成蹊大学工学研究報告
J. Fac. Eng., Seikei Univ.
Vol.41 No.2 (2004) pp.9-20

独立な２つの二項確率の差の信頼区間

岩崎　学[*1]，橋垣　学[*2]

Confidence intervals for the difference between two independent binomial proportions

Manabu Iwasaki[*1] and Satoshi Hashigaki[*2]

ABSTRACT：Interval estimation for the difference of independent binomial proportions is no doubt one of the most fundamental problems in statistical data analysis. Several types of intervals have been introduced in the literature. This paper reviews such intervals and evaluates them in terms of coverage probability (actual confidence coefficient). Our first finding is that if it is required that the coverage probability should be greater than the nominal confidence coefficient then the exact interval is the only method to be used. The second finding is that, if we require that the coverage probability is near the nominal value on average, either the hybrid score interval or the Wald interval with a standard error suggested by the testing problem is easily calculated and also shows good performance.

KEY WORDS：Coverage probability, Exact interval, Normal approximation, Score interval, Wald interval.

(Received October 12. 2004)

１．はじめに

二項分布に関する統計的推測は正規分布と並んで最も多く議論され，その応用分野も多岐に渡っている。特に医薬分野では，薬剤あるいは治療法の有効性が二項分布に基づく推測により評価され，新薬などの承認の可否を含め実際上大きな意味を持ってくる。二項分布は離散型確率分布の中で最も基本的なものであり，どのような初等的な教科書でも必ず扱われているので，その推測法はすでに確立していると考えられがちである。しかし，実際はそうではない。その証拠に，二項分布の推測に関する学術論文が近年になってもなおいくつも書かれ続けている。本論では，特に２つの二項確率の差の信頼区間について，これまでの研究をまとめるとともに，実用上有益な方法は何であるかを探る。

試行回数 n，二項確率 p の二項分布を $B(n, p)$ と書く。２ つの独立な二項分布 $B(m, p_1)$ および $B(n, p_2)$ の場合，成功の確率の差 $\theta = p_1 - p_2$ の信頼区間は，それぞれの群における成功の回数を x および y としたとき，各群の成功の比率を $\hat{p}_1 = x/m$，$\hat{p}_2 = y/n$ として

$$(\hat{p}_1 - \hat{p}_2) \pm z(\alpha/2)\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} \quad (1.1)$$

とすることが初等的な教科書では多い。ここで $z(\alpha/2)$ は標準正規分布 $N(0, 1)$ の上側 $100\alpha/2\%$ 点である（95%信頼区間の場合は $z(0.025) \approx 1.96$）。ところが，最近の研究では特にサンプルサイズがそう大きくない場合 (1.1) は好ましくないというのがほぼ一致した見解である。

２標本での最大の問題点は，未知の二項確率が p_1

*1 情報処理専攻教授 (iwasaki@is.seikei.ac.jp)
Professor, Department of Information Science
*2 ファイザー製薬株式会社

と p_2 と 2 つあるのに対し，差 $\theta = p_1 - p_2$ というひとつのパラメータのみに注目することから今 1 つのパラメータが攪乱母数となり，その扱いをどのようにするかにある。

２．二項分布の基本的性質

ここでは二項分布に関する基本的な性質をまとめておく。詳しくは Johnson, Kotz and Kemp (1992) あるいは竹内・藤野 (1981) などを参照されたい。以下では，離散型確率変数 X が試行回数 n，成功の確率 p の二項分布に従うことを $X \sim B(n, p)$ と書く。X の確率関数は $x = 0, 1, \ldots, n$ に対して

$$f(x \mid p) = \Pr(X = x) = {}_nC_x p^x (1-p)^{n-x}$$

である（ここで ${}_nC_x = n!/\{x!(n-x)!\}$ は二項係数）。X の低次のモーメントは以下のようである：

・期待値：$E[X] = np$

・分散：$V[X] = np(1-p)$

・平均値まわりの 3 次モーメント：

$$\mu_3 = E[(X-np)^3] = np(1-p)(1-2p)$$

・平均値まわりの 4 次モーメント：

$$\mu_4 = E[(X-np)^4]$$
$$= np(1-p)\{1 + 3(n-2)p(1-p)\}$$

・歪度：

$$\beta_1[X] = \frac{\mu_3}{(V[X])^{3/2}} = \frac{1}{\sqrt{n}} \cdot \frac{1-2p}{\sqrt{p(1-p)}} \qquad (2.1)$$

・尖度：

$$\beta_2[X] = \frac{\mu_4}{(V[X])^2} - 3 = \frac{1}{n}\left\{\frac{1}{p(1-p)} - 6\right\} \qquad (2.2)$$

これより X/n の期待値と分散は

・期待値：$E[X/n] = p$

・分散：$V[X/n] = p(1-p)/n$

となり，歪度 $\beta_1[X/n]$ と尖度 $\beta_2[X/n]$ はそれぞれ (2.1), (2.2) と同じである。また，

$$E\left[\frac{X}{n}\left(1-\frac{X}{n}\right)\right] = \frac{n-1}{n}p(1-p) \qquad (2.3)$$

も容易に示される。

X および Y が互いに独立にそれぞれ $B(m, p_1)$, $B(n, p_2)$ に従うとき，同時確率関数は

$$f(x, y \mid p_1, p_2) = \Pr(X = x, Y = y)$$
$$= {}_mC_x p_1^{\,x}(1-p_1)^{m-x}\,{}_nC_y p_2^{\,y}(1-p_2)^{n-y} \qquad (2.4)$$

である。ここで二項確率の差を $\theta = p_1 - p_2$ とすると，(2.4) は

$$f(x, y \mid \theta, p_1) \qquad (2.5)$$
$$= {}_mC_x\,{}_nC_y\,p_1^{\,x}(1-p_1)^{m-x}(p_1-\theta)^y(1-p_1+\theta)^{n-y}$$

となる。さらに，$\alpha = (p_1 + p_2)/2$ と置くと $p_1 = (2\alpha + \theta)/2$, $p_2 = (2\alpha - \theta)/2$ であるので，(2.4) は

$$f(x, y \mid \theta, \alpha)$$
$$= 0.5^{m+n}\,{}_mC_x\,{}_nC_y(2\alpha+\theta)^x \qquad (2.6)$$
$$\times\{2-(2\alpha+\theta)\}^{m-x}(2\alpha-\theta)^y\{2-(2\alpha-\theta)\}^{n-y}$$

となる。また，p_2 のオッズを $\lambda = p_2/(1-p_2)$ とし，オッズ比を

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \qquad (2.7)$$

として，$s = x + y$ と置き (2.4) を変形すると

$$g(x, s \mid \psi, \lambda)$$
$$= {}_mC_x\,{}_nC_y\,\lambda^s\psi^x\left(\frac{1}{1+\lambda\psi}\right)^m\left(\frac{1}{1+\lambda}\right)^n$$

となる。したがって，$X + Y = s$ が与えられたときの X の条件付確率関数は

$$g_1(x \mid s, \psi) = {}_mC_x\,{}_nC_{s-x}\psi^x \Big/ \sum_k {}_mC_k\,{}_nC_{s-k}\psi^k \qquad (2.8)$$

とオッズ比 ψ のみの関数となる。この確率分布は非心超幾何分布とよばれる。特に $\psi = 1$，すなわち $p_1 = p_2$ のときは

$$g_1(x \mid s) = {}_mC_x\,{}_nC_{s-x} / {}_{m+n}C_s$$

と通常の超幾何分布になる。

次節以降の議論のため $X/m - Y/n$ のモーメントを求めておく。結果は以下のようである：

・期待値：$E\left[\frac{X}{m} - \frac{Y}{n} \mid p_1, p_2\right] = p_1 - p_2$

・分散：$V\left[\frac{X}{m} - \frac{Y}{n} \mid p_1, p_2\right] = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}$

・平均値まわりの 3 次モーメント：

$$E\left[\left\{\left(\frac{X}{m}-\frac{Y}{n}\right)-(p_1-p_2)\right\}^3 \mid p_1, p_2\right]$$
$$=\frac{p_1(1-p_1)(1-2p_1)}{m^2}-\frac{p_2(1-p_2)(1-2p_2)}{n^2}$$

・平均値まわりの 4 次モーメント：

$$E\left[\left\{\left(\frac{X}{m}-\frac{Y}{n}\right)-(p_1-p_2)\right\}^4 \mid p_1, p_2\right]$$
$$=3\left\{\frac{p_1(1-p_1)}{m}+\frac{p_2(1-p_2)}{n}\right\}^2$$
$$+\frac{p_1(1-p_1)\{1-6p_1(1-p_1)\}}{m^3}$$
$$+\frac{p_2(1-p_2)\{1-6p_2(1-p_2)\}}{n^3}.$$

これより $p_1=p_2=p$ のときは

$$E\left[\frac{X}{m}-\frac{Y}{n}\mid p\right]=0$$
$$V\left[\frac{X}{m}-\frac{Y}{n}\mid p\right]=\left(\frac{1}{m}+\frac{1}{n}\right)p(1-p)$$
$$E\left[\left\{\left(\frac{X}{m}-\frac{Y}{n}\right)-2p\right\}^3 \mid p\right],$$
$$=\left(\frac{1}{m^2}-\frac{1}{n^2}\right)p(1-p)(1-2p)$$
$$E\left[\left\{\left(\frac{X}{m}-\frac{Y}{n}\right)-2p\right\}^4 \mid p\right]$$
$$=3\left\{\left(\frac{1}{m}+\frac{1}{n}\right)p(1-p)\right\}^2$$
$$+\left(\frac{1}{m^3}+\frac{1}{n^3}\right)p(1-p)\{1-6p(1-p)\}$$

となり，さらに $m=n$ では 3 次モーメントは恒等的に 0 で，

$$E\left[\left\{\left(\frac{X}{n}-\frac{Y}{n}\right)-2p\right\}^4 \mid p\right]$$
$$=3\left\{\frac{2}{n}p(1-p)\right\}^2+\frac{2}{n^3}p(1-p)\{1-6p(1-p)\}$$

となることがわかる。歪度 $b_1[X/m-Y/n]$ と尖度 $b_2[X/m-Y/n]$ は上記の結果より求められるが，応用上重要な $p_1=p_2=p$ の場合の具体的な形は

$$b_1\left[\frac{X}{m}-\frac{Y}{n}\mid p\right]=\frac{n-m}{\sqrt{m+n}}\cdot\frac{(1-2p)}{\sqrt{p(1-p)}} \qquad (2.9)$$

および

$$b_2\left[\frac{X}{m}-\frac{Y}{n}\mid p\right]=\frac{m^3+n^3}{mn(m+n)^2}\left\{\frac{1}{p(1-p)}-6\right\} \qquad (2.10)$$

である。ひとつの二項分布の場合の (2.1) および (2.2) とは標本数で定義される係数部分のみが異なる。(2.9) より m と n がほぼ同じであれば p の値にかかわらずその分布はおおよそ左右対称となり，正規分布への近似がよくなる。また，$m=n$ のとき (2.10) は

$$b_2\left[\frac{X}{n}-\frac{Y}{n}\mid p\right]=\frac{1}{2n}\left\{\frac{1}{p(1-p)}-6\right\}$$

と (2.2) で試行回数が $2n$ になった場合と一致する。図 2.1 は $m=n=20, p_1=p_2=0.9$ の場合の $X/m-Y/n$ の分布とその正規近似である。ひとつの二項分布のときと異なり，二項確率が 1 に近くても正規近似はかなりよい。

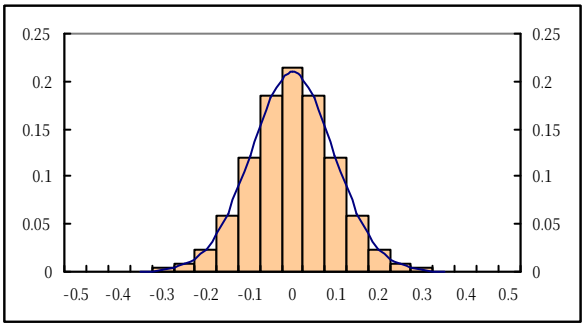


図 2.1 差 X/m　Y/n の分布と正規近似
（$m=n=20$，$p_1=p_2=0.9$）

3．種々の信頼区間とその性質

ここでは，X および Y を互いに独立にそれぞれ $B(m, p_1)$，$B(n, p_2)$ に従う確率変数としたとき，確率の差 $q=p_1-p_2$ の信頼区間の構成法を挙げ，その基本的な性質を吟味する。3.1 節および 3.2 節では正規近似による方法を取り上げる。3.1 節では $\hat{q}\pm z(\alpha/2)SE[\hat{q}]$ の形の区間を比較し，3.2 節ではスコア型および擬似的なデータを加える方法を扱う。さらに，3.3 節では尤度関数に基づく区間，3.4 節では正確な確率計算による方法を議論する。数値例は

3.5節でまとめて示す。なお，Newcombe (1998b) では，本節で取り上げた信頼区間以外のものを含め全部で11種類の区間の比較を行なっている。

3．1　正規近似に基づく方法−1

二項確率 p_1 および p_2 の自然な推定量はそれぞれ $\hat{p}_1 = X/m$，$\hat{p}_2 = Y/n$ であるので，差 $q = p_1 - p_2$ の推定量は $\hat{q} = \hat{p}_1 - \hat{p}_2 = X/m - Y/n$ とするのが自然である（これは最尤推定量でもある）。期待値と分散は

$$E[\hat{q}] = q, \quad V[\hat{q}] = \frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n} \qquad (3.1)$$

であり，m および n が十分大きいとき $\hat{q}$ は近似的に (3.1) を期待値と分散に持つ正規分布に従う。

$$SE[\hat{q}] = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}} \qquad (3.2)$$

とすると $Z = (\hat{q} - q)/SE[\hat{q}] \sim N(0, 1)$ であるので，この Z が pivot となる。これより，$N(0, 1)$ の上側 100α%点を $z(\alpha/2)$ として

$$\hat{q} \pm z(\alpha/2)SE[\hat{q}] \qquad (3.3)$$

の形の信頼区間が求められる。$SE[\hat{q}]$ は (3.2) のように未知パラメータ p_1，p_2 を含むので，何らかの形で推定する必要がある。最も単純に p_1 および p_2 に それぞれの推定値を代入して

$$SE_1[\hat{q}] = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}} \qquad (3.4)$$

とすることにより，第1節の (1.1) が得られる。これはWald型の信頼区間である。

$SE[\hat{q}]$ の推定値は (3.4) だけではない。検定

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2 \qquad (3.5)$$

では，H_0 の下での二項確率を $p_1 = p_2 = p$ とすると $SE[\hat{q}] = \sqrt{(1/m + 1/n)p(1-p)}$ であり，共通の p の推定値は $\hat{p} = (x+y)/(m+n)$ であるので，これを用いて

$$SE_2[\hat{q}] = \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\hat{p}(1-\hat{p})} \qquad (3.6)$$

としてもよい。Miettinen and Nurminen (1985) では $p(1-p)/(m+n)$ の不偏推定量が

$$\{(m+n)/(m+n-1)\}\hat{p}(1-\hat{p})$$

であることから（(2.3) 参照），(3.6) の右辺の根号の中を

$$(1/m + 1/n)\{(m+n)/(m+n-1)\}\hat{p}(1-\hat{p})$$

としている。(3.5) の検定では H_0 の下での標準誤差 (3.6) を採用するのが自然であり，実際，検定統計量 $Z^2 = (\hat{p}_1 - \hat{p}_2)^2/(SE_2[\hat{q}])^2$ は 2×2 分割表の独立性のカイ二乗検定統計量に一致し，多群の二項確率の検定への拡張も容易である（たとえばCollett (1991) を参照）。SE_2 を用いた信頼区間は，近似的に，このカイ二乗検定で棄却されないパラメータの範囲という解釈が成り立つ。

以下では (3.4) の SE_1 と (3.6) の SE_2 の比較を行なう。信頼区間は (3.3) で与えられるので，この比較により各区間の区間幅の情報が得られる。$g = m/(m+n)$ と置く $(0 \le g \le 1)$。このとき $\hat{p} = g\hat{p}_1 + (1-g)\hat{p}_2$ である。若干の式の変形により

$$\frac{mn}{m+n}(SE_2^{\,2} - SE_1^{\,2})$$

$$= g(1-g)(\hat{p}_1 - \hat{p}_2)^2 + (1-2g)\{\hat{p}_2(1-\hat{p}_2) - \hat{p}_1(1-\hat{p}_1)\}$$

が得られる。したがって，$g = 1/2$ すなわち $m = n$ の場合には常に $SE_2 \ge SE_1$ であることがわかる。$m \neq n$ では逆に $SE_2 < SE_1$ となることもあり得る。たとえば $g \to 0$ で

$$\hat{p}_2(1-\hat{p}_2) < \hat{p}_1(1-\hat{p}_1)$$

の場合である。具体的に $m = 6$，$n = 50$ で $x = 3$，$y = 5$ では $\hat{p}_1 = 3/6 = 0.5$，$\hat{p}_2 = 5/50 = 0.1$ および $\hat{p} = 8/56 = 1/7 = 0.143$ であるので $SE_2 - SE_1 = -0.057$ と負になる。

$m = n$ のとき $SE_2 \ge SE_1$ となることは次のようにしても導かれる（数学的に興味深いので示しておく）。2次関数 $g(x) = x(1-x)$ は凹関数であるので $0 \le g \le$

1 より

$$g(g\hat{p}_1+(1-g)\hat{p}_2)\geq gg(\hat{p}_1)+(1-g)g(\hat{p}_2)$$

となる。この左辺は $g(\hat{p})$ である。したがって

$$\hat{p}(1-\hat{p})\geq g\ \hat{p}_1(1-\hat{p}_1)+(1-g)\hat{p}_2(1-\hat{p}_2)$$

より

$$\left(\frac{1}{m}+\frac{1}{n}\right)\hat{p}(1-\hat{p})\geq\frac{1}{n}\hat{p}_1(1-\hat{p}_1)+\frac{1}{m}\hat{p}_2(1-\hat{p}_2)$$

を得る。この式の右辺は SE_1 の係数部分の $1/m$ と $1/n$ を入れ替えた形になっているので，$m = n$ であれば $SE_2 \geq SE_1$ がいえる。

以上をまとめて，$m \approx n$ であれば概ね $SE_2 \geq SE_1$ であり，したがって，(3.6) の SE_2 で標準誤差を計算したほうが区間幅が広くなる。このことは，次節の被覆確率の比較の結果に現れてくる。

次に，$p_1 = p_2$ の下での $Z_1 = (\hat{p}_1 - \hat{p}_2)/SE_1$ および $Z_2 = (\hat{p}_1 - \hat{p}_2)/SE_2$ の分布形の検討を行なう。分母に (3.2) の真値を用いた

$$Z = (\hat{p}_1 - \hat{p}_2)/SE$$

は，特に $m \approx n$ のとき第 2 節で見たように $N(0, 1)$ に近いが，分母に推定値を用いた際の分布形はどうかというのが問題意識である。Z_1 と Z_2 は共に漸近的には $N(0, 1)$ に従うことは示されるが，m および n があまり大きくない場合には理論的な考察は困難であるので，ここでは数値計算の結果を示す．表 3.1 は $m = n = 20$ とした場合のいくつかの p_1 および p_2 に対する Z_1 と Z_2 の分散と尖度である（期待値と歪度は恒等的に 0）．表 3.1 の計算では，$x = y = 0$ および $x = y = 20$ のときは Z_1 および Z_2 の分母分子が共に 0 となるがこのときは $Z_1 = Z_2 = 0$ とした。また，$x = 0$, $y = 20$ および $x = 20$, $y = 0$ では $SE_1 = 0$ となり，Z_1 は分子が 0 でないので無限大に発散するため，便宜上 $Z_1 = \pm 100$ とした（確率が小さいのでいくつにしてもそう大きな影響はない）。

表 3.1 から p_1 および p_2 が極端に 0 または 1 に近くなければ Z_1 および Z_2 の分散は 1 よりも大きく，超過は Z_1 のほうが大きいことが見て取れる。また，尖度も Z_2 のほうが 0 に近く，全体として Z_1 よりも Z_2 が $N(0, 1)$ に近いことが分かる。

表 3.1　Z_1 と Z_2 の分散と尖度 ($m = n = 20$)

p1	0.1	0.3	0.5	0.7	0.9	0.95	0.98
p2	0.1	0.3	0.5	0.7	0.9	0.95	0.98
V[Z1]	1.079	1.113	1.115	1.113	1.079	0.940	0.590
V[Z2]	1.010	1.026	1.026	1.026	1.010	0.894	0.569
Kurt[Z1]	-0.390	0.248	0.335	0.248	-0.390	-0.751	-0.440
Kurt[Z2]	-0.543	-0.102	-0.054	-0.102	-0.543	-0.831	-0.486

上で調べた (3.4) あるいは (3.6) 以外の標準誤差の想定値としては，根号の中の各分散の不偏推定量を用いた

$$SE_1'[\hat{q}] = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m-1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n-1}}$$

を用いる方法，および (3.3) に連続修正 (continuity correction = CC) を加えた区間

$$\hat{q} \pm \{z(\alpha/2)SE[\hat{q}]+CC\}$$

などが比較検討されている。連続修正項 CC としては，Yates の修正 $CC_Y = 1/(2m) + 1/(2n)$ を始めいくつかのものが提案されている。これらは本報告では扱わないので，詳しくは Hauck and Anderson (1986) を見られたい。

3．2　正規近似に基づく方法−2

標本比率の差 $\hat{q} = \hat{p}_1 - \hat{p}_2$ について，$\chi_1^2(\alpha)$ を自由度 1 のカイ二乗分布の上側 100α% 点とした

$$\Pr\left(\frac{\{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)\}^2}{p_1(1-p_1)/m+p_2(1-p_2)/n}<\chi_1^2(\alpha)\right) = 1-\alpha \quad (3.7)$$

を $p_1 - p_2$ について解くことにより，スコア型の信頼区間が得られる。Anbar (1983) は (3.7) で $p_2 = p_1 - q$ とし q と p_1 の関数に変形した上で，q のスコア型の信頼区間を p_1 に推定値を代入することにより導出した。しかし Mee (1984) はこの計算において，$p_1 = q + p_2$ として同様の計算を施すと異なる区間になってしまうことを指摘しその対応策を示した。いずれにしてもこれらの区間は複雑であることから，Newcombe (1998b) は 1 変量の場合のスコア検定の結果を用いたハイブリッド型の区間を提案した（Agresti and Caffo (2000) も参照）。すなわち，両群における

$$\frac{|\hat{p}_1 - p_1|}{\sqrt{p_1(1-p_1)/m}} = z(\alpha/2)$$

および

$$\frac{|\hat{p}_2 - p_2|}{\sqrt{p_2(1-p_2)/n}} = z(\alpha/2)$$

の解をそれぞれ $l_1 \le u_1$ および $l_2 \le u_2$ とするとき，100(1 – α)%信頼区間の両端を

$$q_L = \hat{p}_1 - \hat{p}_2 - z(\alpha/2)\sqrt{\frac{l_1(1-l_1)}{m} + \frac{u_2(1-u_2)}{n}} \quad (3.8)$$

および

$$q_U = \hat{p}_1 - \hat{p}_2 + z(\alpha/2)\sqrt{\frac{u_1(1-u_1)}{m} + \frac{l_2(1-l_2)}{n}} \quad (3.9)$$

とするものである。本稿ではこれをハイブリッドスコア型区間とよぶ。なお，Newcombe (1998b) および Agresti and Caffo (2000) では下限 (3.8) で l_1, u_2 を，上限で u_1, l_2 を一貫して用いているが，この選択はある意味で任意であり，$\hat{p}_1$ と $\hat{p}_2$ の大小関係に応じて使い分けるとの提案もなされている。しかし，次節で述べるような被覆確率の検討を行なったところ，使い分けをしないほうが被覆確率が名目の信頼係数に近かったため，ここでは $\hat{p}_1$ と $\hat{p}_2$ の大小関係によらず (3.8) および (3.9) で区間の両下限を求める方法を採用する。

また，Agresti and Caffo (2000) は，1 変量の場合の Agresti and Coull (1998) の観測度数に仮想的データを加えた Wald 型の区間に類似し，実際の観測度数 x，y，$m-x$，$n-y$ に 1 つずつの擬似データを加え，$\tilde{p}_1 = (x+1)/(m+2)$ および $\tilde{p}_2 = (y+1)/(n+2)$ として $\tilde{q} = \tilde{p}_1 - \tilde{p}_2$ とし，

$$SE_1[\tilde{q}] = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{m+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n+2}}$$

と置いた Wald 型の区間 $\tilde{q} \pm SE_1[\tilde{q}]$ を提案した。この区間は，1 標本の場合のようなスコア型区間の近似といった意味はないが，Agresti and Caffo (2000) は各度数に加えるデータ数として 1 以外にもいくつかの場合を考察した上で種々の理由から 1 を選択している。

3．3　尤度関数に基づく方法

第 2 節で述べたように，2 つの独立な二項分布の同時確率関数に基づく尤度関数は 2 つの二項確率 p_1 および p_2 の 2 変数関数であり，差 $\theta = p_1 - p_2$ の関数として扱うためにはもう 1 つのパラメータ（攪乱母数）の扱いが問題となる。θ でないパラメータの選択法は一意でなく，(2.5) のように (θ, p_1) としても，あるいは (2.6) のように $\chi = (p_1 + p_2)/2$ として (θ, χ) としてもよい。

攪乱母数 γ をもつ尤度関数に基づく推測法には大きく 2 つある。1 つ目は攪乱母数をある値においた条件付き推測，2 つ目は θ をある値としたときの γ_θ の最尤推定値を γ の値とする profile likelihood 法である。条件付き推測では，攪乱母数をある値に置くため，この母数が何らかの実際的な意味を持つものである必要がある。たとえば新薬開発の臨床試験において，既存薬の有効率 p_1 に比べて開発薬の有効率 p_2 がどの程度かを評価する場合には，p_1 は基準であるので p_1 の条件付きで推測を行ない，その条件を動かして推測結果を吟味するというアプローチが現実的であろう。Anbar (1983) への Mee (1984) の討論に対する返答で D. Anbar も同様の見解を述べている。文献上では (2.6) の (θ, χ) のパラメータ化が多く行なわれているが，単なる計算上の便宜のためという感が強い。R. A. Fisher は何というであろうか。

Profile likelihood 法の計算の骨子を述べる。(2.6) より θ と χ に関する対数尤度関数は，定数項を省いて

$$\begin{aligned} l(\theta,\chi) &= \log f(x,y|\theta,\chi) \\ &\propto x\log(2\chi+\theta) + (m-x)\log\{2-(2\chi+\theta)\} \\ &\quad + y\log(2\chi-\theta) + (n-y)\log\{2-(2\chi-\theta)\} \end{aligned}$$

となる。θ および χ の最尤推定値は

$$\hat{\theta} = x/m - y/n, \quad \hat{\chi} = (x/m + y/n)/2$$

である。計算の目的は $l(\theta, \chi)$ を何らかの意味で θ の 1 変数関数と見て，ある定数 d に対し，

$$l(\hat{\theta},\hat{\chi}) - l(\theta,\chi) \le d$$

となる θ の範囲の探索である。計算は次の 2 段階で行なわれる：

(1) 与えられた θ に対し，θ を所与とした x の条件付き最尤推定値 $\hat{x}_\theta$ を求める，

(2) 対数尤度関数の値 $l(\theta, \hat{x}_\theta)$ が $l(\hat{\theta}, \hat{x}) - d$ 以下かどうかをチェックする。

上記の計算を x および y の値が 0 または m, n に等しい場合とそうでない場合に分けて実行する必要がある。計算の詳細は Miettinen and Nurminen (1985) あるいは Newcombe (1998b) を参照されたい。第 4 節の 1 標本の場合にも尤度関数法は数値計算を必要とし，2 標本でも上記 (1) の x の条件付き最尤推定値の計算および θ に関する (2) の数値的な探索は計算上の負荷が大きいため，実用的とはいえない。

3．4　正確な確率評価に基づく法

1 変量の場合の Clopper-Pearson 型区間のように，2 標本の場合も被覆確率（実際の信頼係数）が常に名目の信頼係数以上となる区間を求めようとする試みがなされている。Thomas and Gart (1977) は，$X = x^*$ が観測されたとき (2.7) のオッズ比 ψ の正確な信頼区間 (ψ_L, ψ_U) を (2.8) の非心超幾何分布 $g_1(x \mid s, \psi)$ を用いて

$$\sum_{i=0}^{x^*} g_1(x \mid s, \psi_U) = \alpha / 2$$

および

$$\sum_{i=x^*}^{s} g_1(x \mid s, \psi_L) = \alpha / 2$$

により求め，$\theta = p_1 - p_2$ と ψ との関係より θ の信頼区間を求める方法を提唱した。

しかし，Santner and Snell (1980) はこの方法では必ずしも被覆確率が名目値を上回らないことを指摘し，次の手順を提案した：観測度数を $X = x^*$ および $Y = y^*$ とし，$D(x^*, y^*) = x^*/m - y^*/n$ と定義する。そして確率

$$Q_{p_1,\theta}(x^*, y^*) = \Pr(D(X,Y) \le D(x^*, y^*)) = \sum_{D(x,y) \le D(x^*,y^*)} f(x, y \mid \theta, p_1)$$

を求める。ここで $f(x, y \mid \theta, p_1)$ は (2.5) で与えられる (X, Y) の同時確率である。$-1 \le \theta = p_1 - p_2 \le 1$ であるので，p_1 は θ が与えられるとその取りうる値の範囲は

$$I(\theta) = \{p_1 : \max(0, \theta) \le p_1 \le \min(1, 1+\theta)\}$$

である。攪乱母数である p_1 を除くため $I(\theta)$ の範囲における supremum をとり

$$Q_\theta(x) = \sup_{p_1} \{Q_{p_1,\theta}(x^*, y^*) : p_1 \in I(\theta)\}$$

とする。そして $Q_\theta \le \alpha/2$ を満足する θ を探索することにより $100(1 - \alpha)\%$信頼区間の下限が得られる。上限も同様に求めることができる。

Berger and Boos (1994) は正確な検定の方法を提案したが，彼らの方法は信頼区間の算出に応用でき，現に正確な確率計算のためのソフトウェア StatXact に採用されている。この方法は攪乱母数の p_1 を除く際に，上記の Santner and Snell のアルゴリズムで $I(\theta)$ に関する supremum を取らず，探索範囲を限定するものである。Berger and Boos のアルゴリズムは以下のようである。

まず p_1，p_2 のそれぞれについての $100(1 - \gamma)\%$信頼区間 $A_1 = (L_1, U_1)$ および $A_2 = (L_2, U_2)$ を考える。$(p_1, p_2) \in A_1 \times A_2$ であると仮定して Santner and Snell の手順を行なう。$\theta_{min} = L_2 - U_1$，$\theta_{max} = U_2 - L_1$ であるので，p_1 の範囲は

$$I_r(\theta) = \{p_1 : \max(L_1, L_2 + \theta) \le p_1 \le \min(U_1, U_2 + \theta)\}$$

となり，信頼下限は $Q_\theta(x^*, y^*) \le \alpha/2 - \gamma$ を満たす θ となる。上限も同様である。この方法では γ の値をどう取るかが問題となる。$\gamma = 0$ のときが Santner and Snell の方法に相当する。また常識より γ の最大値は $\alpha/2$ を越えることはできない。StatXact では $\gamma = 0.001$ がデフォルトになっている。

3．5　数値例

ここでは簡単な数値例を用い，上記で挙げた種々の構成法によって求められた信頼区間を示す。以下では，$m = 20$，$n = 18$ とし，$x = 8$ を固定した上で $y = 10$，2，0 とした場合を示す。$m = 20$，$x = 8$ とした 1 標本での信頼区間は第 4 節の数値例 4.1 で示している。ここでの計算例からは，Wald (SE_2) 型とスコア型は類似し，Wald (SE_1) 型は特に観測度数が 0 あるいは n に近い場合には異なる様相を示すことが見て取れる。また，正確型の区間は区間幅がかなり広くなっていることも重要な点である。m および n がある程度大きい場合には構成法による違いは少ないが，m および n がそう大きくない場合には違

いが顕著となる。

・$m = 20, n = 18, x = 8, y = 10$：$\hat{q} = \hat{p}_1 - \hat{p}_2 = -0.156$

Wald (SE_1) 型：(–0.470, 0.159)　幅 = 0.629
Wald (SE_2) 型：(–0.474, 0.162)　幅 = 0.636
スコア型：(–0.424, 0.150)　幅 = 0.574
正確法：(–0.502, 0.174)　幅 = 0.676

・$m = 20, n = 18, x = 8, y = 2$：$\hat{q} = \hat{p}_1 - \hat{p}_2 = 0.289$

Wald (SE_1) 型：(0.030, 0.548)　幅 = 0.518
Wald (SE_2) 型：(0.008, 0.569)　幅 = 0.561
スコア型：(0.006, 0.517)　幅 = 0.511
正確法：(–0.031, 0.609)　幅 = 0.640

・$m = 20, n = 18, x = 8, y = 0$：$\hat{q} = \hat{p}_1 - \hat{p}_2 = 0.4$

Wald (SE_1) 型：(0.185, 0.615)　幅 = 0.429
Wald (SE_2) 型：(0.140, 0.660)　幅 = 0.519
スコア型：(0.147, 0.613)　幅 = 0.466
正確法：(–0.084, 0.564)　幅 = 0.648

４．各信頼区間の比較

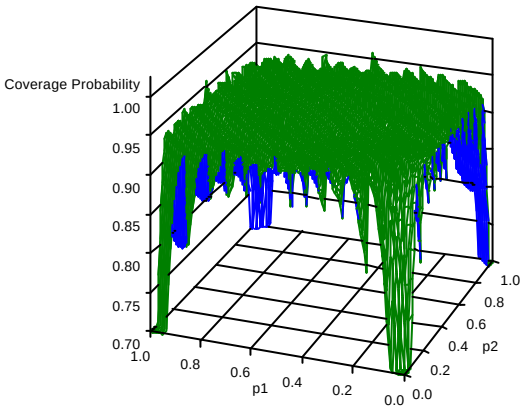
ここでは，第 3 節で議論した信頼区間のうち導出法が過度に複雑でなく実用的なもの，および理論的に興味深いのもとして，SE_1 および SE_2 による Wald 型信頼区間（(1.1) および (3.3), (3.4), (3.6) 参照），ハイブリッドスコア型区間（(3.8), (3.9) 参照）および Santner and Snell (1980) による正確な信頼区間（第 3.4 節参照）を取り上げ，そのパフォーマンスを特に被覆確率の観点から吟味する。

これらを取り上げた（他のものを取り上げなかった）理由は，SE_1 による Wald 型区間は第 1 節でも触れたとおり，多くの初等的な教科書で記載されているものであり，そのパフォーマンスを調べておくことはきわめて重要であるからである。SE_2 による Wald 型区間およびハイブリッドスコア型区間は，計算が比較的簡単でその解釈も容易なためである。特に，SE_2 による Wald 型区間は検定との関係が明白であり，ハイブリッドスコア型区間は 1 標本での結果がそのまま使えるという意味で有意義でもある。また，下で述べるようにそれらのパフォーマンスは類似であると同時に，他の複雑な区間もこれらとあまり大差はないことが過去の文献で調べられている。正確法による区間は，被覆確率が常に名目の信頼係数を下回ってはならないとの要請がある場合には唯一の選択肢であるためここに選んだ。

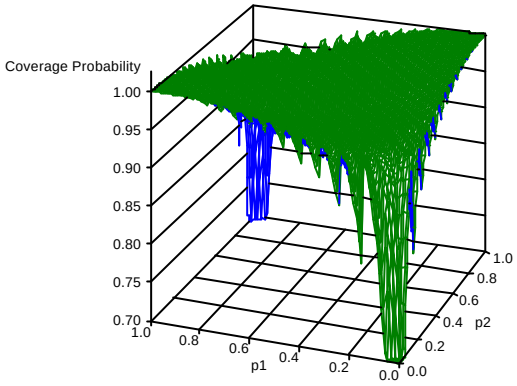
ここで取り上げたものを含めたいくつかの区間に関する比較は多くの文献でなされているので参考にされたい（たとえば Agresti and Caffo (2000), Beal (1987), Chan and Zhang (1999), Coe and Tamhane (1993), Gart and Nam (1990), Hauck and Anderson (1986), Miettinen and Nurminen (1985), Newcombe (1998b), Peskun (1993), Santner and Yamagami (1993), Soms (1989), Wallenstein (1997) など）。

以下に $m = n = 20$ とした場合の被覆確率のグラフを示す。図 4.1 は各信頼区間の被覆確率の 3D 表示である。これにより，それぞれの被覆確率の大雑把な形状が見て取れるであろう。図 4.2 は p_2 をいくつか変えた場合に，横軸に p_1 を取った被覆確率のグラフである。これにより被覆確率の様相がさらに明白に読み取れる。

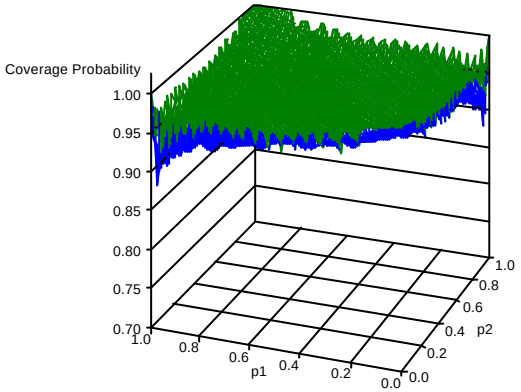
これらの図より，SE_1 による Wald 型信頼区間 (1.1) はほとんどすべての p_1, p_2 において被覆確率が名目値（この場合は 0.95）を下回り，各種文献で望ましくないとされていることが裏付けられている。SE_2 による Wald 型区間およびハイブリッドスコア型区間は，p_1 および p_2 の組み合わせにより被覆確率が名目値の上下に分布するが，「平均的」には名目値に近い。正確法による信頼区間は，被覆確率が名目値を下回らないというものの名目値からの超過が大きく，実際上は広すぎる過度に保守的な区間となっている。$m = n = 20$ 以外のサンプルサイズの設定の場合も吟味したが傾向はほぼ同じであった。



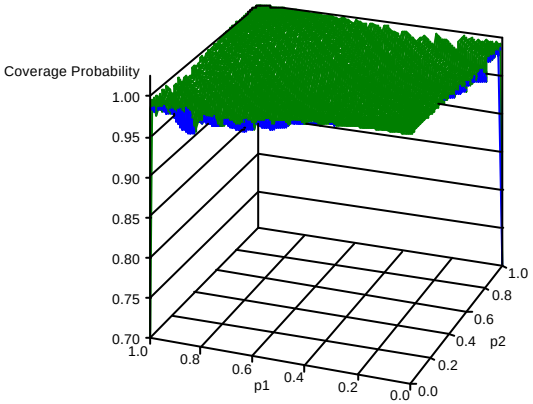
a) SE_1 による Wald 型信頼区間



(b) SE_2 による Wald 型信頼区間

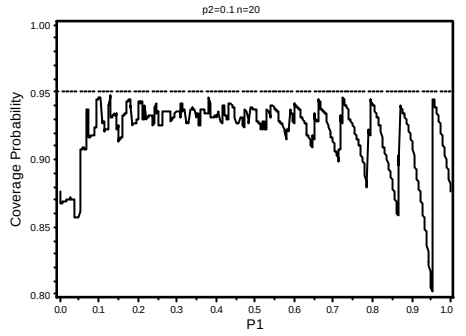


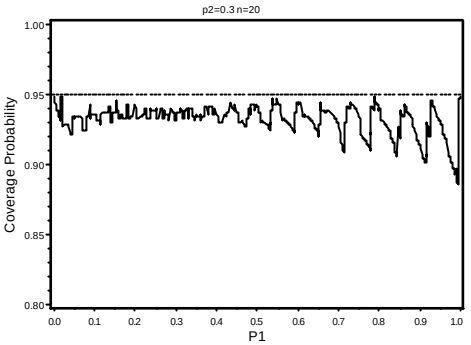
(c) ハイブリッドスコア型信頼区間

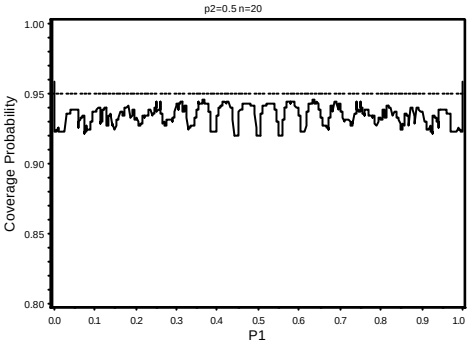


(d) 正確法による信頼区間

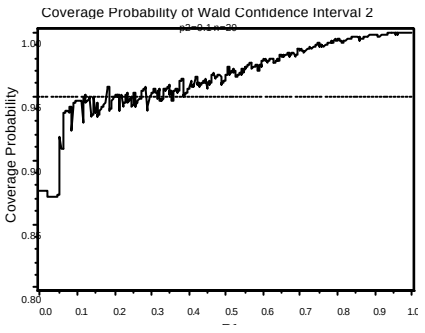
図 4.1　95%信頼区間の被覆確率の 3D 表示

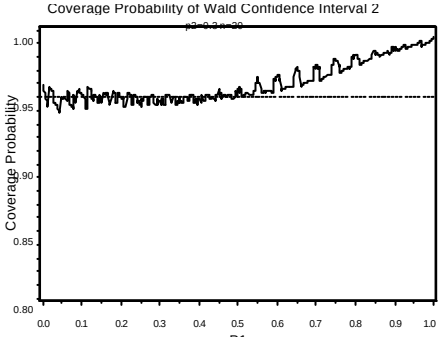


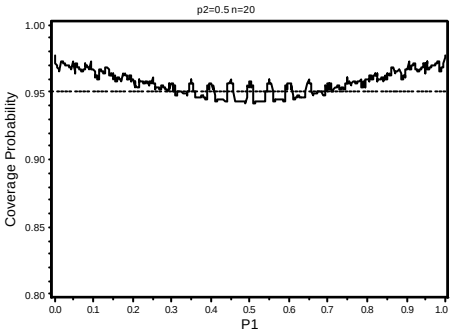




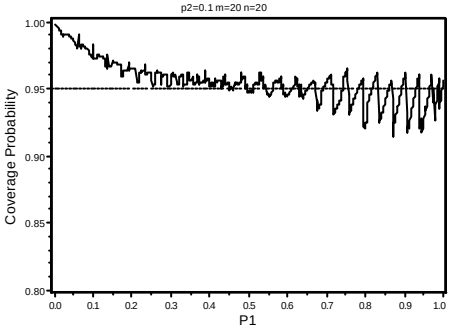
(a) SE_1 による Wald 型信頼区間

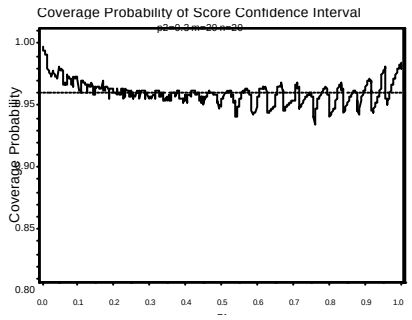


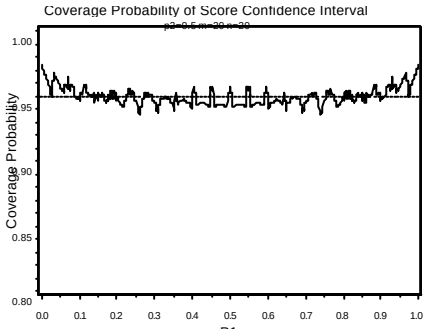




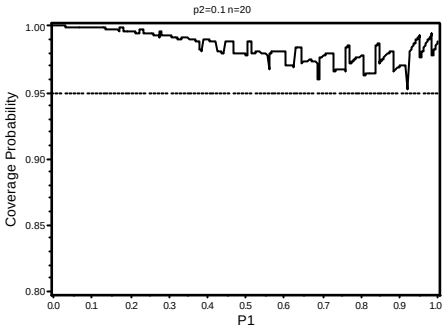
(b) SE_2 による Wald 型信頼区間

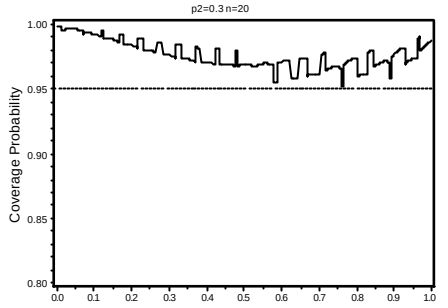


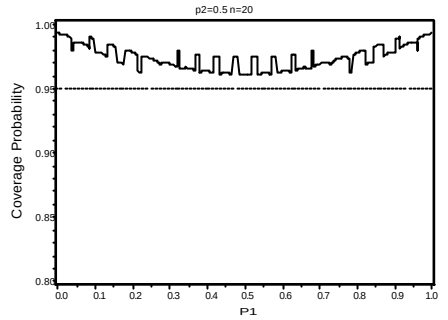




(c) ハイブリッドスコア型信頼区間







(d) 正確法による信頼区間

図 4.2　p_2 =0.1，0.3，0.5 と固定したときの 95%信頼区間の被覆確率（横軸：p_1）

5．おわりに

2 標本での二項確率の差の信頼区間について概観してきた。2 標本での確率の差の場合は，二項確率が 0 あるいは 1 に近くても統計量 $\hat{q} = \hat{p}_1 - \hat{p}_2$ の標本分布の歪みが少ない代わりに攪乱母数の取り扱いが難しい。第 1 節で述べたように，(1.1) の Wald 型の信頼区間は被覆確率（実質の有意水準）が名目の信頼係数を下回るという意味で望ましいものではない。では，多く提案されている信頼区間の中でどれが望ましいものであろうか。選択は区間の使われ方による。被覆確率がすべての母集団確率の下で名目の信頼率を下回ってはならないとの要請がある場合には，正確な信頼区間が唯一の選択肢である。ただし，被覆確率が名目値を上回る程度は想像以上に

大きく，保守的なものとなってしまう点は注意を要する。

被覆確率が「平均的に」名目値に近ければよいのであれば選択肢は広がる。2 標本における確率の差では，(3.6) の SE_2 に基づく Wald 型の区間 (3.3) が検定との関係も明白でしかも計算が簡単である点，推奨に足るであろう。多群への拡張性の点からも有用である。それ以外の区間は，多少被覆確率の面で性能がよくなることで統計家は満足しても，計算法が面倒で実用的とはいい難く教科書には載せにくい。統計手法は，統計の専門家のみならず実務に携わる多くの人々にとって有用かつ分かりやすいものでなくてはならない。統計の専門家がわずかの性能の向上を示す手法を提案しても実務家に受け入れられないのでは意味がない。サンプルサイズがある程度大きければ，最も単純な (1.1) の Wald 型で事足りるが，そうでない場合にはこれらの区間は望ましくない。したがって，教科書の記述を，確率の差では SE_2 に基づくものあるいはハイブリッドスコア型区間に変える必要があると考える。

本報告で議論しなかった話題としてサンプルサイズの設計の問題がある。サンプルサイズを設計するためには信頼区間の構成法を定めておかなければならない。2 変量の場合には Farrington and Manning (1990)，Levin and Chen (1999)，Sahai and Khurshid (1996) などの研究があるが，さらに研究を進める必要がある。

本報告での題材は初めに述べたように古典的なものである。しかし，計算手段の進展に伴い，これまで看過されてきた事柄が新たな問題として認識されたのである。この種の問題は決して少なくない。

参考文献

Agresti, A. and Caffo, B. (2000) Simple and effective intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, **54**, 280- 288.

Agresti, A. and Coull, B. A. (1998) Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, **52**, 119-126.

Anbar, D. (1983) On estimating the difference between two probabilities, with special reference to clinical trials. *Biometrics*, **39**, 257-262.

Beal, S. L. (1987) Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, **43**, 941-950.

Berger, R. L. and Boos, D. D. (1994) *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89, 1012-1016.

Berry, G. and Armitage, P. (1995) Mid-P confidence intervals: a brief review. *The Statistician*, **44**, 417-423.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001) Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101-133 (with discussion).

Chernick, M. R. and Liu, C. Y. (2002) The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods. *American Statistician*, **56**, 149-155.

Chan, I. S. F. and Zhang, Z. (1999) Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*, **55**, 1202-1209.

Clopper, C. J. and Pearson, E. S. (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404-413.

Coe, P. R. and Tamhane, A. C. (1993) Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Communications in Statistics, Simulation and Computation*, **22**, 925-938.

Collett, D. (1991) *Modelling Binary Data*. Chapman & Hall, London.

Farrington, C. P. and Manning, G. (1990) Test statistics and sample size formulae for com- parative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, **9**, 1447-1454.

Gart, J. J. and Nam, J.-M. (1990) Approximate interval estimation of the difference in binomial parameters: correction for skewness and exten- sion to multiple tables. *Biometrics*, **46**, 637-643.

Hauck, W. W. and Anderson, S. (1986) A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *American Statistician*, **40**, 318-322.

Henderson, M. and Meyer, M. C. (2001) Exploring the confidence interval for a binomial parameter in a first course in statistical computing. *American Statistician*, **55**, 337-344.

岩崎　学 (1993) mid-P value：その考え方と特性. 応

用統計学, **22**, 67-80.

岩崎　学 (2001) **非線型方程式の単純反復解法とその統計的応用．応用統計学**, **30**, 107-118.

Iwasaki, M. and Hidaka, N. (2001) Notes on the central and shortest confidence intervals for a binomial parameter. *Japanese Journal of Biometrics*, **22**, 1-13.

Iwasaki, M. and Hidaka, N. (2002) Determination of sample sizes for use in constructing confidence intervals for a binomial parameter. *Journal of the Japanese Society of Computational Statistics*, **15**, 19-32.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992) *Univariate Discrete Distributions, Second Edition*. John Wiley & Sons, New York.

Levin, B. and Chen, X. (1999) Is the one-half continuity correction used once or twice to derive a well-known approximate sample size formula to compare two independent binomial distri- butions? *American Statistician*, **53**, 62-66.

Mee, R. W. (1984) Confidence bounds for the difference between two probabilities. *Biometrics*, **40**, 1175-1176.

Miettinen, O. and Nurminen, M. (1985) Compara- tive analysis of two rates. *Statistics in Medicine*, **4**, 213-226.

Newcombe, R. G. (1998a) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857- 872.

Newcombe, R. G. (1998b) Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, **17**, 873-890.

Peskun, P. H. (1993) A new confidence interval method based on the normal approximation for the difference of two binomial probabilities. *Journal of the American Statistical Association*, **88**, 656-661.

Sahai, H. and Khurshid, A. (1996) Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. *Statistics in Medicine*, **15**, 1-21.

Santner, T. J. and Snell, M. K. (1980) Small-sample confidence intervals for $p_1 - p_2$ and p_1/p_2 in 2×2 contingency tables. *Journal of the American Statistical Association*, **75**, 386-394.

Santner, T. J. and Yamagami, S. (1993) Invariant small sample confidence intervals for the difference of two success probabilities. *Communications in Statistics, Simulation and Computation*, **22**, 33-59.

Soms, A. P. (1989) Some recent results for exact confidence intervals for the difference between two proportions. *Communications in Statistics, Simulation and Computation*, **18**, 1343-1357.

竹内　啓・藤野和建 (1981) **２項分布とポアソン分布．東京大学出版会．**

Thomas, D. G. and Gart, J. J. (1977) A table of exact confidence limits for differences and ratios of two proportions and their odds ratio. *Journal of the American Statistical Association* **72**, 73-76; Corrigenda (1978), **73**, 233.

Wallenstein, S. (1997) A non-iterative accurate asymptotic confidence interval for the difference between two proportions. *Statistics in Medicine*, **16**, 1329-1336.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209-212.