# 2つの比率の差の同等性検定の正確な評価

M2011MM049 水野裕太
指導教員：松田眞一

## 1　はじめに

医薬品の開発にあたり，いくつかの局面で同等性試験を行う必要が生じることがある。同等性試験では積極的に2つの薬剤が同等であることを示すことが要求され，検定で棄却されないことをもって同等性を主張することはできない。そのような関係で同等性検定と呼ばれる統計的方法論が発展してきた。

同等性検定の方法は大きく分けると2つある。それはハンディキャップ（非劣性マージン）方式の検定を用いた方法と信頼区間を用いた方法である。本研究ではそのうち信頼区間を用いた方法についてのみ考察する。

同等性検定を行う方法を調べてみると，平均値の差を利用している方法ばかり紹介されていることが分かった。比率の差を用いた方法もいくつかあったが，正規近似を用いているものばかりで，検定結果と信頼区間が正確に対応している方法を見つけることができなかった。

そこで，水野 [6] では，2つの比率の差の信頼区間を用いて同等性を判定する場合にはどのようにすればよいのかという疑問を持ち，比率の差の同等性検定の研究を進め，検定結果と信頼区間が正確に対応している方法を提案し，シミュレーションを行うことで検出力の評価を行った。

さらに，本論文では，その続きとしてシミュレーションによらない正確な計算の下で，正確な検出力の評価を行い，最終的には正確に検出力が0.8を超える例数を出力する関数を作成することを目的としている。

## 2　同等性検定の方法

比率の差の同等性検定の方法はいくつかの文献（Cesana[1], 広津 [3], Kang and Chen[4], Phillips[7], 丹後 [8], Tango[9]）で提案されている。しかし，これらの方法では，3つの問題点が指摘される。1つ目は，比率の差の検定とその比率の差の信頼区間の対応関係が正確にはないということ，2つ目は，正規近似を用いた比率の差の信頼区間は無意味な範囲になってしまう時があること，そして，3つ目は，比率の差の信頼区間を用いる場合，どのような判定区間の幅（ハンディキャップ $\Delta$）にするのか検討をする必要があるということである。この節では，その3つの問題点を解決した水野 [6] で提案した同等性検定の方法について紹介する。さらに，新たに考えた同等性判定の基準についても述べていく。

### 2.1　Fisher の正確確率検定に対応する比率の差の信頼区間

#### 2.1.1　導入

先ほど示した問題点を解決する為に，松田 [5] の方法を考えていく。この方法は，オッズ比の信頼区間から比率の差の信頼区間を求めるものであり，Fisher の正確確率検定に対応する比率の差の信頼区間を求めることができる。また，この方法は正規近似を用いておらず，正確な比率の差の信頼区間を求めることができる。すなわち，この方法を用いることで前節で挙げた3つの問題点の内，2つが解決できる。

ここで，**表**1のように $2\times 2$ の分割表の記号を導入する。

表 1　分割表の記号法

| | 有効 | 無効 | 計 | 真の成功率 |
|---|---|---|---|---|
| 薬剤 A | $x_A$ | $n_A - x_A$ | $n_A$ | $p_A$ |
| 薬剤 B | $x_B$ | $n_B - x_B$ | $n_B$ | $p_B$ |
| 計 | $k$ | $N-k$ | $N$ | |

また，$x_A$〜B($n_A$, $p_A$), $x_B$〜B($n_A$, $p_B$) とそれぞれ二項分布に従っている二項分布モデルとする。

#### 2.1.2　比率の差の信頼区間の構成法

オッズ比の信頼区間から比率の差の信頼区間を求める方法は松田 [5] で以下の2つの方法が示されている。

1. 周辺度数を真値であるかのように扱い，比率に関する次の条件式を設定し，それを基にオッズ比の信頼区間から変換する方法。

$$n_A p_A + n_B p_B = k$$

(この式は「2つの群を合わせた場合の真の比率は $k/N$ である」ということを定式化したものである。)

2. オッズ比に対応する比率の組 $(p_A, p_B)$ の中で「2項分布モデルにおいて周辺度数が得られる確率」が最大となるものを求める方法でオッズ比の信頼区間から変換する。

前者の方法は，信頼区間を求める方法が簡便であるが，安易な方法であるという印象を受けてしまう。それに対し，後者の方法は最尤法に似た考え方であり受け入れやすいが求め方が複雑となる。しかし，この二つの方法から求められる信頼区間は一致することが，松田 [5] で証明されているため，求め方は前者の方法を，意味は受け入れやすい後者の方法をと考えることができるので，これらの方法から比率の差の信頼区間を求めていく。

### 2.2　オッズ比の信頼区間

オッズ比の信頼区間について述べていく。**表**1のデータを例とすると，オッズ比は

$$\theta = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}$$

と示すことができる。

このオッズ比の信頼区間の求め方を Fagerland et al.[2] を参照し述べていく。**表**1の全ての周辺合計が固定され

るという条件を，有効の合計 $k$ と無効の合計 $N-k$ の数値に付けると，有効数 $y_A$ の観測確率は非心超幾何分布に従い，

$$f(y_A|\theta) = \frac{\left(\begin{array}{c} k \\ y_A \end{array}\right)\left(\begin{array}{c} N-k \\ n_A-y_A \end{array}\right)\theta^{y_A}}{\sum_{i=n_{max}}^{n_{min}}\left(\begin{array}{c} k \\ i \end{array}\right)\left(\begin{array}{c} N-k \\ n_A-i \end{array}\right)\theta^{i}}$$

と表すことができる。ここで，$n_{max} = max(0, k-n_B)$，$n_{min} = min(n_A, k)$ である。片側 $\times 2$ 検定[1]による，$\theta$ の正確な条件付き信頼区間 $(L, U)$ は

$$\sum_{y_A=x_A}^{n_{min}} f(y_A|L) = \alpha/2$$

$$\sum_{y_A=n_{max}}^{x_A} f(y_A|U) = \alpha/2$$

から求めることができる[2]。なお，この信頼区間は，Fisher の正確確率検定と常に対応している[3]。

### 2.3 同等判定の基準

前節で述べたように，同等性を判定する基準を新たに考える必要がある。新たに 2 通りの判定基準を考えた。1 つ目は，判定区間を $[-0.1, 0.1]$ と固定して，この範囲内に入れば同等であると判定する方法である。2 つ目は，1 つ目の固定区間の方法の欠点を補うため二項分布の標準偏差を考慮して，判定区間を $[-\sqrt{p(1-p)}/5, \sqrt{p(1-p)}/5]$ と可変区間にし，この範囲内に入れば同等であると判定する方法である。

先に述べたように固定区間に用いた 0.1 という値は一般的に比率の差の同等性検定でハンディキャップとして用いられる値であり，$p = 0.5$ のときの幅 0.1 は全体の 2 割に当たるため平均値の差の検定の場合と概ね対応するものと考えられている。また，可変区間の幅は $p = 0.5$ のときに固定区間と同じ 0.1 の幅となるように調整したものである。

### 2.4 提案手法の評価と改良

この 2 つの判定基準を基にシミュレーションを行った結果（シミュレーション方法と細かい結果は水野 [6] で述べている），真の成功率 $p$ に関係なく例数設定ができるためには，判定区間を可変区間にすると良いことが分かった。また，この判定区間を利用すると検出力を 0.8 以上に保つためには，例数を約 500 以上に設定する必要があることも分かった。しかし，例数を約 500 以上に設定しなければいけないということは実用的ではないので，必要例数をもっと小さくする課題が残った。

この課題を少しでも改善するために，1 つの対策を考える。それは，ハンディキャップを単純に大きくして同等性判定を行うというものである。一般的に与えられている $\Delta = 0.1$ は非劣性検証を目的としたものと考えると，積極的に同等を検証したい場合は新たなハンディキャップを模索する必要があると考える。そこで，$\Delta = 0.15$ とした場合について考える。先と同様の可変区間を設定する場合はハンディキャップを $\Delta$ とした場合，

$$[-2\sqrt{p(1-p)}\Delta, 2\sqrt{p(1-p)}\Delta]$$

という区間を設定することとなる。

## 3 正確な計算による検出力の評価

$\Delta = 0.15$ として水野 [6] と同様のシミュレーションを行なったところ各群の例数が約 200 より大きくなると検出力が約 0.8 を超えることが分かった。このことから，$\Delta = 0.15$ とすることで必要例数を少なくすることができることは分かったので，かなり膨大な例数を必要とするという問題点を解消するための一つの方法と考えられるが，このハンディキャップ $\Delta$ の値を 0.15 としても問題ないのかということを考えると疑問が残る。

しかし，どのような判定基準にすれば問題ないのかということを勝手に決めることはできないので，自由に真の成功率 $p$ とハンディキャップ $\Delta$，そして，信頼率 $\alpha$ を与えることで，正確に検出力 0.8 を超える必要最低例数を求めるプログラムを作ることを目標にした。

そこで，まずシミュレーションによらない正確な計算方法を確立するために，真の成功率 $p$ とハンディキャップ $\Delta$，そして，信頼率 $\alpha$ を与えることで，自由に検出力を求めることができるプログラムを作成する。この節では，このプログラムについて述べていく。

また，これ以後で同等性検定を行っている場合，判定基準は $[-2\sqrt{p(1-p)}\Delta, 2\sqrt{p(1-p)}\Delta]$ を用いている。

### 3.1 実装した関数について

正確な計算による検出力は，判定区間に入る分割表のみを取り出し，それらの分割表になる確率をそれぞれ求め，それらを全て足すことで求めることができる。しかし，この求め方でそのままプログラムを作成すると，かなり時間がかかると考えられるので，無駄な計算をできるだけ省き，より高速にプログラムを実行できるように工夫する必要がある。

そこで考えられる対策が 2 つある。1 つ目は，判定区間に入り同等であるとされた分割表の確率を求める時に，1 つずつ求めていくのではなく，累積確率を求める関数を使うことで，確率の計算量をより少なくするというものである。

2 つ目は，与えられる全ての分割表で同等性の判別を行うのではなく，判定区間に入る分割表とそれらの隣にある判定区間に入らない分割表のみ同等性判定を行い，無駄に同等性判定を行わないようにする方法である。

---

[1] 片側 $\times 2$ 検定に対応する信頼区間の構成方法とは，信頼区間の外側に出る確率 $\alpha$ を上側と下側で $\alpha/2$ ずつで考えることによって算出する方法のことである。非対称な分布ではいわゆる両側検定の信頼区間とは異なることに注意する。

[2] この信頼区間は Fagerland et al.[2] で Cornfield exact conditional として紹介されている方法で条件付き分布の下で正確な信頼区間を与える方法である。

[3] Fisher の正確確率検定との対応も両側検定ではなく片側検定の $p$ 値を 2 倍したものと対応することになる。R の fisher.test 関数での $p$ 値は前者であるのでそのままでは対応しないことに注意する。

この 2 つの方法を組み合わせてプログラムを作成したところ，この 2 つの方法を使っていない普通のプログラムよりもおよそ 10 倍くらい速く実行することができた。

### 3.2 シミュレーションとの比較

以下に示した**表** 2，**表** 3 が正確な計算から求めた検出力の結果である。

表 2 $\Delta = 0.1$，$\alpha = 0.9$ のときの検出力

| 各群の例数 | $p$=0.5 | $p$=0.4 | $p$=0.3 | $p$=0.2 |
|---|---|---|---|---|
| 100 | 0 | 0 | 0 | 0 |
| 200 | 0.197 | 0.203 | 0.210 | 0.206 |
| 300 | 0.528 | 0.529 | 0.537 | 0.533 |
| 400 | 0.727 | 0.737 | 0.738 | 0.738 |
| 500 | 0.862 | 0.854 | 0.856 | 0.856 |
| 600 | 0.921 | 0.924 | 0.923 | 0.923 |
| 700 | 0.960 | 0.959 | 0.960 | 0.960 |

表 3 $\Delta = 0.15$，$\alpha = 0.9$ のときの検出力

| 各群の例数 | $p$=0.5 | $p$=0.4 | $p$=0.3 | $p$=0.2 |
|---|---|---|---|---|
| 100 | 0.276 | 0.283 | 0.291 | 0.296 |
| 200 | 0.788 | 0.798 | 0.798 | 0.802 |
| 300 | 0.954 | 0.950 | 0.951 | 0.952 |
| 400 | 0.990 | 0.989 | 0.989 | 0.989 |

**表** 2，**表** 3 から，$\Delta = 0.1$ と $\Delta = 0.15$ のどちらも，シミュレーション結果とほぼ同じ結果を得ることができたので，正確な計算から検出力を求める関数の計算に問題は無いと考えられる。

## 4 例数設計について

### 4.1 問題点

前節で述べたように正確な計算から検出力を求める関数を作成することができたので，この関数を用いて正確に検出力 0.8 を超える必要最低例数を求める関数を作成する。しかし，この関数を作成する時に大きな問題点がある。例数 $n$ を大きくすると検出力は大きくなることは，シミュレーション結果や正確な計算による検出力の結果から明らかである。しかし，**表** 4 のように，例数 $n$ を 1 つずつ増やして検出力を出力すると，単調増加でないことが分かる。

この問題点を考慮した上で正確に検出力 0.8 を超える必要最低例数を求めなければならないので，検出力 0.8 を超えた最初の例数が正確に検出力 0.8 を超える必要最低例数ということはできない。さらにその先の検出力も求め，次に検出力の数値が最も小さくなる例数の検出力が 0.8 を超えているかどうかまで調べなければならない。もし，次に検出力の数値が最も小さくなる例数の検出力が 0.8 を超えていなければ，次に検出力が 0.8 を超える例数をその先で見つけて同じことを繰り返して，正確に検出力 0.8 を超える必要最低例数を見つけなければならない。

表 4 $p = 0.5$，$\Delta = 0.1$, $\alpha = 0.9$ の検出力

| n | 検出力 | n | 検出力 | n | 検出力 |
|---|---|---|---|---|---|
| 440 | 0.7877 | 447 | 0.8079 | 454 | 0.8044 |
| 441 | 0.7872 | 448 | 0.8074 | 455 | 0.8039 |
| 442 | 0.7868 | 449 | 0.8069 | 456 | 0.8037 |
| 443 | 0.7890 | 450 | 0.8064 | 457 | 0.8093 |
| 444 | 0.8094 | 451 | 0.8059 | 458 | 0.8245 |
| 445 | 0.8089 | 452 | 0.8054 | 459 | 0.8240 |
| 446 | 0.8084 | 453 | 0.8049 | 460 | 0.8235 |

この問題点から，検出力を求める関数を何度も実行しなければならないので，できるだけ速く実行できるように考慮する必要がある。

### 4.2 必要例数の求め方の提案

先に述べた問題点をふまえた上で探索的に必要例数を求めなければならないので，速く関数を実行させるためには最初の基点となる例数を求めたい必要最低例数にできるだけ近づけたい。

そこで 1 つの方法を提案する。それは $p$ と $\Delta$，そして，$\alpha$ にある一定の条件を設定し，その条件に当てはまる組み合わせから得られる必要最低例数を求め，その結果を関数にあらかじめ代入しておく方法である。この方法は，ある一定の条件に当てはまらない組み合わせのみ，あらかじめ求めた必要例数を基に線形補間などを用いて，基点となる例数を近似的に求め，その基点から求めたい必要例数を探索的に求めることになる。また，ある一定の条件を満たす $p$ と $\Delta$ と $\alpha$ の 3 つの組み合わせであれば，あらかじめ計算して関数に代入されているのですぐに結果を出力することができるという利点がある。

しかし，ある一定の条件を満たす全ての $p$ と $\Delta$ と $\alpha$ の 3 つの組み合わせから求められる例数を 1 つずつ求めなければならないので，関数を実装するのが大変になってしまうという問題点が挙げられるが，この方法が最も速く関数を実行できると考えられるで，この方法で関数を作成していく。

### 4.3 条件の設定について

関数を作成する前に設定しなければならない条件が 2 つある。まず 1 つ目は与えられる $p$ と $\Delta$ の数値の範囲をどのようにするかというものである。できる限り広い範囲で計算可能としたいが，先に述べたように，ある一定の条件を満たす $p$ と $\Delta$ の組み合わせから得られる例数を全て 1 つずつ求めなければならないので限度がある。そこで，$0.1 \leqq p \leqq 0.9$ と $0.1 \leqq \Delta \leqq 0.2$ とし，どちらの値もある程度広めに設定することとする。また，信頼率 $\alpha$ については，$0 < \alpha < 1$ の範囲で実行可能とする。

2 つ目は先程ある一定の条件として述べたあらかじめ求めておく $p$ と $\Delta$ と $\alpha$ の 3 つの組み合わせから得られる例数をどこまで細かく求めるかを設定しなければならないというものである。この条件も細かくしすぎるとあらかじめ求めておく必要例数の数が多くなりすぎてしまい，

求めきれなくなってしまう。しかし，刻み幅が大きくなりすぎてしまうと条件を満たしていない $p$ と $\Delta$ と $\alpha$ の組み合わせから得られる必要例数を求めるための基点の精度が悪くなってしまう。そこで，$p$ の数値による必要例数への影響が小さいことから，$p$ を 0.1 刻みとして少し大きめに設定し，必要例数への影響が大きい $\Delta$ を 0.005 刻みとして，かなり細かく設定することとする。そして，信頼率 $\alpha$ は 0.8 と 0.9 の 2 つのみとし，この条件に当てはまる組み合わせの必要例数を全てあらかじめ求めることとする。また，この条件を「刻み条件」と呼ぶことにする。

## 5 正確に検出力を保つ必要最低例数を求める関数

前節で述べた例数設計を基に正確に検出力を保つ必要最低例数を求める R 上の関数 cal.samp.size() を作成した。

### 5.1 cal.samp.size の引数

引数については前節でも述べたように，

**p** 真の成功率 $p$（$0.1 \leqq p \leqq 0.9$）

**d** ハンディキャップ $\Delta$（$0.1 \leqq \Delta \leqq 0.2$）

**a** 信頼率 $\alpha$（$0 < \alpha < 1$）

の 3 つを与えることになる。

### 5.2 実行例

ここでは，$p = 0.45$,$\Delta = 0.132$,$\alpha = 0.8$ の場合の実行例を以下に示す。

```
> cal.samp.size(0.45,0.132,0.8)
Necessary sample size to keep statistical
power 0.8 exactly
  n   = sample size,
power = statistical power of the sample size.
$n
[1] 203
$power
[1] 0.8011526
```

この場合は $p$ と $\Delta$ の 2 つが「刻み条件」に当てはまっていないことになるが，約 2 分で出力することができた。また，$p$ と $\Delta$ は同値のままで $\alpha$ のみ $\alpha = 0.79$ として，全て「刻み条件」に当てはまらない場合としても，約 2 分で出力することができた。

### 5.3 作成した関数の評価

作成した関数は,「刻み条件」を満たした必要例数をあらかじめ代入しておくということになるので，この条件を満たした $p$ と $\Delta$ と $\alpha$ の 3 つの組み合わせを与えられた時は，この結果をそのまま出力するが，この条件を満たしていない $p$ と $\Delta$ と $\alpha$ の 3 つの組み合わせを与えられた時は,「刻み条件」を満たした必要例数の結果を基に近似的に基点を求め，そこから正確に検出力を保つ必要最低例数を求めて出力するということになる。

そのため,「刻み条件」を満たしていない場合，必要例数を求めるのにかなり時間がかかってしまうと予想していた。特に，信頼率に関しては基準となるのが 0.8 と 0.9 のみしかないので，信頼率が 0.8,0.9 以外で与えられた場合，基点の精度が特に悪くなり，実行時間にかなり差が出るのではないかと考えていた。しかし，必要例数が 300 以下になる場合であれば，どのような条件を与えても約 15 分以内で結果を出力することができることが分かったので，信頼率が 0.8,0.9 以外の場合でもかなり精度の良い基点を求めることができているのではないかと考えられる。

しかし，必要例数が 400 以上になる場合では，検出力を求める関数の計算時間が 1 つ 1 つ長くなり，結果を出力するのに 1 時間以上かかる場合もある。そのため，実用的には $\alpha \leqq 0.95$ が上限となる。

## 6 おわりに

正確に検出力を保つ必要最低例数を自由に求められる関数を作成したことで直接的に「かなり多くの例数が必要」という元々の課題を解決できたわけではないが，提案法を用いる状況に応じて，ハンディキャップ $\Delta$ や信頼率 $\alpha$ を変化させることで，必要例数を減らすことができるようになった。今後，この関数を利用し，新たな評価基準を模索していきたいと思う。

## 参考文献


[1] Cesana, Bruno M.: Sample size for testing and estimating the difference between two paired and unpaired proportions: a‘ two-step ’procedure combining power and the probability of obtaining a precise estimate, *Statist. Med.*, **23**, 2359-2373, 2004.

[2] Fagerland, Morten W., Lydersen, Stain and Laake, Petter: Recommended confidence intervals for two independent binomial proportions, *Stat. Methods Med. Res.*, (DOI: 10.1177/0962280211415469), 2011.

[3] 広津千尋: 「医学・薬学データの統計解析」, 東京大学出版会, 東京, 2004.

[4] Kang, Seung-Ho and Chen, James J.: An approximate unconditional test of non-inferiority between two proportions, *Statist. Med.*, **19**, 2089-2100, 2000.

[5] 松田眞一: 比率の差の信頼区間に関する考察,「計算機統計学」, **18**, 95–105, 2005.

[6] 水野裕太: 「2 つの比率の差の同等性試験に関する研究」, 南山大学数理情報学部卒業論文, 2011.

[7] Phillips, Kem F.: A new test of non-inferiority for anti-infective trials, *Statist. Med.*, **22**, 201-212 (DOI: 10.1002/sim.1122), 2003.

[8] 丹後俊郎: 「新版 医学への統計学」, 朝倉書店, 東京, 1993.

[9] Tango, Toshiro: Equivalence test and confidence interval for the difference in proportions for the paired-sample desgin, *Statist. Med.*, **17**, 891-908, 1998.