非劣性試験における割合の差の信頼区間と例数設計

○魚住 龍史[1]　飯塚 政人[2]　浜田 知久馬[3]
[1]京都大学大学院医学研究科 医学統計生物情報学
[2]田辺三菱製薬株式会社 開発本部 データサイエンス部
[3]東京理科大学大学院 工学研究科 経営工学専攻

Asymptotic confidence intervals and sample size calculations for the difference between independent binomial proportions in non-inferiority clinical trials

Ryuji Uozumi[1] , Masato Iizuka[2] , and Chikuma Hamada[3]
[1]*Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine*
[2]*Data Science Department Development Division Mitsubishi Tanabe Pharma Corporation*
[3] *Department of Management Science, Graduate School of Engineering, Tokyo University of Science*

要旨

V9.2 以降，FREQ プロシジャの大幅な機能拡張により，割合の信頼区間のタイプを指定できるようになった．V9.2 以降 BINOMIAL オプションで 1 標本の割合の信頼区間のタイプを指定できるようになり，V9.3 以降 RISKDIFF オプションで 2 標本の差の信頼区間のタイプも指定できるようになった．そして，V9.4 の RISKDIFF オプションでは，信頼区間のタイプにさらなる拡張がなされた．本発表では，非劣性試験の下，2 標本の割合の差の信頼区間を主解析とする場合を想定する．非劣性試験では，信頼区間に基づく仮説の検証が行われることが一般的であるため，解析計画の段階で信頼区間のタイプを選択することが重要となる．具体的な主解析方法を特定した上，非劣性試験に必要な例数を見積もるためには POWER プロシジャが有用である．しかし，POWER プロシジャの TWOSAMPLEFREQ ステートメントで指定できる検定方法（信頼区間のタイプ）は限られるため，主解析として選択した方法に基づき例数設計を行うためには，シミュレーションによる評価が必要になってしまう．本発表では，正規分布への漸近近似を利用した割合の差の信頼区間に焦点を当て，SAS/STAT 13.1 の POWER プロシジャの計算結果では適切に例数設計を行うことができない場合を取りあげる．そして，飯塚，魚住，浜田 (2014) により推奨された Miettinen-Nurminen 信頼区間を主解析として選択した場合の例数設計法, 及び SAS/STAT 13.2 における機能拡張についても紹介する．

キーワード：非劣性　FREQ　RISKDIFF　割合の差　信頼区間　POWER　TWOSAMPLEFREQ　例数設計　Miettinen-Nurminen　Farrignton-Manning

1 はじめに

医学研究では，ある治療の有効性を評価するために，有効割合や改善割合のような 2 値データを解析することがある．このような 2 値データに対する解析を行うために FREQ プロシジャは有用である．TABLE ステートメントにおけるオプションでカイ二乗検定や Fisher 正確検定，割合の差の信頼区間を出力できる．V9.2 以降，FREQ プロシジャの大幅な機能拡張により，割合の信頼区間のタイプを指定できるようになった．例えば，V9.2 以降 BINOMIAL オプションで 1 標本の割合の信頼区間のタイプを指定できるようになり，V9.3 以降 RISKDIFF オプションで 2 標本の差の信頼区間のタイプも指定できるようになった [17]．そして，V9.4 の RISKDIFF オプションでは，信頼区間のタイプにさらなる拡張がなされた．この拡張により，2 値データに対する主解析の計画時に，多くの選択肢ができたといえる．これらの信頼区間のタイプのうち，V9.1 以前までによく用いられていた Wald 信頼区間の問題点が指摘され, Miettinen-Nurminen 信頼区間 [10] や Newcombe スコア信頼区間 [11] の使用が推奨されるようになった [5, 14, 15].

本発表では，非劣性試験の下，2 標本の割合の差の信頼区間を主解析とする場合を想定する．非劣性試験では，信頼区間に基づく仮説の検証及び報告が行われることが一般的である．このため，解析計画の段階で信頼区間のタイプを選択することが重要となる．さらに，解析計画時に信頼区間のタイプを特定した上で必要例数を見積もらなければならない．例数設計を実施するためには POWER プロシジャが有用である．しかし，POWER プロシジャの TWOSAMPLEFREQ ステートメントで指定できる検定方法（信頼区間のタイプ）は限られるため，主解析として選択した方法によっては POWER プロシジャでは計算できない．そのため，POWER プロシジャでサポートされていない解析方法を主解析とした場合の例数設計を行うためには，擬似乱数を用いたシミュレーションによる評価に頼らなければならない [16].

本稿では，正規分布への漸近近似を利用した割合の差の信頼区間に焦点を当て，SAS/STAT 13.1 の POWER プロシジャの計算結果では適切に例数設計を行うことができない場合がある点を取りあげる. そして, 飯塚, 魚住, 浜田 (2014) により推奨された Miettinen-Nurminen 信頼区間を主解析として選択した場合の例数設計法を示す.

2 想定する 2 値データ

表 1 : 2×2 分割表

	有効 $(Y=1)$	無効 $(Y=2)$	計	有効割合	パラメータ
治療 1 ($X=1$)	n_{11}	n_{12}	$n_{1\cdot}$	$p_1 = n_{11}/n_{1\cdot}$	π_1
治療 2 ($X=2$)	n_{21}	n_{22}	$n_{2\cdot}$	$p_2 = n_{21}/n_{2\cdot}$	π_2
計	$n_{\cdot 1}$	$n_{\cdot 2}$	N	$p_T = n_{\cdot 1}/N$	

本稿で想定する 2 値データを表 1 に示す．表 1 は，治療 1 ($X=1$) と治療 2 ($X=2$) の 2 群を対象とした並行群間比較試験から得られるデータとする．表 1 において，有効割合の差は

$$p_1 - p_2 = \frac{n_{11}}{n_{1\cdot}} - \frac{n_{21}}{n_{2\cdot}}$$

となる．

ここで，以下のように帰無仮説 H_0，対立仮説 H_A をそれぞれ

$$H_0 : \pi_1 - \pi_2 = \Delta$$
$$H_A : \begin{cases} \pi_1 - \pi_2 \neq \Delta & (\text{Two-sided}) \\ \pi_1 - \pi_2 > \Delta & (\text{Upper}) \\ \pi_1 - \pi_2 < \Delta & (\text{Lower}) \end{cases}$$

とする $(\Delta \leq 0)$．優越性試験の場合は $\Delta = 0$ となる．このとき，FREQ プロシジャを用いてプログラム 1 を実行させることにより，有効割合の差の信頼区間を出力できる．複数の種類の構成法に基づく信頼区間を同時に出力したい場合は，CL = (*type*) と指定する．例えば，Wald 信頼区間，Miettinen-Nurminen 信頼区間，Newcombe スコア信頼区間を同時に出力したい場合は，CL = (WALD MN NEWCOMBE) と指定する．

プログラム 1 : FREQ プロシジャによる割合の差の信頼区間の指定

```
proc freq data=data;
  tables x*y / riskdiff(cl = type);
run;
```

RISKDIFF オプションで出力されるデフォルトは Wald 信頼区間である．しかし，Wald 信頼区間は被覆確率が保たれない場合があることが指摘されている [14, 15]．V9.4 (SAS/STAT 13.1) の FREQ プロシジャにおける TABLES ステートメントの RISKDIFF オプションでは，計 11 種類の信頼区間を出力できる [12, 15]．これらの信頼区間のうち，被覆確率及び検出力の観点から，Miettinen-Nurminen 信頼区間と Newcombe スコア信頼区間の使用が推奨されている [15]．

3 信頼区間の構成法の数理

本章では，正規分布への漸近近似を利用した割合の差の信頼区間のうち，Wald 信頼区間，Agresti-Caffo 信頼区間，Hauck-Anderson 信頼区間，Farrignton-Manning 信頼区間，Mee 信頼区間，Miettinen-Nurminen 信頼区間の数理を示す．その他の信頼区間については飯塚，魚住，浜田 (2014) を参照されたい [15]．

3.1 Wald 信頼区間 [2]

Wald 信頼区間は，CL = WALD と指定すると出力させることができ，$\pi_1 - \pi_2$ の漸近正規性より以下のように構成される．

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_{1\cdot}} + \frac{p_2(1-p_2)}{n_{2\cdot}}}$$

ただし，z_κ を標準正規分布の上側 κ %点とする．

3.2 Agresti-Caffo 信頼区間 [1]

Agresti-Caffo 信頼区間は，CL = AC あるいは CL = AGRESTICAFFO と指定すると出力させることができ，各セルに 1 度数足し，Wald 信頼区間を導いた方法である．

$$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1^*(1-p_1^*)}{n_{1\cdot}+2} + \frac{p_2^*(1-p_2^*)}{n_{2\cdot}+2}}$$

ただし，

$$p_1^* = \frac{n_{11}+1}{n_{1\cdot}+2}, \qquad p_2^* = \frac{n_{21}+1}{n_{2\cdot}+2}$$

である．JMP(R) 11 (SAS Institute Inc., Cary, NC, USA) で割合の 2 標本検定を実施した場合，出力される割合の差の信頼区間は Agresti-Caffo 信頼区間である．

3.3 Hauck-Anderson 信頼区間 [8]

Hauck-Anderson 信頼区間は，CL = HA と指定すると出力させることができ，Wald 信頼区間より分散を大きくした上で，連続修正項を加えた方法である．

$$(p_1 - p_2) \pm \left(\mathrm{CC} + z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1\cdot}-1} + \frac{p_2(1-p_2)}{n_{2\cdot}-1}} \right)$$

ただし，

$$\mathrm{CC} = \frac{1}{2\cdot\min(n_{1\cdot}, n_{2\cdot})}$$

である．

3.4 Farrignton-Manning 信頼区間 [6]

Farrignton-Manning 信頼区間は，CL = FM と指定すると出力させることができ，スコア型の信頼区間として以下のように構成される．

$$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}}$$

$\tilde{p}_1$ と $\tilde{p}_2$ は帰無仮説の下での制限付き最尤推定値であり，以下の 3 次方程式を解くことによって得られる．

$$\sum_{k=0}^{3} L_k \tilde{p}_1^k = 0$$

ここで，$L_3 = N$，$L_2 = (n_{2\cdot} + 2n_{1\cdot})\Delta - N - n_{11} - n_{21}$，$L_1 = (n_{2\cdot} + 2n_{1\cdot})\Delta - N - n_{11} - n_{21}$，$L_0 = n_{11}\Delta(1-\Delta)$ であり，

$$\tilde{p}_1 = 2u\cos(w) - b/3a$$
$$\tilde{p}_2 = \tilde{p}_1 - \Delta$$

と求まる．ただし，

$$w = (\pi + \cos^{-1}(v/u^3))/3$$
$$v = b^3/(3a)^3 - bc/6a^2 + d/2a$$
$$u = \mathrm{sign}(v)\sqrt{b^2/(3a)^2 - c/3a}$$

$$a = 1 + n_{2\cdot}/n_{1\cdot}$$
$$b = -(1 + n_{2\cdot}/n_{1\cdot} + p_1 + (n_{2\cdot}/n_{1\cdot})p_2 + \Delta(n_{2\cdot}/n_{1\cdot} + 2))$$
$$c = \Delta^2 + \Delta(2p_1 + n_{2\cdot}/n_{1\cdot} + 1) + p_1 + (n_{2\cdot}/n_{1\cdot})p_2$$
$$d = -p_1\Delta(1+\Delta)$$

である．

なお，Δ は非劣性マージンを表し，CL = FM (NULL = – Δ) と指定する．例えば，$\Delta = -0.05$ として，$H_0 : \pi_1 - \pi_2 \le \Delta$，$H_A : \pi_1 - \pi_2 > \Delta$ を考える場合，CL = FM (NULL = 0.05) と指定する．優越性試験の場合は $\Delta = 0$ となる．$\Delta = 0$ のとき，$\tilde{p}_1 = \tilde{p}_2 = p_T$ となり，このとき Farrignton-Manning 信頼区間は

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_T(1-p_T)}{n_{1\cdot}} + \frac{p_T(1-p_T)}{n_{2\cdot}}}$$

となる．デフォルトは $\Delta = 0$ であり，CL = FM (NULL = 0) に対応する．

3.5 Mee 信頼区間 [9]

Mee 信頼区間は，CL = MN (CORRECT = NO) あるいは CL = MN (MEE) と指定すると出力され，アウトプット画面で Miettinen-Nurminen-Mee と表示される．スコア型の検定統計量として，

$$T_{\mathrm{Mee}} = \frac{p_1 - p_2 - \Delta}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}}}$$

を用いると，Mee 信頼区間は

下限 ： $T_{\mathrm{Mee}} = -z_{\alpha/2}$

上限 ： $T_{\mathrm{Mee}} = z_{\alpha/2}$

を解くことによって，信頼下限，信頼上限がそれぞれ得られる．下限，上限でそれぞれ算出されるため，Farrignton-Manning 信頼区間と異なり，左右非対称のスコア型信頼区間となる．

3.6 Miettinen-Nurminen 信頼区間 [10]

Miettinen-Nurminen 信頼区間は，CL = MN と指定すると出力させることができる．Mee 信頼区間構成のときに考えたスコア統計量 T_{Mee} の分散成分にバイアス補正項を加えた統計量

$$T_{\mathrm{MH}} = \frac{p_1 - p_2 - \Delta}{\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}}} \sqrt{\frac{N-1}{N}}$$

を用いると，Miettinen-Nurminen 信頼区間は

下限 ： $T_{\mathrm{MH}} = -z_{\alpha/2}$

上限 ： $T_{\mathrm{MH}} = z_{\alpha/2}$

を解くことによって，下限，上限がそれぞれ得られる．Mee 信頼区間と同様に，左右非対称の信頼区間である．分散成分のバイアス補正項の影響により，例数が少ない場合は Mee 信頼区間よりも広くなる．

4 例数設計法の数理

本章では，SAS/STAT 13.1 における POWER プロシジャの TWOSAMPLEFREQ ステートメントで利用できる例数設計法を取りあげる. なお, 現在の POWER プロシジャでは信頼区間に基づく例数設計は実施できず, 対応する検定方法を指定して例数設計を行うことになる．POWER プロシジャの ONESAMPLEFREQ ステートメントによる例数設計では，CI オプションがあり，FREQ プロシジャの BINOMIAL オプションで指定可能な信頼区間の構成法を指定することができる．ただし，CI オプションでは信頼区間に基づく精度ベースの例数設計を行うだけであり，これは検出力を 50%として例数設計を行うことに対応する．

SAS/STAT 13.1 の POWER プロシジャにおける TWOSAMPLEFREQ ステートメントで指定できる検定方法は，カイ二乗検定，尤度比検定，Fisher の正確検定である．本稿では，カイ二乗検定に基づく例数設計法を示し，加えて Farrignton-Manning によるスコア検定に基づく例数設計法を示す．

4.1 カイ二乗検定に基づく例数設計 [7]

Pearson のカイ二乗検定の検定統計量は以下のように構成される．

$$T_{\mathrm{P}} = \frac{(p_1 - p_2) - \Delta}{\sqrt{p_T(1-p_T)\left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}\right)}} = \sqrt{Nw_1w_2}\,\frac{(p_1 - p_2) - \Delta}{\sqrt{p_T(1-p_T)}}$$

ただし， w_i は治療 i における例数の割合 $(w_i = n_{i\cdot}/n_{\cdot\cdot})$ であり，POWER プロシジャにおける GROUPWEIGHTS オプションで指定する値に該当する．このとき，検出力$1-\beta$は以下の式

$$1-\beta = \begin{cases} \Phi\left(\dfrac{\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_\alpha\sqrt{p_T(1-p_T)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) & \text{(Upper)} \\ \Phi\left(\dfrac{-\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_\alpha\sqrt{p_T(1-p_T)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) & \text{(Lower)} \end{cases}$$

で与えられ，両側検定の場合は

$$1-\beta = \Phi\left(\frac{\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_{\alpha/2}\sqrt{p_T(1-p_T)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) + \Phi\left(\frac{-\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_{\alpha/2}\sqrt{p_T(1-p_T)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) \quad \text{(Two-sided)}$$

で与えられる．以上より，片側検定の場合，必要例数は以下の式で算出することができる．

$$N = \frac{\{z_\alpha\sqrt{p_T(1-p_T)} + z_\beta\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}\}^2}{w_1w_2\{(p_1 - p_2) - \Delta\}^2} \quad (1)$$

プログラム 2 : POWER プロシジャによるカイ二乗検定に基づく例数設計

```
proc power;
   twosamplefreq test=pchi
      groupproportions = (p1 p2)
      groupweights=(w1 w2)
      nullproportiondiff = Δ
      sides = u
      alpha = 0.025
      ntotal = .
      power = 1-β;
run;
```

$H_0 : \pi_1 - \pi_2 \le \Delta$，$H_A : \pi_1 - \pi_2 > \Delta$とした場合，POWER プロシジャの TWOSAMPLEFREQ ステートメントにおいて，プログラム 2 でカイ二乗検定に基づく例数設計を行うことができる．TEST = PCHI の代わりに，TEST = LRCHI と指定すれば尤度比検定，TEST = FISHER と指定すれば Fisher の正確検定に基づく例数設計も行うことが可能である [12].

4.2 Farrington-Manning 検定に基づく例数設計 [6]

Farrington-Manning のスコア検定の検定統計量は以下のように構成される．

$$T_{\mathrm{FM}} = \frac{(p_1 - p_2) - \Delta}{\sqrt{\dfrac{\tilde{p}_1(1-\tilde{p}_1)}{n_{1\cdot}} + \dfrac{\tilde{p}_2(1-\tilde{p}_2)}{n_{2\cdot}}}} = \sqrt{Nw_1w_2}\,\frac{(p_1 - p_2) - \Delta}{\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)}}$$

3.5 節で述べた Mee 信頼区間の構成のために用いた検定統計量T_{Mee}と同様である．このとき，検出力$1-\beta$は以下の式

$$1-\beta = \begin{cases} \Phi\left(\dfrac{\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_\alpha\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) & \text{(Upper)} \\ \Phi\left(\dfrac{-\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_\alpha\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) & \text{(Lower)} \end{cases}$$

で与えられ，両側検定の場合は

$$\begin{aligned} 1-\beta = {} & \Phi\left(\frac{\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_{\alpha/2}\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) \\ & + \Phi\left(\frac{-\{(p_1 - p_2) - \Delta\}\sqrt{Nw_1w_2} - z_{\alpha/2}\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)}}{\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}}\right) \end{aligned} \quad \text{(Two-sided)}$$

で与えられる．以上より，片側検定の場合，必要例数は以下の式で算出することができる．

$$N = \frac{\{z_\alpha\sqrt{w_2\tilde{p}_1(1-\tilde{p}_1) + w_1\tilde{p}_2(1-\tilde{p}_2)} + z_\beta\sqrt{w_2p_1(1-p_1) + w_1p_2(1-p_2)}\}^2}{w_1w_2\{(p_1 - p_2) - \Delta\}^2} \quad (2)$$

なお，$\tilde{p}_1$と$\tilde{p}_2$は 3.4 節に示した通りである．

3.4 節で述べたように，優越性試験の場合は$\Delta = 0$となる．このとき$\tilde{p}_1 = \tilde{p}_2 = p_T$であるため，(2) 式は (1) 式に帰着し，カイ二乗検定に基づき求めた必要例数と一致する．

Farrington-Manning 検定に基づく例数設計を行う場合，SAS/STAT 13.1 までは POWER プロシジャでサポートされていなかったため，擬似乱数を用いたシミュレーションによる評価に頼らなければならなかった．しかし，SAS/STAT 13.2 の POWER プロシジャの機能拡張により，Farrington-Manning 検定に基づく例数設計も行うことができるようになった [4, 13].

5 数値例

実際の臨床研究の数値例として，皮膚感染症に対する非劣性試験を取りあげる [3]．試験群と対照群を対象とした 2 群間のランダム化比較試験であり，有意水準は片側 2.5%，非劣性マージンは 10%として，主解析方法に Miettinen-Nurminen 信頼区間の信頼下限が –10% を下回らないことで非劣性が検証されるように解析が計画された．計画時に想定した有効割合は 85%であり，検出力の名義水準を 90%として，必要例数は 556 例

と見積もられた．統計解析計画として，必要例数は Farrignton-Manning のスコア検定に基づき求めたことが述べられており，(2) 式より必要例数を見積もると 552 例と算出される．

本臨床試験から得られた結果を表 2 に示す．主要評価項目は 3 つ設定されており，研究 1 及び研究 2 から得られるそれぞれのデータを対象とした解析と 2 つの研究の併合データを対象とした解析の計 3 つが主解析と計画された．差の信頼区間として，いずれも Miettinen-Nurminen 信頼区間を用いて報告されている．表 2 より，信頼下限が –10% を上回っているため，すべての主要評価項目に対して非劣性が検証された．

表 2：非劣性試験から得られた結果 [3]

研究	試験群 $n_{11}/n_{1\cdot}(p_1)$	対照群 $n_{21}/n_{2\cdot}(p_2)$	差 (95% CI)
1	240/288 (83.3)	233/285 (81.8)	1.5 (–4.6, 7.9)
2	285/371 (76.8)	288/368 (78.3)	–1.5 (–7.4, 4.6)
1 + 2	525/659 (79.7)	521/653 (79.8)	–0.1 (–4.5, 4.2)

ここで，POWER プロシジャによる例数設計を考える．$H_0:\pi_1-\pi_2\le\Delta$，$H_A:\pi_1-\pi_2>\Delta$ とした場合，プログラム 3 を実行することによって必要例数が算出されるが，536 例と見積もられてしまい，本臨床試験の論文で報告されている必要例数よりも 20 例少なく算出されてしまう．見積もられた 536 例を必要例数として，擬似乱数を用いたシミュレーションによる評価を行うと，Miettinen-Nurminen 信頼区間が主解析の場合，検出力は名義水準 90%を満たさないことを確認できる．

プログラム 3：POWER プロシジャによるカイ二乗検定に基づく例数設計の数値例

```
proc power;
   twosamplefreq test=pchi
      groupproportions = (.85 .85)
      groupweights=(1 1)
      nullproportiondiff = -0.10
      sides = u
      alpha = 0.025
      ntotal = .
      power = 0.9;
run;
```

SAS/STAT 13.2 からの POWER プロシジャでは，TWOSAMPLEFREQ ステートメントで指定できる検定方法として，カイ二乗検定，尤度比検定，Fisher の正確検定の他に，Farrignton-Manning のスコア検定も指定できる [13]．Farrignton-Manning のスコア検定に基づく例数設計を行うためには，プログラム 4 のように，TEST = FM と指定する [4]．

Mee 信頼区間の下限と上限は片側検定の棄却域から算出され，Mee 信頼区間の構成のために用いるスコア検定統計量 T_{Mee} は Farrignton-Manning 検定のスコア検定統計量 T_{FM} と等しい．Miettinen-Nurminen 信頼区間の

構成のために用いるスコア統計量T_{MN}は，T_{Mee}の分散成分にバイアス補正項を加えたスコア統計量であるため，例数が少なくなければT_{MN}とT_{Mee}はほぼ一致する．したがって，本臨床試験のように例数の多い試験を考える場合は，主解析方法に Miettinen-Nurminen 信頼区間を計画した場合の例数設計法として，Farrignton-Manning のスコア検定に基づく例数設計法が推奨される．

プログラム 4 : POWER プロシジャによる Farrignton-Manning のスコア検定*に基づく例数設計の数値例

```
proc power;
   twosamplefreq test=fm
      grouppproportions = (.85 .85)
      groupweights=(1 1)
      nullproportiondiff = -0.10
      sides = u
      alpha = 0.025
      ntotal = .
      power = 0.9;
run;
```

*SAS/STAT 13.2 より実行可能

6　まとめ

本稿では，V9.4 (SAS/STAT 13.1) の FREQ プロシジャで算出可能な割合の差の信頼区間のうち，正規分布への漸近近似を利用した信頼区間の一部を取りあげた．POWER プロシジャで割合の差の信頼区間を主解析として例数設計を行う場合，カイ二乗検定に基づく例数設計のみ利用可能であり，飯塚, 魚住, 浜田 (2014) により推奨された Miettinen-Nurminen 信頼区間を主解析とした場合の例数設計を行うためにはシミュレーションによる評価が避けられなかった．優越性試験の例数設計の場合，カイ二乗検定に基づき算出する例数と Farrignton-Manning のスコア検定に基づく例数設計の結果は一致する．しかし，非劣性試験の例数設計の場合，カイ二乗検定と Farrignton-Manning のスコア検定に基づき算出する例数では結果が異なる．これは，カイ二乗検定統計量の分母の分散成分が帰無仮説の下での制限付き最尤推定値の分散となっていないためであり，非劣性試験の例数設計の場合，カイ二乗検定に基づく例数設計では適切な例数を見積もることができない．数値例として，非劣性試験の事例による評価を行った結果，Miettinen-Nurminen 信頼区間を主解析とした場合にカイ二乗検定に基づく例数設計を適用すると，例数が過小評価されて見積もられてしまうことを示した．そこで，例数が少なくない場合に Miettinen-Nurminen 信頼区間に近い結果となる Farrignton-Manning のスコア検定に基づく例数設計法を紹介した．Farrignton-Manning のスコア検定に基づく例数設計は，SAS/STAT 13.2 からの POWER プロシジャで実行可能である [4, 13]．今後の POWER プロシジャの機能拡張により，Farrignton-Manning のスコア検定以外の方法に基づく例数設計も実施できるようになることが期待される．

なお，本稿では例数が少なくない場合の検討を行った．例数が少ない場合，割合の差の正確な信頼区間の性能が良いことも報告されている [5, 15, 18]．割合の差の非劣性検定については，武藤，宮島，榊原 (2014) によって FREQ プロシジャでサポートされていない方法のプログラム及び性能評価の報告が行われており [18]，今後 FREQ プロシジャの EXACT ステートメントにおける機能拡張として追加されることが期待される．

参考文献

[1] Agresti A, Caffo B. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*. **54:**280–288, 2000.

[2] Altman DG, Machin D, Bryant TN, et al. *Statistics with confidence* (2nd edn). London: BMJ Books, 2000.

[3] Boucher HW, Wilcox M, Talbot GH, et al. Once-weekly dalbavancin versus daily conventional therapy for skin infection. *New England Journal of Medicine*. **370**:2169–2179, 2014.

[4] Castelloe J, Watts D. Equivalence and Noninferiority Testing Using SAS/STAT® Software. *Proceedings of the SAS Global Forum*. Cary, NC: SAS Institute Inc., 2015. Available at http://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf.

[5] Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*. **24**:224–254, 2015.

[6] Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*. **9**:1447–1454, 1990.

[7] Fleiss JL, Tytun A, Ury SHK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*. **36**:343–346, 1980.

[8] Hauck WW, Anderson S. A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *The American Statistician*. **40**:318–322, 1986.

[9] Mee RW. Confidence bounds for the difference between two probabilities. *Biometrics*. **40**:1175–1176, 1984.

[10] Miettinen OS, Nurminen. Comparative analysis of two rates. *Statistics in Medicine*. **4**:213–226, 1985.

[11] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*. **17**:873–890, 1998.

[12] SAS Institute Inc. *SAS/STAT(R) 13.1 User's Guide*, Cary, NC, USA: SAS Institute Inc; 2013.

[13] SAS Institute Inc. *SAS/STAT(R) 13.2 User's Guide*, Cary, NC, USA: SAS Institute Inc; 2014.

[14] 飯塚政人, 浜田知久馬. 2群の割合の差における信頼区間の構成法の比較. SASユーザー総会 論文集 2013, 461–473.

[15] 飯塚政人, 魚住龍史, 浜田知久馬. FREQプロシジャによる割合の差の信頼区間 –V9.4における機能拡張と性能評価–. SAS ユーザー総会 論文集 2014, 527–538.

[16] 魚住龍史, 浜田知久馬. RAND 関数による擬似乱数の生成. SAS ユーザー総会 論文集 2013, 325–333.

[17] 浜田知久馬. SAS による 2 値データの解析「ここまでできる FREQ プロシジャ V.9.3」. SAS ユーザー総会 論文集 2012, 3–56.

[18] 武藤彬正, 宮島育哉, 榊原伊織. SASによる二項比率の差の非劣性検定の比較. SASユーザー総会 論文集 2014, 464–473.

連絡先

E-mail : *uozumi@kuhp.kyoto-u.ac.jp*