〔総　　説〕

中間解析，サンプルサイズ再計算を伴う生物学的同等性試験における統計的課題

Statistical Issues Concerning Bioequivalence Studies Involving Interim Analysis and Sample Size Re-calculation

棚橋 昌也*，菅波 秀規

MASAYA TANAHASHI*，HIDEKI SUGANAMI

興和株式会社

Summary：The “Guideline for Bioequivalence Studies of Generic Products” has been partially revised and stricter control of the type I error rate is required. According to the revised guideline, if the type I error rate is maintained at 5% overall, interim analysis and sample size re-calculation can be performed in a bioequivalence study. Kieser and Rauch have shown that conventional approaches such as group sequential and adaptive designs can be applied to this type of bioequivalence study. In addition, Maurer et al. modified the Combination Test addressed by Kieser and Rauch and proposed a new algorithm for sample size re-calculation.
This paper explains statistical issues in bioequivalence studies in which interim analysis and sample size re-calculation are performed and summarizes the approach proposed by Maurer et al. as one of the applicable methods.

Key words：bioequivalence study, group sequential design, adaptive design, sample size re-calculation

要旨：「後発医薬品の生物学的同等性試験ガイドライン」が一部改正され，第一種の過誤確率を厳格に制御することが求められるようになった．改正後のガイドラインでは，試験全体の第一種の過誤確率が5%に維持されていれば，生物学的同等性試験において中間解析やサンプルサイズの再計算を行うことができるとされている．Kieser and Rauch は，このような生物学的同等性試験に，従来の群逐次デザインやアダプティブデザインの手法が適用できることを示した．さらに，Maurer et al. は，Kieser and Rauch が取り上げた Combination Test を修正し，サンプルサ

* 〒 103-8433 東京都中央区日本橋本町 3-4-14
TEL：03-3279-7463　FAX：03-3279-7869
E-mail：m-tanahs@kowa.co.jp

［筆者略歴］

棚橋 昌也

・学歴，学位
平成 18 年　3 月　南山大学数理情報学部数理科学科卒業
平成 20 年　3 月　南山大学大学院数理情報研究科
　　数理情報専攻博士前期課程修了

・職歴
平成 20 年　4 月　興和株式会社臨床解析部

・学会活動等
日本製薬工業協会データサイエンス部会推進委員

菅波秀規

・学歴，学位
平成　5 年　3 月　九州工業大学情報工学部卒業
平成　7 年　3 月　九州工業大学大学院情報科学専攻修了
平成 19 年　3 月　東京理科大学大学院経営工学専攻修了
平成 19 年　3 月　工学博士（東京理科大学）

・職歴
平成　7 年　4 月　興和株式会社開発管理部統計解析課
平成 18 年 10 月　興和株式会社臨床解析部統計解析課
　　課長
平成 23 年　4 月　興和株式会社臨床解析部部長
令和　2 年　6 月　興和株式会社執行役員臨床解析部部長

・学会活動等
学会：日本計量生物学会（評議員），日本応用統計学会，日本臨床試験学会，日本薬剤疫学会，日本緑内障学会，International Biometric Society，Society for Clinical Data Management（journal editor）
その他：日本製薬工業協会データサイエンス部会副部会長，ICH E9（R1）expert，ICH E20 topic leader，SAS ユーザー会世話人，東北大学非常勤講師・技術評価委員，東京大学登録研究員，東京理科大学非常勤講師，認定責任試験統計家，SCDM 日本支部 publication committee leader

イズ再計算の新しいアルゴリズムを提案した.
本論文では，中間解析とサンプルサイズ再計算を伴う生物学的同等性試験で対処すべき統計的課題を説明し，適用可能な手法の一つとして，Maurer et al. が提案した手法を概説する.

キーワード：生物学的同等性試験，群逐次デザイン，アダプティブデザイン，サンプルサイズ再計算

1．はじめに

「後発医薬品の生物学的同等性試験ガイドライン」（以降，BE ガイドラインとする）が令和 2 年 3 月に一部改正され，生物学的同等性（BE）試験の評価方法に対する考え方にも変更が加えられた．改正前の BE ガイドラインでは，「本試験」において有意水準 5%の二つの片側検定（もしくは，両側 90%信頼区間）に基づく BE 評価を行い，サンプルサイズ[脚注 1)] の不足により BE が示せない場合には，本試験のサンプルサイズの半分以上の「例数追加試験」を 1 回に限り行うことが許容されていた（ただし，例数追加試験を行う場合は，その旨を本試験開始前にプロトコルに定めておく必要があった）．また，例数追加試験を行った場合には，本試験のデータと併合し，試験を変動要因の一つとして BE 評価を行うことができた[1)]．例数追加試験までを一つの試験として考えたときに，一つの試験の中で検定を複数回繰り返すことによる多重性の問題が生じるが，改正前の BE ガイドラインの Q & A では，「本来生物学的に非同等な製剤の場合には，バイオアベイラビリティの平均値の比が第一段階で生物学的に同等の領域に入り，第二段階の例数追加試験に踏み切る確率は大きく見積もっても 50%であり，例数追加試験による α への寄与はたかだか 2.5%である．バラツキの大きい薬物の第一段階での α は 5%よりも小さいと考えられるので，例数追加試験による α の増大はそれほど危惧しなくてもよい.」と述べられている[2)]．一方，改正後の BE ガイドライン（及びその Q & A）では，本試験の検証試験としての位置づけが明確にされ，厳格な第一種の過誤の確率の制御が必要とされている．本試験とは別に例数追加試験を実施すること，及び，そのデータを本試験と合わせて解析することは，原則として認められていない．しかしながら，事前に計画した中間解析の結果に基づき必要サンプルサイズを追加することは許容されている（ただし，1 回に限る）[3,4)]．中間解析を伴う BE 試験を計画する場合，それぞれの BE 評価に対する有意水準を調整する必要があり，中間解析の結果に基づくサンプルサイズの追加を行う場合には，更なる有意水準の調整，もしくは検定統計量の構成を工夫する必要がある．

本論文では，2 節で，改正後ガイドラインで述べられている本試験において，中間解析及びその結果に基づくサンプルサイズの再計算を計画する場合に考慮すべき統計的諸問題を説明する．3 節では，その統計的諸問題への一つの解決として，Maurer et al.[5)] で述べられている Combination Test について紹介する．4 節では，サンプルサイズ再計算の方法について述べる．5 節で数値例を示し，6 節で簡単なまとめを与える．

2．改正後の BE ガイドラインで考慮すべき統計的諸問題

BE 試験での第一種の過誤は，試験薬が対照薬に対して真には BE でないにもかかわらず，BE と判定してしまうことである．改正後の BE ガイドラインでは，BE 試験におけるこの第一種の過誤の確率を厳格に 5%以下に制御することを要求している．

BE 試験に限らず，臨床試験では，通常，事前に想定されるエンドポイントの群間差とばらつきに基づいて，目標とする検出力を満たすサンプルサイズを計算する．想定が正しければ，試験の検出力は目標とする検出力に等しくなる．しかしながら，例えば，エンドポイントのばらつきが想定よりも大きい場合は，試験の検出力は目標とする検出力よりも低いものとなる．BE 評価で用いられるクロスオーバー試験では，

脚注 1) BE ガイドラインにおける「例数」と同義．その他に，例えば ICH E9 ガイドラインでは，「被験者数」が sample size を表す用語として用いられている．本論文では，表記ゆれを避けるために，一貫して「サンプルサイズ」を用いる（引用を除く）.

PK パラメータ（通常，AUC と Cmax）の試験薬と対照薬の幾何平均比と個体内分散に基づいて，サンプルサイズを計算する．これらの情報は予試験などの他の試験から得ることができるが，その情報の確かさが BE 試験の成否に大きく影響するため，試験途中で得られた情報に基づく BE 評価や必要サンプルサイズの再計算は自然な要求である．改正後の BE ガイドラインは，1 回に限り中間解析とその結果に基づくサンプルサイズの再計算を許容しているが，先にも述べた通り，その試験における第一種の過誤確率を厳格に 5%以下[脚注 2)] に制御することを要求している．このような試験では，大きく分けて二つの第一種の過誤確率の増大を引き起こす可能性のある要因が考えられる．一つは，中間解析と最終解析で BE 評価を繰り返し行うことによるものである．もう一つは，中間解析の結果に基づくサンプルサイズ追加によるものである．

2.1 BE 評価の繰り返しによる第一種の過誤確率の増大

試験終了時の解析（最終解析）に加えて，試験途中にそれまでに集積したデータを用いた解析（中間解析）を行い，何らかの意思決定を行うデザインは，群逐次デザイン（group sequential design）と呼ばれる．群逐次デザインでは，試験中に複数回の検定を繰り返すことによる多重性の問題が生じる．宝くじを例にして多重性の問題を説明する．20 本中 1 本当たりが含まれる宝くじ（つまり，5%の確率で当たりを引くことができる）を一度だけ引くことができるものを中間解析がない試験に対応させると，その宝くじを複数回引くことができるものは群逐次デザインに対応するものと考えることができる．5%の確率で当たる宝くじを複数回引くことによって，5%よりも大きい確率で当たりを引くことができることは想像に難くないだろう．複数回引いてもなお，当たる確率を 5%にするためには，外れくじを増やして 1 回あたりに当たる確率を下げればよい．

群逐次デザインにおける第一種の過誤確率の制御のために用いられる方法として，Pocock の方法や，O'Brien and Fleming の方法などがある．Pocock の方法と O'Brien and Fleming の方法は，いずれも，試験全体の第一種の過誤確率を有意水準 α 以下に制御するために，1 回の検定当たりの有意水準を調整する（下げる）ものである．Pocock の方法では，中間解析と最終解析を同一の水準で検定するように棄却限界値を計算する．最終解析の検定統計量は，その一部に中間解析までのデータを含むため，中間解析の検定統計量との間に正の相関をもつ．例として，分散既知の正規分布に従う連続データの群逐次デザインを考える．2 群のサンプルサイズが等しい時，中間解析までの各群のサンプルサイズを n_1，中間解析から最終解析までの各群のサンプルサイズを n_2 とすると（すなわち，試験全体での各群のサンプルサイズは $n_1 + n_2$ である），中間解析と最終解析の検定統計量間の相関は $\sqrt{n_1/(n_1+n_2)}$ となる．Pocock の方法は，中間解析と最終解析の検定統計量間の相関を考慮した同時分布において，中間解析と最終解析を同一の有意水準で検定するように棄却限界値を決定する．Fig. 1 に幾何学的イメージを示す．

楕円は，x 軸方向に中間解析の検定統計量 Z_1，y 軸方向に最終解析の検定統計量 Z をとる同時分布を表している[脚注 3)]．x 軸上の任意の点 z_{α_1} を中間解析での棄却限界値，y 軸上の任意の点 z_{α_2} を最終解析での棄却限界値とすると，グレーの色付きで示す領域の確率 $P(Z_1 < z_{\alpha_1} \cap Z < z_{\alpha_2})$ が $1 - \alpha$ となるように，z_{α_1} と z_{α_2} を決定すれば第一種の過誤確率を α 以下に制御することができる．Pocock の方法では，$z_{\alpha_1} = z_{\alpha_2}$ として，棄却限界値を決定する．一方，O'Brien and Fleming の方法は，情報量（被験者数）が少ない中間解析での有意水準を抑えて，最終解析に多くの有意水準を割り当てるように棄却限界値を決定する．Table 1 に，Pocock の方法と O'Brien and Fleming の方法において，中間解析を 1 回行う場合の中間解析と最終解析のそれぞれで用いる有意水準を示す．

脚注 2) BE 評価における帰無仮説は，幾何平均比が BE の許容域の下限以下または上限以上である．これらは同時には成り立つことはない．いずれかの帰無仮説が真であるとき，有意水準 5%の二つの片側検定が両方有意になる確率は，5%×もう一方の帰無仮説を棄却する確率となるため，5%以下となる．つまり，第一種の過誤の確率（真には BE ではないにもかかわらず，二つの片側検定が有意になる確率）は 5%以下である．

脚注 3) ここで示したような二変量の同時分布は紙面に垂直方向に高さ（確率密度）を持つため，等高線状に表すことが多いが，ここでは省略している．なお，楕円は，Z_1 と Z の値域がこの領域にしかないことを意味しているわけではない．実際には，いずれも $-\infty$ から ∞ の範囲で値をとり得る．

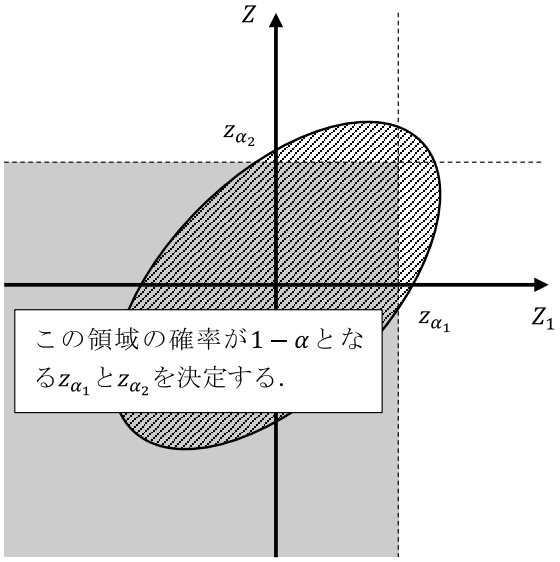

Fig. 1　群逐次デザインで棄却限界値を決定する幾何学的イメージ

Table 1　Pocock の方法と O'Brien and Fleming の方法における有意水準

	中間解析	最終解析
Pocock	0.0304	0.0304
O'Brien and Fleming	0.0088	0.0467

2.2　中間解析の結果に基づくサンプルサイズ調整による第一種の過誤確率の増大

中間解析の結果に基づいて実施中の試験の特徴を調整するデザインはアダプティブデザイン（adaptive design）と呼ばれ，2021 年現在，ICH においても新規ガイドラインを整備中である[15]．中間解析の結果に基づいてサンプルサイズを再計算するとき，中間解析以降のサンプルサイズは，中間解析で得られた情報の関数となる．例として，中間解析で得られた分散の推定値のみをサンプルサイズの再計算に用いることを考える．中間解析以降のサンプルサイズは，中間解析で得られた分散が小さければ少なくなり，その分散が大きければ多くなる．

2.1 節の単純な例で示したように，分散既知の正規分布に従う連続データの群逐次デザインでは，中間解析と最終解析の検定統計量間の相関は $\sqrt{n_1/(n_1+n_2)}$ となる．サンプルサイズを再計算すると，中間解析以降のサンプルサイズ n_2 は中間解析の結果により変動し，さらには中間解析と最終解析の検定統計量間の相関もその影響を受けて変わり得るため，標準的な群逐次法の適用では，第一種の過誤確率を制御することはできない．

この問題に対処する方法の一つとして，逆正規法が知られている．先の例と同様に，分散既知の正規分布に従う連続データについて中間解析と最終解析の 2 回検定する群逐次デザインを考える．ただし，中間解析以降のサンプルサイズ n_2 は中間解析の結果に基づいて決定されるとする．Z_1 を中間解析までのデータを用いて計算した検定統計量，Z_2 を中間解析以降，最終解析までのデータを用いて計算した検定統計量とする．Z_2 は中間解析より前のデータを含まないため，Z_1 と Z_2 は独立である．事前に指定した任意の重み w ($0 < w < 1$）に対して，最終解析の検定統計量

$$Z = \sqrt{w}Z_1 + \sqrt{1-w}Z_2$$

は，帰無仮説の下で標準正規分布に従う．例えば，$w = 1/2$ とすると，最終解析の検定統計量 Z に対して，中間解析までのデータを含む Z_1 と，中間解析以降，最終解析までのデータを含む Z_2 は，中間解析以降のサ

ンプルサイズ n_2 の大きさに関わらず，1/2 ずつ寄与することを意味する．中間解析の結果に依存する n_2 の大きさとは独立な重み w を用いることによって，中間解析の検定統計量 Z_1 と最終解析の検定統計量 Z は，n_2 の大きさに依らず，相関 $\sqrt{w}$ を持つため，2.1 で述べたような標準的な群逐次法を適用することができる．逆正規法では，より一般に，以下のように最終解析の検定統計量を構成する．

$$Z = \sqrt{w}\Phi^{-1}(1 - p_1) + \sqrt{1 - w}\Phi^{-1}(1 - p_2)$$

ここで，$\Phi^{-1}(\cdot)$ は標準正規分布の累積分布関数の逆関数であり，p_1 と p_2 はそれぞれ，中間解析までの検定統計量に対応する p 値と，中間解析以降，最終解析までの検定統計量に対応する p 値を表す．

3. Combination Test (Maurer et al.[5])

ここでは，2 剤 2 期クロスオーバーデザインを想定する．また，中間解析と最終解析の 2 回 BE の評価が行われ，中間解析ではその結果に基づいて必要サンプルサイズの再計算を行うとする．説明のため，中間解析までをステージ 1，中間解析から最終解析までをステージ 2 とする．同等性評価パラメータは対数変換した AUC または Cmax とする．対数変換したパラメータの試験製剤と参照製剤の平均をそれぞれ μ_T，μ_R とする．$\delta = \mu_T - \mu_R$ とするとき，二つの片側検定（Two One-Sided Test，以降，TOST とする）に対応する帰無仮説と対立仮説は以下の通りである．ここで，$-\Delta$ と Δ はそれぞれ BE の下方と上方の許容域であり，通常は $\Delta = \log(1.25) = |\log(0.8)|$ である．

下方の片側検定：

$$H_{01}: \delta \leq -\Delta \text{ vs. } H_{11}: \delta > -\Delta$$

上方の片側検定：

$$H_{02}: \delta \geq \Delta \text{ vs. } H_{12}: \delta < \Delta$$

中間解析を行わない BE 試験では，二つの片側検定が有意水準 5% でいずれも有意であるとき，BE が成立していると判定される．これは，両側 90% 信頼区間が BE の許容域に含まれるか否かの判定と，本質的に同一の評価である．H_{01} と H_{02} に対するステージ i の検定統計量は，それぞれ，

$$T_{i1} = \frac{\Delta + \hat{\delta}_i}{\sqrt{2\hat{\sigma}_i^2/n_i}}$$

$$T_{i2} = \frac{\Delta - \hat{\delta}_i}{\sqrt{2\hat{\sigma}_i^2/n_i}}$$

である[脚注 4)]．$\hat{\delta}_i$，$\hat{\sigma}_i^2$，n_i はそれぞれ，ステージ i における δ の推定値，被験者内分散の推定値，サンプルサイズを表す．T_{i1} と T_{i2} はそれぞれ，帰無仮説 H_{01}，H_{02} の下で，自由度 ν の t 分布に従う．T_{i1} と T_{i2} に対応する p 値をそれぞれ，p_{i1} と p_{i2} とする．

3.1 Standard Combination Test

Kieser and Rauch[6] は，中間解析及びその結果に基づくサンプルサイズ再計算を伴う BE 試験においても，標準的な群逐次法，及び，逆正規法の適用により，第一種の過誤確率を制御可能であることを示した．Maurer et al.[5] は，彼ら自身の提案法との対比のために，Kieser and Rauch の提案法を Standard Combination Test と呼称した．

脚注 4) 上方の片側検定の対立仮説は左側仮説であるが，対応する検定統計量では右側仮説を考えていることに注意する．（後の条件付き検出力の計算のために，下方と上方の対立仮説の方向を揃えている．）

Standard Combination Test では，まず，ステージ 1 とステージ 2 の統計量を逆正規法によって統合する際の重み w $(0 < w < 1)$ を設定する．$Z_{ij} = \Phi^{-1}(1 - p_{ij})$ とするとき，ステージ 2 終了時（最終解析）の検定統計量 Z_{0j} は，

$$Z_{0j} = \sqrt{w}Z_{1j} + \sqrt{1 - w}Z_{2j}$$

となる．2.2 節で説明した通り，Z_{1j} と Z_{0j} は，帰無仮説 H_{0j} の下で，二変量標準正規分布に従う．

$$\begin{pmatrix} Z_{1j} \\ Z_{0j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{w} \\ \sqrt{w} & 1 \end{pmatrix}\right)$$

試験全体で制御したい第一種の過誤確率の水準を α（通常は 0.05），ステージ i 終了時の検定に対する棄却限界値を z_{α_i} とするとき，z_{α_i} は以下の解として得ることができる．

$$P(Z_{1j} < z_{\alpha_1} \cap Z_{0j} < z_{\alpha_2}) = 1 - \alpha$$

Pocock の方法のように，$\alpha_1 = \alpha_2$ の下で解を得ることを考えると，$w = 0.5$ のとき，$z_{\alpha_1} = z_{\alpha_2} = 1.8754$, $\alpha_1 = \alpha_2 = 0.0304$ である．

ステージ 1 終了後の TOST では，$p_{11} < \alpha_1$ かつ $p_{12} < \alpha_1$ であるとき BE を宣言し，試験を終了する．BE が成立しなければ，ステージ 1 の結果に基づきサンプルサイズを再計算し，ステージ 2 に進む（試験を継続する）．ステージ 2 終了後，検定統計量 Z_{1j} と Z_{2j} を重み w により統合した検定統計量 Z_{0j} に基づき，再度 TOST を実施する．Z_{0j} の観測値 z_{0j} が，$z_{01} < z_{\alpha_2}$ かつ $z_{02} < z_{\alpha_2}$ であるとき BE を主張することができる．

3.2 Maximum Combination Test

3.1 節で説明した Standard Combination Test では，実際のサンプルサイズとは独立した単一の重み w を用いて，ステージ 1 とステージ 2 の検定統計量を統合する．理想的には，重み w が，実際のサンプルサイズの比 $n_1/(n_1 + n_2)$ に一致するとき最大の検出力が期待できる．しかしながら，ステージ 2 のサンプルサイズ n_2 は，ステージ 1 の結果に基づき計算されるため，このような重みを試験の実施前に設定することは不可能である．Maurer et al.[5] は，一つの重み w に加えて，ステージ 2 により重みが割り当たるようにした w^*（つまり，$1 - w^* > 1 - w$）を同時に設定することを提案し，それを Maximum Combination Test と呼称した．

ここでは，事前に二つの異なる重み w と w^* $(0 < w^* < w < 1)$ を設定することを考える．それぞれの重みに対応する統計量 Z_{0j} と Z^*_{0j} は以下のようになる．

$$Z_{0j} = \sqrt{w}Z_{1j} + \sqrt{1 - w}Z_{2j}$$
$$Z^*_{0j} = \sqrt{w^*}Z_{1j} + \sqrt{1 - w^*}Z_{2j}$$

Maximum Combination Test では，$Z_{maxj} = \max(Z_{0j}, Z^*_{0j})$ を帰無仮説 H_{0j} に対するステージ 2 終了時の検定統計量とする．Z_{1j}, Z_{0j}, Z^*_{0j} は，帰無仮説 H_{01} の下で，以下の三変量標準正規分布に従う．

$$\begin{pmatrix} Z_{1j} \\ Z_{0j} \\ Z^*_{0j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{w} & \sqrt{w^*} \\ \sqrt{w} & 1 & \left(\sqrt{ww^*} + \sqrt{(1 - w)(1 - w^*)}\right) \\ \sqrt{w^*} & \left(\sqrt{ww^*} + \sqrt{(1 - w)(1 - w^*)}\right) & 1 \end{pmatrix}\right)$$

試験全体で制御したい第一種の過誤確率の水準を α（通常は 0.05），ステージ 1，ステージ 2 終了時の棄却限界値をそれぞれ，$z_{\alpha_{1,max}}$, z_{max} とするとき，$z_{\alpha_{1,max}}$ と z_{max} は，以下の解として得られる．

$$P(\{Z_{1j} < z_{\alpha_{1,max}}\} \cap \{Z_{0j} < z_{max}\} \cap \{Z_{0j}^{*} < z_{max}\}) = 1 - \alpha$$

先と同様に，$z_{\alpha_{1,max}} = z_{max}$ の下で解を得ることを考えると，例えば，$w = 0.5$，$w^{*} = 0.25$ のとき，$z_{\alpha_{1,max}} = z_{max} = 1.9374$ であり，対応する有意水準は 0.02635 である．

以降，Standard Combination Test と同様の手順で，BE を評価する．

4．サンプルサイズ再計算の方法

2.2 節で説明した通り，中間解析の結果に基づくサンプルサイズの再計算を行うときに第一種の過誤確率が増大する可能性があるが，これは，実際のサンプルサイズとは独立した重みを用いることによって制御可能である．言い換えれば，実際のサンプルサイズとは独立した重みを用いる限りは，サンプルサイズの計算方法に依らず，第一種の過誤確率を制御することができる．

δ と σ^2 に対して，計画時に仮定した値をそれぞれ δ_p，σ_p^2 とする．Patterson and Jones[7] の例では，中間解析で得られた σ^2 の推定値 $\hat{\sigma}_1^2$ を計画時に仮定した値 σ_p^2 から置き換えることによってサンプルサイズを再計算している．具体的には，試験開始時に，δ_p，σ_p^2 と試験全体の有意水準 α（BE 試験では，0.05），目標とする検出力 $1 - \beta$（例えば，0.8）の下で，必要サンプルサイズ N を計算する．ここでは，ステージ 1 のサンプルサイズを $n_1 = N/2$ として試験を実施したとする．ステージ 1 終了後，BE が成立しなかった場合は，計画時の δ_p，ステージ 1 で得られた σ^2 の推定値 $\hat{\sigma}_1^2$，最終解析で用いる有意水準（Standard または Maximum Combination Test で計算したもの），目標とする検出力 $1 - \beta$ を用いて，必要サンプルサイズを再計算する．再計算したサンプルサイズを N' とすると，ステージ 2 のサンプルサイズは $n_2 = N' - n_1$ となる．

試験の計画時に仮定した被験者内分散 σ_p^2 の不確実性に対処するためのものとして，このような単純な手順を採り得るかもしれない．この手順では，ステージ 1 のデータはすでに得られている（つまり，確率変数ではなく実現値である）にもかかわらず，これから得られるものとして考えている（つまり，未だ確率変数と考えている）ことに注意が必要である．Maurer et al.[5] はこれに対処するために，conditional error rate と estimated target conditional power をサンプルサイズの再計算に用いることを提案した．

Standard Combination Test と，ステージ 1 のデータを所与とした conditional error rate に基づくステージ 2 の検定は，等価であることが知られている．ステージ 1 の検定統計量の実現値 $z_{1j} = \Phi^{-1}(1 - p_{1j})$，重み w，ステージ 1 と 2 それぞれの有意水準 $\alpha_j, j = 1, 2$ の下で，ステージ 2 の conditional error rate は次のようになる．

$$\alpha_{comb_j}^{c} = 1 - \Phi\left((z_{\alpha_2} - \sqrt{w} z_{1j})/\sqrt{1 - w}\right)$$

このとき，ステージ 1 のデータを所与としたステージ 2 の検定の条件付き検出力は，ステージ 2 の検定統計量 $T_{2j}, j = 1, 2$ をそれぞれ対応する $\alpha_{comb_j}^{c}$ で検定するときの検出力に等しい．同様に，Maximum Combination Test での conditional error rate は次のようになる．

$$\alpha_{max_j}^{c} = 1 - \Phi(min_j),$$
$$min_j = \min\left[(z_{max} - \sqrt{w} z_{1j})/\sqrt{1 - w}, (z_{max} - \sqrt{w^{*}} z_{1j})/\sqrt{1 - w^{*}}\right]$$

条件付き検出力に基づくステージ 2 のサンプルサイズは，$\alpha_{comb_j}^{c}$（Standard Combination Test の場合）または $\alpha_{max_j}^{c}$（Maximum Combination Test の場合）を有意水準とした TOST の検出力が，目標とする検出力を上回るときの n_2 として計算することができる．しかしながら，一般には，$\alpha_{comb_1}^{c} \neq \alpha_{comb_2}^{c}$，$\alpha_{max_1}^{c} \neq \alpha_{max_2}^{c}$，つまり，下方と上方の片側検定の（条件付き）有意水準が等しくならないため，二変量 t 分布の数値積分によって条件付き検出力を計算する必要がある．（本原稿執筆時点において，多くのソフトウェアは，有意水準が等しい場合のサンプルサイズもしくは検出力の計算にしか対応していない．）さらに，conditional error rate が不等であることに関連して，δ_p よりも 0 に近い δ'（つまり，$|\delta_p| > |\delta'|$）において，検出力が低くな

る．これは，通常の有意水準が等しいときの検出力の計算では生じない．言い換えれば，$|\delta_p| > |\delta'|$ であるとき，δ' の検出力は δ_p の検出力よりも大きい．彼らは，この問題に対する解として，以下の δ_{ap}

$$\delta_{ap} = \begin{cases} +|\delta_p| \text{ if } \alpha^c_{comb_1} > \alpha^c_{comb_2} (\text{or } \alpha^c_{max_1} > \alpha^c_{max_2}) \\ -|\delta_p| \text{ if } \alpha^c_{comb_1} \leq \alpha^c_{comb_2} (\text{or } \alpha^c_{max_1} \leq \alpha^c_{max_2}) \end{cases}$$

を δ_p に代えて，サンプルサイズの計算に用いることを導いた．（彼らはこれを adaptive planning と呼んだ．）

Maurer et al.[5] は，さらなる改良として，サンプルサイズの再計算で目標とする検出力に，estimated target conditional power，$(1 - \beta^c)$ を用いることを提案した．

$$(1 - \beta^c) = (\hat{\beta}_1 - \beta)/\hat{\beta}_1$$

ここで，$(1 - \beta)$ は試験全体で目標とする検出力（例えば，0.8），$(1 - \hat{\beta}_1)$ はステージ 1 の有意水準 α_1，サンプルサイズ n_1，計画時に仮定した δ_p と σ_p^2 の下で計算した検出力である．

ここでは，サンプルサイズを再計算する方法として，初期の仮定のうち σ_p^2 を $\hat{\sigma}_1{}^2$ に置き換える単純な方法と，Maurer et al.[5] が提案する conditional error rate と estimated target conditional power を用いる方法を紹介した．Maurer et al.[5] でも述べられている通り，いずれの方法も中間解析で得られた σ^2 や δ の推定値に基づいているため，サンプルサイズ再計算の結果として，試験全体での検出力 $1 - \beta$ が目標の値になることを保証するものではない．また，再計算後のサンプルサイズが実施不可能な規模となることも考えられる．その場合，試験を早期中止することや，検出力（の推定値）が目標の値に満たないまま試験を継続するといったことが考えられる．その他にも，追加の中止基準を設定することも考えられる．このような，中途の情報に基づいてデザインに変更を加える試験（いわゆる，アダプティブデザイン）では，シミュレーションによる動作特性（第一種の過誤確率，検出力，期待サンプルサイズなど）の評価が必要不可欠である．

5. 数値例

ここでは，2 剤 2 期クロスオーバーデザインのバランスデータを想定した仮想データに対して，Combination Test を適用することを考える．計画時の仮定として，$\delta_p = 0$（幾何平均比が 1），$\sigma_p^2 = 0.086$（被験者内 CV が 0.3）とする．有意水準 $\alpha = 0.05$，検出力 $1 - \beta = 0.8$ とすると，通常のサンプルサイズを固定したデザインにおける必要サンプルサイズは，$N = 32$ と計算される．ここでは，その半分である $n_1 = 16$ をステージ 1 のサンプルサイズとする．BE の許容域は，$(-\Delta, \Delta)$，$\Delta = \log(1.25) = |\log(0.8)|$ とする．

5.1 Standard Combination Test

重み $w = 0.5$ のとき，$z_{\alpha_1} = z_{\alpha_2} = 1.8754$，$\alpha_1 = \alpha_2 = 0.0304$ である．ステージ 1 終了後（中間解析時），$\hat{\delta}_1 = -0.046$，$\hat{\sigma}_1{}^2 = 0.1201$ を得た．TOST の下方と上方の検定統計量，t_{11} と t_{12} は，それぞれ，

$$t_{11} = \frac{\log(1.25) + (-0.046)}{\sqrt{2(0.1201)/16}} = 1.4458$$

$$t_{12} = \frac{\log(1.25) - (-0.046)}{\sqrt{2(0.1201)/16}} = 2.1966$$

である．自由度 14（$= 16 - 2$）の t 分布より，対応する p 値はそれぞれ，$p_{11} = 0.0851$（$> \alpha_1$），$p_{12} = 0.0227$（$< \alpha_1$）であるため，BE は成立しない．サンプルサイズ再計算後，ステージ 2 に進む．

$z_{11} = \Phi^{-1}(1 - 0.0851) = 1.3714$，$z_{12} = \Phi^{-1}(1 - 0.0227) = 2.0011$ より，conditional error rate はそれぞれ，

$$\alpha^{c}_{comb_1} = 1 - \Phi\left((1.8754 - \sqrt{0.5}\cdot 1.3714)/\sqrt{1 - 0.5}\right) = 0.1001$$
$$\alpha^{c}_{comb_2} = 1 - \Phi\left((1.8754 - \sqrt{0.5}\cdot 2.0011)/\sqrt{1 - 0.5}\right) = 0.2575$$

となる．Estimated target conditional power は，$(1 - \hat{\beta}_1) = 0.0591$ より，次のようになる．

$$(1 - \beta^{c}) = (0.9409 - 0.2)/0.9409 = 0.7874$$

計算した conditional error rate と，$\delta_p = 0$，$\hat{\sigma}_1{}^2 = 0.1201$ を用いて，estimated target conditional power（$= 0.7874$）を上回るサンプルサイズを計算すると，$n_2 = 26$ を得る．

$n_2 = 26$ のステージ 2 終了後（最終解析時），$\hat{\delta}_2 = 0.0233$，$\hat{\sigma}_1{}^2 = 0.0975$ を得た．ステージ 2 の t_{21} と t_{22} は，それぞれ，

$$t_{21} = \frac{\log(1.25) + (0.0233)}{\sqrt{2(0.0975)/26}} = 2.8457$$
$$t_{22} = \frac{\log(1.25) - (0.0233)}{\sqrt{2(0.0975)/26}} = 2.3076$$

であり，対応する p 値はそれぞれ，自由度 24（$= 26 - 2$）の t 分布より，$p_{21} = 0.0045$，$p_{12} = 0.0150$ と計算される．$z_{21} = \Phi^{-1}(1 - 0.0045) = 2.6148$，$z_{22} = \Phi^{-1}(1 - 0.0150) = 2.1707$ より，ステージ 2 終了後（最終解析時）の検定統計量は，

$$z_{01} = \sqrt{0.5}\cdot 1.3714 + \sqrt{1 - 0.5}\cdot 2.6148 = 2.8187$$
$$z_{02} = \sqrt{0.5}\cdot 2.0011 + \sqrt{1 - 0.5}\cdot 2.1707 = 2.9499$$

となる．z_{01} と z_{02} のいずれも，$z_{\alpha_2} = 1.8754$ より大きい．したがって，下方，上方の両方の帰無仮説を棄却し，BE が成立する．

5.2 Maximum Combination Test

二つの重みを $w = 0.5$，$w^{*} = 0.25$ とするとき，$z_{\alpha_{1,max}} = z_{max} = 1.9374$ であり，対応する有意水準は 0.02635 である．5.1 節と同様に，ステージ 1 終了後（中間解析時），$\hat{\delta}_1 = -0.046$，$\hat{\sigma}_1{}^2 = 0.1201$ を得た．TOST の下方と上方の帰無仮説に対する p 値はそれぞれ，$p_{11} = 0.0851$（> 0.02635），$p_{12} = 0.0227$（< 0.02635）であるため，BE は不成立である．サンプルサイズ再計算後，ステージ 2 に進む．

$z_{11} = 1.3714$，$z_{12} = 2.0011$ より，conditional error rate はそれぞれ，

$$\alpha^{c}_{max_1} = 0.0856, \qquad \alpha^{c}_{max_2} = 0.2300$$

となる．$(1 - \hat{\beta}_1) = 0.0457$ より，estimated target conditional power は，$(1 - \beta^{c}) = 0.7904$ となる．

計算した conditional error rate と，$\delta_p = 0$，$\hat{\sigma}_1{}^2 = 0.1201$ を用いて，estimated target conditional power（$= 0.7904$）を上回るサンプルサイズを計算すると，$n_2 = 28$ を得る．

$n_2 = 28$ のステージ 2 終了後（最終解析時），$\hat{\delta}_2 = -0.0123$，$\hat{\sigma}_2{}^2 = 0.0930$ を得た．ステージ 2 の t_{21} と t_{22} は，それぞれ，

$$t_{21} = 2.5869, \qquad t_{22} = 2.8887$$

であり，対応する p 値はそれぞれ，自由度 26（$= 28 - 2$）の t 分布より，$p_{21} = 0.0078$，$p_{22} = 0.0038$ と計算される．$z_{21} = 2.4174$，$z_{22} = 2.6651$ より，ステージ 2 終了後（最終解析時）の検定統計量は

$$z_{max_1} = \max(2.6791, 2.7792) = 2.7792$$
$$z_{max_2} = \max(3.2995, 3.3086) = 3.3086$$

となる．z_{max_1} と z_{max_2} のいずれも，$z_{max} = 1.9374$ より大きい．したがって，下方，上方の両方の帰無仮説を棄却し，BE が成立する．

6. まとめ

改正後の BE ガイドラインに関連して，BE 試験の途中での中間解析，及び，中間解析の結果に基づくサンプルサイズの再計算に伴って生じる統計的諸問題について説明した．これは，BE 試験に固有のものではなく，近代盛んに研究されている群逐次デザインやアダプティブデザインでの課題と同一のものである．これらの課題については，Bretz et al.[8]，平川・五所[9]，手良向・大門[10] や上村[11] などが参考になるだろう．

Kieser and Rauch[6] は，群逐次デザインやアダプティブデザインでの方法が BE 試験においても適用可能であることを示した．Maurer et al.[5] は，Combination Test と条件付き検出力に基づくサンプルサイズ再計算の方法を提案した．彼らの提案法を適用するためには，途中に確率密度関数の数値積分を必要とすることから，多少の煩雑さを伴うかもしれない．その他の方法としては，Potvin et al.[12] が提案した Method A, B, C, D の四つの方法がある．Health Canada のガイドライン[13] は，Method C を推奨法として挙げている．Potvin et al. の方法は，適用が簡便である一方で，Kieser and Rauch が指摘するように，有意水準の設定や，その手順の構成が（おそらくは）ヒューリスティックに決定されているため，必ずしも第一種の過誤確率を制御できることを保証しない．いくつかの条件で第一種の過誤確率が増大することは，彼ら自身のシミュレーションや，Montague et al.[14] の追加のシミュレーションでも示されている．改正後の BE ガイドラインの要求に対応するためには，厳格に第一種の過誤確率を制御可能な方法を適用することが必要だろう．

謝辞

本論文を作成するにあたり，ご確認，コメントを頂いたファイザー株式会社 土綿慎一氏に感謝する．

利益相反（COI）の開示

本稿作成に関し，開示すべき利益相反関係はない．

引用文献

1） 緒方宏泰．医薬品の生物学的同等性試験―ガイドライン対応―．付録 後発医薬品の生物学的同等性試験ガイドライン（平成 24 年 2 月 29 日付薬食審査発 0229 第 10 号，別紙 1）．東京：株式会社じほう；2013. pp.247-66.
2） 緒方宏泰．医薬品の生物学的同等性試験―ガイドライン対応―．付録 後発医薬品の生物学的同等性試験ガイドライン Q & A（平成 24 年 2 月 29 日付事務連絡，別紙 1）．東京：株式会社じほう；2013. pp.267-92.
3） 後発医薬品の生物学的同等性試験ガイドライン（薬生薬審発 0319 第 1 号令和 2 年 3 月 19 日，別紙 1）
4） 後発医薬品の生物学的同等性試験ガイドライン Q & A（令和 2 年 3 月 19 日事務連絡，別紙 1）
5） Maurer W, Jones B, Chen Y. Controlling the type I error rate in two-stage sequential adaptive designs when testing for average bioequivalence. *Statistics in Medicine*. 2018; 37: 1587-607.
6） Kieser M, Rauch G. Two-stage designs for cross-over bioequivalence trials. *Statistics in Medicine*. 2015; 34: 2403-16.
7） Patterson S, Jones B. Bioequivalence and statistics in clinical pharmacology, 2nd edition. 6. Adaptive bioequivalence trials. London: CRC Press/Chapman and Hall; 2016. pp.141-87.
8） Bretz F, Koenig F, Brannath W. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*. 2009; 28: 1181-217.
9） 平川晃弘，五所正彦監訳．臨床試験のためのアダプティブデザイン．東京：朝倉書店；2018.
10） 手良向聡，大門貴志訳．臨床試験デザイン ベイズ流・頻度流の適応的方法．東京：株式会社メディカル・パブリケーションズ；2014.
11） 上村鋼平．臨床試験における被験者数再設定―方法論の概説と統計学的留意点―．計量生物学．2012；33：77–99.
12） Potvin D, DiLiberti CE, Hauck WW et al. Sequential design approaches for bioequivalence studies with crossover

designs. *Pharmaceut. Statist*. 2008; 7: 245-62.
13） Health Canada. Guidance document: Conduct and analysis of comparative bioavailability studies. Date adopted: 2012/02/08, Revised date: 2018/06/08, Effective date: 2018/09/01 (for submissions filed on or September 1, 2018)
14） Montague TH, Potvin D, DiLiberti CE et al. Additional results for 'Sequential design approaches for bioequivalence studies with crossover designs'. *Pharmaceut. Statist*. 2012; 11: 8-13.
15） ICH official website. E20 adaptive clinical trials concept paper. https://database.ich.org/sites/default/files/E20_FinalConceptPaper_2019_1107_0.pdf（参照 2021-04-30）