二项分布参数的区间估计

朱永生

（高能物理研究所、中国科学院大学）

《粒子物理数据分析基础与前沿》研讨会

2016.9.11~14，　北京

1. 动机

• 二项分布 ---- 随机试验的结果只有“成功”“失败”两种，

n 次随机试验中，“成功”次数为 s，s 为二项分布变量，概率分布为

$$\binom{n}{s} p^s (1-p)^{n-s}$$

• 二项分布参数 p ---- 表示一个随机事件被判定为“成功”的概率，

对应于探测器探测效率或事例判选效率。需要确定其误差：

$$p = \hat{p}_{-\sigma_-}^{+\sigma_+}. \tag{1.1}$$

$\left[\hat{p}-\sigma_-, \hat{p}+\sigma_+\right]$包含 p 真值的概率为 0.6827。

• 不对称性参数 A 及误差：

$$A = \frac{f-b}{f+b} = \frac{f-b}{n} = 2\frac{f}{n} - 1. \tag{1.2}$$

$$E(A) = 2p-1, \qquad \hat{A} = 2\hat{p} - 1 \tag{1.3}$$

$$\sigma_{\hat{A}} = 2\sigma_{\hat{p}}. \tag{1.4}$$

2. 试验总数 *n* 为固定常数时 *p* 的区间估计

用 MC 模拟样本确定 p 时 n 为固定常数。

2.1 Wald 区间

》 二项分布变量考虑为正态近似（n 充分大，p 不接近 0 或 1 的情形下是较好的近似）

- 点估计 $\hat{p}=\hat{s}/n$

  性质 $E(\hat{p})=p,\quad V(\hat{p})=p(1-p)/n.$

  令 $$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}\sim N(0,1)$$

- 由标准正态变量 u 的性质： $P\left(-z_{\alpha/2}\le u\le z_{\alpha/2}\right)=1-\alpha$

  $Z_{\alpha/2}$ 是 N(0,1) 双侧 α 分位数

  求得 p 的 $CL=\gamma=1-\alpha$ 的置信区间为

$$\begin{aligned}p(\gamma)&=\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\\&=\frac{1}{n}\left(\hat{s}\pm Z_{\alpha/2}\sqrt{\frac{\hat{s}(n-\hat{s})}{n}}\right).\end{aligned}\tag{2.1}$$

- $\gamma=1-\alpha=0.6827$ 时，对应于 $Z_{\alpha/2}=1$。上式简化为一倍标准误差区间

$$p(1\sigma)=\frac{1}{n}\left[\hat{s}\pm\sqrt{\hat{s}-\frac{\hat{s}^2}{n}}\right].\tag{2.2}$$

》 优缺点：计算简单。　涵盖概率不足。

当 $\hat{p}\sim 0$ 或 ~ 1，涵盖概率明显低于名义置信水平。

当 $\hat{p}=0$ 或 $=1$，置信区间长度为 0。

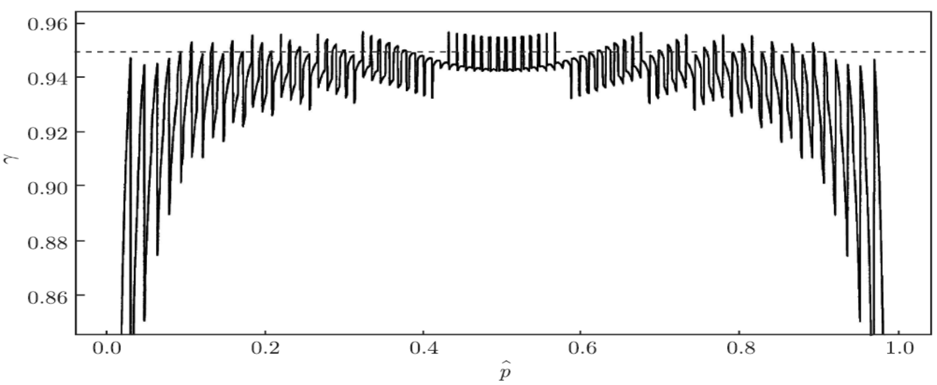

图 1　名义置信水平 95%的 Wald 区间的涵盖概率 γ 与 $\hat{p}$ 的关系（n = 100）

2.2 Wilson 区间

Wilson E B, J. Am. Stat. Assoc. 1927(22):209, http://www.jstor.org/stable/2276774

》 认为 $$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\sim N(0,1)$$ （注意与 Wald 区间的差别）

》 参数 p 的置信水平 $\gamma=1-\alpha$ 的 Wilson 区间为

$$p(\gamma)=\frac{\tilde{p}+T/2}{1+T}\pm\frac{\sqrt{\tilde{p}(1-\tilde{p})T+T^2/4}}{1+T},\tag{2.3a}$$

$$\hat{p}=\frac{\tilde{p}+T/2}{1+T}=\frac{\hat{s}+(Z_{\alpha/2})^2/2}{n+(Z_{\alpha/2})^2},\quad \tilde{p}=\frac{\hat{s}}{n},\quad T=\frac{(Z_{\alpha/2})^2}{n}.$$

或直接写为

$$p(\gamma)=\frac{\left[\tilde{p}+\frac{z_{\alpha/2}^2}{2n}\right]\pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}+\frac{z_{\alpha/2}^2}{4n^2}}}{1+z_{\alpha/2}^2/n},\tag{2.3b}$$

》当置信水平 $\gamma=1-\alpha=0.6827$ 时， $z_{\alpha/2}=1$，上式简化为一倍标准误差区间

$$p(1\sigma)=\frac{1}{n+1}\left[(\hat{s}+0.5)\pm\sqrt{\hat{s}-\frac{\hat{s}^2}{n}+0.25}\right].\tag{2.4}$$

》特点

当 $\hat{s}=0,n$ 的极端情形下，Wilson 区间的参数 p 的估计值 $\hat{p}$ 分别为 >0 和 <1 的值

（当 n 次试验中成功的次数 $\hat{s}=0$， n+1 次试验中成功的次数 $\hat{s}$ 有可能不等于 0;

当 n 次试验中成功的次数 $\hat{s}=n$,n+1 次试验中成功的次数 $\hat{s}$ 有可能不等于 n+1。）

当置信水平 $\gamma=1-\alpha=0.6827$ 时

$\hat{s}=0$： $\hat{p}=\dfrac{0.5}{n+1}$， $\sigma=\dfrac{0.5}{n+1}$; （区间长度不为 0）

$\hat{s}=n$： $\hat{p}=\dfrac{n+0.5}{n+1}$， $\sigma=\dfrac{0.5}{n+1}$. （区间长度不为 0）

》优缺点：计算简单。几乎可用于 p ~ 0 和 p ~ 1 附近之外的一切 p 值和一切 n 值。

涵盖概率高于 Wald 区间的涵盖概率，平均地接近于名义置信水平。

在 p ~ 0 和 p ~ 1 附近，涵盖概率低于名义置信水平。

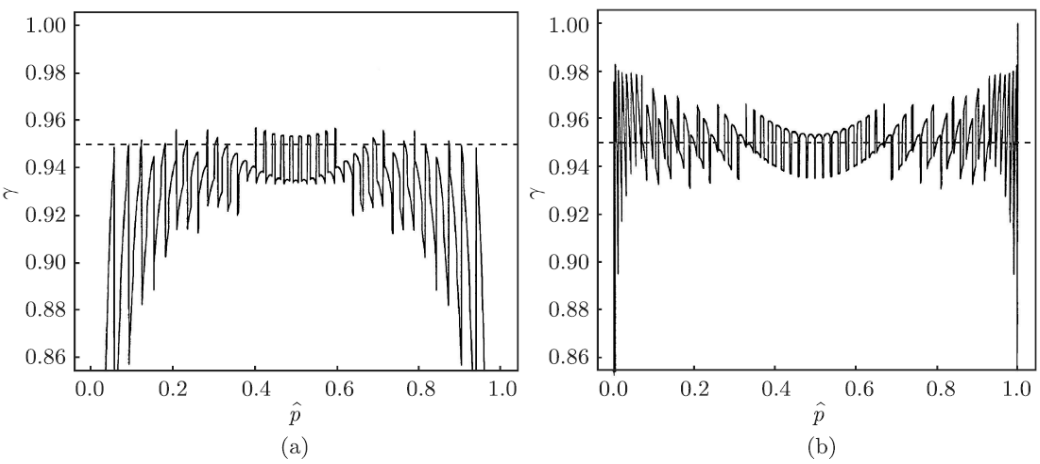

图 2　名义置信水平 95%的 Wald 区间（左图）和 Wilson 区间（右图）的涵盖概率 γ 与 $\hat{p}$ 的关系（$n=50$）

2.3 Agresti-Coull（A-C）区间

Brown L D,　Cai T T,　DasGupta a,　Stat. Science 2001(16):101;

Agresti A,　Coull B A,　Am. Stat.　1998(52):119

》也称为校正 Wald 区间（Adjusted Wald interval）

即 Wald 公式中的 $\hat{s}$ 用 $\tilde{s}=\hat{s}+z_{\alpha/2}^2/2$ 代替，

n 用 $\tilde{n}=n+z_{\alpha/2}^2/2$ 代替。

$CL=\gamma=1-\alpha$ 的置信区间为

$$p(\gamma)=\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{\tilde{n}}}, \tag{2.5a}$$

$$\hat{p}=\frac{\tilde{s}}{\tilde{n}}\equiv\frac{\hat{s}+z_{\alpha/2}^2/2}{n+z_{\alpha/2}^2}=\frac{\tilde{p}+\dfrac{z_{\alpha/2}^2}{2n}}{1+\dfrac{z_{\alpha/2}^2}{n}},\quad \tilde{p}=\frac{\hat{s}}{n}.$$

或直接写为

$$p(\gamma)=\frac{\tilde{p}+\dfrac{z_{\alpha/2}^2}{2n}}{1+\dfrac{z_{\alpha/2}^2}{n}}\pm\frac{z_{\alpha/2}}{1+\dfrac{z_{\alpha/2}^2}{n}}\sqrt{\frac{\left(\tilde{p}+\dfrac{z_{\alpha/2}^2}{2n}\right)\left(1-\tilde{p}+\dfrac{z_{\alpha/2}^2}{2n}\right)}{n+\dfrac{z_{\alpha/2}^2}{2}}}. \tag{2.5b}$$

》置信水平 $\gamma=1-\alpha=0.6827$ 时， $z_{\alpha/2}=1$，上式简化为一倍标准误差区间

$$p(1\sigma)=\frac{1}{n+1}\left[(\hat{s}+0.5)\pm\sqrt{\frac{(\hat{s}+0.5)(n-\hat{s}+0.5)}{n+0.5}}\right]. \tag{2.6}$$

》优缺点：计算简单。适用于几乎一切 n 值和一切 p 值。

中心值 $\hat{p}$ 与 Wilson 区间中心值相等。

区间宽度比 Wilson 区间略宽（保守）。

涵盖概率略高于 Wilson 区间的涵盖概率，且接近于名义置信水平。

特别在 p ~ 0 和 p ~ 1 附近，涵盖概率高于 Wilson 区间。

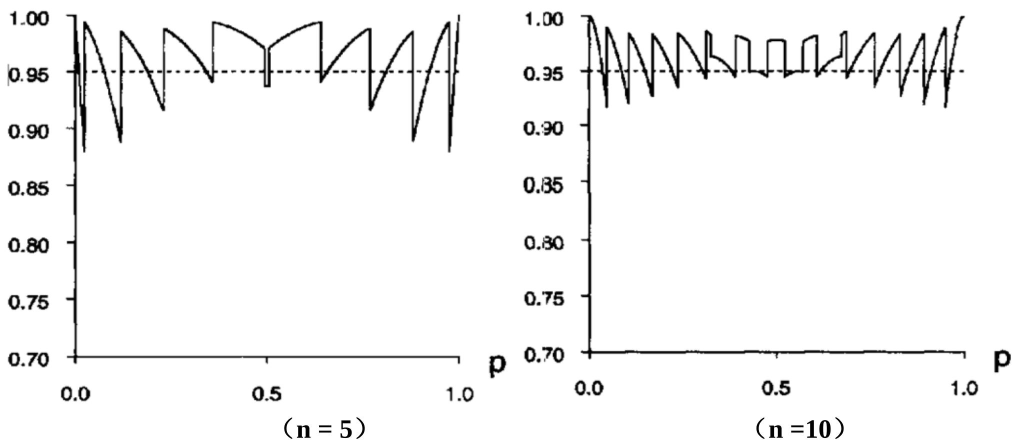

图 3　名义置信水平 95%的 A-C 区间的涵盖概率 γ 与 $\hat{p}$ 的关系

》特点

- $\hat{s}=0,n$ 的极端情形下，A-C 区间参数 p 的估计值 $\hat{p}$ 分别为 >0 和 <1 的值
- 置信水平 $\gamma=1-\alpha=0.6827$ 时

》 $\hat{s}=0$： $\hat{p}=\dfrac{0.5}{n+1}$， $\sigma=\dfrac{\sqrt{0.5}}{n+1}$；

为避免 $p_l(1\sigma)$ 为负，改为

$\hat{s}=0$： $\hat{p}=\dfrac{0.5}{n+1}$， $\sigma_+=\dfrac{\sqrt{0.5}}{n+1}$， $\sigma_-=\dfrac{0.5}{n+1}$． （区间长度不为 0）

》 $\hat{s}=n$： $\hat{p}=\dfrac{n+0.5}{n+1}$， $\sigma=\dfrac{\sqrt{0.5}}{n+1}$．

为避免 $p_u(1\sigma)>1$，改为

$\hat{s}=n$： $\hat{p}=\dfrac{n+0.5}{n+1}$， $\sigma_+=\dfrac{0.5}{n+1}$， $\sigma_-=\dfrac{\sqrt{0.5}}{n+1}$． （区间长度不为 0）

2.4 Clopper-Pearson（C-P）区间

Clopper C J, Pearson E S, Biometrika 1934 (26):404-413

》构建置信水平 $\gamma=1-\alpha$ 的**严格中心置信区间** $[p_l, p_u]$ 的方法

- 给定 n 和 p 值对应的二项分布变量 s 的概率分布为 $\binom{n}{s}p^s(1-p)^{n-s}$
- 给定 n 值，对每个 p 值，选择具有给定积分概率（例 γ =1- α = 68.3%）的 s 值的一个区间 $[s_l, s_u]$：

$$\sum_{s=0}^{s_l}\binom{n}{s}p^s(1-p)^{n-s}\le\frac{\alpha}{2}\ge\sum_{s=s_u}^{n}\binom{n}{s}p^s(1-p)^{n-s}$$

- 给定 n 值，将一切 p 值对应的这一区间连接起来形成 p--s 置信带。
- 该置信带与直线 $s=\hat{s}$ 的两个交点对应于置信概率 γ 的区间 [p_l, p_u]。

》置信水平 $\gamma=1-\alpha$ 的中心置信区间为 $[p_l, p_u]$：

$$p_l(\gamma)=B_{\alpha/2}(\hat{s}, n-\hat{s}+1),\quad p_u(\gamma)=B_{1-\alpha/2}(\hat{s}+1, n-\hat{s}), \tag{2.7a}$$

其中 $B_\alpha(\nu_1,\nu_2)$ 表示自由度 ν_1,ν_2 的贝塔 (Beta) 分布的 α 分位数。

》利用 $F(\nu_1,\nu_2)$ 分布的 α **分位数（quantile）** $q_\alpha(\nu_1,\nu_2)$ 来表示：

贝塔分布和 F 分布的 α 分位数之间存在以下关系：

$$B_\alpha(\nu_1,\nu_2)=\left(1+\frac{\nu_2}{\nu_1}q_{1-\alpha}(2\nu_2, 2\nu_1)\right)^{-1},$$

故有

$$p_l(\gamma)=\left(1+\frac{n-\hat{s}+1}{\hat{s}}q_{1-\alpha/2}\big(2(n-\hat{s}+1), 2\hat{s}\big)\right)^{-1},$$
$$p_u(\gamma)=\left(1+\frac{n-\hat{s}}{\hat{s}+1}q_{\alpha/2}\big(2(n-\hat{s}), 2(\hat{s}+1)\big)\right)^{-1}. \tag{2.7b}$$

》利用 $F(\nu_1,\nu_2)$ 分布的**上侧 *a* 分位数** $f_\alpha(\nu_1,\nu_2)$ 来表示：

上侧 *a* 分位数 $f_\alpha(\nu_1,\nu_2)$ 与 *a* 分位数 $q_\alpha(\nu_1,\nu_2)$ 的关系为

$$f_\alpha(\nu_1,\nu_2)=q_{1-\alpha}(\nu_1,\nu_2),$$

故有

$$p_l(\gamma)=\left(1+\frac{n-\hat{s}+1}{\hat{s}}f_{\alpha/2}\left(2(n-\hat{s}+1),2\hat{s}\right)\right)^{-1},$$
$$p_u(\gamma)=\left(1+\frac{n-\hat{s}}{\hat{s}+1}f_{1-\alpha/2}\left(2(n-\hat{s}),2(\hat{s}+1)\right)\right)^{-1}. \tag{2.7c}$$

》 $\gamma=0.6827$ 的一倍标准误差区间

$$p_l(1\sigma)=\left(1+\frac{n-\hat{s}+1}{\hat{s}}f_{0.15865}\left(2(n-\hat{s}+1),2\hat{s}\right)\right)^{-1},$$
$$p_u(1\sigma)=\left(1+\frac{n-\hat{s}}{\hat{s}+1}f^{-1}_{0.15865}\left(2(\hat{s}+1),2(n-\hat{s})\right)\right)^{-1}. \tag{2.8}$$

》 当 $\hat{s}=0$， $p_l(\gamma)=0$；　　当 $\hat{s}=n$， $p_u(\gamma)=1$。

》 （C-P）区间并不要求有点估计值 $\hat{p}$，可采用 2.2 节 Wilson 区间的 $\hat{p}$（**式(2.3a)**）。

》 结果： $\hat{p}^{+\sigma_+}_{-\sigma_-}$， $\sigma_+=p_u(1\sigma)-\hat{p}$, $\sigma_-=\hat{p}-p_l(1\sigma)$.

》 优缺点：涵盖概率过度，安全；　　计算相对简单。

2.5 似然比（L-R）区间

Cousins R D et al., Nucl. Instr. Meth. in Phys. Resear. 2010(A612):388

》构建置信区间的 FC 方法。

- 给定 n 值，对任一特定的 p 值，定义似然比

$$R(p,s)=\frac{B(s;n,p)}{B\left(s;n,\left(\frac{s}{n}\right)\right)}=\frac{p^s(1-p)^{n-s}}{\left(\frac{s}{n}\right)^s\left(1-\frac{s}{n}\right)^{n-s}}. \tag{2.9}$$

- 给定 n 值，对任一特定的 p 值，其置信水平 γ 的置信区间 $[s_l,s_u]$ 的确定与 FC 方法相同（似然比顺序求和）。由此构成 p--s 置信带。
- 该置信带与直线 $s=\hat{s}$ 的两个交点对应于 $[p_l,p_u]$。

》（L-R）区间并不要求有点估计值 $\hat{p}$，可采用 2.2 节 Wilson 区间的 $\hat{p}$（**式(2.3a)**）。

》优缺点：涵盖概率过度，安全；但比（C-P）区间短；区间的构建步骤较为繁复。

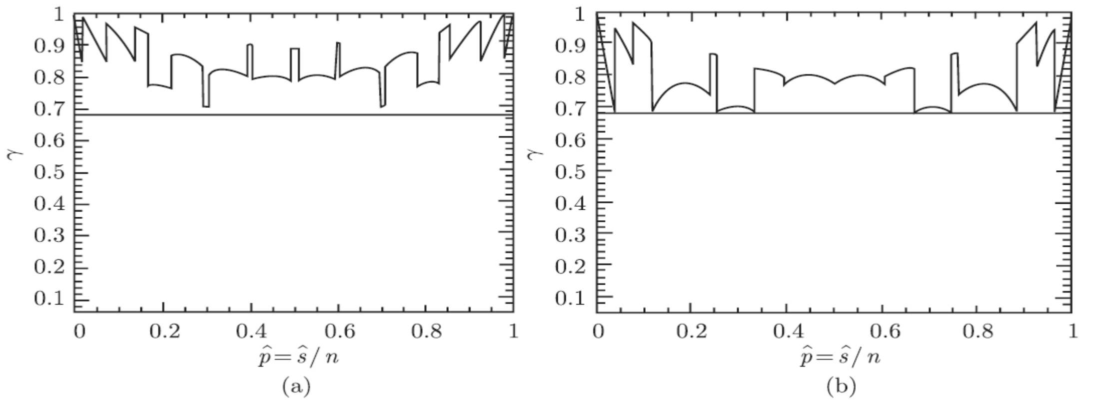

图 4　名义置信水平 68.27%的 C-P 区间（左图）和 L-R 区间（右图）的涵盖概率 γ（$n=10$）

3. 试验总数 *n* 为泊松变量时 *p* 的区间估计

》事例总数 n 为期望值 μ 的泊松变量时，

- 利用实验数据确定事例判选效率属于此情形。
- 利用实验数据确定共振态衰变分支比可能属于此情形。

》3 个变量 s、b、n 的联合概率可表示为

s 和 b 为两个独立的、期望值为 $\mu p \equiv \mu_s$ 和 $\mu q \equiv \mu(1-p) \equiv \mu_b$ 的泊松概率的乘积

$$P(s,b,n) = B(s;n,p)\cdot P(n;\mu) = P(s;\mu p)\cdot P(b;\mu q). \tag{3.1}$$

- p 表示为随机变量 s 和 n 的期望值之比：

$$p = \mu_s/\mu. \tag{3.2}$$

- 定义 s 和 b 两个泊松期望值比 λ 为

$$\lambda = \mu_s/\mu_b. \tag{3.3}$$

- 考虑到

$$\mu = \mu_s + \mu_b, \tag{3.4}$$

得到二项分布参数 p 与泊松期望值比 λ 之间的关系：

$$p = \frac{\mu_s}{\mu} = \frac{\lambda}{\lambda+1} = 1-\frac{1}{\lambda+1}. \tag{3.5}$$

》 实验得到的测量值是两个相互独立的泊松变量 s 和 b 的观测值 $\hat{s}$ 和 $\hat{b}$

（或泊松变量 n 的观测值 $\hat{n}$， 因为 $\hat{n}=\hat{s}+\hat{b}$ ）。

通过 $\hat{s}$ 、$\hat{b}$ 、$\hat{n}$ 构建参数 λ 的置信区间，然后利用关系式(3.5)求得 p 的置信区间。

》 Price 和 Bonett **（P-B 区间）**

Price R M, Bonett D G, Comput. Stat. Data Anal., 2000(34):345

调整 Wald 对数-线性模型（adjusted Wald log-linear model）确定参数 λ 的置信区间。

- λ 置信水平 $\gamma = 1-\alpha$ 置信区间 $[\lambda_-, \lambda_+]$ 的公式：

$$\lambda_\pm = \frac{\hat{s}+0.5}{\hat{b}+0.5}\exp\left(\pm Z_{\alpha/2}\sqrt{\frac{1}{\hat{s}+0.5}+\frac{1}{\hat{b}+0.5}}\right). \tag{3.6}$$

- 利用变换式(3.5)求得 p 的上下限：

$$p_u = \frac{\lambda_+}{\lambda_+ + 1}, \quad p_l = \frac{\lambda_-}{\lambda_- + 1}. \tag{3.7}$$

- p 的估计值 $\hat{p}$ 可利用 Wilson 区间的 $\hat{p}$ 值（见(2.2a)式）。

》优缺点：计算简单。

涵盖概率很少低于名义置信水平。区间宽度比 C-P 区间要窄。

》特点

当 $\hat{s}=0, \hat{n}$ 的极端情形下，参数 p 的估计值 $\hat{p}$ 分别为 >0 和 <1 的值

（与 Wilson 区间的估计值相等。）

$$\hat{s}=0：\ \hat{p} = \frac{0.5}{\hat{n}+1}; \qquad \hat{s}=\hat{n}：\ \hat{p} = \frac{\hat{n}+0.5}{\hat{n}+1}.$$

4. 比较和小结

不同方法的 p 值的置信区间比较（CL = 0.6827 区间），n =10

数据	方法	$[p_l, p_u]$	区间长度	结果表述
n = 10 s = 0	Wald	[0.00, 0.00]	0.00	0.00 ± 0.00
	Wilson	[0.00, 0.09]	0.09	0.05 ± 0.05
	A - C	[0.00, 0.16]	0.16	$0.05^{+0.11}_{-0.05}$
	C - P	[0.00, 0.17]	0.17	$0.05^{+0.12}_{-0.05}$
	P - B	**[0.01, 0.17]**	**0.16**	$0.05^{+0.11}_{-0.04}$
n = 10 s = 1	Wald	[0.00, 0.20]	0.20	0.10 ± 0.10
	Wilson	[0.04, 0.24]	0.20	0.14 ± 0.10
	A - C	[0.03, 0.24]	0.21	0.14 ± 0.11
	C - P	[0.02, 0.29]	0.27	$0.14^{+0.15}_{-0.12}$
	P - B	**[0.06, 0.27]**	**0.21**	$0.14^{+0.13}_{-0.08}$
n = 10 s = 5	Wald	[0.34, 0.66]	0.32	0.50 ± 0.16
	Wilson	[0.35, 0.65]	0.30	0.50 ± 0.15
	A - C	[0.35, 0.66]	0.31	0.50 ± 0.16
	C - P	[0.31, 0.69]	0.38	0.50 ± 0.19
	P - B	**[0.35, 0.65]**	**0.29**	0.50 ± 0.15
n = 10 s = 9	Wald	[0.80, 1.00]	0.20	0.90 ± 0.10
	Wilson	[0.76, 0.96]	0.20	0.86 ± 0.10
	A - C	[0.76, 0.97]	0.21	0.86 ± 0.11
	C - P	[0.71, 0.98]	0.27	$0.86^{+0.12}_{-0.15}$
	P - B	**[0.73, 0.94]**	**0.21**	$0.86^{+0.08}_{-0.13}$
n = 10 s = 10	Wald	[1.00, 1.00]	0.00	1.00 ± 0.00
	Wilson	[0.91, 1.00]	0.09	0.95 ± 0.05
	A - C	[0.89, 1.00]	0.11	$0.96^{+0.04}_{-0.07}$
	C - P	[0.83, 1.00]	0.17	$0.96^{+0.04}_{-0.13}$
	P – B	**[0.83, 0.99]**	**0.16**	$0.96^{+0.03}_{-0.13}$

不同方法的 p 值的置信区间比较（CL = 0.6827 区间），n =100

数据	方法	$[p_l, p_u]$	区间长度	结果表述
n = 100 s = 0	Wald	[0.00, 0.00]	0.00	0.00 ± 0.00
	Wilson	[0.000, 0.010]	0.010	0.005 ± 0.005
	A - C	[0.000, 0.012]	0.012	$0.005^{+0.007}_{-0.005}$
	C - P	[0.000, 0.018]	0.018	$0.005^{+0.013}_{-0.005}$
	P - B	**[0.001, 0.020]**	**0.019**	$0.005^{+0.015}_{-0.004}$
n = 100 s = 1	Wald	[0.00, 0.02]	0.02	0.01 ± 0.01
	Wilson	[0.004, 0.026]	0.022	0.015 ± 0.011
	A - C	[0.003, 0.027]	0.024	0.015 ± 0.012
	C - P	[0.001, 0.033]	0.032	$0.015^{+0.018}_{-0.014}$
	P - B	**[0.007, 0.033]**	**0.026**	$0.015^{+0.018}_{-0.008}$
n = 100 s = 10	Wald	[0.07, 0.13]	0.06	0.10 ± 0.03
	Wilson	[0.074, 0.134]	0.060	0.104 ± 0.030
	A - C	[0.074, 0.134]	0.060	0.104 ± 0.030
	C - P	[0.070, 0.137]	0.067	$0.104^{+0.033}_{-0.034}$
	P - B	**[0.077, 0.139]**	**0.062**	$0.104^{+0.035}_{-0.027}$
n = 100 s = 90	Wald	[0.87, 0.93]	0.06	0.90 ± 0.03
	Wilson	[0.866, 0.926]	0.060	0.896 ± 0.030
	A - C	[0.866, 0.926]	0.060	0.896 ± 0.030
	C - P	[0.863, 0.930]	0.067	$0.896^{+0.034}_{-0.033}$
	P - B	**[0.862, 0.923]**	**0.062**	$0.896^{+0.027}_{-0.035}$
n = 100 s = 100	Wald	[1.00, 1.00]	0.00	1.00 ± 0.00
	Wilson	[0.990, 1.000]	0.010	0.995 ± 0.005
	A - C	[0.988, 1.000]	0.012	$0.995^{+0.005}_{-0.007}$
	C - P	[0.982, 1.000]	0.018	$0.995^{+0.005}_{-0.013}$
	P – B	**[0.980, 0.999]**	**0.019**	$0.995^{+0.004}_{-0.015}$

结论：》 *n* 为小量 ：用 C - P 方法。

》 *n* 不为小量

- ***n* 为固定常数时：** $\hat{s}$ 不接近 $0, n$ 时，一般用 **A-C**（或 **Wilson**）方法；
  $\hat{s}$ 接近 $0, n$ 时，用 **C – P** 方法。
- ***n* 为泊松变量时：** 用 **P - B** 方法。

（谢谢！）