Statistics and Application 统计学与应用, 2018, 7(4), 400-406
Published Online August 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.74046

Hans汉斯

A Modified Beal Interval for the Difference between Two Independent Binomial Proportions

Yanke Wu*, Huajin Chai

School of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang Guangdong
Email: *yanke.wu@163.com, 13809753553@163.com

Received: Jul. 18th, 2018; accepted: Aug. 3rd, 2018; published: Aug. 10th, 2018

Abstract

This paper proposes a modified Beal interval for the difference between independent binomial proportions. The new method with the optimal weight has better coverage probabilities and shorter intervals than most of the often used ones. Most importantly, the optimal weight value has a simple non-iterative form, therefore there is no much additional computation compared with the existing methods. A real extreme case is analysed to show the claimed properties and practical usability.

Keywords

Beal's Interval, Binomial Proportion, Difference, Independent

两个独立二项分布比例差的提升Beal区间估计

吴延科*，柴华金

广东海洋大学数学与计算机学院，广东 湛江
Email: *yanke.wu@163.com, 13809753553@163.com

收稿日期：2018年7月18日；录用日期：2018年8月3日；发布日期：2018年8月10日

摘　要

本文对两个独立二项分布的比例差提出了一种提升Beal区间估计方法，这种方法具有较好的覆盖率和最小覆盖率，区间长度也比较短，并且，新方法使用的最优权值具有显式表达式，计算简单，增加的计算量很小。模拟结果和实例分析表明了这种方法具有良好的有效性和稳定性。

*通讯作者。

文章引用: 吴延科, 柴华金. 两个独立二项分布比例差的提升 Beal 区间估计[J]. 统计学与应用, 2018, 7(4): 400-406.
DOI: 10.12677/sa.2018.74046

关键词

Beal区间，二项比例，比例差，独立的

1. 研究背景

区间估计是统计推断的基本任务之一，很多常用的区间估计方法都是基于正态近似。令 X_1, X_2 是两个独立的二项分布变量，分别服从二项分布 $B(n_1, p_1), B(n_2, p_2)$。两个二项分布的比例差定义为 $\theta = p_1 - p_2$，θ 的最大似然估计是 $\hat{\theta} = \hat{p}_1 - \hat{p}_2$，其中 $\hat{p}_1 = \frac{X_1}{p_1}, \hat{p}_2 = \frac{X_2}{p_2}$ 分别是 p_1, p_2 的最大似然估计。给定 p_1 和 p_2，$\hat{\theta}$ 的方差是 $\mathrm{var}(\hat{\theta}; p_1, p_2) = \frac{1}{n_1} p_1(1-p_1) + \frac{1}{n_2} p_2(1-p_2)$，用 $\hat{p}_1, \hat{p}_2$ 分别替换 p_1, p_2 得到著名的 Wald 区间

$$CI_{Wald} = \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)},$$

其中 $z_{\frac{\alpha}{2}}$ 是标准正态分布的 $\frac{\alpha}{2}$ 分位数。Fleiss [1]对 Wald 区间做了一个连续校正，得到

$$CI_{Wald,cc} = \hat{\theta} \pm \left\{ z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1}\hat{p}_1(1-\hat{p}_1) + \frac{1}{n_2}\hat{p}_2(1-\hat{p}_2)} + \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \right\}.$$

Mee [2]使用 p_1, p_2 的限制最大似然估计 $\tilde{p}_1, \tilde{p}_2$ 代替最大似然估计 $\hat{p}_1, \hat{p}_2$，得到

$$CI_{Mee} = \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_1}\tilde{p}_1(1-\tilde{p}_1) + \frac{1}{n_2}\tilde{p}_2(1-\tilde{p}_2)}.$$

Beal [3]引入了一个讨厌参数 $\eta = \frac{1}{2}(p_1 + p_2)$，对 p_1, p_2 重参数化得到 $p_1 = \eta + \frac{\theta}{2}, p_2 = \eta - \frac{\theta}{2}$，给定 η 和 θ，可以得到 $\hat{\theta}$ 的方差是 $\mathrm{var}(\hat{\theta}; \eta, \theta) = u\left[4\eta(1-\eta) - \theta^2\right] + 2v(1-2\eta)\theta$，其中 $u = \frac{1}{4}\left(\frac{1}{n_1} + \frac{1}{n_2}\right), v = \frac{1}{4}\left(\frac{1}{n_1} - \frac{1}{n_2}\right)$。取 η 的一个估计 $\tilde{\eta}$，求解 $(\hat{\theta} - \theta)^2 \le z_{\frac{\alpha}{2}}^2 \mathrm{var}(\hat{\theta}; \tilde{\eta}, \theta)$ 得到 Beal 区间

$$CI_{Beal} = \theta^* \pm w,$$

其中

$$\theta^* = \frac{\hat{\theta} + z_{\alpha/2}^2 v(1-2\tilde{\eta})}{1 + z_{\alpha/2}^2 u},$$

$$w = \frac{z_{\alpha/2}}{1 + z_{\alpha/2}^2 u} \sqrt{u\left[4\tilde{\eta}(1-\tilde{\eta}) - \hat{\theta}^2\right] + 2v(1-2\tilde{\eta})\hat{\theta} + 4z_{\alpha/2}^2 u^2 \tilde{\eta}(1-\tilde{\eta}) + z_{\alpha/2}^2 v^2 (1-2\tilde{\eta})^2}.$$

Beal [3]使用贝叶斯方法得到 $\tilde{\eta}=\frac{1}{2}\left(\frac{X_1+\mu}{n_1+2\mu}+\frac{X_2+\mu}{n_2+2\mu}\right)$，其中 $\mu\geq 0$。Beal 研究了不同 μ 值得到的区间的小样本行为后建议使用 $\mu=0$ 或 $\mu=\frac{1}{2}$，这两个值对应的 Beal 区间分别又称为 Haldane 区间和 Jeffreys-Perks 区间。Roths & Tebbs [4]发现，细心选择 μ 值可以提升 Beal 区间的表现，他们给出了 μ 值的最大似然估计和矩估计，使用这两个 μ 值的区间分别记为 $CI_{Beal-MLE}$ 和 $CI_{Beal-MOM}$。

本文我们关注 Beal 区间中 η 值的非对称性问题，提出一种提升 Beal 区间。

2. 提升 Beal 区间

Newcomb [5]通过大量的模拟计算发现，Haldane 区间的覆盖率在实际中可以接近 0，而 Jeffreys-Perks 区间虽然可以在一定程度上改善这种情况，但仍然不能彻底避免覆盖率过小的现象。基于此，我们需要改进 Beal 区间的表现。

记

$$r=\frac{\mu}{2}\left(\frac{1}{n_1+2\mu}+\frac{1}{n_2+2\mu}\right),f_i=\frac{n_i}{n_i+2\mu},i=1,2 \tag{2.1}$$

其中 $\mu\geq 0$，则 $\tilde{\eta}=\frac{1}{2}\left(f_1\hat{p}_1+f_2\hat{p}_2\right)+r$。对 Haldane 区间，$\mu=0$，对 Jeffreys-Perks 区间，$\mu=\frac{1}{2}$。Beal [3]取 η 为 p_1,p_2 的算术平均，但是 p_1,p_2 对 η 的影响可能是不同的，因此，我们取

$$\eta_\lambda=\lambda p_1+(1-\lambda)p_2,\lambda\in[0,1]$$

作为讨厌参数，重参数化得到 $p_1=\eta_\lambda+(1-\lambda)\theta,p_2=\eta_\lambda-\lambda\theta$，则

$$\operatorname{var}\left(\hat{\theta};\eta_\lambda,\theta\right)=\frac{\left[\eta_\lambda+(1-\lambda)\theta\right]\left[1-\eta_\lambda-(1-\lambda)\theta\right]}{n_1}+\frac{(\eta_\lambda-\lambda\theta)(1-\eta_\lambda+\lambda\theta)}{n_2}.$$

假设 $\tilde{\eta}_\lambda$ 是 η_λ 的一个估计，求解

$$\left(\hat{\theta}-\theta\right)^2-z_{\alpha/2}^2\operatorname{var}\left(\hat{\theta};\tilde{\eta}_\lambda,\theta\right)=0 \tag{2.2}$$

即得到以两个根为端点的 θ 的一个置信区间。经过繁琐的计算可以得到提升的 Beal 区间

$$CI_{Beal-M}=\theta_\lambda^*\pm w_\lambda,$$

其中

$$\theta_\lambda^*=\frac{\hat{\theta}+z_{\alpha/2}^2u_2\left(1-2\tilde{\eta}_\lambda\right)}{1+z_{\alpha/2}^2u_1},$$

$$w_\lambda=\frac{z_{\alpha/2}}{1+z_{\alpha/2}^2u_1}\sqrt{2u_2\left(1-2\tilde{\eta}_\lambda\right)\hat{\theta}+z_{\alpha/2}^2u_2^2\left(1-2\tilde{\eta}_\lambda\right)^2+u_3\tilde{\eta}_\lambda\left(1-\tilde{\eta}_\lambda\right)-u_1\hat{\theta}^2+z_{\alpha/2}^2u_1u_3\tilde{\eta}_\lambda\left(1-\tilde{\eta}_\lambda\right)},$$

$$u_1=\frac{(1-\lambda)^2}{n_1}+\frac{\lambda^2}{n_2},u_2=\frac{1-\lambda}{2n_1}-\frac{\lambda}{2n_2},u_3=\frac{1}{n_1}+\frac{1}{n_2}. \tag{2.3}$$

使用 Beal [3]的贝叶斯方法可以得到

$$\tilde{\eta}_\lambda=\lambda\frac{X_1+\mu}{n_1+2\mu}+(1-\lambda)\frac{X_2+\mu}{n_2+2\mu}. \tag{2.4}$$

$\tilde{\eta}_\lambda$ 的值影响区间的端点和中点，我们期望 $\tilde{\eta}_\lambda$ 的均方误差(MSE)达到最小。容易计算得到 $\tilde{\eta}_\lambda$ 的偏差和方差分别为

$$\text{bias}\left(\tilde{\eta}_\lambda\right)=\lambda\frac{\mu\left(1-2p_1\right)}{n_1+2\mu}+\left(1-\lambda\right)\frac{\mu\left(1-2p_2\right)}{n_2+2\mu},$$

$$\text{var}\left(\tilde{\eta}_\lambda\right)=\lambda^2 f_1^2\frac{p_1\left(1-p_1\right)}{n_1}+\left(1-\lambda\right)^2 f_2^2\frac{p_2\left(1-p_2\right)}{n_2},$$

其中 f_1,f_2 的定义见(2.1)式。最小化 $MSE\left(\tilde{\eta}_\lambda\right)=\text{bias}^2\left(\tilde{\eta}_\lambda\right)+\text{var}\left(\tilde{\eta}_\lambda\right)$ 可以得到最优的 λ：

$$\begin{aligned}\lambda_{opt}\left(p_1,p_2\right)&=\frac{f_2^2p_2\left(1-p_2\right)/n_2-r_2\left(r_1-r_2\right)}{f_1^2p_1\left(1-p_1\right)/n_1+f_2^2p_2\left(1-p_2\right)/n_2+\left(r_1-r_2\right)^2}\\&=\frac{MSE\left(\tilde{p}_2\right)-\text{bias}\left(\tilde{p}_1\right)\text{bias}\left(\tilde{p}_2\right)}{MSE\left(\tilde{p}_1\right)+MSE\left(\tilde{p}_2\right)-2\text{bias}\left(\tilde{p}_1\right)\text{bias}\left(\tilde{p}_2\right)}\end{aligned}$$

其中 $r_i=\mu\frac{1-2p_i}{n_i+2\mu}=\text{bias}\left(\tilde{p}_i\right),\tilde{p}_i=\frac{X_i+\mu}{n_i+2\mu},i=1,2$。实际中，使用 $\hat{p}_1,\hat{p}_2$ 替换 p_1,p_2 可以得到可用的最优调节参数 $\lambda_{opt}\left(\hat{p}_1,\hat{p}_2\right)$。

3. 模拟

我们使用两个模拟试验验证提升 Beal 区间的效果，第一个用于检验覆盖率和最小覆盖率，第二个用于检验区间长度。作为对比，我们同时给出 Wald 方法、Mee 方法和 Beal 方法(包含 Roths & Tebbs [4]改良的两种方法)的模拟结果。

3.1. 检验覆盖率

给定 $n_1,n_2,z_{\alpha/2}$ 条件下，方法 Δ 的覆盖率定义为 $CP\left(p_1,p_2;n_1,n_2,z_{\alpha/2}\right)=P\left(\theta=p_1-p_2\in CI_\Delta\right)$，最小覆盖率定义为 $\min_{(p_1,p_2)\in(0,1)^2}CP\left(p_1,p_2;n_1,n_2,z_{\alpha/2}\right)$。取定 $\alpha=0.05$，分别取 $\left(n_1,n_2\right)=\left(10,15\right)$ 和 $\left(n_1,n_2\right)=\left(30,50\right)$，计算 9 种方法的覆盖率和最小覆盖率，结果见图 1~图 2 和表 1。

从图 1~图 2 和表 1 可以看出，Mee 方法和提升的 Jeffreys-Perks 方法具有较高的最小覆盖率。

3.2. 检验置信区间长度

我们来评估 9 种方法的平均区间长度。给定 $n_1,n_2,p_1,p_2,z_{\alpha/2}$，平均区间长度定义为

$$ML=\sum_{x_1=0}^{n_1}\sum_{x_2=0}^{n_2}\prod_{i=1}^{2}\binom{n_i}{x_i}p_i^{x_i}\left(1-p_i\right)^{n_i-x_i}\left|CI_\Delta\right|,$$

其中 $\left|\cdot\right|$ 表示区间长度。取定 $\alpha=0.05$，$p_1=0.9$，分别取 $\left(n_1,n_2\right)=\left(10,15\right)$ 和 $\left(n_1,n_2\right)=\left(30,50\right)$，取 $p_2=0.05,0.45,0.85$，计算 9 种方法的平均区间长度，结果见表 2。我们发现，Wald、Haldane、Beal-MOM、Haldane M 和 Jeffreys-Perks M 方法都具有相对较小的平均区间长度。结合 3.1 节的覆盖率和最小覆盖的结果，我们推荐使用 Jeffreys-Perks M 方法，即提升的 Jeffreys-Perks 方法。

4. 实例分析

我们使用 Wallenstein [6]的数据，这是一个有关种族歧视的法律案例，详情见原文。这里，$n_1=379,n_2=6,X_1=379,X_2=1$。我们之所以选择这个案例，是因为这里的 $X_2=1$ 属于极端情况。判断一种区间估计方法的好坏，其中一个标准就是看这个方法能否恰当地处理这种极端数据。我们使用提升的

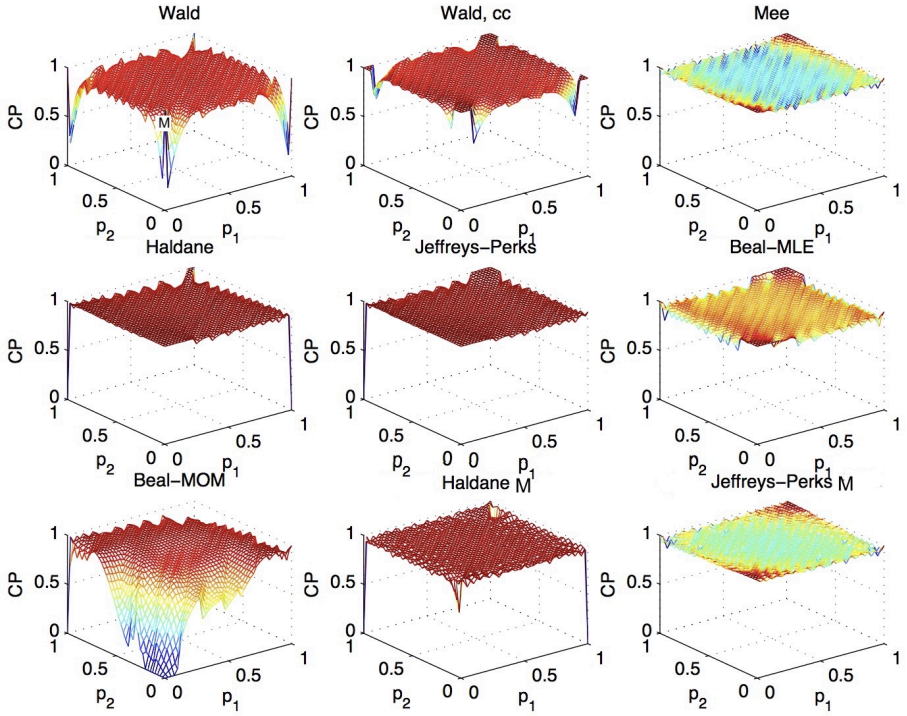

Figure 1. CP for nine methods, $\alpha = 0.05$, $(n_1, n_2) = (10,15)$

图 1. 九种方法的覆盖率，$\alpha = 0.05$，$(n_1, n_2) = (10,15)$

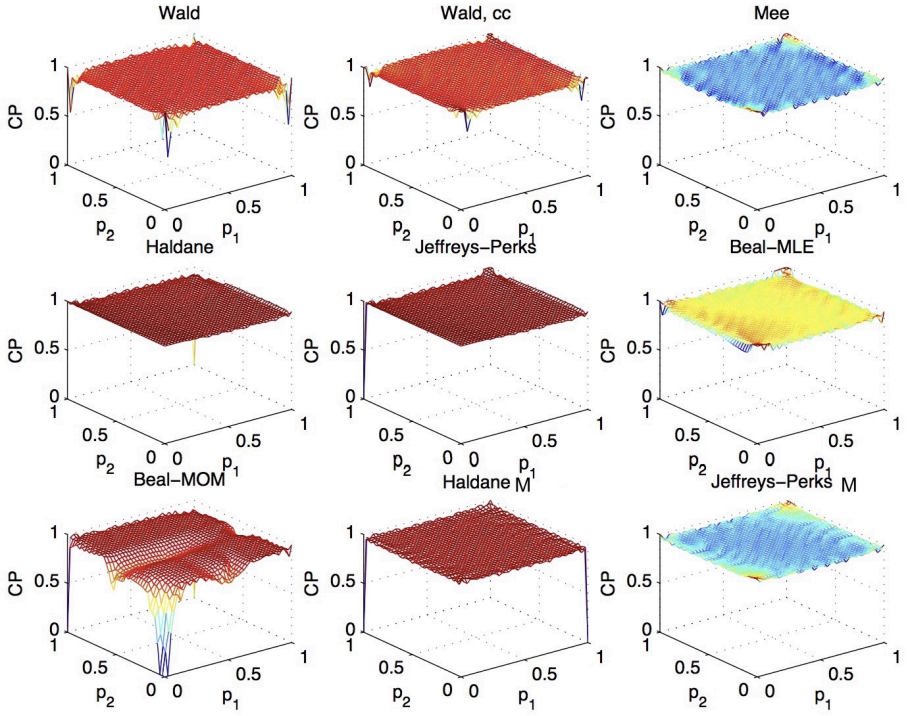

Figure 2. CP for nine methods, $\alpha = 0.05$, $(n_1, n_2) = (30,50)$

图 2. 九种方法的覆盖率，$\alpha = 0.05$，$(n_1, n_2) = (30,50)$

Beal 方法估计比例差的 95%和 99%置信区间。作为对比，我们也给出前述几种方法的估计结果，见表 3。我们发现，对于这种极端情况，除了 Mee 方法和提升的 Beal 方法，其它方法估计出的区间都超出了[−1,1]的合理范围，这种现象称为 overshoot 现象，而 Mee 方法和提升的 Beal 方法可以避免这种现象的发生。此外，Mee 方法和提升的 Haldane 方法具有相同的估计结果，但是在 3.2 节的模拟中，Mee 方法的平均区

Table 1. Min CP for nine methods
表 1. 九种方法的最小覆盖率

Method	$(n_1,n_2)=(10,15)$	$(n_1,n_2)=(30,50)$
Wald	0.2336	0.5313
Wald, cc	0.6497	0.7848
Mee	0.9139	0.9298
Haldane	0	0
Jeffreys-Perks	0	0
Beal-MLE	0.8290	0.8760
Beal-MOM	0	0
Haldane M	0	0
Jeffreys-Perks M	0.9026	0.9298

Table 2. Mean confidence interval length for nine methods ($\alpha=0.05$, $p_1=0.9$)
表 2. 九种方法的平均置信区间长度($\alpha=0.05$, $p_1=0.9$)

	Method	$p_2=0.05$	$p_2=0.45$	$p_2=0.85$
$(n_1,n_2)=(10,15)$	Wald	0.3654	0.5937	0.4741
	Wald, cc	0.5321	0.7604	0.6408
	Mee	0.4255	0.6078	0.5864
	Haldane	0.4375	0.5728	0.4473
	Jeffreys-Perks	0.4402	0.5883	0.5068
	Beal-MLE	0.4379	0.5906	0.5078
	Beal-MOM	0.4364	0.5657	0.4207
	Haldane M	0.4325	0.5656	0.4481
	Jeffreys-Perks M	0.4348	0.5875	0.5066
$(n_1,n_2)=(30,50)$	Wald	0.2369	0.3436	0.2847
	Wald, cc	0.2902	0.3969	0.3380
	Mee	0.2465	0.3477	0.3122
	Haldane	0.2465	0.3394	0.2788
	Jeffreys-Perks	0.2472	0.3427	0.2896
	Beal-MLE	0.2464	0.3428	0.2886
	Beal-MOM	0.2467	0.3344	0.2745
	Haldane M	0.2462	0.3384	0.2791
	Jeffreys-Perks M	0.2457	0.3426	0.2896

Table 3. The estimated 95% and 99% confidence intervals for the selected data
表 3. 实际数据的 95%和 99%置信区间

Method	95%		99%	
	CI	Length	CI	Length
Wald	[0.5351, 1.1315]	0.5964	[0.4414, 1.2252]	0.7838
Wald, cc	[0.4504, 1.2162]	0.7657	[0.3568, 1.3099]	0.9531
Mee	[0.4365, 0.9699]	0.5334	[0.3365, 0.9801]	0.6437
Haldane	[0.4473, 1.0315]	0.5842	[0.3097, 1.0623]	0.7526
Jeffreys-Perks	[0.4420, 1.0492]	0.6072	[0.3068, 1.0849]	0.7781
Beal-MLE	[0.4481, 1.0289]	0.5808	[0.3102, 1.0591]	0.7489
Beal-MOM	[0.4424, 1.0480]	0.6056	[0.3070, 1.0833]	0.7763
Haldane M	[0.4365, 0.9699]	0.5334	[0.3365, 0.9801]	0.6437
Jeffreys-Perks M	[0.4362, 0.9695]	0.5332	[0.3357, 0.9796]	0.6439

间长度比提升的 Haldane 方法要大。综上所述，实际中我们推荐使用 Mee 方法和我们提出的提升 Jeffreys-Perks 方法。

5. 结论

本文我们通过改良 Beal 区间中的讨厌参数的选取，提出了一种提升 Beal 区间方法，最优调节参数可以通过一个显式表达式给出，计算简单。实验模拟显示我们的方法具有大的覆盖率和最小覆盖率，平均区间长度也比较短。实际中，我们推荐使用 Mee 方法和我们提出的提升 Jeffreys-Perks 方法。

基金项目

本文为“广东海洋大学人文社会科学项目：二项抽样下两独立总体的比例差的统计推断”项目成果。

参考文献

[1] Fleise, J.L. (1981) Statistical Methods for Rates and Proportions. A Wiley Publication in Applied Statistics, 16(2), 326-327.

[2] Mee, R.W. (1984) Confidence Bounds for the Difference between Two Probabilities. *Biometrics*, **40**, 1175-1176.

[3] Beal, S.L. (1987) Asymptotic Confidence Intervals for the Difference between Two Binomial Parameters for Use with Small Samples. *Biometrics*, **43**, 941-950. https://doi.org/10.2307/2531547

[4] Roths, S.A. and Tebbs, J.M. (2006) Revisiting Beal's Confidence Intervals for the Difference of Two Binomial Proportions. *Communications in Statistics-Theory and Methods*, **35**, 1593-1609. https://doi.org/10.1080/03610920600683622

[5] Newcombe, R.G. (1998) Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods. *Statistics in Medicine*, **17**, 873-890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

[6] Wallenstein, S. (1997) A Non-Iterative Accurate Asymptotic Confidence Interval for the Difference between Two Proportions. *Statistics in Medicine*, **16**, 1329-1336. https://doi.org/10.1002/(SICI)1097-0258(19970630)16:12<1329::AID-SIM567>3.0.CO;2-I

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org