关于两个独立二项总体推断的 Bayes 样本容量的确定*

赵占平[1]，苗永旺[2,3]**，蔡 威[4]
（1. 黄淮学院，河南 驻马店 463000；2. 云南农业大学动物科学技术学院，云南 昆明 650201；
3. 云南大学，云南省生物资源保护与利用重点实验室，云南 昆明 650091；
4. 云南大学数学与统计学院，云南 昆明 650091）

摘要：样本容量的确定在现代生物医学研究以及在对两个独立的二项实验进行统计分析时，是经常遇到的一个问题。在实验设计阶段，往往需要计算最佳样本容量，目的是为了保证两个二项参数差的估计值与真实值的误差在所要求的范围内概率最大。巧妙地利用先验信息是实验设计的一个关键环节，目前正在广泛应用的样本量的计算公式在利用先验信息时通常采用点估计的形式。本文提出了确定样本容量的 Bayes 风险准则，给出了样本容量计算的 Monte Carlo 方法，并把这些方法应用到估计两个二项比例差的实验设计上。最后考虑了 0－1 损失函数和平方损失函数下计算样本容量的 Bayes 方法。

关键词：Bayes 样本容量；Beta 先验分布；Bayes 风险；边际后验分布
中图分类号：O 212 **文献标识码**：A **文章编号**：1004－390X（2008）05－0836－04

Bayesian Sample Size Determination for Inference of Independent Two Binomial Populations

ZHAO Zhan-ping[1], MIAO Yong-wang[2,3], CAI Wei[4]
（1. Huanghuai University, Zhumadian 463000, China;
2. Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming 650201, China;
3. Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming 650091, China;
4. Department of Mathematics and Statistics, Yunnan university, Yunnan Kunming 650091, China）

Abstract: Sample Size Determination is commonly encountered in modern medical studies and statistical analysis for two independent binomial experiments. During experimental design for estimating the difference between two binomial proportions, sample size is often calculated to ensure that the estimation will be within a desired distance from the true value with sufficiently high probability. Prudent use of the prior information is a crucial step of experimental planning. Most of sample size formulae in current use employ this information only in the form of point estimates, even though it is usually more accurately expressed as a distribution over a range of values. In this paper, Bayesian posterior risk criteria and Monte carlo method to determine sample size were proposed and these approaches to the design of an experiment to estimate the difference between two binomial proportions were applied. Finally, Bayesian methods for sample size calculation under 0～1 loss function and quadratic loss function were considered.

Key words: bayesian sample size; Beta prior distribution; Bayesian risk; marginal posterior distribution

收稿日期：2007－12－18 修回日期：2008－01－31 **通讯作者 E－mail：yongwangmiao999@yahoo.com.cn
*基金项目：国家自然科学基金项目（30660024）；云南省应用基础研究重点项目（2007C0003Z）；云南省应用基础研究计划面上项目（2006C0034M）；国际自然科学基金项目（10626048）。
作者简介：赵占平（1965－），男，河南遂平人，副教授，在读博士，主要从事数理统计方向的研究。

考虑试验对象的最佳样本容量（简称样本量）的大小是生物医学实验设计的关键环节。人们习惯于根据功效或后验置信区间的长度求样本量，利用功效或后验置信区间长度所得的样本容量计算公式其形式相似，但由于确定样本容量时所站的角度不同，经常导致样本容量大小的显著差异[1~2]。近年来，生物医学杂志上有许多文章在解决样本容量计算时基本上都采用区间估计方法或采用与功效有关的方法[3]。利用 Bayes 风险确定样本容量和其他方法相比不仅精确度较高，而且计算过程也较简单。随着计算机技术的快速发展及其在统计学中的广泛应用，该方法越来越显示其巨大的优越性。本文的研究目标是讨论在对两个二项比例之差做 Bayes 估计时如何确定样本容量才能使风险达到最小，并提出新的样本容量计算方法。

本文依据确定样本容量的 Bayes 风险准则，给出了样本容量计算的 Monte Carlo algorithm（迭代法）[4]。

1　确定样本量的 Bayes 风险准则

设 $X_1,X_2,\cdots,X_n$ 为独立同分布的样本，$X_1 \sim f(x_i,\theta), i = 1,2,\cdots,n, \theta \in \Theta$，$\theta$ 的先验分布为 $\pi(\theta)$，于是样本 $X = (X_1,X_2,\cdots,X_n)$ 的先验预测分布是：

$$f(x) = \int_\Theta f(x \mid (\theta)\pi(\theta)d(\theta) \tag{1}$$

给出样本值 x 时 θ 的后验分布密度为：

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta} \tag{2}$$

其中 $f(x \mid \theta)$ 是样本的联合密度函数[5~6]。若选择统计决策函数 $\delta(x)$，在损失函数 $L[\theta,\delta(x)]$ 下，$\delta(x)$ Bayes 后验风险为：

$$R_\pi(\delta \mid x) = E_{\theta \mid x}\{L[\theta,\delta(x)]\} = \begin{cases} \int_\Theta L[\theta,\delta(x)]\pi(\theta \mid x)\mathrm{d}(\theta), \theta \text{ 为连续} \\ \sum_i L[\theta_i,\delta(x)]\pi(\theta_i \mid x), \theta \text{ 为离散} \end{cases}$$

其中 $\pi(\theta \mid x)$ 为参数 θ 的后验分布密度，其 Bayes 风险为：

$$R_\pi(\delta) = \begin{cases} \int_X R_\pi(\delta \mid x) f(x)\mathrm{d}x, x \text{ 为连续} \\ \sum_x R_\pi(\delta \mid x) f(x), x \text{ 为离散} \end{cases} \tag{3}$$

在决策函数类 $\Delta = \{\delta(x)\}$ 中若存在这样一个决策函数 $\delta^*(x)$ 满足：

$$R_\pi(\delta^*) = \min R_\pi(\delta) \tag{4}$$

则称 δ^* 为决策函数类 $\Delta = \{\delta(x)\}$ 在 Bayes 风险准则下的最优决策函数，简称 Bayes 决策函数或 Bayes 解，在估计问题中，它就称为 Bayes 估计。当 Bayes 决策函数为 $\delta^*(x)$ 时，所求的样本量就是满足下列不等式的最小的正整数 n：

$$R_\pi(\delta^*) \leq \eta \tag{5}$$

这里的 η 是一个正常数[6]。

2　估计两个二项参数差时 Bayes 样本容量的确定

如果一种疾病在两个不同地区的发病率分别是 p_1 和 p_2，现分别从这两个地区的人群中随机抽取个 n 个体，x_1 和 x_2 分别表示这两个地区被抽人员的患病人数。于是 $X_1 \sim B(n,p_1), X_2 \sim B(n,p_2)$，并且 X_1 和 X_2 相互独立，二项参数 p_1 和 p_2 亦相互独立。假设 p_i 具有参数 (α_i,β_i) 的 beta 先验分布，即 $p_i \sim Beta(\alpha_i,\beta_i), i = 1,2$，这里感兴趣是当 n 取何值时对该疾病在这两个地区的发病率之差 $\theta = p_1 - p_2$ 实施 Bayes 估计其风险最小[7]。根据上述假设：

$$f(x_i \mid p_i) = \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i}, i = 1,2 \tag{7}$$

$$\pi(p_i) = \frac{1}{B(\alpha_i,\beta_i)} p_i^{\alpha_i-1}(1-p_i)^{\beta_i-1}, i = 1,2 \tag{8}$$

由(7)(8)两式可推出 $X = (X_1,X_2)$ 的先验预测分布：

$$\begin{aligned} f(x_1,x_2) &= \prod_{i=1}^{2}\int_0^1 f(x_i \mid p_i)\pi(p_i)\mathrm{d}p_i \\ &= \binom{n}{x_1}\int_0^1 p_1^{x_1}(1-p_1)^{n-x_1}\frac{p_1^{\alpha_1-1}(1-p_1)^{\beta_1-1}}{B(\alpha_1,\beta_1)}\mathrm{d}p_1 \\ &\quad \times \binom{n}{x_2}\int_0^1 p_2^{x_2}(1-p_2)^{n-x_2}\frac{p_2^{\alpha_2-1}(1-p_2)^{\beta_2-1}}{B(\alpha_2,\beta_2)}\mathrm{d}p_2 \\ &= \prod_{i=1}^{2}\binom{n}{x_i}\frac{B(\alpha_i+x_i,n-x_i+\beta_i)}{B(\alpha_i,\beta_i)} \end{aligned} \tag{9}$$

$X = (X_1,X_2)$ 的后验分布为：

$$f(p_1,p_2 \mid x_1,x_2) = \frac{f(x_1,x_2 \mid p_1,p_2) \times \pi(p_1)\pi(p_2)}{f(x_1,x_2)} = \prod_{i=1}^{2}\frac{p_i^{x_i+\alpha_i-1}(1-p_i)^{n-x_i+\beta_i-1}}{B(\alpha_i+x_i,n-x_i+\beta_i)} \tag{10}.$$

2.1 损失函数为0－1损失函数

$$令\ L(\theta,\delta) = \begin{cases} 0, & |\delta-\theta| \le \varepsilon \\ 1, & |\delta-\theta| > \varepsilon \end{cases}$$

ε 是一充分小的正数，在此损失函数之下 $\delta(x)$ 的Bayes后验风险函数为：

$$\begin{aligned} R(\delta \mid x) &= E_{\theta \mid x}\{L[\theta,\delta(X)]\} \\ &= \int_{\Theta} L[\theta,\delta(X)]\pi(\theta \mid x)\mathrm{d}\theta \\ &= 1 - \int_{\delta-\varepsilon}^{\delta+\varepsilon}\pi(\theta \mid x)\mathrm{d}\theta \end{aligned} \tag{11}$$

由(11)可知，欲使 $R(\delta \mid x)$ 达到最小值，只有使 $\int_{\delta-\varepsilon}^{\delta+\varepsilon}\pi(\theta \mid x)\mathrm{d}\theta$ 达到最大。显然当 $\delta(x)$ 取 $\pi(\theta \mid x)$ 的最大值点(即 θ 最大后验估计)时，$\int_{\delta-\varepsilon}^{\delta+\varepsilon}\pi(\theta \mid x)\mathrm{d}\theta$ 可以达到最大，从而 $R(\delta \mid x)$ 的值达到最小。因此参数 $\theta = p_1 - p_2$ 的Bayes估计为的最大后验估计 $\hat{\theta}$ 即 $\pi(\theta \mid x)$ 的最大值点[4]。

为得到 θ Bayes估计，须先求 $\pi(\theta \mid x)$ 的最大值点。现作简单线性变换 $p_1 = p_1, p_2 = p_1 - \theta, J = \frac{\partial(p_1,p_2)}{\partial(p_1,\theta)} = 1$，所以有：

$$f(p_1,\theta \mid x) = f(p_1,p_1-\theta \mid x)\,|J| = \frac{p_1^{x_1+\alpha_1-1}(1-p_1)^{n-x_1+\beta_1-1}(p_1-\theta)^{x_2+\alpha_2-1}(1-p_1+\theta)^{n-x_2+\beta_2-1}}{B(\alpha_1+x_1,n-x_1+\beta_1)B(\alpha_2+x_2,n-x_2+\beta_2)} \tag{12}$$

此密度函数只有在下列直线所围成的区域内不等于零 $p_1 = 0, p_1 = 1, \pi_1 = \theta$ 和 $\pi_1 = \theta + 1$，由此可得到参数 θ 的后验边际分布：

$$\begin{aligned} \pi(\theta \mid x_1,x_2) &= \int_{\max(0,\theta)}^{\min(\theta+1,1)} f(p_1,\theta \mid x_1,x_2)\mathrm{d}p_1 \\ &= \int_{\frac{\theta+|\theta|}{2}}^{1+\frac{\theta}{2}-\frac{|\theta|}{2}} f(p_1,\theta \mid x_1,x_2)\mathrm{d}p_1 \end{aligned} \tag{13}$$

求出此积分是比较困难的，因为被积函数 $f(p_1,\theta \mid x_1,x_2)$ 是一个关于 p_1 的多项式其阶数比较高时计算起来比较麻烦。可以用一种比较合适的渐进方法如高斯求积法求出此积分，即 θ 的后验分布 $\pi(\theta \mid x_1,x_2)$，然后求出 θ 的最大后验估计 $\hat{\theta}$ 其最小Bayes风险为：

$$R(\hat{\theta}) = 1 - \sum_{x_1=0}^{n}\sum_{x_2=0}^{n}\int_{\hat{\theta}-\varepsilon}^{\hat{\theta}+\varepsilon}\pi(\theta \mid x_1,x_2)\mathrm{d}\theta f(x_1,x_2)$$

2.2 损失函数为平方损失函数

令 $L(\theta,\delta) = (\theta-\delta^2)$，则 θ 的Bayes估计 $\hat{\theta}$ 为后验分布 $\pi(\theta \mid x)$ 的均值，即 $\hat{\theta} = E(\theta \mid x)$，此时后验风险为：

$$R_{\pi}(\hat{\theta} \mid x) = E_{\theta \mid x}[L(\theta,\hat{\theta})] = E_{\theta \mid x}[\theta - E(\theta \mid x)]^2$$

这里 $x = (x_1,x_2)$，由上式可知 $\hat{\theta}$ 的后验风险即为 θ 的后验方差 $\mathrm{var}(\theta \mid x)$，根据 $x = (x_1,x_2)$ 的先验预测分布 $f(x_1,x_2)$ 可求出 $\hat{\theta}$ 的Bayes风险 $R_{\pi}(\hat{\theta}) = \sum_{x_1=0}^{n}\sum_{x_2=0}^{n}R_{\pi}(\hat{\theta} \mid x)f(x_1,x_2)$. 于是所求的样本量就是满足下列不等式的最小正整数 n.

$$R_{\pi}(\hat{\theta}) = \sum_{x_1=0}^{n}\sum_{x_2=0}^{n}R_{\pi}(\hat{\theta} \mid x)f(x_1,x_2) \le \eta \tag{14}$$

3 样本容量的计算

上一节所给的确定样本容量的准则一般来讲没有精确的计算公式和计算方法。因此，为了求得最佳Bayes样本容量必须进行随机模拟计算[8~10]。当Bayes准则关于 n_0 满足，而 $n = n_0 - 1$ 不满足时，则 n_0 就是所求样本容量. 对于每一个可能的 n 值都必须验证是否满足Bayes风险准则，相邻的两个 n 值后一个的选取依赖于前一个的结果。简单说来这个迭代过程就是利用Monte Carlo algorithm求来满足Bayes风险准则的样本容量。

当损失函数为0－1损失函数时所要求的样本容量就是满足下式的最小整数 n：

$$\sum_{x_1=0}^{n}\sum_{x_2=0}^{n}p\{\theta \in (\hat{\theta}-\varepsilon,\hat{\theta}+\varepsilon) \mid (x_1,x_2)\}f(x_1,x_2) \ge 1-\eta \tag{15}$$

其中：

$$\begin{aligned} &p\{\theta \in (\hat{\theta}-\varepsilon,\hat{\theta}+\varepsilon) \mid (x_1,x_2)\} \\ &= \int_{\hat{\theta}-\varepsilon}^{\hat{\theta}+\varepsilon}\pi(\theta \mid x_1,x_2)\mathrm{d}\theta \end{aligned}$$

$\pi(\theta \mid x_1,x_2)$ 是参数 θ 的后验分布可由(2)式给出，ε 是一个充分小的正数，其大小由研究者根据实际需要而事先设定，$f(x_1,x_2)$ 是 $x = (x_1,x_2)$ 的先验预测分布可通过(1)式给出。

求样本容量的Monte Carlo algorithm过程可概括如下：

(1) 取 n 的初始值为 n_0，首先从总体 $X[X \sim f(x)]$ 中抽取一个容量为 M 的样本 $X' = (X'_1, X'_2, \cdots, X'_M)$，对于样本值中的每一个 x'_j，$j = 1,2,\cdots,M$，求出相应的积分 $\int_{\hat{\theta}-\varepsilon}^{\hat{\theta}+\varepsilon}\pi(\theta \mid x'_j)\mathrm{d}\theta$，然后对所求出的 M 个积分值求平均值，若平均值大于或等于1

$-\eta$,而 $n = n_0 - 1$ 时,所得的平均值小于 $1-\eta$,则 $n - n_0$ 即为所求样本量,否则转入 2.

(2) 令 $n = n_0 + 1$, 重复 1 所进行的运算, 若所求的平均值大于或等于 $1-\eta$,则 $n = n_0 + 1$ 即为所求的样本量。否则,令 $n = n_0 + 2$ 重复进行以上过程直到积分平均值不小于 $1-\eta$ 为止, 这时所对应的 n 值即为所求样本量。

当损失函数取平方损失函数,即 $L(\theta,\delta) = (\theta - \delta)^2$ 时, Bayes 最小风险准则可表为:

$$R_\pi(\hat{\theta}) = \sum_{x_1=0}^{n}\sum_{x_2=0}^{n} \mathrm{var}(\hat{\theta} \mid x)\, f(x_1,x_2)$$

$$= \sum_{x_1=0}^{n}\sum_{x_2=0}^{n} \{E(\theta^2 \mid x) - [E(\theta \mid x)]^2\}\, f(x_1,x_2)$$

$$= \sum_{x_1=0}^{n}\sum_{x_2=0}^{n} \{\int_\Theta \theta^2 \pi(\theta \mid x)\mathrm{d}\theta - [\int_\Theta \theta\pi(\theta \mid x)\mathrm{d}\theta]^2\} \cdot f(x_1,x_2) \quad (16)$$

由上式可知, 利用Bayes风险准则求样本容量时采用 Monte Carlo algorithm 迭代法可如下进行:

(1) 取 n 的初始值 n_0, 首先从总体 $X[X \sim f(x_1,x_2)]$ 中随机抽取一个容量为 M 的样本 $X' = (X'_1, X'_2, \cdots, X'_M)$,对于样本值 $x' = (x'_1, x'_2, \cdots, x'_M)$ 中的每一个 x'_j, 求出 $\int_\Theta \theta^2\pi(\theta \mid x'_j)\mathrm{d}\theta$ 和 $\int_\Theta \theta\pi(\theta \mid x'_j)\mathrm{d}\theta$ 及其差 $\int_\Theta \theta^2\pi(\theta \mid x'_j)\mathrm{d}\theta - [\int_\Theta \pi(\theta \mid x'_j)\mathrm{d}\theta]^2$。

对这 M 个差求平均数, 若平均值小于或等于 η, 而 $n = n_0 - 1$ 时,平均值大于 η, 则 n_0 即为所求样本量。否则, 转入 2。

(2) 取 $n = n_0 + 1$,重复 1 中的计算过程, 若 $n = n_0 + 1$ 时上述平均值不超过 η, 则 $n_0 + 1$ 即为所求样本量, 否则再取 $n = n_0 + 2$, 再转入 1。这样反复迭代直到找到满足 Bayes 风险准则的正整数 n 为止。

4　结论

在生物医学实验中, 选择合适的样本容量直接关系到实验结果的可靠性。因为样本容量太小会妨碍获得必要的抽样信息, 样本量过大会造成资源浪费。本文针对两个独立的二项总体的 Bayes 样本容量的确定在传统的 ACC 准则、ALC 准则及 WOC 准则的基础上提出了 Bayes 风险准则, 给出了 0－1 损失和平方损失函数下最小 Bayes 风险的计算公式。直接利用此公式进行计算尽管面临繁重的计算负担甚至很难得到准确的计算结果, 但是利用计算机采用 Monte Carlo 方法计算结果却很容易得到。利用本文给出的规则和方法求样本容量不仅计算简便更重要的是可以保证风险达到最小。本文所提出问题要得到完全解决还需模拟研究和实例分析, 因篇幅所限这部分内容在另外一篇文章中做详细研究和讨论。总之, 本文给出的确定样本容量 Bayes 方法对于两个独立二项总体的 Bayes 推断是一种简单且行之有效的计算方法。

[参考文献]

[1] BRISTOL D R. Sample Size for Constructing Confidence Intervals and Testing Hypotheses [J]. Statistic in Medicine, 1989, 8: 803－811.

[2] GRIEVE A P. Confidence Intervals and Sample sizes [J]. Biometrics, 1991, 47: 1597－1603.

[3] EVANS S J W, MILLS P, DAWSON J. The End of Thevalues? [J]. British Heart Journal, 1988, 60: 177－180.

[4] 茆诗松, 王静龙, 濮晓龙, 等. 高等数理统计 [M]. 北京: 高等教育出版社, 2004.

[5] LAWRENCE J. Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination [J]. Statistics in medicine, 1997, 16: 769－781.

[6] BERGER J O. Statistical Decision Theory and Bayesian Analysis [M]. New York: Springer-Verlag, 1985.

[7] JOHNSON N, KOTS S. Continuous Univariate Distribution－2 [M]. New York: Wiley, 1988.

[8] ADOCK C J. A Bayesian Approach Tocalculating Sample size for Multinomial Sampling [J]. Satatistician, 1987, 36: 155－159.

[9] ADOCK C J. A Bayesian Approach to Calculating Sample size [J]. Satatistician, 1988, 37: 433－439.

[10] GOULD A L. Sample Size for Evevt Rate Equivalence trails Using Prior Information [J]. Satatistics in Medicine, 1993, 12: 2009－2023.