesen, incl. 95%-Konfidenzintervall – für die untersuchten Szenarien 1-3.

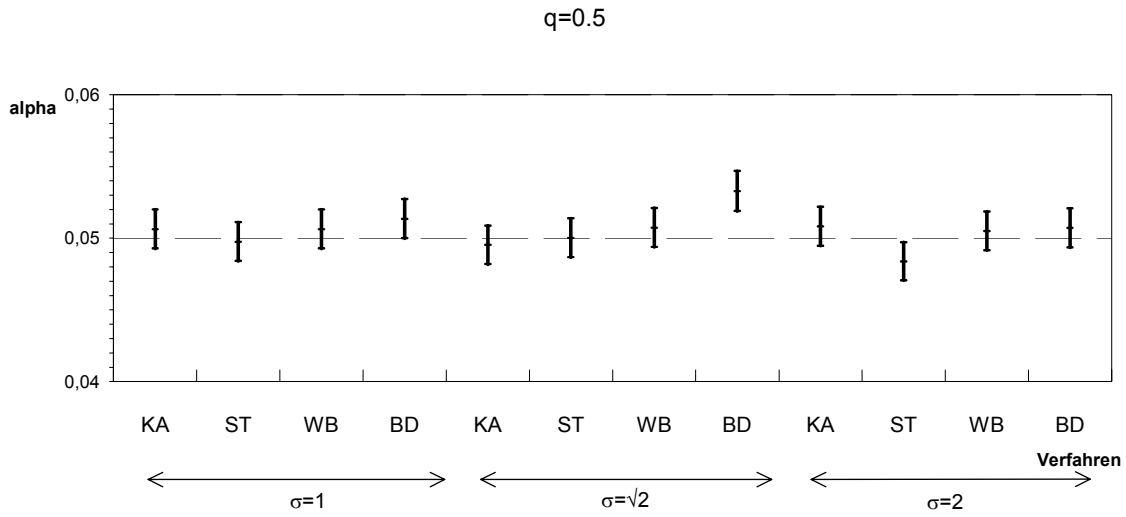


Abbildung 5.1: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI – für die Szenarien 1-3 (q=0.5)*

Es zeigt sich somit, dass nur bei dem Verfahren BD eine nennenswerte Inflation des alphas stattfindet. Alle anderen Verfahren halten das vorgegebene alpha ein. Dies deckt sich mit den Ergebnissen von (Kieser und Friede 2000) (table 1), die für eine ähnliche Konstellation von n_1 ein maximales alpha von 0.0526 bei einer (für das alpha) maximalen Fallzahl von n=76 berechnen. Dies kommt der Simulation mit $\sigma=\sqrt{2}$ und der daraus resultierenden Fallzahl von n=86 am nächsten.

Die Ergebnisse von WB bezüglich eines Anstiegs von alpha auf 0.052 für die Situation $\sigma=2$ konnten bei diesen Simulationen nicht bestätigt werden; hier lag das alpha bei 0.05049. Dafür könnte die unterschiedliche Anzahl von Simulationen (40.000 bei WB / 100.000 bei mir) verantwortlich sein.

Für q=0.2 (Szenarien 7-9) ergibt sich folgendes Bild:

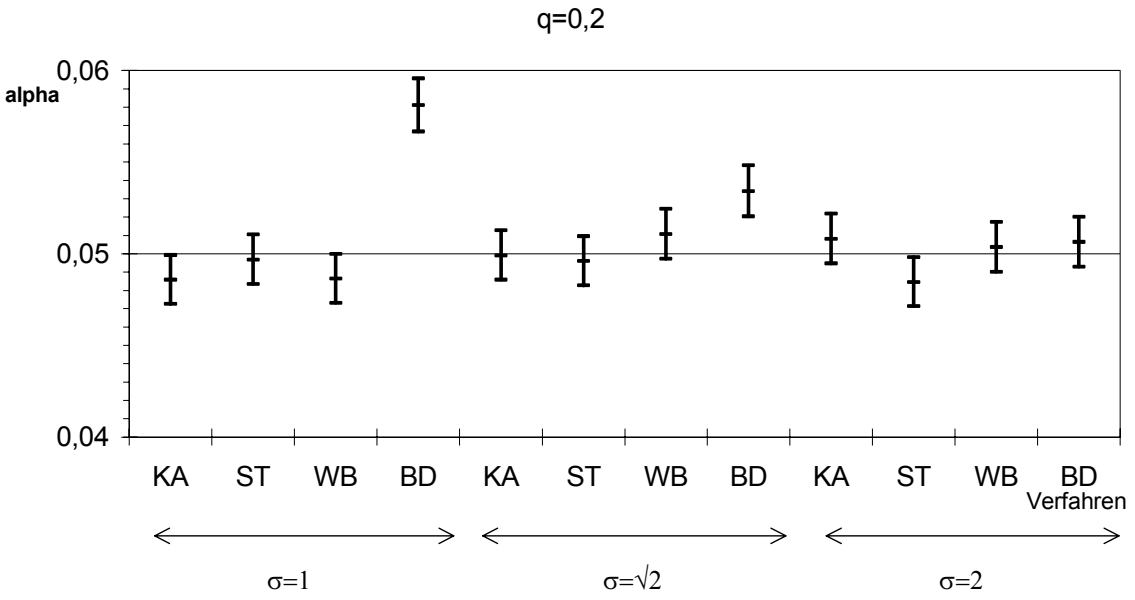


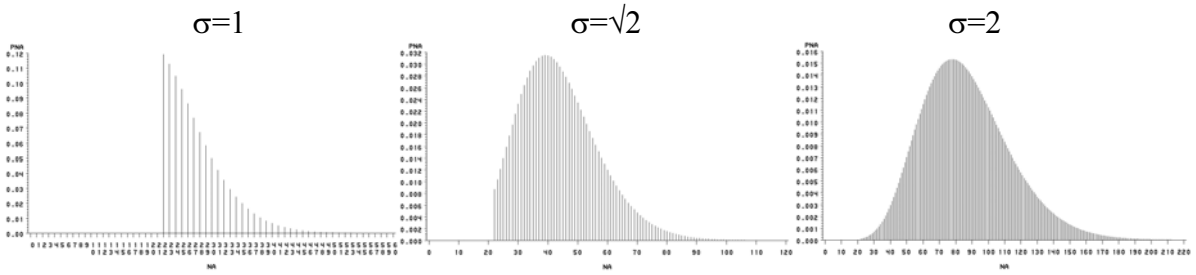
Abbildung 5.2: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95% KI – für die Szenarien 7-9 (q=0.2)*

Auch hier zeigt sich, dass alle Verfahren außer BD die alpha-Vorgabe einhalten. Wiederum decken sich die Ergebnisse bezüglich BD bei den Szenarien $\sigma=1$ und $\sigma=\sqrt{2}$ mit den Resultaten von (Kieser und Friede 2000) (table 1). Hier wird eine maximale Inflation auf $\alpha=0.0564$ bei einer Fallzahl von n=38 beobachtet – dies entspricht am ehesten dem Szenario $\sigma=1$, da hier n=44 resultiert.

Abschließend bleibt festzustellen, dass lediglich das Verfahren BD zu einer alpha-Inflation führt – diese ist am stärksten ausgeprägt bei einem frühen Adjustierungszeitpunkt und bei gleichzeitiger Überschätzung der Standardabweichung.

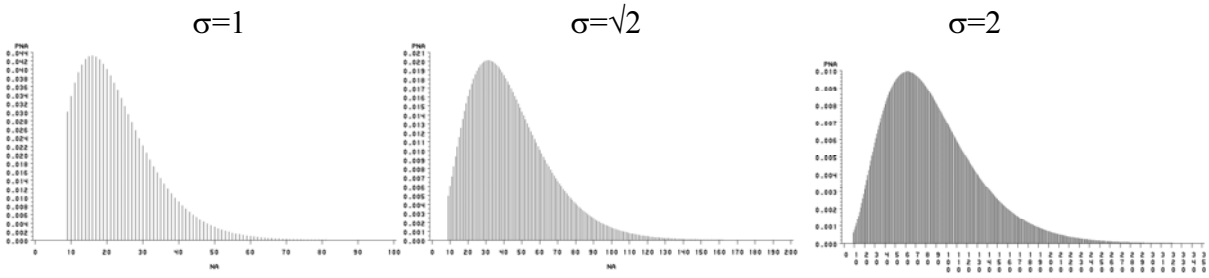
5.1.2 Auswirkungen auf die Verteilung von N

Wie in Kapitel 3.1 beschrieben, lässt sich die exakte Verteilung der adjustierten Fallzahl mit Hilfe einer gestutzten χ_{n1-2}^2- Verteilung herleiten. Für die im Rahmen der Simulationen betrachteten Szenarien für n_1 und σ resultieren demnach folgende Verteilungen:



Mittelwert:	**27.51**	**44.21**	**86.51**
Std:	**4.99**	**13.11**	**27.17**
Median:	**26**	**43**	**84**
Q1:	**24**	**34**	**67**
Q3:	**30**	**52**	**103**

Abbildung 5.3: *Verteilung von N bei den Szenarien 1-6 (q=0.5, n_1=22)*



Mittelwert:	**23.38**	**43.93**	**86.59**
Std:	**10.95**	**22.76**	**45.87**
Median:	**21**	**40**	**79**
Q1:	**15**	**27**	**53**
Q3:	**29**	**56**	**112**

Abbildung 5.4: *Verteilung von N bei den Szenarien 7-12 (q=0.2, n_1=9)*

Beim Vergleich der Standardabweichungen und Interquartilsabstände für die beiden Fälle zeigt sich, dass bei kleinerem n_1 eine größere Streuung vorliegt – als Maß für die stärkere

Unsicherheit bezüglich der endgültigen Fallzahl. Bezüglich der Mittelwerte und Mediane zeigt sich, dass insbesondere im Fall einer initialen Überschätzung der Standardabweichung ($\sigma=1$, $s_0=\sqrt{2}$) der Lageparameter bei einer frühen Adjustierung kleiner ist als bei einer späteren Adjustierung – hier ist also ein früherer Zeitpunkt von Vorteil, da man rechtzeitig die Fallzahl nach unten reduzieren kann.

Im Folgenden sind nun die aus den Simulationen resultierenden Verteilungen von N_a pro Gruppe dargestellt. Dabei führen die Verfahren ST und BD zu den gleichen Ergebnissen, da die Adjustierungsregeln identisch sind („Adjustierung nach oben oder unten“). Außerdem unterscheiden sich die Verteilungen nicht für die Szenarien H_0 / H_1.

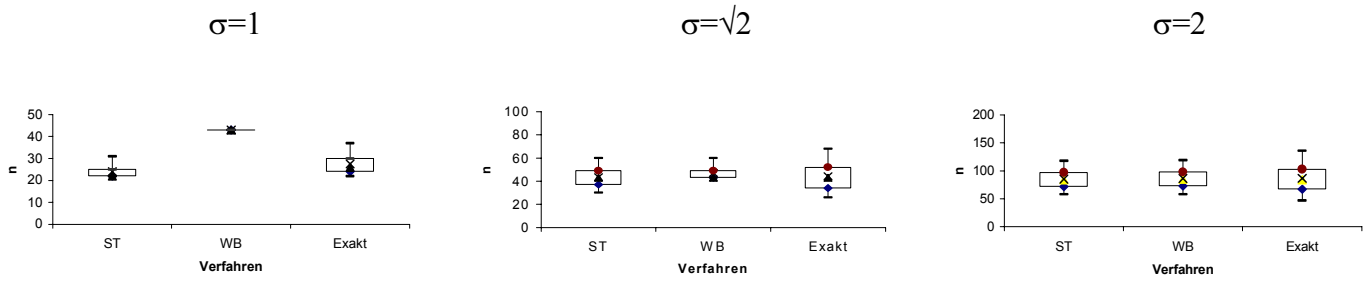


Abbildung 5.5: *Simulationsstudie: Verteilungen von N bei q=0.5 - Box-Whisker-Plot mit: X: Mittelwert, Dreieck: Median, Whisker: 5%-Perzentil / 95%-Perzentil*

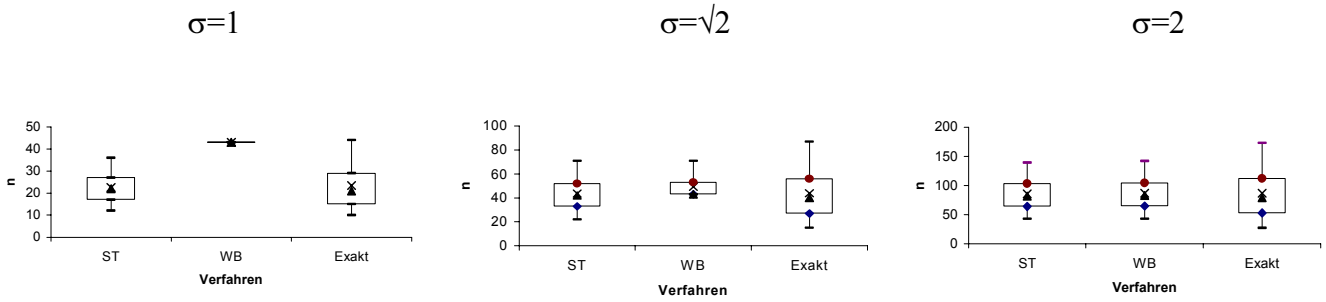


Abbildung 5.6: *Simulationsstudie: Verteilungen von N bei q=0.2 - Box-Whisker-Plot mit: X: Mittelwert, Dreieck: Median, Whisker: 5%-Perzentil / 95%-Perzentil*

Als weiteres Maß zur Beurteilung der Güte der beiden Verfahren, habe ich den Mean Square Error (MSE) berechnet. Dabei habe ich die aus den Simulationen resultierenden Mittelwerte ($\bar{n}_{a,V}$ - mit: V(erfahren) $\in$ {ST, WB}) und Standardabweichungen von n_a (s_V) mit dem aus der wahren Verteilung hergeleiteten Mittelwert $\bar{n}_{a,exakt}$ in folgende Beziehung gesetzt:

$$MSE_V = \left(\bar{n}_{a,V} - \bar{n}_{a,exakt}\right)^2 = (Bias_V)^2 + s_V^2$$

Hier resultieren folgende Ergebnisse:

q	σ	Exakte Verteilung	WB				ST			
		$\bar{n}_a$	$\bar{n}_a$	s	Bias	MSE	$\bar{n}_a$	s	Bias	MSE
0.5	1	27.51	43.00	0.03	15.49	239.94	24.07	3.07	-3.44	21.26
	√2	44.21	47.00	6.14	2.79	45.48	43.56	9.17	-0.65	84.51
	2	86.51	86.45	18.72	-0.06	350.44	85.45	18.30	-1.06	336.01
0.2	1	23.38	43.05	0.62	19.67	387.29	22.56	7.44	-0.82	56.03
	√2	43.93	49.29	10.28	5.36	134.41	43.60	14.90	-0.33	222.12
	2	86.59	86.76	29.78	0.17	886.88	85.42	29.64	-1.17	879.90

Tabelle 5.2: *Simulationsstudie: Bias und MSE für die untersuchten Verfahren / Szenarien*

Daraus lassen sich folgende Schlüsse ziehen:

- Für den Fall, dass σ bei der Fallzahlplanung überschätzt wird, liefert WB ein nach oben verzerrtes n (da nie nach unten adjustiert wird).
- WB liefert geringere Streuungen als ST, für die Fälle $\sigma <= s_0$ – da hier durch die Beschränkung auf die Adjustierung nach oben, die Verteilung nach unten abgeschnitten wird.
- Beide Verfahren liefern ähnliche Ergebnisse für den Fall $\sigma > s_0$.
- Eine frühe Adjustierung (q=0.2) führt zu schlechteren Eigenschaften der geschätzten Fallzahl – unabhängig vom gewählten Verfahren.

5.1.3 Auswirkungen auf die Power

Wird die wahre Standardabweichung σ zur Fallzahlplanung stark unterschätzt ($s_0 << \sigma$), so hat dies einen Verlust an Power zur Folge. So reduziert sich die (geplante) Power von 90% auf die tatsächliche Power von 63% für das Simulationsszenario 6 ($s_0=\sqrt{2}$, $\sigma=2$, $q=0.5$). Mit Hilfe der Adjustierungsstrategien ST, WB und BD kann dieses Absinken verhindert werden. Für den Fall, dass σ bei der Fallzahlplanung überschätzt bzw. genau geschätzt wird, führen alle untersuchten Strategien zur Einhaltung der geforderten Power. Aufgrund der Simulationen ergibt sich für den Fall q=0.5 folgendes Bild für den Anteil abgelehnter Nullhypothesen:

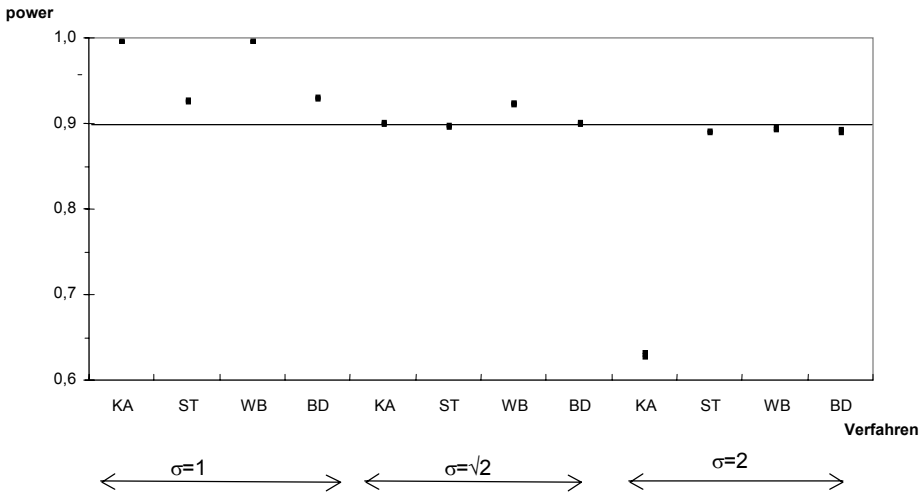


Abbildung 5.7: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI – für die Szenarien 4-6 – alle Verfahren*

Falls keine Adjustierung durchgeführt wird, sinkt die Power auf weniger als 65% - bei unterschätzter Standardabweichung ($s_0=\sqrt{2}$, $\sigma=2$). Dies ist die Motivation aller Fallzahladjustierungsverfahren (s. dazu auch Abbildung 3.1).

Für den Fall „$\sigma=1$“ führt WB zu einer ähnlich hohen Power wie KA – da bei WB die Fallzahl nur nach oben adjustiert wird. In dem vorliegenden Fall ist dies gleichbedeutend mit der Beibehaltung der geplanten Fallzahl.

Beschränkt man den Vergleich auf die Adjustierungsmethoden ST, WB und BD, so ergibt sich folgendes Bild:

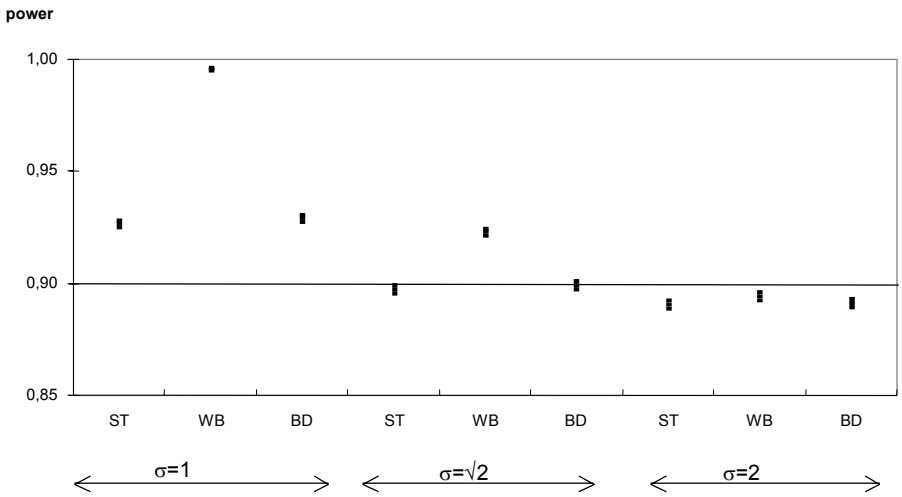


Abbildung 5.8: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI – für die Szenarien 4-6 – Verfahren ST, WB, BD*

Im Fall σ=1, führt die Adjustierung nach WB zu einer extrem hohen Power, da in dieser Situation eine Adjustierung nach unten angemessen wäre, jedoch die ursprüngliche Fallzahl beibehalten wird. Dieses „Überschiessen“ der Power fällt für die beiden anderen Verfahren wesentlich geringer aus.
In der Situation $\sigma = s_0$ wird die vorgegebene Power von 90% von keinem Verfahren wesentlich unterschritten – hier schneidet WB am besten ab.
Schließlich ist bei einer initialen Unterschätzung von σ keines der Verfahren in der Lage die gewünschte Power einzuhalten – jedoch liegt bei allen Verfahren eine gute Annäherung auf ca. 88% vor.

5.1.4 Auswirkungen auf die Weite des Konfidenzintervalls

Nach Beendigung einer Studie ist neben der Testdurchführung auch die Schätzung des Behandlungsunterschieds von Interesse. Dazu dient zur der mittlere Unterschied zwischen den Behandlungsgruppen, nebst 95%-Konfidenzintervall.
Auf der Basis der durchgeführten Simulationen wurden folgende Berechnungen vorgenommen:

- Schätzung des Behandlungsunterschiedes durch Differenzenbildung der Mittelwerte der beiden Behandlungsgruppen (pro Simulation): $\left(\bar{X}_B - \bar{X}_K\right)$
- Berechnung des zugehörigen 95% Konfidenzintervalls (pro Simulation):

$$\left(\bar{X}_B - \bar{X}_K\right) \pm s_p \bullet \sqrt{\frac{4}{n_a}} \bullet t_{0.975,n_a-2}$$

  (mit: s_p – gepoolte Standardabweichung; beim Verfahren nach ST geht in s_p nur die Information aus der 1. Studienphase ein)
- Berechnung der Länge des 95% Konfidenzintervalls (pro Simulation):

$$\frac{4 \bullet s_p}{\sqrt{n_a}} \bullet t_{0.975,n_a-2}$$

- Berechnung des mittleren 95%-Konfidenzintervalls (über alle Simulationen)
- Berechnung der mittleren Länge der 95%-Konfidenzintervalle (über alle Simulationen)

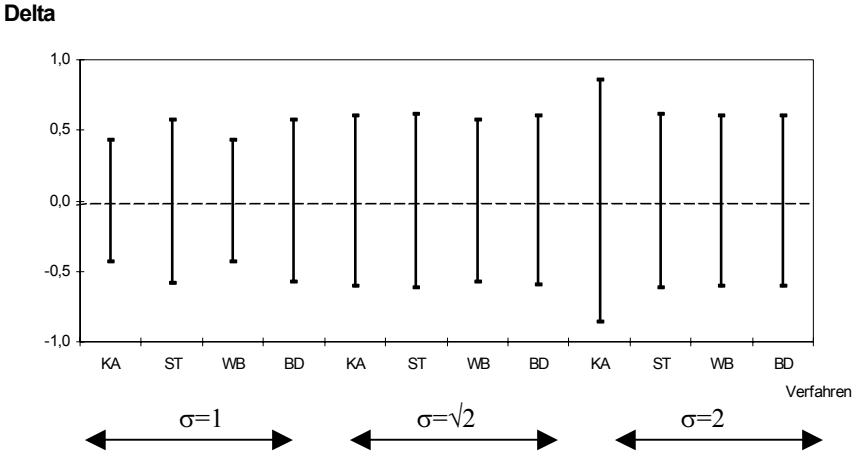


Abbildung 5.9: *Simulationsstudie: Mittlere Länge der 95%-KI für die untersuchten Szenarien / Verfahren – H_0 trifft zu, q=0.5*

Für den Fall der initialen Unterschätzung von σ, führt die Strategie „Keine Adjustierung" somit zu einem großen Konfidenzintervall – diese Vergrößerung kann durch alle untersuchten Adjustierungsverfahren verhindert werden. Im entgegengesetzten Fall $\sigma < s_0$ führen die Strategien KA und WB zu kleineren Konfidenzintervallen, da hier eine Power erreicht wird, die die Planung übersteigt – s. dazu auch Kapitel 5.1.3.
Insgesamt zeigen sich bei den Verfahren ST und BD bei allen Szenarien die konstantesten Verhältnisse.

Ähnliche Schlussfolgerungen ergeben sich bei den ebenfalls untersuchten Konstellationen: H_1 trifft zu / q=0.2.

5.2 Verfahren für nicht entblindete Daten

Bei den im vorigen Kapitel vorgestellten Verfahren müssen die Daten entblindet werden, d.h. die Information, welcher Patient zu welcher Gruppe (1 / 2) gehört, muss verfügbar sein. Nur mit dieser Entblindung ist es möglich, die über die Behandlungsgruppen gepoolte Standardabweichung zu schätzen. Konkret werden dazu die Standardabweichungen für die jeweilige Gruppe benötigt; in diese geht wiederum der gruppenspezifische Mittelwert ein.
Die Entblindung von Studiendaten sollte erst dann durchgeführt werden, wenn es unbedingt notwendig ist – um sicher zu stellen, dass der doppelblinde Charakter der Studie auch nach der Fallzahladjustierung gewährleistet bleibt. Wenn man schon entblindet, möchte man in der Regel auch eine Zwischenauswertung durchführen. Daher ist es naheliegend, zum Zwecke der Fallzahladjustierung auf eine Entblindung zu verzichten, auch wenn es sich „nur“ um eine partielle Entblindung handelt (s. Kapitel 5.1).

Einfache Strategien zur „blinden“ Schätzung der Standardabweichung

Es geht also darum, die Standardabweichung zu schätzen, ohne auf die Information zurückzugreifen, welche Beobachtung zu welcher Behandlungsgruppe gehört. Die einfachste Vorgehensweise besteht darin, die Standardabweichung über alle Beobachtungen zu schätzen, also:

$$s_{gesamt} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{5.19}$$

Diese Methode kann durch Ausnutzung des folgenden Zusammenhangs für die Gesamt-Varianz verfeinert werden. Hierbei wird wiederum von zwei gleich großen Gruppen mit $n_1 = n_2 = n/2$ ausgegangen:

$$\begin{aligned} s^2_{gesamt} &= \frac{1}{n-1}\left(\sum_{i=1}^{n_1}\left(x_{1i} - \bar{x}\right)^2 + \sum_{i=1}^{n_2}\left(x_{2i} - \bar{x}\right)^2\right) \\ &= \frac{1}{n-1}\left((n-2)\bullet s_p^2 + \left(\frac{n}{4}\bullet\left(\bar{x}_1 - \bar{x}_2\right)^2\right)\right) \end{aligned} \tag{5.20}$$

mit:

$$s_p^2 = \frac{\left(\frac{n}{2} - 1\right) \bullet \left(s_1^2 + s_2^2\right)}{n-2}$$

gepoolte Standardabweichung

Somit gilt für das gesuchte s_p^2:

$$s_p^2 = \frac{1}{n-2}\left((n-1) \bullet s_{gesamt}^2 - \frac{n}{4} \bullet \left(\bar{x}_1 - \bar{x}_2\right)^2\right) \tag{5.21}$$

Dabei ist der einzige Term, der getrennte Informationen aus beiden Gruppen und somit eine Entblindung benötigt: $\left(\bar{x}_1 - \bar{x}_2\right)^2$. Die Idee besteht nun darin, diese Größe durch den bei der Fallzahlplanung benutzten relevanten Unterschied δ zu ersetzen. Damit ergibt sich als verfeinerte Möglichkeit der unblinden Schätzung der gepoolten Standardabweichung:

$$s_\delta = \sqrt{\frac{1}{n-2}\left((n-1) \bullet s_{gesamt}^2 - \frac{n}{4} \bullet \delta^2\right)} \tag{5.22}$$

Somit hängen sowohl s_δ als auch s_{Gesamt} von dem tatsächlichen Behandlungsunterschied $\delta_{real} = \left(\bar{x}_1 - \bar{x}_2\right)$ ab. Für die (bereits bei der Simulation benutzten) Konstellation:

- $n_1 = 2 \bullet 22$
- $s_0 = \sqrt{2}$
- $\mu_1 = 0$, $\mu_2 = 1 \Rightarrow \delta = 1$

ergibt sich für die Szenarien bezüglich $\delta_{real} = 0 / 1 / 2$ und bezüglich $\sigma = 1 / \sqrt{2} / 2$ folgendes Bild:

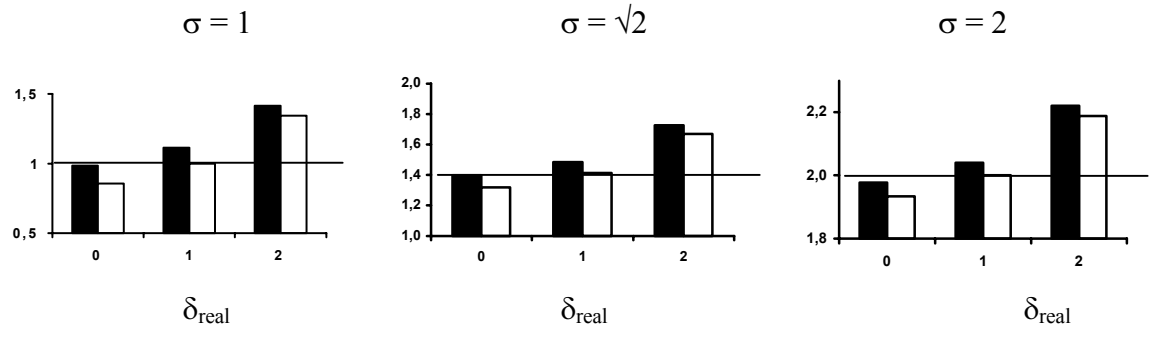


Abbildung 5.10: *Zusammenhang zwischen δ_{real}, σ und: s_{gesamt} (■), s_δ (□), s_P (Referenzlinie)*

Für den Fall $\delta_{real} = \delta$ stimmt s_δ konstruktionsgemäß mit der wahren gepoolten Standardabweichung s_p überein.
Für den Fall, dass der tatsächliche Behandlungsunterschied kleiner ist, als bei der Fallzahlplanung angenommen, erzielt man hingegen mit s_{gesamt} eine bessere Annäherung an s_p als mit s_δ. Somit gibt es durchaus Situationen, bei denen diese Einfach-Strategie gute Ergebnisse erzielt.
Schließlich überschätzen beide Methoden die tatsächliche Standardabweichung, falls der tatsächliche Unterschied das bei der Planung benutzte δ übersteigt. Hier liegt man mit s_δ etwas näher an der Realität als mit s_{gesamt} .

Ein EM-Algorithmus-basiertes Verfahren

Eine weitere Verfeinerung bezüglich der „blinden“ Schätzung der Standardabweichung wurde bei (Gould und Shih 1992), (Gould 1995) und (Shih 1993) vorgenommen – diese basiert auf dem EM-Algorithmus. Die grundsätzliche Annahme ist dabei, dass die Information bezüglich der Gruppenzugehörigkeit der Beobachtung x_i „Missing at Random“ ist, d.h. :

Sei

$$x = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix}$$

der Vektor aller beobachteten Messwerte, und

$$\tau = \begin{pmatrix} \tau_1 \\ \tau_2 \\ . \\ . \\ \tau_n \end{pmatrix}$$

der zugehörige Vektor der Gruppenzugehörigkeit, d.h.

$$\tau_i = \begin{cases} 1, \textit{falls } x_i \textit{ zur Gruppe } 1 \textit{ gehört} \\ 0, \textit{falls } x_i \textit{ zur Gruppe } 2 \textit{ gehört} \end{cases}$$

Dann sind die Gründe für das Fehlen von τ abhängig von x aber nicht von τ.

Es gilt nun:

$$P(\tau_i=1)=1/2.$$

Unter den üblichen Voraussetzungen für den t-Test gilt für die bedingte Verteilung von x_i:

$$f\left(x_i \mid \tau_i, \mu_1, \mu_2, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2\sigma^2}\left(\tau_i(x_i-\mu_1)^2+(1-\tau_i)(x_i-\mu_2)^2\right)\right)} \tag{5.23}$$

Nach geeigneter Wahl von Startwerten $\hat{\mu}_1^0$, $\hat{\mu}_2^0$, $\hat{\sigma}^0$ besteht der EM-Algorithmus aus der Iteration (j = 0, 1,...) des E- und des M-steps, mit:

<u>E-Step</u>

$$P\left(\tau_i^j = 1 \mid x_i\right) = \left[1+\exp\left(\frac{\left(\hat{\mu}_1^j-\hat{\mu}_2^j\right)\left(\hat{\mu}_1^j+\hat{\mu}_2^j-2x_i\right)}{2\hat{\sigma}^{j\,2}}\right)\right]^{-1} \tag{5.24}$$

<u>M-Step</u>

$$\begin{aligned}
\hat{\mu}_1^{j+1} &= \frac{\sum_{i=1}^{n}\tau_i^j x_i}{\sum_{i=1}^{n}\tau_i^j} \\
\hat{\mu}_2^{j+1} &= \frac{\sum_{i=1}^{n}(1-\tau_i^j)x_i}{\sum_{i=1}^{n}(1-\tau_i^j)} \\
\hat{\sigma}^{j+1\,2} &= \frac{1}{n}\sum_{i=1}^{n}\left(\tau_i^j\left(x_i-\hat{\mu}_1^j\right)^2+\left(1-\tau_i^j\right)\left(x_i-\hat{\mu}_2^j\right)^2\right)
\end{aligned} \tag{5.25}$$

Die effektive Durchführung und somit die Konvergenz des EM-Algorithmus hängt im Wesentlichen von der geeigneten Wahl der Startwerte ab. Dazu gibt es folgende Möglichkeiten:

1. Startwert für $\hat{\sigma}$ mittels „simple adjustment“ (nach (Gould und Shih 1992))

$$\hat{\sigma}^{0}_{SA1} = \sqrt{\left(\frac{n-1}{n-2}\right)\left(s^2_{gesamt} - \frac{1}{4}\delta^2\right)} \tag{5.26}$$

Dieser Startwert lässt sich nach dem oben beschriebenen Verfahren geringfügig ändern zu:

2. Startwert für $\hat{\sigma}$ mittels modifiziertem „simple adjustment“ (s.o., bzw. (Zucker, Wittes et al. 1999) und (Kieser und Friede 2000))

$$\hat{\sigma}^{0}_{SA2} = \sqrt{\frac{1}{n-2}\left((n-1)\bullet s^2_{gesamt} - \frac{n}{4}\bullet\delta^2\right)} \tag{5.27}$$

3. Startwert für $\hat{\mu}_1, \hat{\mu}_2$ mittels der bei der Fallzahlplanung benutzten Parameter

Hier werden einfach die für die Planung benutzten Größen bezüglich des mittleren Effektes unter der Kontrollbehandlung und bezüglich des relevanten Unterschiedes benutzt, also:

$$\hat{\mu}^{0}_{1,Pl} = \hat{\mu}_{1,Pl} \qquad \hat{\mu}^{0}_{2,Pl} = \hat{\mu}_{1,Pl} + \delta$$

Weitere, von (Gould und Shih 1992) vorgeschlagene Möglichkeiten gehen von folgender Situation aus:

Seien X_1 und X_2 die Zielvariablen in den beiden Behandlungsgruppen, mit:

$$X_1 \sim N(\mu_1, \sigma^2) \qquad X_2 \sim N(\mu_2, \sigma^2)$$

Dann wird bei nicht entblindeten Daten eine Mischverteilung beobachtet mit bekannter Mischungsrate (=1/2). Mit Hilfe der Regressionsgeraden innerhalb eines QQ-Plots:

$$\left[\underbrace{\Phi^{-1}\left((i-1/2)/n\right)}_{q_i}, x_{(i)}\right]$$

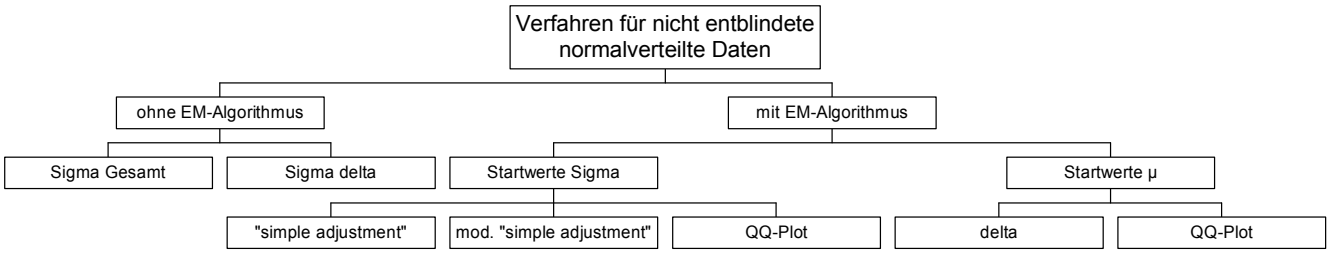
können folgende Rückschlüsse auf die gemeinsame Standardabweichung und auf die Gruppenmittelwerte gezogen werden:

Sei: $x_{(i)} = \alpha_1 + \beta_1 q_i + \varepsilon_{i\,d}$, dann:

$$\hat{\sigma}^{0}_{QQ} = \hat{\beta}_1$$
$$\hat{\mu}^{0}_{1,QQ} = \frac{\hat{\alpha}_1 - \hat{\beta}_1}{5.71} \qquad (5.28)$$
$$\hat{\mu}^{0}_{2,QQ} = \frac{\hat{\alpha}_1 + \hat{\beta}_1}{5.71}$$

Zusammenfassung

Zusammenfassend gibt es also folgende Strategien:



Ein Vergleich einiger dieser Strategien findet sich bei (Zellner, Zellner et al. 2001).

5.2.1 Auswirkungen auf alpha

In Anlehnung an die im Kapitel 5.1.1 beschriebenen Simulationen, wurden für die beiden Einfach-Strategien:

- Sigma Gesamt
- Sigma Delta

jeweils eine Adjustierung nur nach oben durchgeführt. Die Adjustierung wurde dabei nach der Hälfte der geplanten Fallzahl durchgeführt – also: $n_1=2 \bullet 22=44$

Ferner wurden folgende 9 Szenarien berücksichtigt – pro Szenario wurden 100.000 Simulationen durchgeführt:

Szenario		**σ**		
		1	$\sqrt{2}$ $=s_0$	2
δ_{real}	0	1	2	3
	1 = δ	4	5	6
	2	7	8	9

Tabelle 5.3: *Überblick der in den Simulationen berücksichtigten Szenarien*

Bei dieser Simulationsstudie wurden also zusätzliche Szenarien bezüglich δ_{real} berücksichtigt: Neben den beiden Situationen H_0 ($\delta_{real} = 0$) und H_1 an der Stelle $\delta_{real} = 1$ wurde ein weiterer Punkt unter der Alternativhypothese in die Betrachtungen mit einbezogen.

Auf die Anwendung des EM-Algorithmus wurde im Rahmen der Simulationsstudie verzichtet, da bereits die Einfach-Strategien sehr zufriedenstellende Ergebnisse erzielt haben. Somit bestand keine Notwendigkeit ein so komplexes, von zahlreichen Voraussetzungen abhängendes Verfahren wie den EM-Algorithmus anzuwenden. Außerdem haben (Friede und Kieser 2002) auf zahlreiche Mängel dieses Verfahrens hingewiesen.

Szenario		**Strategie**		
δ_{real}	σ	Keine Adjustierung	S_{gesamt}	S_δ
0	1	**0.04942**	**0.04942**	**0.04942**
		0.048077	0.048077	0.048077
		0.050763	0.050763	0.050763
	$\sqrt{2}$	**0.04987**	**0.04976**	**0.04989**
		0.048521	0.048412	0.048541
		0.051219	0.051108	0.051239
	2	**0.04982**	**0.05097**	**0.05108**
		0.048471	0.049607	0.049715
		0.051169	0.052333	0.052445

Tabelle 5.4: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI für die Szenarien 1-3*

Somit zeigt sich aufgrund der Simulationen, dass das geforderte Signifikanzniveau bei allen Strategien eingehalten wird.

5.2.2 Auswirkungen auf die Verteilung von N

Wie bereits im vorigen Kapitel hergeleitet, ist die exakte Verteilung der Fallzahl für die 3 Szenarien bezüglich σ durch folgende Kenngrößen charakterisiert:

σ	Mittelwert (Mw)	Standard-abweichung (Std)	Median (Med)	Unteres Quartil (Q1)	Oberes Quartil (Q3)
1	27.5	4.99	26	24	30
√2	44.2	13.11	43	34	52
2	86.5	27.17	84	67	103

Tabelle 5.5: *Verteilung von N in Abhängigkeit von σ*

Die entsprechenden Kenngrößen für die untersuchten Strategien / Szenarien lauten nun:

Szenario		Strategie													
δ_{real}	σ	S_{gesamt}							S_δ						
		Mw	Std	Med	Q1	Q3	Bias	MSE	Mw	Std	Med	Q1	Q3	Bias	MSE
0	1	43.0	0.02	43	43	43	15.5	239.9	43.0	0.00	43	43	43	15.5	239.9
	√2	46.9	6.01	43	43	49	2.7	43.3	45.2	4.62	43	43	45	0.9	22.3
	2	86.5	18.53	85	73	98	0.0	343.3	83.0	18.93	82	70	95	-3.6	371.0
1	1	43.0	0.33	43	43	43	15.5	240.7	43.0	0.13	43	43	43	15.5	240.1
	√2	50.7	8.43	48	43	56	6.5	112.8	48.0	7.24	44	43	51	3.8	66.74
	2	92.1	19.61	91	78	104	5.5	415.6	88.6	20.06	87	74	101	2.1	407.0
2	1	46.8	5.38	44	43	49	19.2	399.1	44.9	3.96	43	43	45	17.4	317.9
	√2	65.6	12.97	65	56	74	21.4	626.6	61.8	12.91	61	52	70	17.6	475.1
	2	108.5	22.83	107	92	123	21.9	1002.6	105.4	23.37	104	89	120	18.9	902.7

Tabelle 5.6: *Simulationsstudie: Verteilungsparameter der Fallzahl N für die untersuchten Strategien / Szenarien*

Im Fall der initialen Überschätzung der Standardabweichung führen beide Strategien konstruktionsgemäß zu einer Überschätzung der erforderlichen Fallzahl – da in keinem Fall nach unten adjustiert wird. Diese Überschätzung wird bei beiden Verfahren noch weiter verstärkt, falls ein Unterschied zwischen den Behandlungsgruppen vorliegt, der den relevanten Unterschied übersteigt. In dieser Situation weist S_{gesamt} etwas schlechtere Eigenschaften (größerer Bias und größerer MSE) als S_δ auf.

Liegt man mit der Schätzung der Standardabweichung bei der Fallzahlplanung richtig, so schätzen beide Verfahren die Fallzahl fast unverzerrt für den Fall, dass die Nullhypothese

zutrifft. Die Verzerrung und auch der MSE werden umso größer, je weiter man sich von H_0 entfernt – mit besseren Eigenschaften für S_{gesamt}.

Ähnliches gilt für den Fall, dass die Standardabweichung bei der Planung unterschätzt wurde. Hier führt S_{gesamt} sogar zu einer unverzerrten Schätzung bei Zutreffen von H_0, während die S_δ-Methode hier zu einer Unterschätzung der wahren Fallzahl führt.

Beide Verfahren werden dann problematisch, wenn δ_{real} das bei der Planung benutzte $\delta_{Planung}$ übersteigt. In diesem Fall führen beide Verfahren zu einer unnötigen Erhöhung der Fallzahl. Insgesamt zeigen also beide Verfahren sehr ähnliche Eigenschaften, mit leichten Vorteilen für S_δ.

5.2.3 Auswirkungen auf die Power

Hier wurden folgende Ergebnisse bei der Simulationsstudie erzielt:

Szenario		**Strategie**		
δ_{real}	σ	Keine Adjustierung	S_{gesamt}	S_δ
1	1	**0.99562** 0.99521 0.99603	**0.99562** 0.99521 0.99603	**0.99562** 0.99521 0.99603
	√2	**0.89976** 0.89790 0.90162	**0.92549** 0.92386 0.92712	**0.91522** 0.91349 0.91695
	2	**0.62958** 0.62659 0.63257	**0.89775** 0.89587 0.89963	**0.88455** 0.88257 0.88653
2	1	**1** 1 1	**1** 1 1	**1** 1 1
	√2	**0.99998** 0.99995 1	**1** 1 1	**0.99999** 0.99997 1
	2	**0.99546** 0.99504 0.99588	**0.99997** 0.99994 1	**0.99994** 0.99989 0.99999

Tabelle 5.7: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI für die Szenarien 4-9*

Die Überschätzung der Standardabweichung bei der Studienplanung führt zwangsläufig zu einer „überpowerten“ Studie – ganz egal ob adjustiert wird oder nicht. Schließlich wird bei

den beschriebenen Verfahren nicht nach unten adjustiert, was in diesem Fall angemessen wäre.

Ist der tatsächliche Unterschied zwischen den Behandlungsgruppen doppelt so groß wie der relevante Unterschied, der bei der Fallzahlplanung benutzt wurde, führt selbst die initiale Unterschätzung der Standardabweichung zu keinem Problem mit der Power. Die geforderte Power wird selbst ohne Adjustierung bei weitem überschritten.

Im interessanten Fall, dass der tatsächliche Unterschied mit dem relevanten Unterschied übereinstimmt und zugleich die Standardabweichung unterschätzt wird, führt die Beibehaltung der Fallzahl dazu, dass die Power auf ca. 63% reduziert wird. Beide Verfahren können dieses Manko vermeiden. Bei S_{gesamt} wird die geforderte Power von 90% nur sehr knapp verfehlt, auch S_δ reicht fast an diese „Referenzmarke“ heran.

Somit sind beide Verfahren geeignet, einen Powerverlust zu vermeiden – mit leichten Vorteilen für S_{gesamt} .

5.2.4 Auswirkungen auf die Weite des Konfidenzintervalls

Angegeben ist wiederum die mittlere Weite des Konfidenzintervalls zur Schätzung des Unterschiedes zwischen den beiden Behandlungsgruppen:

Szenario		**Strategie**		
δ_{real}	σ	Keine Adjustierung	S_{gesamt}	S_δ
0	1	0.85536	0.85536	0.85536
	√2	1.20889	1.15478	1.17722
	2	1.71081	1.20870	1.23635
1	1	0.85529	0.85500	0.85524
	√2	1.20997	1.11358	1.14408
	2	1.71072	1.17246	1.19696
2	1	0.85534	0.82007	0.83667
	√2	1.20983	0.98220	1.01353
	2	1.70953	1.08108	1.09826

Tabelle 5.8: *Simulationsstudie: Mittlere Länge des 95%-KI zur Schätzung des Gruppenunterschiedes – für die untersuchten Strategien*

Bei allen Strategien nimmt erwartungsgemäß die Länge des Konfidenzintervalls mit steigender Standardabweichung zu. Die beiden Adjustierungsstrategien führen jedoch zu einer

Verkleinerung der Konfidenzintervalle im Vergleich zu keiner Adjustierung, da hier mehr Patienten bei der Schätzung berücksichtigt werden. Da S_{gesamt} grundsätzlich zu einer höheren adjustierten Fallzahl als S_δ führt (s. Kapitel 5.2.2), resultieren hier auch kürzere Konfidenzintervalle.

Abschließend bleibt festzustellen, dass beide hier untersuchten Strategien das Signifikanzniveau einhalten und insbesondere in der Lage sind, ein Absinken der Power zu vermeiden. Mit der konservativen Strategie S_{gesamt} ist man immer auf der sicheren Seite, was die adjustierte Fallzahl und die Weite des Konfidenzintervalls angeht.

5.3 Vergleich der Verfahren für entblindete / nicht entblindete Daten

Bei den blinden Methoden traten hinsichtlich der Einhaltung des vorgegebenen Signifikanzniveaus keine Probleme auf, während bei den unblinden Methoden das Verfahren von (Birkett und Day 1994) zu einer nennenswerten Inflation geführt hat.

Beim Einfluss auf die Verteilung von N_a ergibt sich folgende Situation:

Szenario			**Strategie**							
			Unblind				Blind			
			WB		ST		S_{gesamt}		S_δ	
q	**δ_{real}**	**σ**	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
0.5	0	1	15.5	239.9	-3.4	21.3	15.5	239.9	15.5	239.9
		√2	2.8	45.5	-0.7	84.5	2.7	43.3	0.9	22.3
		2	-0.1	350.4	-1.1	336.0	0.0	343.3	-3.6	371.0
	1	1	15.5	239.9	-3.4	21.3	15.5	240.7	15.5	240.1
		√2	2.8	45.5	-0.7	84.5	6.5	112.8	3.8	66.7
		2	-0.1	350.4	-1.1	336.0	5.5	415.6	2.1	407.0

Tabelle 5.9: *Simulationsstudien: Verteilung von N_a – unblinde / blinde Strategien*

Das Verfahren nach Stein schneidet hier insgesamt am besten ab, da hier auch nach unten adjustiert werden kann. Beim Vergleich der Strategien, die nur nach oben adjustieren, unterscheiden sich die blinden und unblinden Verfahren nur unwesentlich.

Auch bezüglich der Power gibt es nur marginale Unterschiede zwischen den blinden und unblinden Methoden – insbesondere wird in beiden Fällen der Powerverlust bei initialer Unterschätzung der Standardabweichung vermieden. Schließlich können die blinden Methoden auch bei der Länge des Konfidenzintervalls mit den unblinden Methoden mithalten. Somit bleibt festzustellen, dass die beiden näher untersuchten blinden Methoden S_{gesamt} und S_{δ} am besten geeignet sind, eine effektive Fallzahladjustierung bei normalverteilten Daten durchzuführen.

6 Fallzahladjustierung beim Vergleich von Ereignisraten

Werden zwei Patientengruppen hinsichtlich eines dichotomen Merkmals mittels des Chi-Quadrat-Tests verglichen, so werden bei der Fallzahlplanung folgende Parameter benötigt: Signifikanzniveau, Power, erwartete Rate unter der Kontrollbehandlung, relevanter Unterschied bzgl. der Rate. Anders als beim Vergleich von Mittelwerten, bei dem die Standardabweichung den „Zielparameter" für die Fallzahladjustierung darstellt, kann hier lediglich die erwartete Rate unter der Kontrollbehandlung als entsprechender Zielparameter herangezogen werden. Die grundsätzliche Idee aller diesbezüglichen Adjustierungsverfahren besteht nun wieder darin, nach geeigneter Schätzung dieser Rate zu einem Zwischenzeitpunkt die Fallzahl mittels dieser Schätzung zu adjustieren - unter Beibehaltung der bei der initialen Fallzahlplanung benutzten Parameter Signifikanzniveau, Power und relevanter Unterschied.
Im Folgenden werden die Adjustierungsstrategien dahingehend unterschieden, ob für die Schätzung eine Entblindung notwendig ist oder nicht.

6.1 Verfahren für entblindete Daten

Die erstmals von (Herson und Wittes 1993) vorgestellte Strategie (kurz: HW) besteht aus folgenden Schritten:

1. Fallzahlplanung auf Grundlage von α, β, $p_{0,K}$ und $p_{0,B}$ – dabei kann p_B entweder in der Form:
$$p_{0,B} = p_{0,K} + \delta$$
oder:
$$p_{0,B} = p_{0,K} \bullet r$$
oder:
$$p_{0,B} = (OR \bullet p_{0,K}) / (1- p_{0,K} + (OR \bullet p_{0,K}))$$
angegeben werden (s. Kapitel 2.2). Daraus resultiert nach (2.20) die geplante Fallzahl n_P.
2. Rekrutierung eines ersten Anteils der geplanten Patienten: $n_1 = q \bullet n_P$ – mit $0 < q < 1$.
3. Bestimmung der Rate in der Kontrollgruppe: $p_{1,K} = e_{1,K} / n_{1,K}$ – mit: $e_{1,K}$ - Anzahl der Ereignisse bei den $n_{1,K}$ Patienten.
4. Neue Fallzahlplanung auf Grundlage von α, β, $p_{1,K}$ und $p_{1,B} = p_{1,K}+\delta$ oder $p_{1,B}=p_{1,K} \bullet r$ oder $p_{1,B}=(OR \bullet p_{1,K}) / (1-p_{1,K} + (OR \bullet p_{1,K}))$ – analog zur Fallzahlplanung: $n_{neu,HW}$.

5. Adjustierte Fallzahl:

$$n_{a,HW} = \begin{cases} n_{min} & - \text{ falls } n_{neu,HW} < n_{min} \\ n_{neu,HW} & - \text{ falls } n_1 \le n_{neu,HW} \le n_{max} \\ n_{max} & - \text{ falls } n_{neu,HW} > n_{max} \end{cases} \quad (6.1)$$

Dabei sind n_{min} und n_{max} vor Studienbeginn festzulegende Unter- bzw. Obergrenzen für die Fallzahl, z.B. $n_{min}=n_p$ → Adjustierung nur nach oben, $n_{max}= 2\bullet n_p$→ maximal wird bis zur doppelten der ursprünglich geplanten Fallzahl adjustiert.

Wie bereits in Kapitel 5 erwähnt, kann man diese Strategie verändern bzw. ergänzen zu einer Adjustierung in Abhängigkeit einer Ober/Untergrenze für n_{neu} :

$$n_{a,HW} = \begin{cases} n_p & - \text{ falls } a \bullet n_p \le n_{neu,HW} \le b \bullet n_p \\ \max(n_1, n_{neu,HW}) & - \text{ sonst} \end{cases} \quad (6.2)$$

mit geeignet zu wählenden a <1 und b>1

Um die Verblindung nicht zu gefährden, empfehlen die Autoren, den beteiligten Prüfärzten n_{neu} nicht mitzuteilen. Schließlich deutet ein hoher Wert von n_{neu} im Vergleich zu n_p auf eine sehr viel geringere Ereignisrate unter der Kontrollbehandlung hin, als bei der Planung angenommen (s. dazu auch die Abbildung im folgenden Kapitel).

6.2 Verfahren für nicht entblindete Daten

6.2.1 Das Verfahren von Gould

Ein von (Gould 1992) und (Gould 1995) vorgestelltes Verfahren (kurz: G) sieht vor, dass zum Zeitpunkt der Adjustierung (nach n_1 Patienten) die Gesamt-Ereignisrate (über alle Patienten) geschätzt wird, also

$$p_1 = \frac{e_1}{n_1}$$

Mithilfe dieser Schätzung wird eine erneute Fallzahlplanung durchgeführt: Falls die ursprüngliche Fallzahlplanung auf p_K und einem (additiven) δ beruhte, also: $p_{0,B} = p_{0,K} - \delta$, lauten die adjustierten Größen:

$$p_{1,K} = p_1 + \frac{\delta}{2}$$
$$p_{1,B} = p_1 - \frac{\delta}{2} \tag{6.3}$$

Das ursprüngliche δ wird somit auch weiterhin berücksichtigt.

Analog dazu wird bei einer initialen Fallzahlplanung der Form p_B/p_K= r mit den Größen

$$p_{1,K} = \frac{2 \bullet p_1}{r+1}$$
$$p_{1,B} = p_{1,K} \bullet r \tag{6.4}$$

eine aktualisierte Fallzahlplanung durchgeführt.

Bei einer Fallzahlplanung auf Grundlage eines vorgegebenen relevanten Odds ratios (ROR), lautet der Zusammenhang zwischen p_1, $p_{1,K}$ und $p_{1,B}$:

$$p_{1,K} = \frac{\left(2 \bullet p_1 - 2 \bullet ROR \bullet p_1 + 1 + ROR\right) - \sqrt{\left(2 \bullet p_1 - 2 \bullet ROR \bullet p_1 + 1 + ROR\right)^2 - \left(8 \bullet \left(1 - ROR\right) \bullet p_1\right)}}{2 \bullet (1 - ROR)}$$
$$p_{1,B} = \frac{ROR \bullet p_{1,K}}{\left(1 - p_{1,K}\right) + \left(ROR \bullet p_{1,K}\right)} \tag{6.5}$$

Die adjustierte Fallzahl $n_{neu,G}$ geht dann aus (2.20) hervor – mit den bei der ersten Fallzahlplanung benutzten α und ß. Ähnlich wie bei der unblinden Adjustierung nach HW schlagen auch hier die Autoren als adjustierte Fallzahl vor:

$$n_{a,G} = \begin{cases} n_p & - \text{ falls } n_{neu,G} < n_p \\ n_{neu,G} & - \text{ falls } n_p \le n_{neu,G} \le 2 \bullet n_p \\ 2 \bullet n_p & - \text{ falls } n_{neu,G} > 2 \bullet n_p \end{cases} \tag{6.6}$$

Es handelt sich somit um eine Adjustierung „nur nach oben“ mit dem doppelten der geplanten Fallzahl als Obergrenze.

6.2.2 Das Verfahren von Shih&Zhao

Eine andere Form der blinden Fallzahladjustierung wird bei (Shih und Zhao 1997) vorgeschlagen (kurz: SZ): Zunächst muss das Design der Studie dahingehend geändert werden, dass die Patienten in zwei Strata (S und T) gleichmäßig aufgeteilt werden. Im ersten Stratum erfolgt dann die randomisierte Zuteilung zu der Behandlungsgruppe mit Wahrscheinlichkeit $\pi \neq 0.5$ (die Autoren schlagen $\pi = 0.2$ vor); als Zuteilungswahrscheinlichkeit zur Kontrollgruppe wird dann $(1-\pi)$ gewählt. Im zweiten Stratum sind diese Wahrscheinlichkeiten gerade vertauscht, so dass insgesamt die Wahrscheinlichkeit in eine der beiden Gruppen randomisiert zu werden bei 0.5 liegt. Zum Zeitpunkt der Fallzahladjustierung werden nun für beide Strata die Gesamt-Ereignisraten (über die beiden Behandlungsgruppen) beobachtet: $p_{1,S}$ in Stratum S und $p_{1,T}$ in Stratum T. Diese beiden stratum-spezifischen Schätzer werden nun dazu benutzt, die Ereignisraten für die Behandlungs- und die Kontrollgruppe zu schätzen, und zwar in der Form:

$$\begin{aligned} p_{1,B} &= \frac{\pi \bullet p_{1,S} - (1-\pi) \bullet p_{1,T}}{2\pi - 1} \\ p_{1,K} &= \frac{\pi \bullet p_{1,T} - (1-\pi) \bullet p_{1,S}}{2\pi - 1} \end{aligned} \tag{6.7}$$

Mit diesen Raten wird dann die in (2.20) beschriebene Fallzahl neu berechnet. Die endgültige adjustierte Fallzahl kann dann analog zu den vorigen Methoden aus $n_{neu,SZ}$ wie folgt resultieren:

$$n_{a,SZ} = \begin{cases} n_{min} & - \text{ falls } n_{neu,SZ} < n_{min} \\ n_{neu,SZ} & - \text{ falls } n_1 \le n_{neu,SZ} \le n_{max} \\ n_{max} & - \text{ falls } n_{neu,SZ} > n_{max} \end{cases} \tag{6.8}$$

Dabei kann folgendes Problem auftauchen: Liegen die beiden Gruppenraten $p_{1,B}$ und $p_{1,K}$ sehr eng beieinander, so resultiert eine sehr hohe Fallzahl, da bei der Fallzahlformel (2.20) dieser Unterschied quadratisch in den Nenner eingeht. Es empfiehlt sich daher das beschriebene Verfahren wie folgt zu modifizieren (kurz: SZM):

Es wird nur die Rate in der Kontrollgruppe ($p_{1,K}$) gemäß obiger Formel geschätzt. Anschließend wird $p_{1,B}$ aus $p_{1,K}$ und der bei der Fallzahlplanung benutzten Relation zwischen p_B und p_K berechnet.

6.2.3 Ein Konfidenzintervall-basiertes Verfahren

Der Zusammenhang zwischen der Fallzahl und der aktualisierten Schätzung der Ereignisrate in der Kontrollgruppe stellt sich für den multiplikativen Unterschied r = 0.5 (Halbierung der Ereignisrate in der Behandlungsgruppe) wie folgt dar:

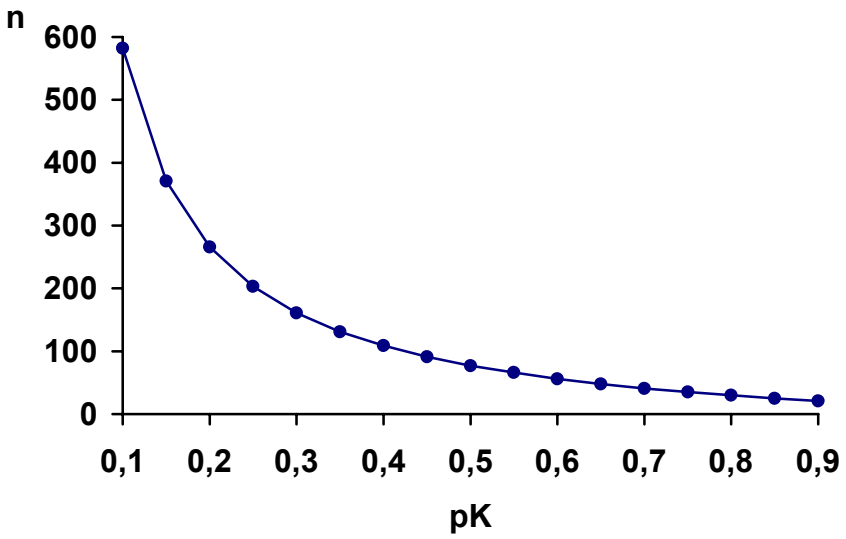


Abbildung 6.1: *Zusammenhang zwischen n und p_K – bei einer Fallzahlplanung mit: α=0.05, 1-β=0.9, $p_{0,B} = r \bullet p_{0,K}$, r=0.5*

Die Fallzahl muss somit nur für den Fall, dass die Ereignisrate in der Kontrollgruppe kleiner ist als bei der Fallzahlplanung angenommen wurde, nach oben adjustiert werden. Insbesondere bei sehr kleinem p_K fällt diese Adjustierung groß aus, da ein quadratischer Zusammenhang besteht. Ähnliche Bilder ergeben sich für andere Werte von r. Je kleiner das r (großer relativer Effekt, z.B. bedeutet r=0.1, dass die Ereignisrate in der Behandlungsgruppe auf ein Zehntel der Rate in der Kontrollgruppe reduziert wird) desto geringer ist die resultierende Fallzahl. Ebenso gilt: Je größer r, desto größer die Fallzahl. In der folgenden Abbildung ist r = 0.7:

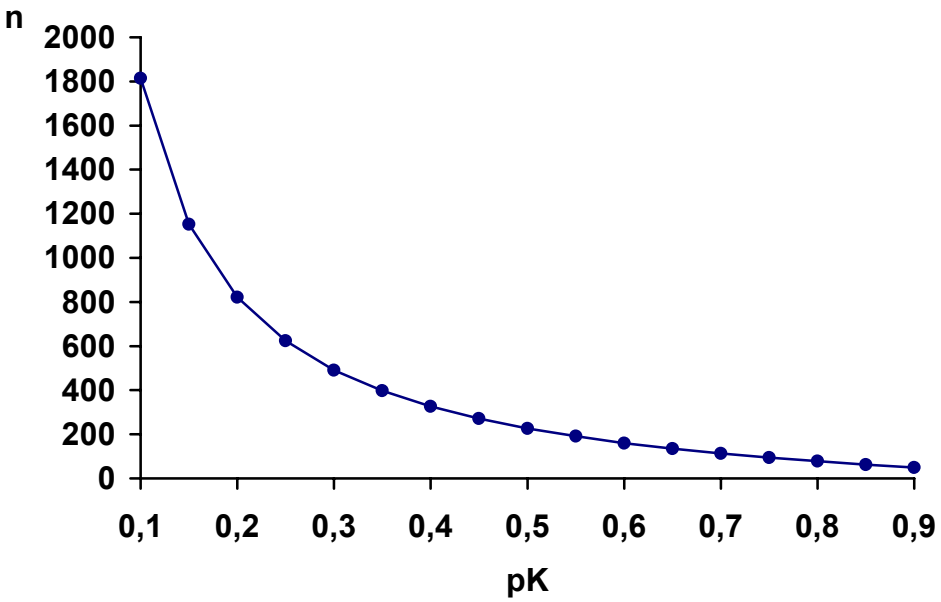


Abbildung 6.2: *Zusammenhang zwischen n und p_K – bei einer Fallzahlplanung mit: α=0.05, 1-β=0.9, $p_{0,B} = r \bullet p_{0,K}$, r=0.7*

Aus diesen Überlegungen heraus ergibt sich eine weitere Adjustierungsmöglichkeit: Die Schätzung der Ereignisrate mittels der unteren Grenze eines geeigneten Konfidenzintervalls. In Kombination mit dem Gould'schen Verfahren ergibt sich somit als entsprechende Schätzung der unteren Grenze eines einseitigen (1-γ)-Konfidenzintervalls:

$$p_{1,\gamma} = \frac{e_1}{n_1} - z_{1-\gamma} \bullet \sqrt{\frac{\frac{e_1}{n_1} \bullet \left(1 - \frac{e_1}{n_1}\right)}{n_1}} \qquad (6.9)$$

Daraus resultieren, analog zu Gould als Schätzung für die Rate in der Kontrollgruppe:

$$p_{1,K,\gamma} = \frac{2 \bullet p_{1,\gamma}}{r+1}$$

$$p_{1,B} = p_{1,K} \bullet r$$

– für den Fall, dass bei der Fallzahlplanung mit einem multiplikativen Zusammenhang zwischen p_B und p_K gearbeitet wurde.

6.3 Vergleich der Methoden

Im Folgenden werden die vorgestellten Methoden hinsichtlich der Auswirkungen auf α, $(1-\beta)$, auf die Verteilung der resultierenden Fallzahl und auf die Weite des Konfidenzintervalls zur Schätzung des Gruppenunterschieds untersucht. Dabei bilden wiederum Simulationsstudien die Basis dieser Evaluationen.

6.3.1 Simulationen

Ausgehend von einer Fallzahlplanung mit folgenden Parametern:

- $\alpha = 5\%$ (zweiseitig)
- $1-\beta = 90\%$
- $p_{0,K} = 60\%$
- $r = 2/3$
- $p_{0,B} = 40\%$

resultiert die Fallzahl:

- $n_p = 2 \bullet 130 = 260$

Mittels SAS® habe ich Simulationen bezüglich folgender „unblinder“ und „blinder“ Adjustierungsmethoden durchgeführt:

- nach dem Verfahren von Herson&Wittes (**HW**)
- nach dem Verfahren von Gould (**G**)
- nach dem Verfahren von Shih&Zhao (**SZ**)
- nach dem modifizierten Verfahren von Shih&Zhao (**SZM**)
- nach dem Konfidenzintervall-basierten Verfahren mit $(1-\gamma)=90\%$ (**KI90**)

Dabei wurde bei allen Verfahren mit der Obergrenze $n_{max}=2 \bullet n_p$ und der Untergrenze $n_{min} = n_p$ gearbeitet. Es fand somit eine Adjustierung nur nach oben statt, bei der die adjustierte Fallzahl pro Gruppe $n_{a,B}$ aus der berechneten Fallzahl n_{neu} folgendermaßen bestimmt wurde.

$$n_{a,B} = \begin{cases} 130 & - \text{ falls } n_{neu} < 130 \\ n_{neu} & - \text{ falls } 130 \le n_{neu} \le 260 \\ 260 & - \text{ falls } n_{neu} > 260 \end{cases}$$

Als Vergleich wurde dazu die Strategie „keine Adjustierung (**KA**)“ durchgeführt.

Dabei habe ich folgende Szenarien bzgl. p_B und p_K untersucht:

Szenario	$\mathbf{p_K}$	$\mathbf{p_B}$
1	0.6	0.4
2	0.6	0.6
3	0.2	0.2
4	0.8	0.8
5	0.4	0.6
6	0.6	0.8
7	0.8	0.5
8	0.4	0.2
9	0.3	0.2
10	0.6	0.5

Tabelle 6.1: *Simulationsstudie: Überblick Szenarien*

Einige der untersuchten Szenarien (0.2/0.2, 0.8/0.8 – 0.6/0.4, 0.4/0.6 – 0.6/0.8, 0.4/0.2) weisen zwar eine paarweise Symmetrie zu 0.5 auf und scheinen auf den ersten Blick überflüssig, jedoch zeigten die Simulationen durchaus Unterschiede zwischen diesen Paaren auf.

Als Zeitpunkt der Adjustierung wurde die Hälfte der geplanten Fälle gewählt ($n_1=2 \bullet 65=130$). Insgesamt wurden pro Szenario 100.000 Simulationen durchgeführt – zur Motivation für diese Anzahl s. Kapitel 5.1.1.

6.3.2 Auswirkungen auf alpha

Angegeben ist jeweils der Anteil abgelehnter Nullhypothesen:

Szenario		**Strategie**					
p_K	p_B	**KA**	**HW**	**G**	**SZ**	**SZM**	**KI90**
0.6	0.6	**0.04967**	**0.04838**	**0.04938**	**0.05468**	**0.04911**	**0.04959**
		0.048323	0.047050	0.048037	0.053271	0.047771	0.048244
		0.051017	0.049710	0.050723	0.056089	0.050449	0.050936
0.2	0.2	**0.04958**	**0.05004**	**0.04952**	**0.05325**	**0.05066**	**0.04961**
		0.048235	0.048689	0.048175	0.051858	0.049301	0.048264
		0.050925	0.051391	0.050865	0.054642	0.052019	0.050956
0.8	0.8	**0.04958**	**0.05005**	**0.04924**	**0.05325**	**0.05049**	**0.04905**
		0.048235	0.048699	0.047899	0.051858	0.049133	0.047711
		0.050925	0.051401	0.050581	0.054642	0.051847	0.050389

Tabelle 6.2*: Anteil abgelehnter Nullhypothesen incl. 95%-KI*

Es stellt sich heraus, dass nur das Verfahren SZ zu einer Inflation des vorgegebenen Signifikanzniveaus führt. Alle anderen Verfahren, inklusive des modifizierten Verfahrens nach SZ, halten α ein.

Hier stimmen die Ergebnisse für die beiden „symmetrischen“ Szenarien (s.o.) 0.2/0.2 und 0.8/0.8 lediglich (und erwartungsgemäß) für die Strategie „Keine Adjustierung“ überein.

6.3.3 Auswirkungen auf die Verteilung von N

Grundsätzlich handelt es sich bei der Verteilung von N_a um eine auf den Adjustierungsbereich (130-260) gestutzte Verteilung. Die Verteilung von N_{neu} besitzt dagegen eine sehr viel größere Streuung, da hier keine Stutzung vorgenommen wird.

Abgetragen ist jeweils der Median, das untere und obere Quartil der Verteilung der adjustierten Fallzahl n_a pro Gruppe – für die jeweiligen Szenarien. Abweichungen von diesem Darstellungstyp werden entsprechend hervorgehoben:

Szenarien, die zu einer geringen Adjustierung der Fallzahl führen

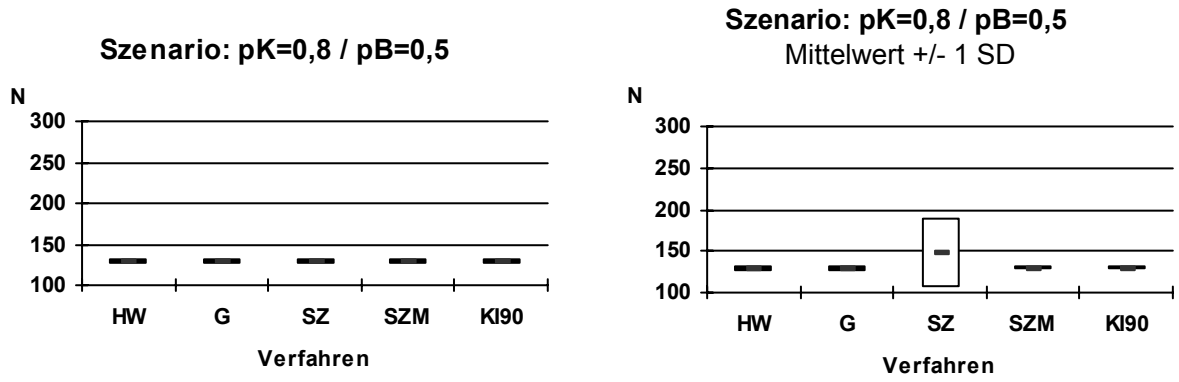
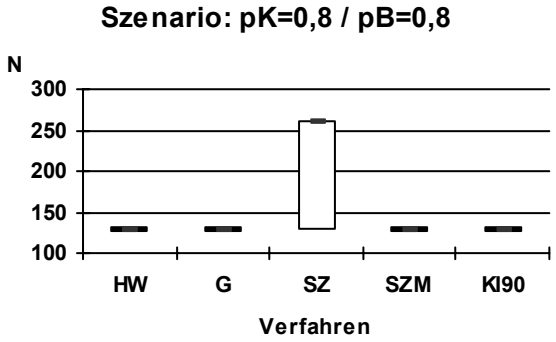


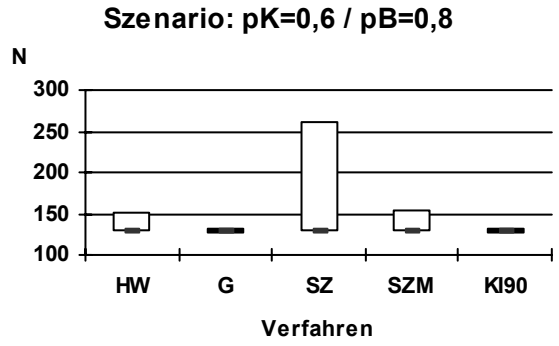
Abbildung 6.3: *Simulationsstudie: Verteilung von N (Median, Quartile bzw. Mittelwert, Standardabweichung) – Szenario: p_K=0.8 / p_B=0.5*

Für den Fall, dass p_K bei der Fallzahlplanung stark unterschätzt wurde, führen alle Verfahren im Wesentlichen zu einer Beibehaltung der geplanten Fallzahl. Ohne eine Untergrenze würden alle Verfahren die Fallzahl weit nach unten adjustieren. Betrachtet man den Mittelwert und die Standardabweichung der Verteilungen (rechte Grafik), wird deutlich, dass das Verfahren SZ zu einer starken Streuung der Fallzahl führt.



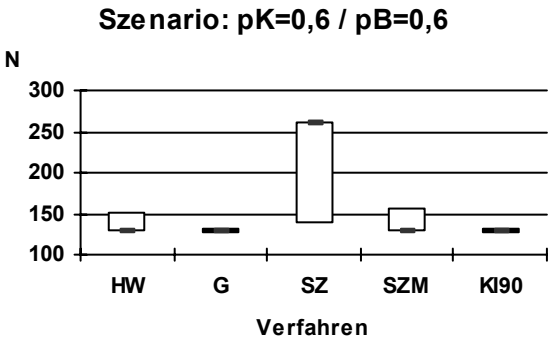
Bei diesem Szenario wird ein weiteres Manko des Verfahrens SZ deutlich: Hier führt die Schätzung beider Raten logischerweise zu einem minimalen Unterschied zwischen den Gruppen. In ca. 9% der Simulationen war dieser Unterschied gleich Null, so dass keine Fallzahlberechnung durchgeführt werden konnte. Als Konsequenz steigt die Fallzahl unnötigerweise an – im Median auf 260, im Mittel auf 215. Durch die Modifikation (Verfahren SZM) wird dieses Manko behoben.

Abbildung 6.4: *Simulationsstudie: Verteilung (Median, Quartile) von N Szenario: p_K=p_B=0.8*



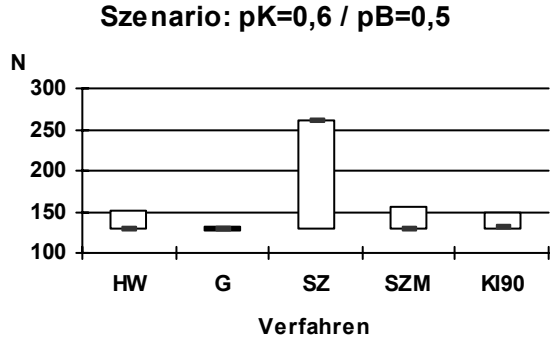
Bei diesem Szenario zeigt sich ein ähnliches Bild wie bei p_K=0.8 / p_B=0.8: SZ weist eine sehr große Streuung auf – während hier die Unterschiede in der Lage der Verteilung geringer sind (der Median liegt bei allen Verfahren bei 130, die Mittelwerte liegen zwischen 130 (G, KI) und 173 (SZ).

Abbildung 6.5 *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K=0.6 / p_B=0.8



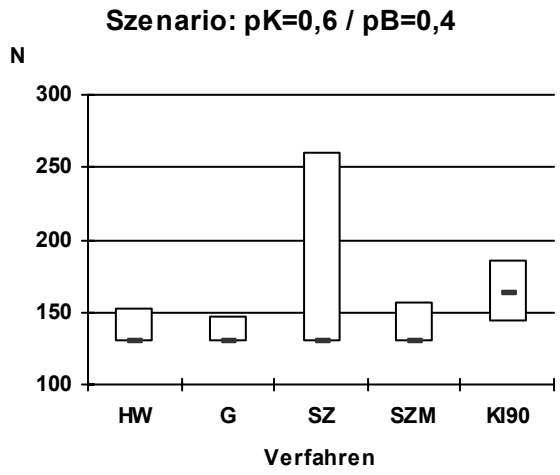
Der einzige wesentliche Unterschied zum vorigen Szenario besteht darin, dass nun beim Verfahren SZ die mediane Fallzahl von 130 auf 260 hochschnellt – hervorgerufen durch die Tatsache, dass nun p_K gleich p_B ist.

Abbildung 6.6: *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K= p_B=0.6



Ein nahezu identisches Bild ergibt sich bei diesem Szenario.

Abbildung 6.7: *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K=0.6 / p_B=0.5



Dieses Szenario entspricht der bei der Fallzahlplanung getroffenen Annahmen.

Es kommt bei allen Verfahren zu einer leichten Adjustierung nach oben – diese fällt bei KI bezüglich des Medians am deutlichsten aus.

Da sich HW nur an p_K orientiert, führt diese Strategie zu dem gleichen Ergebnis wie beim vorigen Szenario. Bei SZ reduziert sich die mediane Fallzahl wieder auf 130, da dieses Szenario von der Nullhypothese entfernt ist.

Abbildung 6.8: *Simulationsstudie: Verteilung (Median, Quartile) von N Szenario: p_K=0.6 / p_B=0.4*

Szenarien, die zu einer starken Adjustierung führen

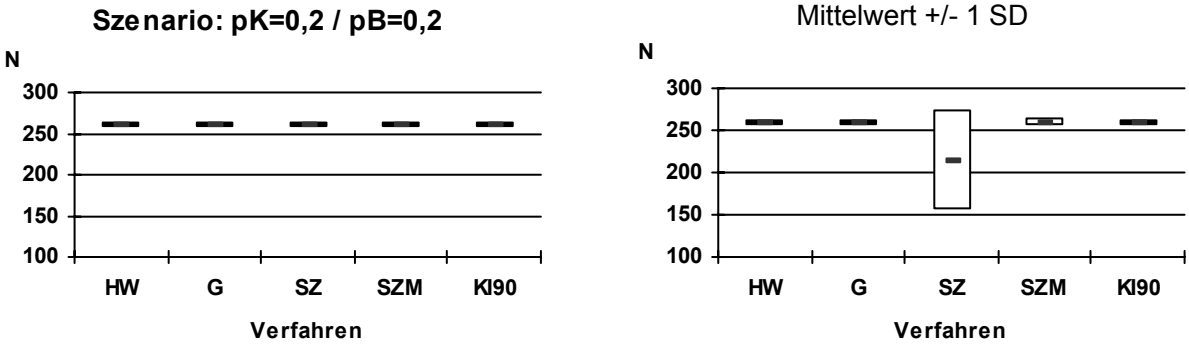


Abbildung 6.9: *Simulationsstudie: Verteilung von N - Szenario: p_K= p_B=0.2*

Im Fall der starken Überschätzung wird bei allen Verfahren bis zur Obergrenze nachadjustiert. Ohne eine Obergrenze würde die mediane Fallzahl bis auf 700 „explodieren“. Die Sonderrolle von SZ bezüglich der Streuung wird auch hier deutlich.

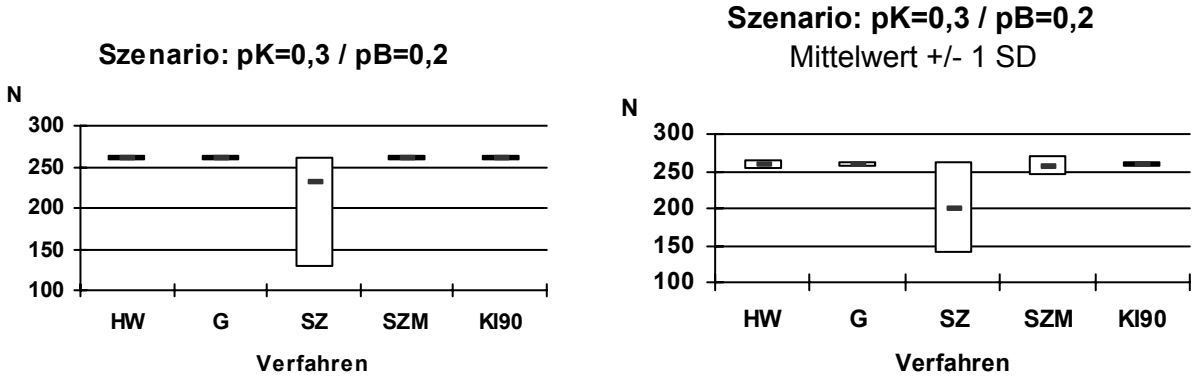
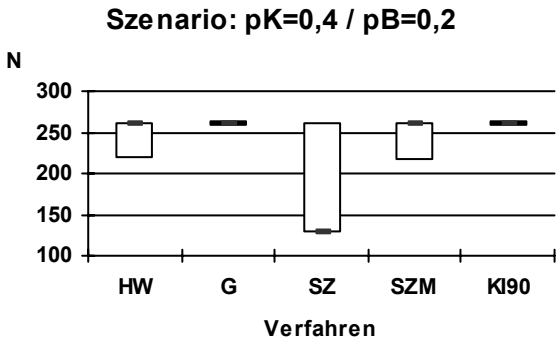


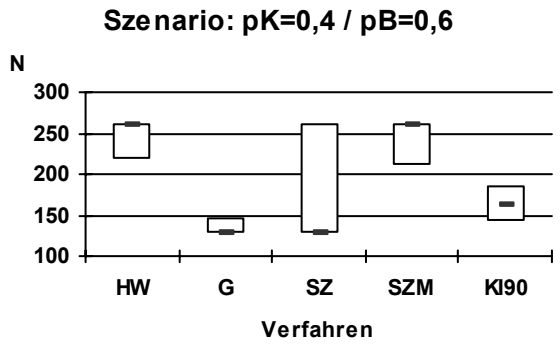
Abbildung 6.10: *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K=0.3 / p_B=0.2

Bei diesem Szenario führt das Verfahren SZ zu einer geringeren Adjustierung als die anderen Verfahren, die alle konstant fast bis zur Obergrenze adjustieren. Außerdem wird hier wieder deutlich, dass das Verfahren SZ eine deutlich höhere Streuung der Fallzahl bewirkt.



Der einzige Unterschied zum vorherigen Szenario liegt darin, dass sich die Verteilungen auf einem etwas niedrigeren Niveau befinden – hervorgerufen durch das größere p_K.

Abbildung 6.11: *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K=0.4 / p_B=0.2



Hier zeigt sich ein uneinheitliches Bild: Während HW und SZM zu einer maximalen Adjustierung führen, fällt die Adjustierung bei den anderen Verfahren deutlich geringer aus. Bei beiden Verfahren liegt der Grund dafür in der Fixierung auf p_K. Bei HW wird p_K unblind aus den Daten geschätzt, während bei SZM diese Schätzung blind erfolgt. Da p_K mit 0.4 sehr viel geringer ausfällt, als bei der Fallzahlplanung angenommen wird in den meisten Fällen bis zur Obergrenze adjustiert.

Abbildung 6.12: *Simulationsstudie: Verteilung (Median, Quartile) von N*
Szenario: p_K=0.4 / p_B=0.6

6.3.4 Auswirkungen auf die Power

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.6	0.4	**0.90825**	**0.94339**	**0.92044**	**0.96559**	**0.94477**	**0.95067**
		0.90646	0.94196	0.91876	0.96446	0.94335	0.94933
		0.91004	0.94482	0.92212	0.96672	0.94619	0.95201

Tabelle 6.3: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI*
Szenario: p_K=0.6 / p_B=0.4

Für die bei der Fallzahlplanung benutzte Lage der Alternative führt die Einfach-Strategie „Keine Adjustierung“ erwartungsgemäß zu einer Einhaltung der geforderten Power von 90%. Alle Adjustierungsstrategien führen in diesem Fall zu einer Erhöhung der Power, da die Fallzahl nur nach oben adjustiert wird. Der größte Powergewinn resultiert aus der Strategie SZ, da hier die Fallzahl im Mittel am höchsten nach oben adjustiert wird.

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.4	0.6	**0.90825**	**0.98047**	**0.92002**	**0.96559**	**0.97673**	**0.94949**
		0.90646	0.97961	0.91834	0.96446	0.97580	0.94813
		0.91004	0.98133	0.92170	0.96672	0.97766	0.95085

Tabelle 6.4: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI*
Szenario: p_K=0.4 / p_B=0.6

Bei diesem Szenario ist die Ereignisrate in der Kontrollgruppe niedriger als in der Behandlungsgruppe. Da der absolute Unterschied zwischen den Gruppen jedoch unverändert zu den bei der Fallzahlplanung getroffenen Annahmen ist, wird auch hier bei keiner Adjustierung die Power eingehalten. Der größte Powergewinn resultiert hier aus der Strategie HW.

Hier stimmen die Ergebnisse für die beiden „symmetrischen“ Szenarien (s.o.) 0.6/0.4 und 0.4/0.6 lediglich (und erwartungsgemäß) für die Strategie „Keine Adjustierung“ überein.

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.6	0.8	**0.94584**	**0.94680**	**0.94602**	**0.98050**	**0.95045**	**0.94630**
		0.94444	0.94541	0.94462	0.97964	0.94910	0.94490
		0.94724	0.94819	0.94742	0.98136	0.95180	0.94770
0.8	0.5	**0.99939**	**0.99954**	**0.99951**	**0.99986**	**0.99949**	**0.99940**
		0.99924	0.99941	0.99937	0.99979	0.99935	0.99925
		0.99954	0.99967	0.99965	0.99993	0.99963	0.99955

Tabelle 6.5: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI*
Szenarien: p_K=0.6 / p_B=0.8, p_K=0.8 / p_B=0.5

Bei diesen Szenarien ist der tatsächliche Unterschied zwischen den Gruppen mindestens genauso groß wie bei der Fallzahlplanung berücksichtigt. Dadurch wird die Power eingehalten ohne adjustieren zu müssen. Jede Adjustierung führt zu einer zusätzlichen Erhöhung der Power.

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.4	0.2	**0.94584**	**0.99893**	**0.99874**	**0.98050**	**0.99847**	**0.99914**
		0.94444	0.99873	0.99852	0.97964	0.99823	0.99896
		0.94724	0.99913	0.99896	0.98136	0.99871	0.99932

Tabelle 6.6: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI Szenario: $p_K=0.4 / p_B=0.2$*

Auch hier wird die Power beim Festhalten an der geplanten Fallzahl nicht verringert. Im Gegenteil – sie steigt auf ca. 94% an. Dies resultiert aus der Tatsache, dass bei dieser Konstellation von p_K und p_B bereits eine Fallzahl von n=109 (pro Gruppe) ausreichend gewesen wäre, um eine Power von 90% zu garantieren.
Hier stimmen wieder die Ergebnisse für die beiden „symmetrischen“ Szenarien (s.o.) 0.6/0.8 und 0.4/0.2 für die Strategie „Keine Adjustierung“ überein.

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.3	0.2	**0.46382**	**0.75543**	**0.75201**	**0.64937**	**0.75261**	**0.75537**
		0.46073	0.75277	0.74933	0.64641	0.74994	0.75271
		0.46691	0.75809	0.75469	0.65233	0.75528	0.75803

Tabelle 6.7: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI Szenario: $p_K=0.3 / p_B=0.2$*

Für den Fall, dass der tatsächliche Unterschied zwischen p_K und p_B geringer ausfällt, als bei der Fallzahlplanung angenommen, sinkt die Power auf 46% ab – wenn nicht adjustiert wird. Die Adjustierungsverfahren können diesen Powerverlust verringern, dabei schneidet das Verfahren SZ am schlechtesten ab. Alle anderen Verfahren erhöhen die Power auf ca. 75%.

Szenario		Strategie					
p_K	p_B	KA	HW	G	SZ	SZM	KI90
0.6	0.5	**0.37959**	**0.40252**	**0.38062**	**0.54111**	**0.41678**	**0.40058**
		0.37658	0.39948	0.37761	0.53802	0.41372	0.39754
		0.38260	0.40556	0.38363	0.54420	0.41984	0.40362

Tabelle 6.8: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI Szenario: $p_K=0.6 / p_B=0.5$*

Die Kompensation des Powerverlusts fällt hier bei allen Strategien außer SZ deshalb geringer aus, da sich diese Verfahren an p_K orientieren – dieses unterscheidet sich nicht von dem bei der Fallzahlplanung. Der tatsächliche Unterschied zwischen p_K und p_B wird bei den Adjustierungsstrategien HW, G, SZM und KI90 nicht berücksichtigt – hier geht der relevante Unterschied in der Form $r = p_B / p_K$ ein.

6.3.5 Auswirkungen auf die Weite des Konfidenzintervalls

Die Länge des 95%-Konfidenzintervalls ist:

$$2 \bullet 1.96 \bullet SE(p_K - p_B).$$

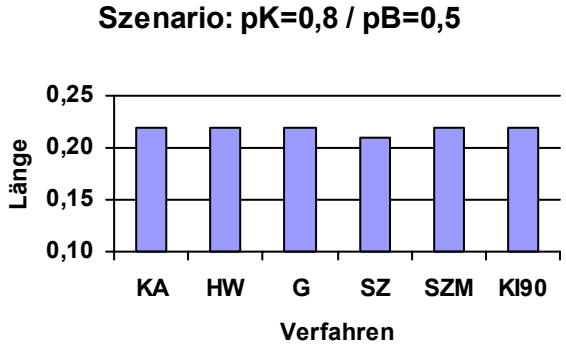
Da:

$$SE(p_K - p_B) = \sqrt{\frac{p_K(1-p_K)}{n_K} + \frac{p_B(1-p_B)}{n_B}} \qquad (6.10)$$

und:

$$n_K = n_B = n_a/2$$

wird die Länge des Konfidenzintervalls durch die adjustierte Fallzahl n_a bestimmt. Je größer n_a, desto kleiner wird das Konfidenzintervall. Somit lassen sich die im Kapitel „Auswirkungen auf die Verteilung von n“ getroffenen Beobachtungen entsprechend übertragen.



Hier unterscheiden sich die Verfahren nicht nennenswert, da hier die Fallzahlen nur sehr geringfügig adjustiert werden – es bleibt im wesentlichen bei der geplanten Fallzahl.

Abbildung 6.13: *Simulationsstudie: Mittlere Länge des 95%-KI – Szenario: p_K=0.8 / p_B=0.5*

Hier führt nur das Verfahren nach SZ zu einer relevanten Erhöhung der Fallzahl und somit zu einer Verkleinerung des Konfidenz-intervalls.

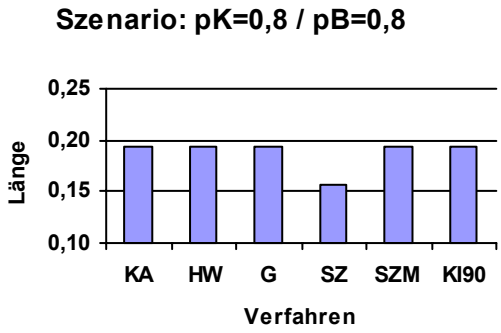
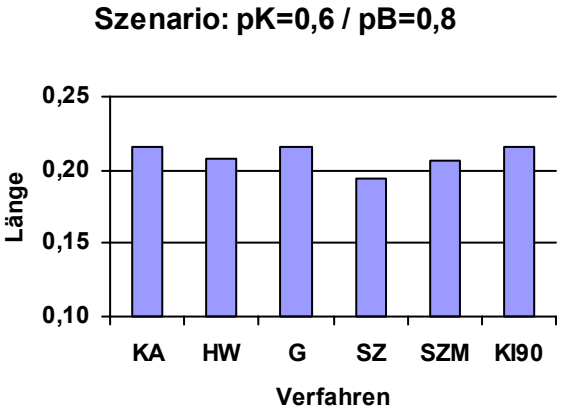


Abbildung 6.14: *Simulationsstudie: Mittlere Länge des 95%-KI – Szenario: $p_K = p_B = 0.8$*



Da hier nur bei SZ die Fallzahl adjustiert wird, kommt es zu der Reduktion in der Länge des Konfidenzintervalls bei diesem Verfahren.

Abbildung 6.15: *Simulationsstudie: Mittlere Länge des 95%-KI – Szenario: $p_K = 0.6 / p_B = 0.8$*

Es folgen 3 ähnliche Situationen:

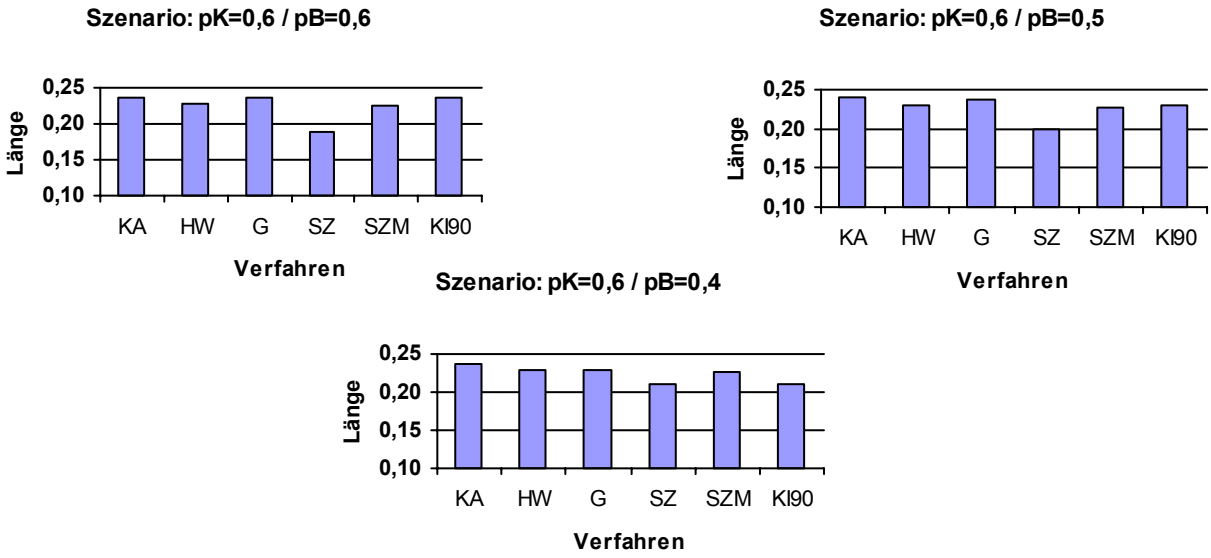


Abbildung 6.16: *Simulationsstudie: Mittlere Länge des 95%-KI Szenarien: $p_K = 0.6 / p_B = 0.6$, $p_K = 0.6 / p_B = 0.5$, $p_K = 0.6 / p_B = 0.4$*

Bei den folgenden Szenarien wird die Überlegenheit der Verfahren HW, G, SZM und KI90 gegenüber dem Festhalten an der geplanten Fallzahl deutlich. Es kommt zu einer ausgeprägten Reduzierung der Länge des Konfidenzintervalls:

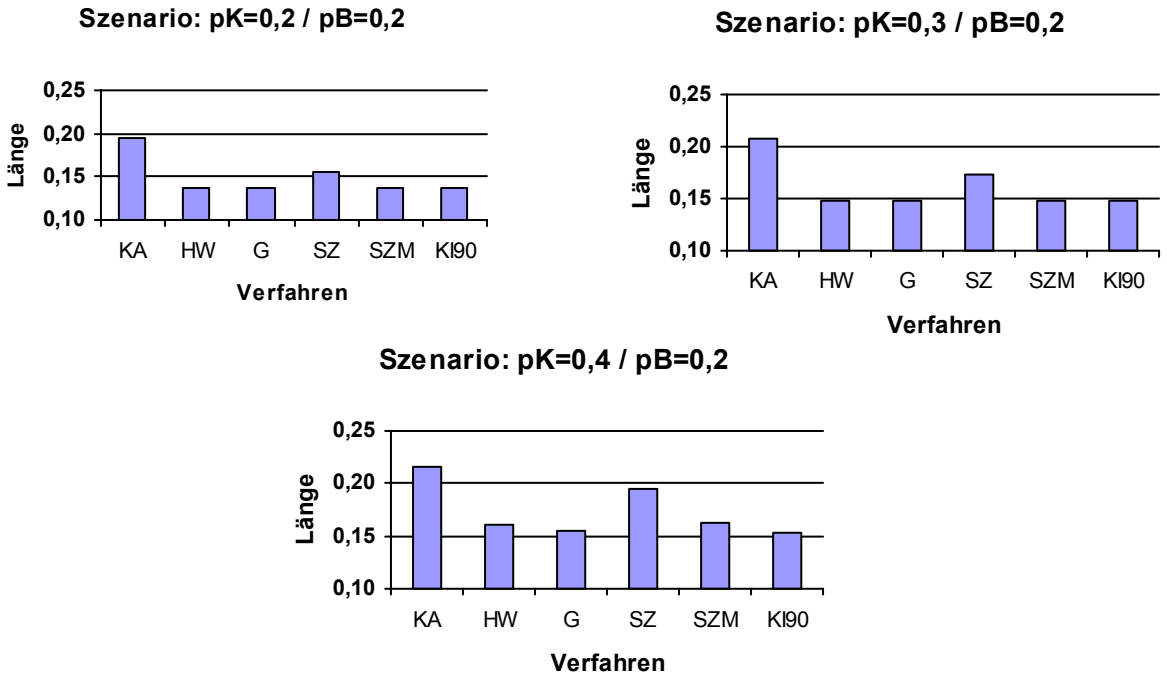


Abbildung 6.17: *Simulationsstudie: Mittlere Länge des 95%-KI*
Szenarien: p_K=0.2 / p_B=0.2, p_K=0.3 / p_B=0.2, p_K=0.4 / p_B=0.2

Analog zur Verteilung der adjustierten Fallzahl haben hier nur die Verfahren HW und SZM einen relevanten Effekt.

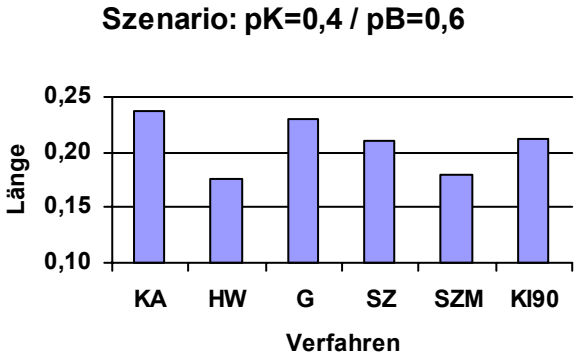


Abbildung 6.18: *Simulationsstudie: Mittlere Länge des 95%-KI – Szenario: p_K=0.4 / p_B=0.6*

6.4 Zusammenfassung

Abschließend bleibt festzustellen, dass das Verfahren nach SZ aus mehreren Gründen problematisch ist: Es hält das vorgegebene Signifikanzniveau nicht ein, führt insbesondere bei Zutreffen der Nullhypothese zu unnötig großen Fallzahlen und weist eine sehr große Streuung auf. Somit ist von diesem Verfahren abzuraten.

Bei den verbleibenden blinden und unblinden Verfahren zeigen sich nur geringe Unterschiede, so dass grundsätzlich eines der blinden Verfahren angewandt werden sollte. Hierbei weisen die Verfahren nach Gould und das Konfidenzintervall-basierte Verfahren den Vorteil im Vergleich zum modifizierten Verfahren nach SZ auf, dass keine Veränderung am ursprünglichen Studiendesign (stratifizierte Randomisierung) durchgeführt werden muss. Dabei bietet das Verfahren KI den zusätzlichen Vorteil, dass eine grundsätzlich höhere Power erzielt werden kann.

7 Fallzahladjustierung bei Repeated Measurements

Zu diesem Thema findet sich in der Literatur lediglich das Verfahren von (Shih und Gould 1995). Dabei wird davon ausgegangen, dass als „summary statistic“ der Mehrfachmessungen die Steigung benutzt wird. Eine weitere Voraussetzung besteht darin, dass zum Zeitpunkt der Fallzahladjustierung die Studie für einen Teil der Patienten bereits abgeschlossen sein muss. Somit können alle vorliegenden Messungen zur Schätzung der benötigten Steigungsparameter (Streuung innerhalb der Patienten, Streuung zwischen den Patienten) benutzt werden. Jedoch ist dann sehr häufig die Rekrutierungsphase bereits abgeschlossen, so dass die Fallzahladjustierung in diesem Fall zu spät kommt. Im Folgenden werde ich daher ein neues Verfahren vorstellen, welches dieses Manko behebt.

7.1 Das Verfahren

In vielen klinischen Studien ist die Veränderung eines klinischen Parameters im Laufe einer Therapie die primäre Zielgröße. Für die Fallzahlplanung einer solchen Studie werden Annahmen bezüglich der Standardabweichung der Differenz benötigt. Diese Annahmen, die oft aus ähnlichen Vorstudien stammen, können sich im Laufe der Studie als zu niedrig herausstellen, was einen Verlust der Power der Studie nach sich zieht.
Eine klassische Fallzahladjustierung arbeitet mit Zielgröße-Daten der zuerst rekrutierten Patienten und nimmt dann eine Schätzung der Standardabweichung vor. Jedoch liegen bei diesem Studientyp die ersten Beobachtungen der Zielgröße erst dann vor, wenn die ersten Patienten die Studie abgeschlossen haben. Zu diesem Zeitpunkt ist in der Regel die Rekrutierung bereits beendet. Somit ist eine eventuelle Nachrekrutierung von Patienten aus logistischen und ökonomischen Gründen nur sehr schwer möglich.
Die hier vorgestellte Methode erlaubt es, eine Fallzahladjustierung während der Rekrutierungsphase durchzuführen. Dazu werden die Baseline-Daten und die ersten Follow-up-Daten benutzt - mit dem Ziel eine Schätzung der Standardabweichung der Differenz (Therapieende – Therapieanfang) vorzunehmen.

Sei x_{ijk} die k-te Messung des j-ten Patienten in Gruppe i – mit $i \in \{B ; K\}$; $j=1, ..., n_i$, $n_B = n_K = n/2$; $k=1, ..., t$ (t-1: Dauer der Studie).

Die für die Fallzahlplanung benutzte Zielvariable ist die Differenz der letzten zur ersten Messung:

$$d_{ij}=x_{ijt}-x_{ij1}$$

mit:

$$D_i \sim N(\mu_i,\sigma)$$

Dann ergibt sich nach (2.14) als geplante Fallzahl:

$$n_p = \left(4 \bullet \left[\left(\frac{s_0}{\delta} \right)^2 \left(z_{1-\alpha/2} + z_{1-\beta} \right)^2 \right] \right) + 1$$

mit: δ: klinisch relevante Differenz $| \mu_K - \mu_B |$

s_0: Schätzung von σ

Eine „klassische“ Fallzahladjustierung beim Vergleich von Mittelwerten, wie in Kapitel 5 vorgestellt, ist bei diesem Studientyp nicht anwendbar, da zur Durchführung Daten vom Studienende (X_E) vorliegen müssen - die Rekrutierung ist dann i.d.R. bereits abgeschlossen. Die Idee besteht nun darin, die Daten für die Fallzahladjustierung zu nutzen, die während der Rekrutierungsphase gesammelt wurden; s. dazu auch folgendes Schema:

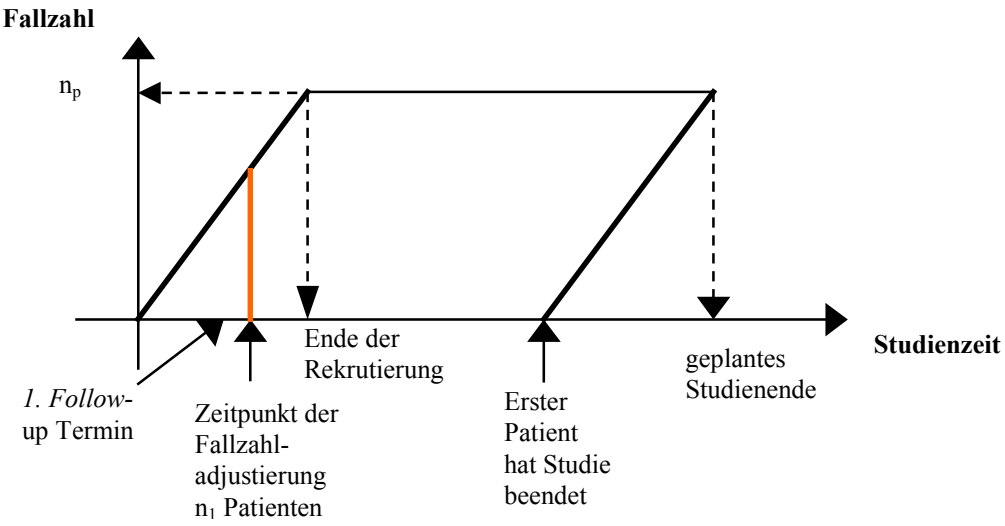


Abbildung 7.1: *Schema zur Fallzahladjustierung bei Repeated Measurements*

Zum Zeitpunkt der Adjustierung liegen dann folgende Daten vor:

X_{ij1}, mit $i \in \{B ; K\}$; $j=1, ..., n_{i1}$; $n_{B1} = n_{K1} = n_1/2$;

X_{ij2}, mit $i \in \{B ; K\}$; $j=1, ..., n_{i2}$; $n_{B2} = n_{K2} = n_2/2$;

mit: n_1 : Anzahl der Patienten mit Anfangsmessung

n_2 : Anzahl der Patienten mit 1. Follow-up-Messung

Unter der Voraussetzung:

$$X_{ik} \sim N(\mu_{ik},\sigma_k)$$

gilt für die Differenz zwischen erster und letzter Messung

$$d_{ij}=x_{ijt}-x_{ij1}:$$

$$D_i \sim N(\mu_{Di},\sigma_D)$$

mit:

$$\sigma_D^2 = \sigma_1^2 + \sigma_t^2 - 2 \bullet \rho_{1t} \bullet \sigma_1 \bullet \sigma_t \tag{7.1}$$

Die einzelnen Komponenten, die sich zu σ_D^2 zusammensetzen, werden nun – nach Entblindung - wie folgt geschätzt:

$$s_{i1} = \sqrt{\frac{1}{n_{i1}-1} \sum_{j=1}^{n_{i1}} \left(x_{ij1} - \overline{x}_{i.1} \right)^2} \tag{7.2}$$

Standardabweichung am Anfang der Studie – pro Gruppe

$$s_{i2} = \sqrt{\frac{1}{n_{i2}-1} \sum_{j=1}^{n_{i2}} \left(x_{ij2} - \overline{x}_{i.2} \right)^2} \tag{7.3}$$

Standardabweichung bei der ersten Follow-up-Messung – pro Gruppe

$$s_{it} = \max \left\{ s_{i1} , s_{i1} + \left(t \bullet \left(s_{i2} - s_{i1} \right) \right) \right\} \tag{7.4}$$

lineare Extrapolation der beobachteten Standardabweichungen – pro Gruppe, mit s_{i1} als Untergrenze

$$r_{1t,i} = \left(r_{12,i}\right)^{t-1} \tag{7.5}$$

mit $r_{12,i}$ – Pearson-Korrelationskoeffizient für Messzeitpunkte 1 und 2 in Gruppe i
Autoregression erster Art

$$s_{iD} = \sqrt{s_{i1}^2 + s_{it}^2 - 2 \bullet r_{1t,i} \bullet s_{i1} \bullet s_{it}} \tag{7.6}$$

Schließlich wird als Schätzung für σ_D benutzt:

$$s_D = \sqrt{\frac{\left(\frac{n_1}{2} - 1\right)(s_{BD}^2 + s_{KD}^2)}{n_1 - 2}} \tag{7.7}$$

Bei Anwendung einer „Adjustierung nach unten oder oben“ (vgl. Kapitel 5) ergibt sich somit als adjustierte Fallzahl:

$$n_{a,UO} = \max\left\{\left[\left(\frac{s_D}{s_0}\right)^2 n_p\right] + 1, n_1\right\} \tag{7.8}$$

Bei Anwendung einer „Adjustierung nur nach oben“ (vgl. Kapitel 5) resultiert als adjustierte Fallzahl:

$$n_{a,O} = \begin{cases} n_p & , \text{falls } s_D \le s_0 \\ \left[\left(\frac{s_D}{s_0}\right)^2 n_p\right] + 1 & , \text{falls } s_D > s_0 \end{cases} \tag{7.9}$$

7.2 Simulationsstudie

Zur Beurteilung dieser beiden Strategien der Fallzahladjustierung habe ich Simulationen durchgeführt. Dabei bin ich von folgender Situation ausgegangen:

- Anzahl der wiederholten Messungen: 5 – zu den Zeitpunkten: k=0, 1, 2, 3, 4 Jahre – t=4
- Fallzahlplanung:
  - α=5% (zweiseitig)
  - 1-β=90%
  - μ_B=1, μ_K=0 $\rightarrow$ δ=1
  - s_0=2
  - 20% drop-outs
  - $\Rightarrow n_p$= 2 x 108 =216
- Rekrutierungszeit: 2 Jahre
- Zeitpunkt der Fallzahladjustierung: 1,5 Jahre
- Anzahl der Patienten, die zur Fallzahladjustierung berücksichtigt werden: n_1= 2 • 81, n_2= 2 • 27
- 3 Szenarien bzgl. σ: σ=1 (< s_0) / σ=2 (= s_0) / σ=3 (> s_0)
- 2 Szenarien bzgl. H_0: H_0 trifft zu / H_0 trifft nicht zu
- Vergleich der beiden Adjustierungsverfahren „nach oben“ (**O**) und „nach unten oder oben“ (**UO**) mit „keiner Adjustierung“ (**KA**)

Es wurden 100.000 Simulationen pro Szenario mittels SAS® durchgeführt.

7.2.1 Auswirkungen auf alpha

Im Folgenden sind die Anteile der abgelehnten Nullhypothesen incl. des zugehörigen 95% Konfidenzintervalls für die 3 Szenarien bezüglich σ dargestellt – bei Zutreffen von H_0:

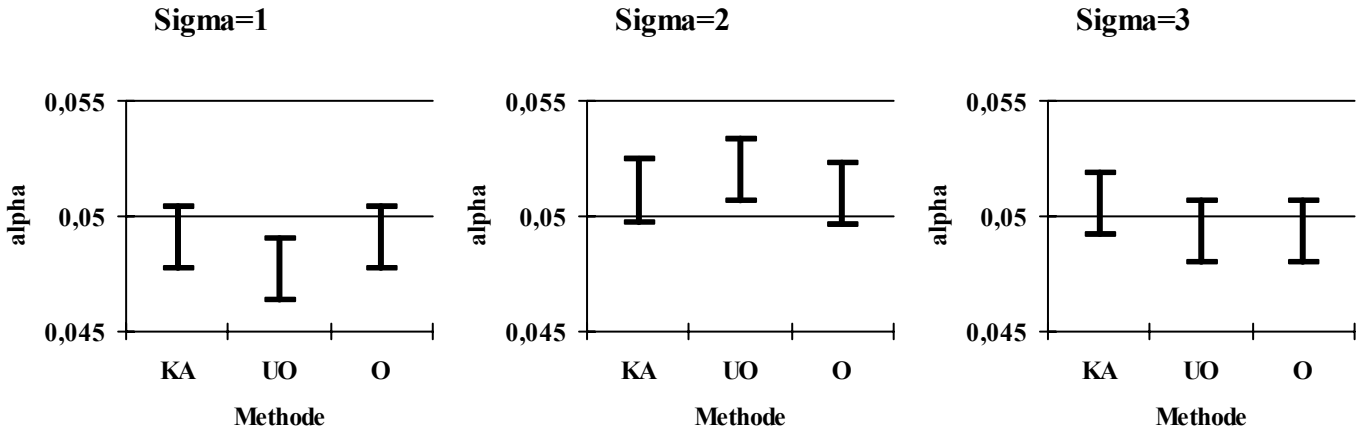


Abbildung 7.2: *Simulationsstudie: Anteil der abgelehnten Nullhypothesen incl. 95%-KI*

Es ergeben sich somit keine gravierenden Abweichungen vom geforderten Signifikanzniveau von 0,05 – bei allen Verfahren. Die Methode UO führt zu einer leichten Inflation des alphas beim Szenario σ=2.

7.2.2 Auswirkungen auf die Verteilung von N

Im Folgenden sind für die Fallzahlen – pro Gruppe – jeweils der Median und die Quartile angegeben. Als Referenzlinie wurde zusätzlich die „wahre“ Fallzahl eingezeichnet – resultierend aus dem „wahren“ σ.

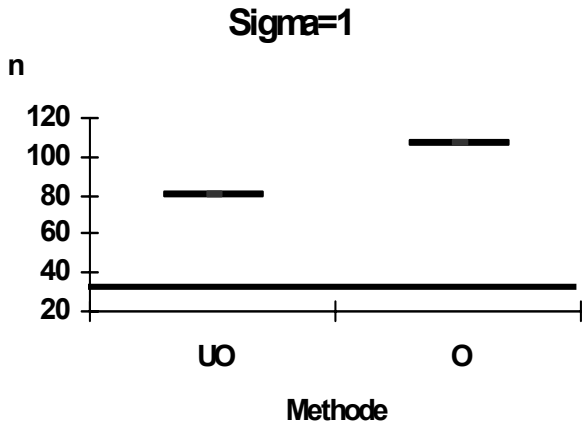


Abbildung 7.3: *Simulationsstudie: Verteilung (Median, Quartile) der adjustierten Fallzahl ($\sigma = 1$)*

Da die Fallzahladjustierung nach 81 rekrutierten Patienten durchgeführt wird, kann die „wahre“ Fallzahl von n=29 natürlich nicht erreicht werden. Bei der Adjustierung UO wird dafür gesorgt, dass kein weiterer Patient rekrutiert wird, während die Strategie O konstruktionsgemäß zu einer Weiterführung der Rekrutierung bis zu den geplanten 107 Patienten führt.

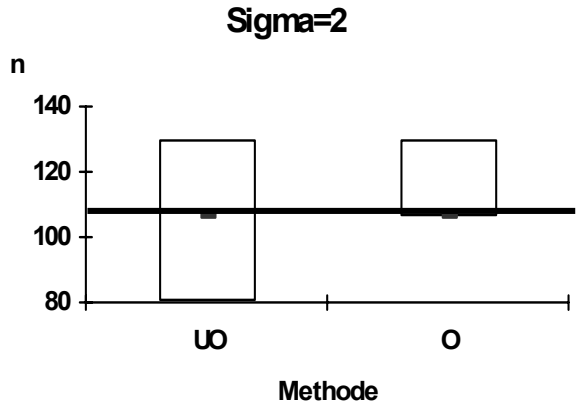


Abbildung 7.4: *Simulationsstudie: Verteilung (Median, Quartile) der adjustierten Fallzahl ($\sigma = 2$)*

Die Strategie UO kann in diesem Fall zu einer Unterschätzung der wahren Fallzahl führen, während O hier nach unten abgesichert ist. Beide Verfahren führen im Median zu korrekten Schätzungen der Fallzahl.

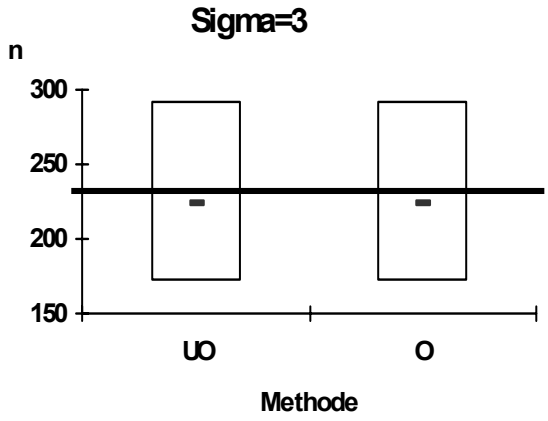


Abbildung 7.5: *Simulationsstudie: Verteilung (Median, Quartile) der adjustierten Fallzahl ($\sigma = 3$)*

Hier führen beide Verfahren zu einer leichten Unterschätzung der tatsächlichen Fallzahl. Die Unterschiede zwischen den Verfahren sind nur marginal.

7.2.3 Auswirkungen auf die Power

Im Folgenden sind wiederum die Anteile der abgelehnten Nullhypothesen incl. des zugehörigen 95% Konfidenzintervalls für die 3 Szenarien bezüglich σ dargestellt – bei Zutreffen von H_1:

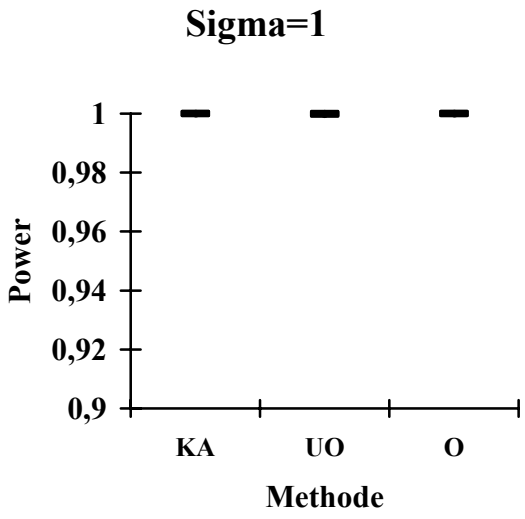


Abbildung 7.6: *Anteil abgelehnter Nullhypothesen incl. 95%-KI ($\sigma = 1$)*

Bei allen Verfahren ist die tatsächliche Power nahe bei 1 – da die „wahre“ Fallzahl (n=29) weit überschritten wird.

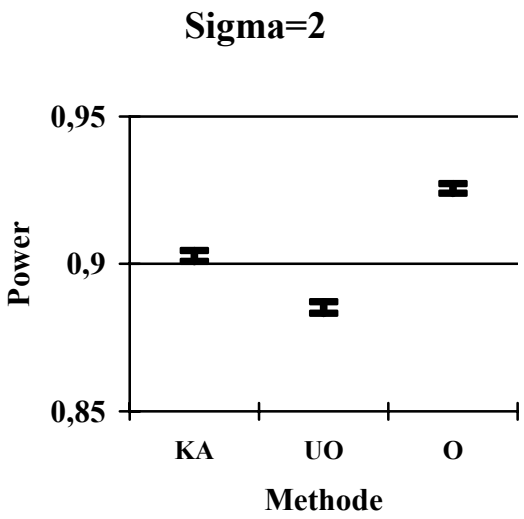


Abbildung 7.7: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI ($\sigma = 2$)*

Hier wird die geforderte (Mindest)-Power von den beiden Verfahren KA und O erreicht – während bei UO hier der Preis dafür bezahlt werden muss, dass die Fallzahl in einigen Fällen fälschlicherweise nach unten adjustiert wird.

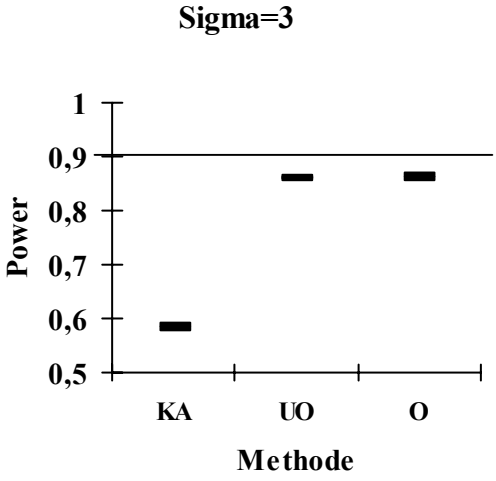


Abbildung 7.8: *Simulationsstudie: Anteil abgelehnter Nullhypothesen incl. 95%-KI ($\sigma = 3$)*

Beide Adjustierungs-Verfahren gewährleisten im Fall „$\sigma > s_0$“ eine deutlich höhere Power als KA - jedoch wird die geplante Power nicht ganz erreicht.

7.2.4 Auswirkungen auf die Weite des Konfidenzintervalls

Im Folgenden ist die mittlere Länge des 95%-Konfidenzintevalls zur Schätzung des Gruppenunterschiedes angegeben.

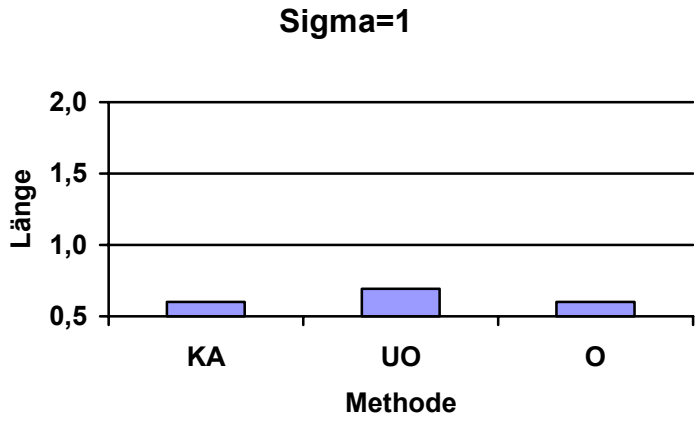


Abbildung 7.9: *Simulationsstudie: mittlere Länge des 95%-KI ($\sigma = 1$)*

Analog zu den Ergebnissen, die bezüglich der Verteilung von n erzielt wurden, liefert UO ein leicht größeres Konfidenzintervall. Die Weite der Konfidenzintervalle stimmen bei O und KA überein – da bei diesem Szenario beide Strategien zum gleichen Resultat führen: die geplante Fallzahl wird beibehalten.

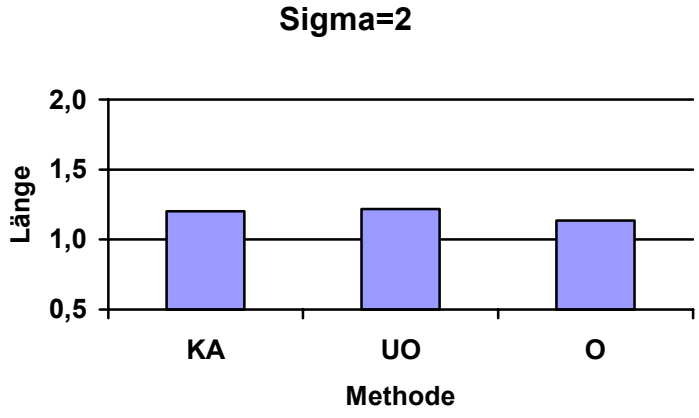


Abbildung 7.10: *Simulationsstudie: mittlere Länge des 95%-KI ($\sigma = 2$)*

Hier kommt es im Vergleich zum vorigen Szenario zu einer deutlichen Vergrößerung der Konfidenzintervalle – analog zu einer Verkleinerung der Power (s.o.). O liefert insgesamt kleinere Konfidenzintervalle, was wiederum die Power-Resultate widerspiegelt.

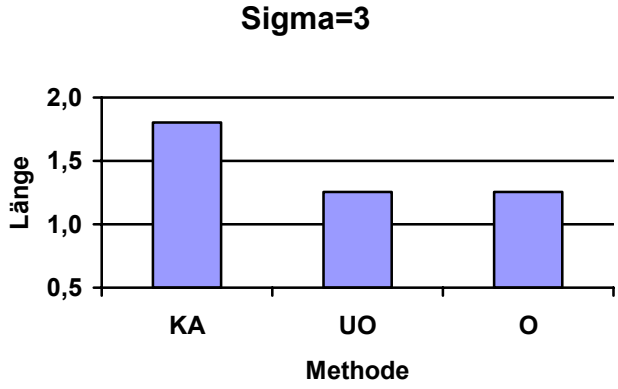


Abbildung 7.11: *Simulationsstudie: mittlere Länge des 95%-KI ($\sigma = 3$)*

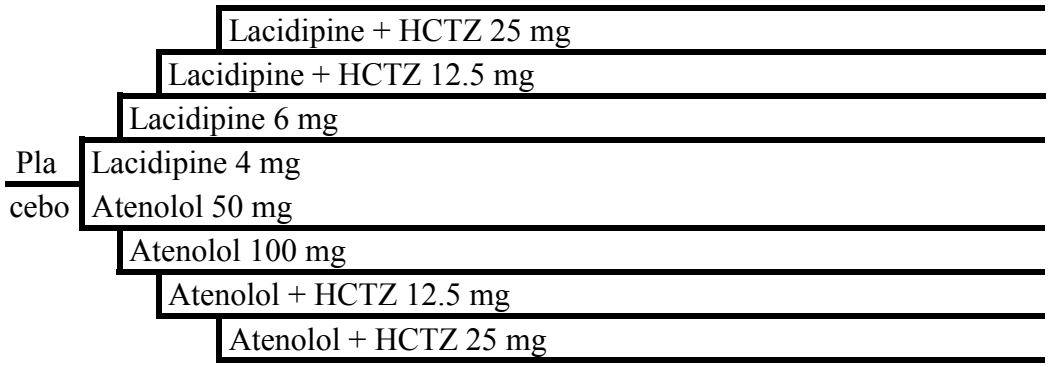
Hier kommt es zu einer weiteren Vergrößerung der Konfidenzintervalle – analog zu der kleiner werdenden Power über die 3 Szenarien. Ebenso kann hier die starke Diskrepanz der beiden Adjustierungsverfahren im Vergleich zu KA beobachtet werden.

Abschließend bleibt festzustellen, dass die Adjustierungsstrategie O hier über alle Beurteilungskriterien hinweg die besten Resultate erzielt.

7.3 Anwendung des Verfahrens bei der ELSA-Studie

Bei der ELSA-Studie (**E**uropean **L**acidipine **S**tudy on **A**therosclerosis) wurde der Einfluss des Calcium-Antagonisten „Lacidipine“ auf die Arteriosklerose untersucht (Zanchetti, Bond et al. 1998). In dieser doppelblinden, randomisierten, multizentrischen Studie wurden die Patienten in eine der beiden Studienarme (Behandlung: Lacidipine, aktive Kontrolle: Atenolol) randomisiert und über einen Zeitraum von 4 Jahren behandelt.

Die folgende Grafik stellt den Studienablauf schematisch dar:



Visit	0	1	2	3	4	5	6	7	8	9	10	11	Follow up
Monat	-1	0	1	3	6	12	18	24	30	36	42	48	48+7d
Ultraschall-Messungen		X				X		X		X		X	
Klinischer Blutdruck	X	X	X	X	X	X	X	X	X	X	X	X	X
24h Blutdruck		X				X		X		X		X	

Abbildung 7.12: *Ablaufschema ELSA*

Die Zielvariable „CBMMax“ beschreibt die mittels B-Mode-Ultraschall gemessene Intima-Media-Dicke der Arteria Carotis – s. dazu folgende Skizze:

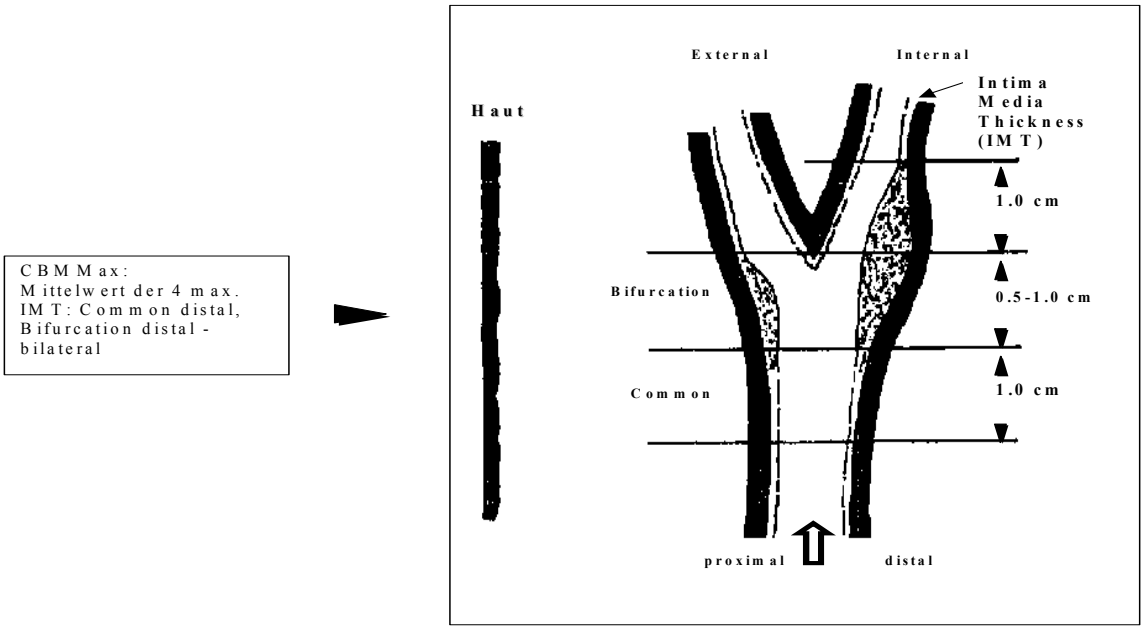


Abbildung 7.13: *Definition der Zielvariablen CBMMax*

Bei der Fallzahlplanung ist man vom Vergleich der CBMMax-Veränderungen (4-Jahres-Messung – Baseline-Messung) mittels t-Test ausgegangen. Folgende Parameter gingen in die Fallzahlplanung ein:

- α=5% (zweiseitig)
- 1-β=90%
- δ=0.04 mm
- s_0=0.236 mm
- drop-out-Rate: 35%

Daraus resultierte die geplante Fallzahl von n_p= 2250. Zur Erreichung dieser Fallzahl war ein Zeitraum von 2 Jahren vorgesehen.

Die im Folgenden vorgestellte Fallzahladjustierung wurde post-hoc, nach Beendigung der Studie durchgeführt. Im Rahmen von ELSA war keine Fallzahladjustierung vorgesehen.

In der folgenden Grafik ist die beobachtete Rekrutierungskurve dargestellt. Die zweite Kurve gibt die kumulierte Anzahl an Jahr-1-Messungen an – bis zu dem Zeitpunkt, an dem die Fallzahladjustierung durchgeführt wurde. Zu diesem Zeitpunkt – ca. 18 Monate nach Rekrutierungsbeginn - waren 1882 Patienten, also ca. 84% der geplanten Fallzahl in die Studie aufgenommen.

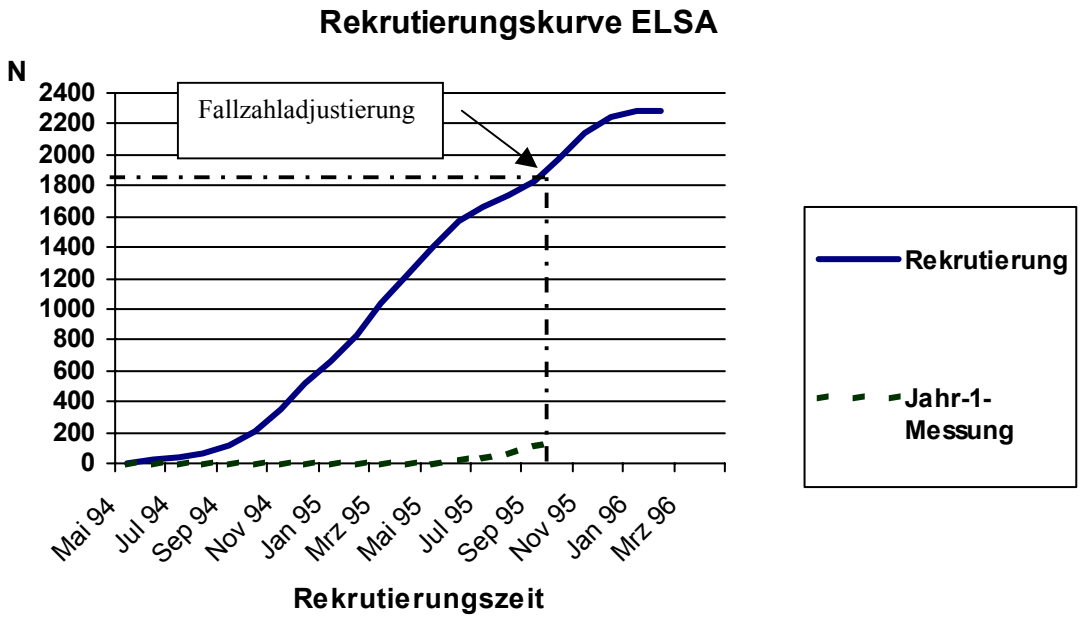


Abbildung 7.14: *Rekrutierungskurve ELSA*

Zum Zeitpunkt der Adjustierung lagen folgende Daten vor:

	Gruppe	
	Atenolol	Lacidipine
n_1	920	962
n_2	73	72
s_1	0.2456199	0.2422018
s_2	0.2122925	0.2325997
r_{12}	0.85571	0.81776

Tabelle 7.1*: relevante Daten zum Zeitpunkt der Fallzahladjustierung*

Daraus wurden folgende Größen berechnet:

	Gruppe	
	Atenolol	Lacidipine
s_t	0.246	0.242
r_{1t}	0.541	0.447
s_D	0.235	0.255

Tabelle 7.2: *Berechnete Größen für die Fallzahladjustierung*

Als Schätzung für σ_D ergab sich schließlich:

$$s_D = \sqrt{\frac{919 \bullet 0.235^2 + 961 \bullet 0.255^2}{920 + 962 - 2}} = 0.245$$

Da $s_D > s_0 = 0.236$ führen beide Adjustierungsstrategien zu einer Erhöhung der geplanten Fallzahl, und zwar auf:

$$n_a = (0.245/0.236)^2 \bullet 2250 = 2433$$

8 Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden zahlreiche Methoden zur Fallzahladjustierung bei kontrollierten klinischen Studien vorgestellt. Diese Methoden haben nicht nur aufgrund der einschlägigen Guidelines ihre Berechtigung, da sie ein sehr effektives Mittel sind, um die Fallzahlplanungsannahmen im Laufe der Studie zu kontrollieren und bei Bedarf die Fallzahl entsprechend anzupassen. Durch die Adjustierungsmethoden wird somit die Power einer Studie wirkungsvoll abgesichert.

Darüber hinaus stellen die meisten Methoden auch die Einhaltung des geforderten Signifikanzniveaus sicher – die Inflation des α's hängt dabei im Wesentlichen von der Größe der Stichprobe ab, die zur Fallzahladjustierung berücksichtigt wird. Der Zeitpunkt der Adjustierung sollte somit so gewählt werden, dass ungefähr die Hälfte der geplanten Fallzahl bzw. mindestens 20 Beobachtungen pro Gruppe vorliegen. Obere Grenzen für das tatsächlich erreichte Signifikanzniveau finden sich in der Arbeit von (Kieser und Friede 2000) für den Vergleich von zwei Mittelwerten mittels t-Test. Andere Arbeiten und auch die in dieser Arbeit vorgestellten Simulationsstudien kommen zu dem Schluss, dass im Zusammenhang mit einer zum adäquaten Zeitpunkt (s.o.) durchgeführten Fallzahladjustierung das Signifikanzniveau korrigiert werden muss. Um ein reales Signifikanzniveau von 5% zu erreichen muss mit einem adjustierten alpha von maximal 4.5% gearbeitet werden.

Eine Fallzahladjustierung, die auf die Entblindung der Daten angewiesen ist, hat immer einen Zwischenauswertungscharakter, da Unterschiede zwischen den Gruppen quantifiziert werden – auch wenn nur eine partielle Entblindung (Gruppe 1 vs. Gruppe 2) durchgeführt wird. Wenn also schon zu einem Zwischenzeitpunkt entblindet wird, so liegt es bei einer Studie nahe, eine Zwischenauswertung durchzuführen, die es eventuell erlaubt, die Studie wegen nachgewiesener Überlegenheit einer Therapie schon frühzeitig zu beenden. Aus dieser Motivation heraus sind Adjustierungsverfahren, die ohne Entblindung arbeiten, besonders interessant. In dieser Arbeit wurde gezeigt, dass sowohl für den Vergleich von Mittelwerten als auch für den Vergleich von Ereignisraten leistungsfähige Methoden zur Fallzahladjustierung bestehen.

Bei der Wahl einer entsprechenden Methode ist darüber hinaus zu berücksichtigen, ob die geplante Fallzahl nur nach oben adjustiert werden darf, oder ob auch eine Reduktion der Fallzahl in Frage kommt. Da bei zulassungsrelevanten klinischen Studien auch sekundäre Wirksamkeitsparameter und Sicherheitsaspekte eine wichtige Rolle spielen, ist man hier mit

der Strategie „Adjustierung nur nach oben“ auf der sicheren Seite. Außerdem hat die Strategie „Adjustierung nach unten oder oben“ bei einigen Simulationen zu unbefriedigenden Ergebnissen geführt, so dass die Adjustierung nach oben die Strategie der Wahl sein sollte.

Es existieren zahlreiche Adjustierungsverfahren für den Vergleich von Mittelwerten (s. Kapitel 5). Auch für den Vergleich von Ereignisraten hat man die Wahl zwischen einigen wenigen Methoden (s. Kapitel 6) – Weiterentwicklungen wurden in dieser Arbeit vorgenommen. Für Repeated-Measurements-Designs existiert das Verfahren von (Shih und Gould 1995), ein weiteres Verfahren habe ich in Kapitel 7 vorgestellt. Für komplexere Auswertungsstrategien wie der Analyse von Überlebenszeiten oder dem Nachweis von Äquivalenz gibt es erste Ansätze (s. z.B. (Tavare, Sobel et al. 1995), (Gould 1993)). Hier sehe ich den größten Bedarf für Weiterentwicklungen, insbesondere für Methoden, die keine Entblindung benötigen.

Notation und Bezeichnungen

Zufallsvariablen werden mit lateinischen Großbuchstaben und die Ausprägungen einer Zufallsvariablen mit lateinischen Kleinbuchstaben bezeichnet – eventuelle Abweichungen von diesem Prinzip dienen der besseren Lesbarkeit und erschließen sich aus dem jeweiligen Kontext.

Fallzahlen

$\mathbf{n_p}$ – vor Studienbeginn geplante Gesamtfallzahl

$\mathbf{n_{p,K}}$ – entsprechender Umfang der Kontrollgruppe

$\mathbf{n_{p,B}}$ – entsprechender Umfang der Behandlungsgruppe

$\mathbf{n_P = n_{p,K} + n_{p,B}}$

$\mathbf{n_1}$ – zum Zeitpunkt der Adjustierung vorliegende Gesamtfallzahl ($n_1 = q \bullet n_p$, $0 < q < 1$) -

$\mathbf{n_{1,K}}$ – entsprechender Umfang der Kontrollgruppe

$\mathbf{n_{1,B}}$ – entsprechender Umfang der Behandlungsgruppe

$\mathbf{n_1 = n_{1,K} + n_{1,B}}$

$\mathbf{n_a}$ – adaptierte Fallzahl

$\mathbf{n_{a,K}}$ – entsprechender Umfang der Kontrollgruppe

$\mathbf{n_{a,B}}$ – entsprechender Umfang der Behandlungsgruppe

$\mathbf{n_a = n_{a,K} + n_{a,B}}$

$\mathbf{n_2}$ – nach Adjustierung noch zu rekrutierende Patienten ($\mathbf{n_2 = n_a - n_1}$)

$\mathbf{n_{2,K}}$ – entsprechender Umfang der Kontrollgruppe

$\mathbf{n_{2,B}}$ – entsprechender Umfang der Behandlungsgruppe

$\mathbf{n_2 = n_{2,K} + n_{2,B}}$

$\mathbf{n_{max}}$ – aus praktischen / wirtschaftlichen Überlegungen maximal mögliche Fallzahl

$\mathbf{n_{min}}$ – Untergrenze für die Fallzahl

Varianzen

$\boldsymbol{\sigma^2}$ – wahre Varianz - Parameter

$\mathbf{s_0^2}$ – bei Studienplanung benutzte Schätzung der Varianz

$\mathbf{s_1^2}$ – zum Zeitpunkt der Adjustierung beobachtete gepoolte Varianz

$\mathbf{s_{1,K}^2}$ – entsprechende Varianz in der Kontrollgruppe

$\mathbf{s_{1,B}^2}$ – entsprechende Varianz in der Behandlungsgruppe

$$s_1^2 = \frac{\left(n_{1,B} - 1\right) \bullet s_{1,B}^2 + \left(n_{1,K} - 1\right) \bullet s_{1,K}^2}{n_{1,B} + n_{1,k} - 2}$$

$\mathbf{s_2^2}$ – nach der Adjustierung (in zweiter Studienhälfte) beobachtete gepoolte Varianz

$\mathbf{s_{2,K}^2}$ – entsprechende Varianz in der Kontrollgruppe

$\mathbf{s_{2,B}^2}$ – entsprechende Varianz in der Behandlungsgruppe

$$s_2^2 = \frac{\left(n_{2,B} - 1\right) \bullet s_{2,B}^2 + \left(n_{2,K} - 1\right) \bullet s_{2,K}^2}{n_{2,B} + n_{2,k} - 2}$$

$\mathbf{s_a^2}$ – beobachtete gepoolte Varianz für die gesamte Studie mit adaptierter Fallzahl n_a

$\mathbf{s_{a,K}^2}$ – entsprechende Varianz in der Kontrollgruppe

$\mathbf{s_{a,B}^2}$ – entsprechende Varianz in der Behandlungsgruppe

$$s_a^2 = \frac{\left(n_{a,B} - 1\right) \bullet s_{a,B}^2 + \left(n_{a,K} - 1\right) \bullet s_{a,K}^2}{n_{a,B} + n_{a,k} - 2}$$

Ereignisraten

$\boldsymbol{\pi_K}$ – wahre Rate Kontrollgruppe – Parameter

$\boldsymbol{\pi_B}$ – wahre Rate Behandlungsgruppe – Parameter

$\boldsymbol{p_{0,K}}$ – bei Studienplanung benutzte Schätzung der Rate der Kontrollgruppe – vorgegebene Größe

$\boldsymbol{e_1}$ – zum Zeitpunkt der Adjustierung beobachtete Gesamt-Anzahl der Ereignisse

$\boldsymbol{p_1}$ – zum Zeitpunkt der Adjustierung beobachtete Gesamtrate

$p_{1,K}$ – entsprechende Rate in der Kontrollgruppe

$p_{1,B}$ – entsprechende Rate in der Behandlungsgruppe

$$p_1 = \frac{n_{1,K} p_{1,K} + n_{1,B} p_{1,B}}{n_1}$$

$$p_1 = \frac{e_1}{n_1}$$

$\boldsymbol{p_2}$ – nach der Adjustierung (in zweiter Studienhälfte) beobachtete Gesamtrate

$p_{2,K}$ – entsprechende Rate in der Kontrollgruppe

$p_{2,B}$ – entsprechende Rate in der Behandlungsgruppe

$$p_2 = \frac{n_{2,K} p_{2,K} + n_{2,B} p_{1,B}}{n_2}$$

$\boldsymbol{p_a}$ – Gesamtrate für die gesamte Studie mit adaptierter Fallzahl n_a

$p_{a,K}$ – entsprechende Rate in der Kontrollgruppe

$p_{a,B}$ – entsprechende Rate in der Behandlungsgruppe

$$p_a = \frac{n_{a,K} p_{a,K} + n_{a,B} p_{a,B}}{n_a}$$

Sonstiges

$a \bullet b$	kennzeichnet (bei Bedarf) die Multiplikation von a mit b
$\mathbf{N(\mu,\sigma^2)}$	Normalverteilung mit Erwartungswert μ und Varianz σ^2 zugehörige Dichtefunktion: $f(x) = \frac{1}{\sqrt{2\pi}} \bullet e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$\mathbf{z_p}$	p-Quantil der Standardnormalverteilung – $\Phi^{-1}(p)$, mit $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \bullet e^{-\frac{t^2}{2}} dt$
$\mathbf{t_{p,n}}$	p-Quantil der t-Verteilung mit n Freiheitsgraden
$\mathbf{F_{t,nzp,n}(x)}$	Verteilungsfunktion der nicht-zentralen t-Verteilung mit Nichtzentralitätsparameter nzp und n Freiheitsgraden – an der Stelle x
$F_{\chi^2_n}(x)$	Verteilungsfunktion der χ^2-Verteilung mit n Freiheitsgraden – an der Stelle x
$\mathbf{B(n, p)}$	Binomialverteilung mit Parametern n und p $P(X = k) = \binom{n}{k} \bullet p^k \bullet (1-p)^{n-k} \quad k = 0,...,n$
$\mathbf{[x]}$	Gauß-Klammer von x: größte ganze Zahl <= x
$\Gamma(x)$	Gammafunktion: $\Gamma(x) = \int_0^{\infty} e^{-t}\, t^{x-1}\, dt$, für $x > 0$
$\mathbf{x_{(i)}}$	i-te Ordnungsstatistik der Beobachtungen $x_1, \ldots, x_n$
α	Wahrscheinlichkeit für den Fehler 1. Art, Signifikanzniveau
β	Wahrscheinlichkeit für den Fehler 2. Art, 1-Power
δ	klinisch relevanter Gruppenunterschied
ρ	Korrelation
$\mathbb{R}$	Menge der reellen Zahlen
KI	Konfidenzintervall
Mw	Mittelwert
Std	Standardabweichung
Med	Median
Q1	unteres Quartil
Q3	oberes Quartil
MSE	Mean Square Error

Anhang: Simulationsprogramm

Mit dem folgenden SAS-Programm wurde die im Kapitel 5.1 (Vergleich von Mittelwerten, Verfahren für entblindete Daten) vorgestellte Simulationsstudie durchgeführt. Die in den Kapiteln 5.2 (Vergleich von Mittelwerten, Verfahren für nicht entblindete Daten), 6.1 (Vergleich von Ereignisraten, Verfahren für entblindete Daten), 6.2 (Vergleich von Ereignisraten, Verfahren für nicht entblindete Daten) und 7 (Repeated Measurements) präsentierten Simulationen haben sich stark an diesem Prototypen orientiert. Auf eine Wiedergabe des Quelltextes für diese Simulationsstudien wurde daher verzichtet.

```
*************************************************************************;
*                                                                       *;
* Programm name:    simu_nv_unblind                                     *;
* Program Output:   adj. sample size, alpha, power,                     *;
*                   Länge des Konfidenzintervalls                       *;
* Description:      Simulationen fuer folgende Ansaetze zur             *;
*                   Fallzahladjustierung bei normalverteilten Daten:    *;
*                   Stein 1945, Wittes&Brittain 1990, Birkett&Day 1994, *;
*                   vs. fixed design                                    *;
* Operating System: WINDOWS NT                                          *;
* SAS-Version:      6.12 (Orlando)                                      *;
* Programmer:       Michael Hennig                                      *;
*                                                                       *;
*************************************************************************;

options pageno=1 nomprint nomlogic;

proc format ;
 value te
  0='NOT reject H0'
  1='reject H0';
run;


%macro simnv1(alpha=0.05,beta=0.1,mean1s=, mean2s=, sigmas=, mean1t=,
mean2t=, sigmat=,p=0.5, outn=, sim=, outsum=);

***** Sample size planning ***********;
data fix;
 alpha=&alpha;
 beta=&beta;
 ** Mean 1,2 used for sample size calc.**;
 mean1s=&mean1s;
 mean2s=&mean2s;
 delta=abs(mean1s-mean2s);

 ** Mean 1,2: true **;
 mean1t=&mean1t;
 mean2t=&mean2t;

 sigmas=&sigmas;** sigma assumed for sample size calculation **;
 sigmat=&sigmat;  ** "true" sigma **;

 ** fixed sample size (n per group) - using sigmas**;
 nfixi=ceil((2*sigmas**2*(probit(alpha/2)+probit(beta))**2)/(delta**2));

 ** fixed sample size (n per group) - using sigmat**;
 ntrue=ceil((2*sigmat**2*(probit(alpha/2)+probit(beta))**2)/(delta**2));
```

```
 ** time point of sample size adjustment**;
 p=&p;
 n1=ceil(nfixi*&p);
 %let
nfixi=ceil((2*&sigmas**2*(probit(&alpha/2)+probit(&beta))**2)/((abs(&mean1s
-&mean2s))**2));
 * for p=0.5, 0.75*;
 *%let
maxngr=ceil(2.5*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((a
bs(&mean1s-&mean2s))**2)));

*maxngr=ceil(2.5*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((
abs(&mean1s-&mean2s))**2)));
 * for p=0.2*;
 %let
maxngr=ceil(3.5*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((a
bs(&mean1s-&mean2s))**2)));

maxngr=ceil(3.5*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((a
bs(&mean1s-&mean2s))**2)));
 *for p=0.1*;
 *%let
maxngr=ceil(4*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((abs
(&mean1s-&mean2s))**2)));

*maxngr=ceil(4*ceil((2*&sigmat**2*(probit(&alpha/2)+probit(&beta))**2)/((ab
s(&mean1s-&mean2s))**2)));

label
       alpha='Alpha used for sample size calc.'
       beta='Beta used for sample size calc.'
       delta='Delta'
       sigmas='Sigma assumed for sample size calc.'
       sigmat='True sigma'
       mean1s='Mean 1 - used for sample size calc.'
       mean2s='Mean 2 - used for sample size calc.'
       mean1t='True Mean 1'
       mean2t='True Mean 2'
       nfixi='Sample size (per group) - sigmas'
       ntrue='Sample size (per group) - true sigma'
       n1='Time point of adjustment'
       maxngr='Maximum n per group (for simu)';

run;

proc print label data=fix noobs;
 title 'Sample size adjustment acc. to Wittes&Brittain and Stein -
Simulations';
 title2 "alpha=&alpha, beta=&beta, mean1s=&mean1s, mean2s=&mean2s,
sigmas=&sigmas";
 title3 "mean1t=&mean1t, mean2t=&mean2t, sigmat=&sigmat, p=&p, outn=&outn,
outsum=&outsum, sim=&sim";
 title4 'Initial setting';
var alpha beta mean1s mean2s sigmas nfixi mean1t mean2t sigmat ntrue p n1
maxngr;
 run;
```

```
*****************************;
*** Simulation         *******;
*****************************;
data simu; *(compress=YES) - bei Bedarf;
 set fix;

do simu=1 to &sim;  ** Anzahl der Simulationen**;
 do i=1 to &maxngr;  ** max. ad. Stichprobengroesse per group = 2.5 * ntrue
(s.o.)**;
  x=mean1t+sigmat*normal(1202);
  group=1;
  ****SAMPLE SIZE ADJUSTMENT*****;
  if (i le n1) then interim=1;
    else interim=0;
  output;
 end;
 do i=&maxngr+1 to (2*&maxngr);
  x=mean2t+sigmat*normal(1202);
  group=2;
  ****SAMPLE SIZE ADJUSTMENT*****;
  if (((i le &maxngr+n1)) ) then interim=1;
   else interim=0;
  output;
 end;
end;
length simu 4 i group interim 3;  ** um Datei "klein" zu halten **;
drop alpha beta mean1s mean2s delta mean1t mean2t sigmas sigmat nfixi ntrue
maxngr p n1;


proc univariate data=simu noprint;
 where interim=1;
 var x;
 by simu group;
 output out=nawb std=sg n=ng;
run;

data gr1;
 set nawb;
 if group=1;
 rename ng=na sg=sa;
 drop group;

data gr2;
 set nawb;
 if group=2;
 rename ng=nb sg=sb;
 drop group;

***Tricks um die beiden Dateien fix (1 Observation) und wb1 (#Sim.
observations) ***;
*** mergen zu koennen ***;
data wb1;
 merge gr1 gr2 ;
 k=1;

data fix;
 set fix;
  k=1;
```

```
data out1;
 merge fix wb1 ;
  by k;
  spool1=sqrt( ( ((na-1)*sa**2) +((nb-1)*sb**2) )/(na+nb-2) );
 if spool1 le sigmas then nawb=nfixi;
    else nawb=ceil((nfixi*(spool1**2))/(sigmas**2));
  z=(sigmas**2)/(nfixi-1);
  nst=max(n1, ceil(spool1**2/z)+1);
  maxn=max(nawb,nst);
  keep simu spool1 nawb nfixi sigmas na nb sa sb z nst nawb;



/*
proc print data=out1 ;
 title4 'Adjustment according to Wittes&Brittain and Stein';
 run;
*/

proc univariate data=out1 ;
 title4 'Verteilung spool1 nawb nst';
 var spool1 nawb nst;
 output out=sum2  min=mins minnwb minnst median=meds mednwb mednst max=maxs
maxnwb maxnst
        mean=means meannwb meannst std=stds stdnwb stdnst;
run;



**** WB / ST ************************************************;
**** Ziehung der Gesamtstichprobe - nawb/nst Beobachtungen ***;
************************************************************;
data simu;* (compress=YES) - bei Bedarf;
 merge simu(keep=i simu x group) out1(keep=simu nawb nst);
  by simu;
 if (i le nawb) or (i gt &maxngr and i le &maxngr+nawb) then w=1; ** Stipro
fuer WB ***;
 if (i le nst) or (i gt &maxngr and i le &maxngr+nst) then s=1;   ** Stipro
fuer ST/BD **;
 length simu 4 i group nawb nst 3;  ** um Datei "klein" zu halten **;
run;


************************;
** Testdurchfuehrung ****;
************************;

********** fixed design ****;
proc means noprint data=simu;
 where (i le &nfixi) or (i gt &maxngr and i le &maxngr+&nfixi);
 var x;
  by simu group;
 output out=fd mean=mean std=std n=n;
run;

data mean1;
 set fd;
  if group=1;
 rename mean=mean1fd std=std1fd n=n1fd;
 keep mean std n simu;
run;
```

```
data mean2;
 set fd;
  if group=2;
 rename mean=mean2fd std=std2fd n=n2fd;
 keep mean std n simu;
run;

data testfd;
 merge mean1 mean2;
  by simu;
 difffd=mean1fd-mean2fd;
 spool2fd=sqrt( ( ((n1fd-1)*std1fd**2) +((n2fd-1)*std2fd**2) )/(n1fd+n2fd-
2) );
 tfd=difffd/(spool2fd*sqrt(2/n1fd));
 abstfd=abs(tfd);
 ffd=tfd**2;
 tq975fd=tinv(0.975,n1fd+n2fd-2);
 cilfd=difffd-(spool2fd*tinv(0.975,n1fd+n2fd-2)*sqrt(2/n1fd));
 ciufd=difffd+(spool2fd*tinv(0.975,n1fd+n2fd-2)*sqrt(2/n1fd));
 lcifd=ciufd-cilfd;
 pfd=2*(1-probt(abstfd,n1fd+n2fd-2));
run;

/*
proc print;
 title3 'FD - per Hand';
 run;
*/


********** WB design ****;

*** "per Hand" *************;

proc means noprint data=simu;
 where w=1;
 var x;
  by simu group;
 output out=wb mean=mean std=std n=n;
run;

data mean1;
 set wb;
  if group=1;
 rename mean=mean1wb std=std1wb n=n1wb;
 keep mean std n simu;
run;

data mean2;
 set wb;
  if group=2;
 rename mean=mean2wb std=std2wb n=n2wb;
 keep mean std n simu;
run;
```

```
data testwb;
 merge mean1 mean2;
  by simu;
 diffwb=mean1wb-mean2wb;
 spool2wb=sqrt( ( ((n1wb-1)*std1wb**2) +((n2wb-1)*std2wb**2) )/(n1wb+n2wb-
2) );
 twb=diffwb/(spool2wb*sqrt(2/n1wb));
 abstwb=abs(twb);
 fwb=twb**2;
 tq975wb=tinv(0.975,n1wb+n2wb-2);
 cilwb=diffwb-(spool2wb*tinv(0.975,n1wb+n2wb-2)*sqrt(2/n1wb));
 ciuwb=diffwb+(spool2wb*tinv(0.975,n1wb+n2wb-2)*sqrt(2/n1wb));
 lciwb=ciuwb-cilwb;
 pwb=2*(1-probt(abstwb,n1wb+n2wb-2));
run;

/*
proc print;
title3 'WB - per Hand';
 run;
*/


********** Stein design ****;

** analog zu  Proc TTest ***;
*** 1. Mittelwert ueber alle Beobachtungen ***;
proc means noprint data=simu;
 where s=1;
 var x;
   by simu group;
 output out=meanst mean=mean;
run;

data mean1;
 set meanst;
  if group=1;
 rename mean=mean1st;
 keep mean simu;
run;

data mean2;
 set meanst;
  if group=2;
 rename mean=mean2st;
 keep mean simu;
run;

data mean12;
 merge mean1 mean2;
  by simu;
 diffst=mean1st-mean2st;
run;

*** 2. Std aus ersten Beob. ****;
data std;
 set out1(keep=spool1 nst simu na nb);
run;
```

```
*** 3. Teststatistik aus 1. und 2. ****;
data testst;
 merge mean12 std;
  by simu;
 rename na=n1st nb=n2st;
 tst=diffst/(spool1*sqrt(2/nst));
 abst=abs(tst);
 tq975st=tinv(0.975,na+nb-2);
 cilst=diffst-(spool1*tinv(0.975,na+nb-2)*sqrt(2/nst));
 ciust=diffst+(spool1*tinv(0.975,na+nb-2)*sqrt(2/nst));
 lcist=ciust-cilst;
 pst=2*(1-probt(abst,na+nb-2));
run;

/*
proc print;
 title3 'ST - per Hand';
 run;
*/


******* Birkett/Day-design (adjusting down/up - wie bei Stein)***;
***      Test mit allen Beob. fuer Std - anders als bei Stein ***;

*** "per Hand" *************;

proc means noprint data=simu;
 where s=1;
 var x;
  by simu group;
 output out=bd mean=mean std=std n=n;
run;

data m1bd;
 set bd;
  if group=1;
 rename mean=mean1bd std=std1bd n=n1bd;
 keep mean std n simu;
run;

data m2bd;
 set bd;
  if group=2;
 rename mean=mean2bd std=std2bd n=n2bd;
 keep mean std n simu;
run;

data testbd;
 merge m1bd m2bd;
  by simu;
 diffbd=mean1bd-mean2bd;
 spool2bd=sqrt( ( ((n1bd-1)*std1bd**2) +((n2bd-1)*std2bd**2) )/(n1bd+n2bd-
2) );
 tbd=diffbd/(spool2bd*sqrt(2/n1bd));
 abstbd=abs(tbd);
 fbd=tbd**2;
 tq975bd=tinv(0.975,n1bd+n2bd-2);
 cilbd=diffbd-(spool2bd*tinv(0.975,n1bd+n2bd-2)*sqrt(2/n1bd));
 ciubd=diffbd+(spool2bd*tinv(0.975,n1bd+n2bd-2)*sqrt(2/n1bd));
 lcibd=ciubd-cilbd;
 pbd=2*(1-probt(abstbd,n1bd+n2bd-2));
run;
```

```
/*
proc print;
title3 'BD - per Hand';
 run;
*/

*** alle Testergebnisse zusammen **;


data testall;
 merge testfd testwb testst testbd;
  by simu;
 if pwb ge 0.05 then tewb=0;
   else tewb=1;
 if pfd ge 0.05 then tefd=0;
   else tefd=1;
 if pst ge 0.05 then test=0;
   else test=1;
 if pbd ge 0.05 then tebd=0;
   else tebd=1;

 label pwb='p-Value WB'
       pfd='p-Value Fixed Design'
       pst='p-Value Stein'
       pbd='p-Value BD'
       tst='Test-statistic Stein'
       twb='Test-statistic WB'
       tbd='Test-statistic BD'
       tfd='Test-statistic fixed design'
       tewb='Test result WB'
       tebd='Test result BD'
       tefd='Test result Fixed Design'
       test='Test result Stein'
       cilst='Lower Limit 95% CI - Stein'
       ciust='Upper Limit 95% CI - Stein'
       cilfd='Lower Limit 95% CI - FD'
       ciufd='Upper Limit 95% CI - FD'
       cilwb='Lower Limit 95% CI - WB'
       ciuwb='Upper Limit 95% CI - WB'
       cilbd='Lower Limit 95% CI - BD'
       ciubd='Upper Limit 95% CI - BD'
       lcist='Length 95% CI - Stein'
       lcifd='Length 95% CI - FD'
       lciwb='Length 95% CI - WB'
       lcibd='Length 95% CI - BD'
;
 format tewb tefd test tebd te.;

data outall;
 merge out1 testall;
  by simu;
```

```
/*
proc print label;
 var simu spool1 nfixi sigmas na nb sa sb z
          n1fd n2fd mean1fd mean2fd difffd cilfd ciufd std1fd std2fd
spool2fd pfd tfd tefd
          nawb n1wb n2wb mean1wb mean2wb diffwb cilwb ciuwb std1wb std2wb
spool2wb pwb twb tewb
          nst n1st n2st nst mean1st mean2st diffst cilst ciust pst tst
test;
 title4 'All results - data outall';
*/


proc univariate data=outall noprint;
 var cilfd ciufd lcifd cilwb ciuwb lciwb cilst ciust lcist cilbd ciubd
lcibd;
 output out=sum1 mean=mcilfd mciufd mlcifd mcilwb mciuwb mlciwb mcilst
mciust mlcist mcilbd mciubd mlcibd;
 title4 'Summary Confidence Intervals for estimating the difference between
groups';
run;

proc univariate data=outall noprint;
 var nawb nst;
 output out=&outn mean=mwnwb mwnst median=mednwb mednst min=minnwb minnst
max=maxnwb maxnst
        q1=q1nwb q1nst q3=q3nwb q3nst p5=p5nwb p5nst p95=p95nwb p95nst  ;
 title4 'Summary Distribution NWB / NST';
run;

proc freq data=testall;
 tables tewb*tefd;
 tables test*tefd;
 tables tebd*tefd;
 tables test*tewb;
 tables test*tebd;
 title4 'Test results';
run;

**** CI fuer alpha und beta ************;
proc freq data=testall noprint;
 tables test / out=fost ;
 tables tewb / out=fowb;
 tables tefd / out=fofd;
 tables tebd / out=fobd;
run;

data ci1;
 set fost;
 a=lag(count);
 mst=count;
 n=mst+a;
 if test=1;
 keep mst n test;

data ci2;
 set fowb;
 mwb=count;
 if tewb=1;
 keep mwb tewb;
```

```
data ci3;
 set fofd;
 mfd=count;
 if tefd=1;
 keep mfd tefd;

data ci4;
 set fobd;
 mbd=count;
 if tebd=1;
 keep mbd tebd;

data ci;
 merge ci1 ci2 ci3 ci4;
 nenner=2*(n+(probit(0.975)**2));
 pst=mst/n;
 z1st=(2*mst)+(probit(0.975)**2);
 z2st=probit(0.975)*sqrt( (probit(0.975)**2)+(4*mst*(1-(mst/n))));
 plowst=(z1st-z2st)/nenner;
 pupst=(z1st+z2st)/nenner;

 pwb=mwb/n;
 z1wb=(2*mwb)+(probit(0.975)**2);
 z2wb=probit(0.975)*sqrt( (probit(0.975)**2)+(4*mwb*(1-(mwb/n))));
 plowwb=(z1wb-z2wb)/nenner;
 pupwb=(z1wb+z2wb)/nenner;

 pfd=mfd/n;
 z1fd=(2*mfd)+(probit(0.975)**2);
 z2fd=probit(0.975)*sqrt( (probit(0.975)**2)+(4*mfd*(1-(mfd/n))));
 plowfd=(z1fd-z2fd)/nenner;
 pupfd=(z1fd+z2fd)/nenner;

 pbd=mbd/n;
 z1bd=(2*mbd)+(probit(0.975)**2);
 z2bd=probit(0.975)*sqrt( (probit(0.975)**2)+(4*mbd*(1-(mbd/n))));
 plowbd=(z1bd-z2bd)/nenner;
 pupbd=(z1bd+z2bd)/nenner;


/*
proc print;
 title4 'Anteile abgelehnter Nullhypothesen incl. 95%-CI
(Normalapproximation)';
 title5 '(FD = Fixed Design, ST = Stein, WB=Wittes&Brittain)';
 var tefd test tewb pfd plowfd pupfd pst plowst pupst pwb plowwb pupwb;
run;
*/
```

```
data &outsum;
 merge sum1 sum2 ci fix;
 bst=meannst-ntrue;
 bwb=meannwb-ntrue;
 bbd=bst;
 msest=(bst**2)+(stdnst**2);
 msewb=(bwb**2)+(stdnwb**2);
 msebd=msest;
 label bst='Bias - Stein'
       bwb='Bias - WB'
       bbd='Bias - BD'
       msest='MSE - Stein'
       msebd='MSE - BD'
       msewb='MSE - WB';
run;

proc print label data=&outsum;
 title4 'Summary';
 var pfd plowfd pupfd
     pst plowst pupst
     pwb plowwb pupwb
     pbd plowbd pupbd
     minnst mednst maxnst meannst stdnst bst msest
     minnwb mednwb maxnwb meannwb stdnwb bwb msewb
     mcilfd mciufd mlcifd
     mcilst mciust mlcist
     mcilwb mciuwb mlciwb
     mcilbd mciubd mlcibd;
run;

%mend simnv1;


*** p=0.5 ***;

 ** H0 true ***;
  %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2), mean1t=0,
mean2t=0,sigmat=1, p=0.5, outn=outn01, sim=10000, outsum=outsum01 );

  * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=0,sigmat=sqrt(2), p=0.5, outn=outn02, sim=100000,
outsum=outsum02 );

  %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2), mean1t=0,
mean2t=0,sigmat=2, p=0.5, outn=outn03, sim=100000, outsum=outsum03 );

 ** H1 true ***;
  * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=1, p=0.5, outn=outn04, sim=100000,
outsum=outsum04 );

   * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=sqrt(2), p=0.5, outn=outn05, sim=100000,
outsum=outsum05 );

   %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=2, p=0.5, outn=outn06, sim=100000,
outsum=outsum06 );
```

```
*** p=0.2 ***;

 ** H0 true ***;
  %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2), mean1t=0,
mean2t=0,sigmat=1, p=0.2, outn=outn11, sim=10000, outsum=outsum11 );

  * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=0,sigmat=sqrt(2), p=0.2, outn=outn12, sim=100000,
outsum=outsum12 );

  %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2), mean1t=0,
mean2t=0,sigmat=2, p=0.2, outn=outn13, sim=100000, outsum=outsum13 );

 ** H1 true ***;
  * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=1, p=0.2, outn=outn14, sim=100000,
outsum=outsum14 );

   * %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=sqrt(2), p=0.2, outn=outn15, sim=100000,
outsum=outsum15 );

   %simnv1(alpha=0.05,beta=0.1,mean1s=0, mean2s=1, sigmas=sqrt(2),
mean1t=0, mean2t=1,sigmat=2, p=0.2, outn=outn16, sim=100000,
outsum=outsum16 );


************************************;
* Zusammenfuegen aller Simulationen **;
************************************;
data sim01sum;
 set outsum01(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=1;
run;

data sim02sum;
 set outsum02(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=2;
run;

data sim03sum;
 set st.outsum03(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st
z2st z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t
sigmas nfixi k);
 sim=3;
run;

data sim04sum;
 set outsum04(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=4;
run;

data sim05sum;
 set outsum05(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=5;
```

```
run;

data sim06sum;
 set outsum06(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=6;
run;

data sim11sum;
 set outsum11(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=11;
run;

data sim12sum;
set outsum12(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=12;
run;

data sim13sum;
 set outsum13(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=13;
run;

data sim14sum;
 set outsum14(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=14;
run;

data sim15sum;
 set outsum15(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=15;
run;

data sim16sum;
 set outsum16(drop=test mst n tewb mwb tefd mfd tebd mbd nenner z1st z2st
z1wb z2wb z1fd z2fd z1bd z2bd alpha beta mean1s mean2s delta mean1t sigmas
nfixi k);
 sim=16;
run;

data sum;
 set sim01sum sim02sum sim03sum sim04sum sim05sum sim06sum sim11sum
sim12sum sim13sum sim14sum sim15sum sim16sum;
run;
```

```
data sim01n;
 set outn01(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=1;
run;

data sim02n;
 set outn02(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=2;
run;

data sim03n;
 set outn03(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=3;
run;

data sim04n;
 set outn04(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=4;
run;

data sim05n;
 set outn05(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=5;
run;

data sim06n;
 set outn06(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=6;
run;

data sim11n;
 set outn11(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=11;
run;

data sim12n;
 set outn12keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=12
run;


data sim13n;
 set outn13(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=13;
run;

data sim14n;
 set outn14(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=14;
run;

data sim15n;
 set outn15(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=15;
run;

data sim16n;
 set outn16(keep=q3nwb q3nst q1nwb q1nst p95nwb p95nst p5nwb p5nst);
 sim=16;
run;
```

```
data n;
 set sim01n sim02n sim03n sim04n sim05n sim06n sim11n sim12n sim13n sim14n
sim15n sim16n;
run;

data allsimus;
 merge sum n;
  by sim;
run;

proc print;
 title2 'Alle Simulationen zusammen';
 run;
```

Literaturverzeichnis

Bauer P. und Köhne K. (1994): "Evaluation of experiments with adaptive interim analyses." *Biometrics* **50**: 1029-1041.

Betensky R. A. und Tierney C. (1997): "An examination of methods for sample size recalculation during an experiment." *Statistics in Medicine* **16**: 2587-2598.

Birkett M. A. und Day S. J. (1994): "Internal pilot studies for estimating sample size." *Statistics in Medicine* **13**: 2455-2463.

Bock J. (1998): *Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte Studien*. München, Wien, R. Oldenbourg Verlag.

Bolland K., Sooriyarachchi M., Whitehead J. (1998): "Sample size review in a head injury trial with ordered categorical responses." *Statistics in Medicine* **17**: 2835-2847.

Bristol D. R. (1989). "Sample sizes for constructing confidence intervals and testing hypotheses." *Statistics in Medicine* **8**: 803-811.

Browne R. H. (1995): "On the use of a pilot sample for sample size determination." *Statistics in Medicine* **14** : 1933-1940.

Casagrande J. T., Pike M. C., Smith P. G. (1978): "An improved approximative formula for calculating sample sizes for comparing two binomial distributions." *Biometrics* **34** : 483-486.

Case L. D., Morgan T. M., Davis C. E. (1987): "Optimal restricted two-stage designs." *Controlled Clinical Trials* **8** : 146-156.

Coffey C. S. und Muller K. E. (1999): "Exact test size and power of a gaussian error linear model for an internal pilot study." *Statistics in Medicine* **18** : 1199-1214.

Coffey C. S. und Muller K. E. (2001): “Controlling test size while gaining the benefits of an internal pilot design.” *Biometrics* **57**: 625-631.

CPMP, Working Party on Efficacy of Medicinal Products (1990): “Good clinical practice for trials in the European community.” *Pharmacology and Toxicology* **67**: 361-372.

Crowder M. J. und Hand D. J. (1990): *Analysis of Repeated Measurements*. London, Chapman & Hall.

Denenberg V. H. (1987): *Statistical Power Analysis in Research,* SAGE Publications.

Denne J. S. und Jennison C. (1999): “Estimating the sample size for a t-test using an internal pilot.” *Statistics in Medicine* **18** : 1575-1585.

Desu M. M. und Raghovarao D. (1990): Sample Size Methodology. *Statistical Modelling and decision science*. G. J. Lieberman und I. Olkin. Boston, San Diego, New York, u. a., Academic Press, Inc. : 94-108.

Donner A. (1984): “Approaches to sample size estimation in the design of clinical trials - a review.” *Statistics in Medicine* **3** : 199-214.

Fisher L. D. (1998): “Self-designing clinical trials.” *Statistics in Medicine* **17** : 1551-1562.

Friede T. und Kieser M. (2002): “On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation.” *Statistics in Medicine* **21**: 165-176.

Gardner M. J. und Altman D. G. (1989): *Statistics with confidence - Confidence Intervals and statistical guidelines.* London, British Medical Journal.

Gould A. L. (1992): “Interim analysis for monitoring clinical trials that do not materially affect the type I error rate.” *Statistics in Medicine* **11** : 55-66.

Gould A. L. (1993): “Sample sizes for event rate equivalence trials using prior information.” *Statistics in Medicine* **12** : 2009-2023.

Gould A. L. (1995): “Planning and revising the sample size for a trial.” *Statistics in Medicine* **14** : 1039-1051.

Gould A. L. (2001): “Sample size re-estimation: recent developments and practical considerations.” *Statistics in Medicine* **20**: 2625-2643.

Gould A. L. und Shih W. J. (1992): “Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance.” *Communication in Statistics* **21**(10): 2833-2853.

Gould A. L. und Shih W. J. (1998): “Modifying the design of ongoing clinical trials without unblinding.” *Statistics in Medicine* **17**: 89-100.

Guggerli U. S., Maurer W., Mellein B.. (1993): “Internally adaptive designs for parallel group trials.” *Drug Information Journal* **27** : 721-732.

Hartung J. (1985): *Statistik - Lehr- und Handbuch der angewandten Statistik*. München, R. Oldenbourg Verlag.

Herson J. H. und Wittes J. (1993): “The use of interim analysis for sample size adjustment.” *Drug Information Journal* **27** : 753-760.

ICH-E6 (1996): ICH E6: Guideline for Good Clinical Practice, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use**:** 53.

ICH-E9 (1998): ICH E9: Statistical Principles for Clinical Trials. London, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use**:** 36.

Jennison C. und Turnbull B. W. (2000): *Group sequential methods with applications to clinical trials*. Boca Raton, Chapman & Hall / CRC.

Kieser M. und Friede T. (2000): "Blinded sample size reestimation in multiarmed clinical trials." *Drug Information Journal* **34**: 455-460.

Kieser M. und Friede T. (2000): "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. und Wassmer G. (1996): "On the use of the upper confidence limit from a pilot sample for sample size determination." *Biometrical Journal* **38 (8)**: 941-949.

Kolman J., Meng P., Scott G. (1998): *Good Clinical Practice Standard Operating Procedures for clinical researchers*. Chichester, New York, Weinheim u.a., John Wiley & Sons.

Kraemer H. C. und Thiemann S. (1987): *How many subjects? Statistical power analysis in research*. Newbuy Park, C. A., SAGE Publications.

Lachin J. M. (1981): "Introduction to sample size determination and power analysis for clinical trials." *Controlled Clinical Trials* **2** : 93-113.

Laird N. M. und Ware J. H. (1982): "Random-effects models for longitudinal data." *Biometrics* **38**: 963-974.

Machin D., Campbell M., Fayers P. (1997): *Sample size tables for clinical studies - second edition*. Oxford, Blackwell Science.

Phillips A. (1998): "Sample size estimation when comparing more than two treatment groups." *Drug Information Journal* **32**: 193-199.

Proschan M. A. und Hunsberger S. A. (1995): "Designed extension of studies based on conditional power." *Biometrics* **51** : 1315-1324.

Rochon J. (1991): “Sample size calculations for two-group repeated-measures experiments.” *Biometrics* **47** : 1383-1398.

Sandvik L., Erikssen J., Mowinckel, P., et al. (1996): “A method for determining the size of internal pilot studies.” *Statistics in Medicine* **15** : 1587-1590.

Saville D. J. (1990): “Multiple comparison procedures: The practical solution.” *The American Statistician* **44** : 174-180.

Schouten H. J. A. (1999a): “Sample size formula with a continuous outcome for unequal group sizes and unequal variances.” *Statistics in Medicine* **18** : 87-91.

Schouten H. J. A. (1999b): “Planning group sizes in clinical trials with a continuous outcome and repeated measures.” *Statistics in Medicine* **18** : 255-264.

Shih W. J. (1993): “Sample Size reestimation for triple blind clinical trials.” *Drug Information Journal* **27** : 761-764.

Shih W. J. und Gould A. L. (1995): “Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change.” *Statistics in Medicine* **14** : 2239-2248.

Shih W. J. und Zhao P. L. (1997): “Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes.” *Statistics in Medicine* **16** : 1913-1923.

Singer J. (1999): “Letter to the Editor: A method for determining the size of internal pilot studies.” *Statistics in Medicine* **18**: 1151-1153.

Stein C. (1945): “A two-sample test for a linear hypothesis whose power is independent of the variance.” *Annals of Mathematical Statistics* **16** : 243-258.

Tavare C. J., Sobel E. L., Gilles F. H. (1995): “Misclassification of a prognostic dichotomous variable: sample size and parameter estimate adjustment.” *Statistics in Medicine* **14** : 1307-1314.

Wittes J. und Brittain E. (1990): “The role of internal pilot studies in increasing the efficiency of clinical trials.” *Statistics in Medicine* **9** : 65-72.

Wittes J., Schabenberger O., Zucker D., et al. (1999): “Internal pilot studies I: Type I error rate of the naive t-Test.” *Statistics in Medicine* **18**: 3481-3491.

Zanchetti A., Bond M. G., Hennig M., et al. (1998): “Risk factors associated with alterations in carotid intima-media thickness in hypertension: baseline data from the European Lacidipine Study on Atherosclerosis.” *Journal of Hypertension* **16 (7)**: 949-961.

Zellner D., Zellner G. E., Frankewitsch T., et al. (2001): “Internal pilot studies for determining the sample size without unblinding.” *Drug Information Journal* **35**: 399-405.

Zucker D. M., Wittes J., Schabenberger O., et al. (1999): “Internal pilot studies II: comparison of various procedures.” *Statistics in Medicine* **18**: 3493-3509.