

# Exact sample size determination for binomial experiments

Elham Rahme<sup>a</sup>, Lawrence Joseph<sup>b,\*</sup>

<sup>a</sup> *Department of Mathematics and Statistics, McGill University, Burnside Hall, 805 Sherbrooke Street West, Montreal, Que., Canada, H3A 2K6*

<sup>b</sup> *Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada*

Received 5 December 1996; received in revised form 31 January 1997

---

## Abstract

In experiments designed to estimate a binomial parameter, sample sizes are often calculated to ensure that the point estimate will be within a desired distance from the true value with sufficiently high probability. Since exact calculations resulting from the standard formulation of this problem can be difficult, “conservative” and/or normal approximations are frequently used. In this paper, some problems with the current formulation are given, and a modified criterion that leads to some improvement is provided. A simple algorithm that calculates the exact sample sizes under the modified criterion is provided, and these sample sizes are compared to those given by the standard approximate criterion, as well as to an exact conservative Bayesian criterion. © 1998 Elsevier Science B.V.

*Keywords:* Binomial; Exact confidence intervals; Normal approximation; Sample size

---

## 1. Introduction

Let  $X$  be a binomial random variable with unknown probability of success  $\theta$ . What should the sample size  $n$  be so that the point estimate will be sufficiently close to  $\theta$  with high probability?

Many recent textbooks on sample size determination (for example, Desu and Raghavarao, 1990; and Lemeshow, et al., 1990) suggest basing sample size calculations for binomial experiments on criteria such as

$$P(|X/n - \theta| \leq d) \geq 1 - \alpha. \quad (1)$$

This formulation ensures that the sample size will be sufficient to estimate the true binomial parameter  $\theta$  by the usual unbiased point estimator  $\hat{\theta} = X/n$ , in the sense that

---

\* Correspondence address: Division of Clinical Epidemiology, Montreal General Hospital, Department of Medicine, 1650 Cedar Avenue, Montreal, Quebec, Canada H3G 1A4.

$|\hat{\theta} - \theta| \leq d$  with probability at least  $1 - \alpha$ . For suitably chosen  $x_1$  and  $x_2$ , the left-hand side of (1) is equal to

$$\sum_{x=x_1}^{x_2} p(x; \theta) = \sum_{x=x_1}^{x_2} \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad (2)$$

where  $n$  is the sample size. However, both the summand as well as  $x_1$  and  $x_2$  depend on the unknown value of  $\theta$ , making direct use of (2) and therefore (1) almost impossible in practice. One “conservative” solution (which we will show to not always be conservative), suggested by Desu and Raghavarao (1990) and others, is to assume that (2) is minimized when  $\theta = 0.5$ . More generally, if it is known that  $\theta \leq m < 0.5$  or  $\theta \geq m > 0.5$ , for some  $m$ , then an alternative solution would be to substitute  $\theta = m$  in (2). This would still be conservative, but guard against the possibility of using an unnecessarily large sample size if  $\theta = 0.5$  is used when in fact  $0 \leq \theta < 0.5$  or  $\theta \geq 0.5$ . The intuition behind labelling these substitutions “conservative” is that the variance function of a binomial random variable,  $n\theta(1 - \theta)$ , is maximized over the interval  $(a, b) \subset [0, 1]$  by the value in  $(a, b)$  closest to 0.5. However, this reasoning is only partially correct, since the effects of  $\theta$  on  $x_1$  and  $x_2$  are ignored in focussing only on the binomial variance.

It is also often suggested that the exact calculation in (2) can be replaced by that given by the normal approximation to the binomial distribution. Letting

$$y_i = \frac{x_i - n\theta}{\sqrt{n\theta(1 - \theta)}},$$

then

$$\sum_{x=x_1}^{x_2} \binom{n}{x} \theta^x (1 - \theta)^{n-x} \approx \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy. \quad (3)$$

The limits  $y_1$  and  $y_2$  are unknown, since  $\theta$  is unknown. However, conservative sample sizes are available by substituting  $\theta = 0.5$  or  $\theta = m$  as above, and using quantiles of the normal distribution to approximate  $y_1$  and  $y_2$ . This leads to the sample size formula

$$n = \left\lceil \frac{Z_{1-(\alpha/2)}^2 \theta(1 - \theta)}{d^2} \right\rceil, \quad (4)$$

where  $Z_{1-\alpha/2}$  is the usual standard normal upper  $100(1 - \alpha/2)\%$  quantile, and  $\lceil a \rceil$  denotes the smallest integer larger than  $a$ . In the case where  $\theta = 0.5$ , (4) reduces to  $n = \lceil Z_{1-\alpha/2}^2 / 4d^2 \rceil$ . These conservative solutions are correct only to the extent that the normal distribution approximates the exact underlying binomial probabilities. However, the degree to which this approximation affects the sample sizes is usually unknown.

In summary, criterion (1) is exact, but difficult or impossible to implement. Criterion (4) is trivial to implement, but provides only an approximate solution. This raises the

question of how much the sample size given by (4) may differ from an exact solution. It would therefore be useful to find an exact method that is easy to implement.

The plan of this paper is as follows. In Section 2, several anomalies associated with the criterion given by (1) are illustrated via examples. The examples also demonstrate that “conservative” approximations are sometimes not truly conservative. A modified exact criterion is presented in Section 3 along with an algorithm that illustrates its feasibility in practice. Sections 4 and 5 provide examples comparing the sample sizes given by the modified criterion to those provided by the most commonly used formula (4), and a conservative Bayesian criterion, respectively. A discussion is found in Section 6. Mathematical details are deferred to an appendix.

## 2. Anomalies

The main problem with using the standard formulation (1) is that  $\theta$  is unknown, and it is difficult or impossible to ascertain which value of  $\theta$  is the most conservative. Consider the following example.

**Example 1.** Let  $d = 0.1$ ,  $1 - \alpha = 0.6$ ,  $n = 5$ , and  $\theta = 0.4$ . Then (2) becomes  $p(x = 2) = 0.3456$  while for  $\theta = 0.5$ , (2) reduces to  $p(x = 2) + p(x = 3) = 0.625$ . Therefore, the minimum probability is not always attained by substituting  $\theta = 0.5$ , that is, the value that provides the maximum variance is not always the most conservative in the sense of minimizing (2). There are also other problems associated with the use of criterion (1).

**Example 2.** Suppose  $d = 0.1$ ,  $1 - \alpha = 0.6$  and  $n = 5$ . As above, for  $\theta = 0.5$ , (2) gives 0.625. There are several anomalies associated with this situation. First, consider the same calculation, but replace  $\theta = 0.5$  by  $\theta = 0.50000001$ . In this case, (1) becomes  $p(x = 3) = 0.3125$ . Thus, the discrete nature of the binomial distribution is such that a little disturbance in  $\theta$  reduces the probability by half. Since we will rarely know  $\theta$  a priori with a high degree of accuracy, this may be a serious concern. Second, restore  $\theta = 0.5$ , but let  $d = 0.0999999999$ . Then (2) becomes 0! Hence, a small decrease in  $d$  costs all of the probability. If strict inequality is considered in Eq. (1), that is,

$$p(|X/n - \theta| < d) \geq 1 - \alpha,$$

the probability is again 0. Furthermore, if  $\theta = 0.5$ , the smallest  $n$  that gives

$$p(|X/n - 0.5| < 0.1) \geq 0.6$$

is  $n = 5$ , but if we take  $n = 6$  then (2) becomes  $p(x = 3) = 0.3125$ , that is, half of the probability is lost when considering a *larger* sample.

While for ease of exposition the above examples featured only small values of  $n$ , Table 1 of Section 4 indicates that similar problems persist for much larger  $n$ . In the next section a modified criterion is suggested to replace (1) that improves upon some

of the undesirable features illustrated above. In particular, sample sizes from the modified criterion can be calculated exactly via an easy program algorithm, and the problems due to the anomalies illustrated by the above examples are diminished.

### 3. Modified criterion

For any given  $\theta$ ,  $\alpha$  and  $d$ , criterion (1) ensures that

$$\Pr\{-d \leq \hat{\theta} - \theta \leq +d\} \geq 1 - \alpha,$$

where  $\hat{\theta} = X/n$  is the usual binomial maximum likelihood estimator. However, one could also consider

$$\Pr\{-a \leq \hat{\theta} - \theta \leq +b\} \geq 1 - \alpha,$$

with  $|a + b| \leq 2d$ . Therefore, instead of the interval of length  $2d$  centered at  $\hat{\theta}$ , the highest density interval of length  $\leq 2d$  containing  $\hat{\theta}$  is considered. This is similar to switching to exact binomial confidence intervals rather than those based on the normal approximation (3), commonly used when  $n$  is small or  $\theta$  is near 0 or 1. Let

$$D_\theta = \{\text{all intervals } I \text{ such that } \theta \in I \text{ and } l(I) \leq 2d\},$$

where  $l(I)$  denotes the length of the interval  $I$ . Then the sample size can be defined as the minimum  $n$  satisfying

$$\inf_{\theta} \left\{ \sup_{I \in D_\theta} \sum_{(k/n) \in I} p(k/n; \theta) \right\} \geq 1 - \alpha, \quad (5)$$

where  $k$  is an integer, and the infimum is over the range of possible values for  $\theta$ .

An algorithm to calculate exact sample sizes using this modified criterion is given below. The algorithm is motivated by Theorem 1, which is proved in the appendix.

**Theorem 1.** Let  $d$ ,  $\alpha$  and  $n$  be given. Let  $i = \lfloor 2nd \rfloor$ , that is,  $i$  is the largest integer smaller than  $2nd$ . For  $j \in \{1, \dots, n - i\}$ , define

$$r_j = \frac{\binom{n}{j-1}}{\binom{n}{i+j}} \quad \text{and} \quad \theta_j = \frac{r_j^{1/(i+1)}}{(1 + r_j^{1/(i+1)})}.$$

Denote by  $H(\theta)$  the probability content of a highest-density interval corresponding to  $\theta$ . Then

$$\inf\{H(\theta); 0 \leq \theta \leq 1\} = \min(\{H(\theta_j); 1 \leq j \leq n - i\}).$$

Theorem 1 states that in order to calculate the minimum highest-density region over  $\theta \in [0, 1]$ , it suffices to consider only the  $n - i$  values of  $\theta_j$ . Similarly, if  $\theta \leq m$ ,

only  $H(\theta_j)$ ,  $\theta_j < m$  need to be computed, and the sample size is the smallest  $n$  such that

$$\min(\{H(\theta_j): \theta_j < m\}, H(m)) \geq 1 - \alpha.$$

This suggests the following algorithm:

1. Given  $d, m \leq 0.5$  and  $\alpha$ , select an initial guess for the sample size  $n$ . (If  $m > 0.5$ , use with  $m' = 1 - m$  in place of  $m$ . The standard formula (4) with  $\theta = m$  could be used to obtain the initial guess.)
2. Calculate  $i = \lfloor 2nd \rfloor$ .

$$r_j = \frac{\binom{n}{j-1}}{\binom{n}{i+j}}, \quad \text{and} \quad \theta_j = \frac{r_j^{1/(i+1)}}{(1 + r_j^{1/(i+1)})}, \quad j = 1, 2, \dots, n - i.$$

3. Calculate  $H(\theta_j) = \sum_{k=j}^{j+i} \binom{n}{k} \theta_j^k (1 - \theta_j)^{n-k}$ .
4. Letting  $s = \max\{j: \theta_j \leq m\}$ , calculate  $H(m) = \sum_{k=s}^{s+i} \binom{n}{k} m^k (1 - m)^{n-k}$ .
5. (a) If there is no bound for  $\theta$ , calculate  $p_{\min} = \min(\{H(\theta_j): 1 \leq j \leq n - i\})$ .  
 (b) If  $\theta < m < 0.5$ , calculate  $p_{\min} = \min(\{H(\theta_j): \theta_j \leq m\}, H(m))$ .
6. Repeat steps 2-5 with a new values for  $n$ , until  $p_{\min} \geq 1 - \alpha$  for  $n$  but not for  $n - 1$ .  
 For example, subsequent values for  $n$  can be selected via a bisectional search algorithm.

The above algorithm is straightforward to program in most programming languages.

#### 4. Examples

Consider again Example 1 from Section 2. This example illustrated that substituting  $\theta = 0.5$  for the unknown  $\theta$  does not guarantee a conservative probability calculation. It is also true that  $\theta = 0.5$  is not necessarily conservative when using the modified criterion. However, Theorem 1 states that the minimum highest-density probability interval occurs when  $\theta = \theta_j$  for some  $j \leq n - i$ , so that the exact minimum probability can easily be found, which is not, in general, the case when using (1).

Under the modified criterion, small disturbances in  $\theta$  do not greatly affect the probabilities, as was the case for the standard formulation in Example 2. For example, using  $\theta = 0.50000001$ , the highest-density interval remains at 0.625. Small decreases in  $d$  also do not affect the highest-density interval probabilities when  $k/n < d < (k + 1)/n$  for some integer  $k$ . However, when  $d = k/n$ , a small decrease in  $d$  produces a loss of one of the end points of the interval. In contrast, under Eq. (1) both end points are lost. When the sample size increases to 6, the probability under the modified criterion is 0.5469 while under Eq. (1) it is 0.3125.

Table 1

Sample sizes (SS) for various values of  $\alpha$ ,  $d$ , and  $m$ , using the normal approximation (4) and the modified criterion (5)

Example #	$(1 - \alpha)$	$d$	$m$	SS using (4)	SS using (5)	% difference
3	0.625	0.1	0.5	20	10	50
4	0.8	0.1	0.5	42	35	17
5	0.9	0.025	0.1	390	360	8
6	0.9	0.05	0.1	98	80	18
7	0.95	0.005	0.02	3012	2800	7
8	0.95	0.03	0.1	385	350	9
9	0.95	0.05	0.01	1825	1750	4
10	0.95	0.05	0.1	139	120	14
11	0.95	0.05	0.5	385	370	4
12	0.95	0.1	0.5	97	90	7

Table 1 provides ten additional examples, for a selection of values for  $1 - \alpha$ ,  $d$ , and  $m$ . The examples illustrate that in experiments requiring small samples, such as when  $\alpha$  and  $d$  are relatively large, the difference between the sample size computed exactly and the one computed using the normal approximation can be as much as 50%. More interestingly, the differences can still approach 20% even when  $1 - \alpha$  takes on the usual 0.9 or 0.95 values, and the sample sizes near 100. Furthermore, discrepancies of 4–10% can be found for sample sizes near 400, and persist up to sample sizes of 3000. In many cases, these differences may be of practical importance.

It is important to recognize that there are two distinct contributing factors to the differences in sample sizes observed in Table 1. The first is that in using a normal approximation to the exact binomial probabilities, one introduces some error. The second is that there has been a subtle but non-negligible change in the sample size criterion, in that the standard criterion is based on a symmetric interval, while the modified criterion employs highest-density intervals. Therefore, even if the normal distribution provided exactly correct probabilities, one would expect the modified criterion to provide lower sample sizes, owing to the efficiency of highest-density intervals.

## 5. An exact conservative Bayesian approach

Problems such as those described in Section 2 arise from the discrete nature of the binomial distribution. From a Bayesian point of view, if a continuous prior distribution such as the one from the standard conjugate Beta family is used to estimate a binomial parameter, then the posterior density is also continuous, and similar problems do not arise.

An exact Bayesian approach to sample size determination for a binomial parameter is given by Joseph et al. (1995). Let the prior distribution of  $\theta$  follow a Beta density

with parameters  $a$  and  $b$ , so that

$$f(\theta) = \frac{1}{B(a, b)} \theta^{(a-1)} (1 - \theta)^{(b-1)}, \quad 0 < \theta < 1,$$

where  $B(a, b)$  is the Beta function with parameters  $(a, b)$ . The posterior density for  $\theta$  given observed data  $x$  out of  $n$  binomial trials is

$$f(\theta|x, n, a, b) = \frac{1}{B(x + a, n - x + b)} \theta^{(x+a-1)} (1 - \theta)^{(n-x+b-1)}, \quad 0 < \theta < 1.$$

The experimenter typically would specify that  $\theta$  should fall in a highest posterior density interval of length  $2d$  with probability  $(1 - \alpha)$ . However, the posterior distribution depends on  $x$ , which is unknown at the planning stage of the experiment. A conservative sample size would be to ensure that the highest posterior density intervals have length at most  $2d$  and converge probability at least  $(1 - \alpha)$  over all possible data  $x$  that may arise. This criterion has been termed the “Worst Outcome Criterion” (WOC) by Joseph et al. (1995).

Since the WOC is also based on highest-density intervals, one would expect it to provide similar sample sizes to those given by the modified criterion of Section 3 when a uniform ( $a = b = 1$ ) prior density is used. Indeed, applying the WOC to example 4 gives a sample size of 39, while a size of 93 results in example 12. These values lie roughly midway between those given by the modified criterion and Eq. (4). This suggests that approximately half of the differences observed in Table 1 may result from using highest posterior densities, with the rest arising from using exact discrete binomial probabilities. While the WOC is exact in the sense that exact Beta quantiles are used to calculate highest posterior density intervals, it of course does not directly use binomial probabilities.

See Joseph et al. (1995) or Joseph et al. (1997) for more details on Bayesian sample size determination.

## 6. Discussion

The purpose of this article was to shed some light on widely used criteria for binomial sample size determination. In particular, calculating conservative sample sizes from the standard exact formulation is problematic, since the conservative value of  $\theta$  is difficult to determine. Assuming that the conservative value  $\theta = 0.5$  is often not correct, while conservative sample sizes from the normal approximation often differ substantially from exact solutions. An algorithm to compute exact sample sizes from a modified criterion is provided.

As might have been expected, for small sample sizes, or when  $m$  is near 0 or 1, the resulting sample sizes can differ substantially from approximate solutions. The examples in Table 1 provide some idea of the potential gains in efficiency that result from switching to the new criterion. In some applications these gains may be worthwhile,

while in others, the normal approximation will suffice. In any case, it at least appears that the conservative normal approximate sample sizes are truly conservative, in the sense that the sample sizes provided are greater than those provided by the exact criterion (5). See Kupper and Hafner (1989) for a similar critique of the most widely used sample size criteria for the normal distribution.

## Appendix

Two preliminary remarks and three lemmas will be useful in proving Theorem 1.

**Remark A.1.** For given  $n$  and  $\theta \leq 0.5$ ,  $p(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$  takes its highest value on the point  $r$ , where  $r/(n+1) \leq \theta < (r+1)/(n+1)$ . For  $\theta = r/(n+1)$ ,  $p(r-1; \theta) = p(r; \theta)$ . Also  $p(k; \theta) > p(k-1; \theta)$  if and only if  $k/(n+1) < \theta$ . A similar argument can be made when  $\theta > 0.5$ . See Rohatgi (1984) for the proofs of these statements.

**Remark A.2.** Fix an integer  $k$  such that  $1 \leq k \leq n/2$ . Then  $p(k/n; \theta)$  is differentiable with respect to  $\theta$ , and it is easy to see that it increases if and only if  $\theta \leq k/n$ .

**Definition A.3.** For given  $d$  and  $n$ , define  $i$  to be the integer such that  $i/n \leq 2d < (i+1)/n$ .

In what follows, it is assumed that  $i \geq 1$ . For  $i = 0$  all of the results proven below will be trivially true.

**Lemma A.4.** Given  $d$ ,  $\alpha$ , and  $n$ , a point  $\theta$  can have at most two highest-density intervals of length  $i$ , namely  $[s/n, (s+i)/n]$  and  $[(s+1)/n, (s+i+1)/n]$ , for some integer  $s$ .

**Proof.** Suppose  $[s/n, (s+i)/n]$  and  $[u/n, (u+i)/n]$  are two highest-density intervals of length  $i$  corresponding to  $\theta$ , and suppose, without loss of generality, that  $s < u$ . It suffices to prove that  $u = s + 1$ . Since  $p(s; \theta) \geq p(s+i+1; \theta)$ ,  $p(s-1; \theta) \leq p(s+i; \theta)$ ,  $p(u; \theta) \geq p(u+i+1; \theta)$ , and  $p(u-1; \theta) \leq p(u+i; \theta)$ , by Remark 1,  $s \leq r < s+i+1$  and  $u \leq r < u+i+1$ , where  $r$  is the point of maximum probability. Hence,  $s \neq u-1$  implies  $s < u-1$  and  $s+i < u+i-1$ , which implies that  $p(u-1; \theta) > p(s; \theta) \geq p(s+i+1; \theta) > p(u+i; \theta)$ , which is a contradiction.  $\square$

**Lemma A.5.** Given  $d$ ,  $\alpha$ ,  $n$  and  $\theta_1 < \theta_2$ , let  $[s/n, (s+i)/n]$  and  $[u/n, (u+i)/n]$  be highest-density intervals corresponding to  $\theta_1$  and  $\theta_2$ , respectively. Then  $u \geq s$ .

**Proof.**  $\theta_1 < \theta_2$  implies that  $r_1 \leq r_2$ , where  $r_1$  and  $r_2$  are the points of maximum probability corresponding to  $\theta_1$  and  $\theta_2$  respectively, defined in Remark 1. Suppose that  $u < s$ , then by Remarks 1 and 2,  $p(u+i+1; \theta_2) > p(u+i+1; \theta_1) > p(s-1; \theta_1) \geq p(u; \theta_1) > p(u; \theta_2)$ , which is a contradiction.  $\square$



**Lemma A.6.** Given  $d, \alpha$ , and  $n$ , an interval  $I$  is a highest-density interval of length  $i$  corresponding to some  $\theta \in [0, 1]$  if and only if it is an interval from the set  $\{[0, i/n], [1/n, (i + 1)/n], \dots, [(n - i)/n, 1]\}$ .

**Proof.** Let  $I_k$  denote the interval  $[k/n, (k + i)/n]$ . To prove the necessary condition, just note that if  $j/n < a < (j + 1)/n \leq (n - i)/n$ , then the probability of the interval  $I_j$  is larger than the probability of the interval  $[a, a + i/n]$ . The proof of sufficiency will proceed by induction. It is clear that  $I_0$  is the highest-density interval corresponding to  $\theta = 0$ . Suppose  $I_{j-1}$  is a highest-density interval, then it suffices to prove that  $I_j$  is also a highest-density interval. Let

$$\theta_j = \max \{ \theta : I_{j-1} \text{ is a highest-density interval for } \theta \}.$$

We will prove that both  $I_{j-1}$  and  $I_j$  are highest-density intervals for  $\theta_j$ . Suppose  $I_{j-1}$  is not a highest-density interval for  $\theta_j$  and let  $I$  be such an interval. Set  $\varepsilon = p(I; \theta_j) - p(I_{j-1}; \theta_j)$ . For every  $k$ , there exists  $\delta_k > 0$  such that  $|\theta_j - \theta| < \delta_k$  implies

$$\left| \binom{n}{k} \theta_j^k (1 - \theta_j)^{n-k} - \binom{n}{k} \theta^k (1 - \theta)^{n-k} \right| < \varepsilon/3(n + 1).$$

Let  $\delta = \min \delta_k$ , and take  $\theta$  such that  $0 < \theta_j - \theta < \delta/2$  and such that  $I_{j-1}$  is a highest-density interval for  $\theta$ . Then

$$|p(I_{j-1}; \theta) - p(I_{j-1}; \theta_j)| < \varepsilon/3$$

and

$$|p(I; \theta) - p(I; \theta_j)| < \varepsilon/3.$$

Therefore,  $p(I_{j-1}; \theta) < p(I; \theta)$ , which is a contradiction. To prove that  $I_j$  is a highest density interval for  $\theta_j$ , let

$$S = \{s : I_s \text{ is a highest-density interval corresponding to some } \theta > \theta_j\}.$$

$S$  is not empty since  $n - i \in S$ . Letting

$v = \min S, [v/n, (v + i)/n]$  is a highest-density interval for  $\theta_j$ . To prove this, let  $\theta$  be such that  $0 < \theta - \theta_j < \delta/2$ , let  $I_j$  be a highest-density interval for  $\theta$ , and proceed as above. Hence, by Lemmas 1 and 2,  $v = j$  and  $I_j$  is a highest-density interval for  $\theta_j$ .  $\square$

**Proof of Theorem 1.** From Lemmas 1 and 2, and the definition of  $\theta_j, j \in \{1 \dots n - i\}$ , the highest-density interval of a point  $\theta \in (\theta_j, \theta_{j+1})$  is unique and equal to  $I_j$ . At  $\theta_j, p(I_{j-1}) = p(I_j)$  which reduces to

$$\binom{n}{j-1} \theta_j^{j-1} (1 - \theta_j)^{n-j+1} = \binom{n}{i+j} \theta_j^{i+j} (1 - \theta_j)^{n-i-j}.$$

Thus,

$$\left(\frac{\theta_j}{(1-\theta_j)}\right)^{i+1} = \frac{\binom{n}{j-1}}{\binom{n}{i+j}} \quad \text{and} \quad \theta_j = \frac{(r_j^{1/(i+1)})}{(1+r_j^{1/(i+1)})}.$$

Let  $\theta \in [\theta_j, \theta_{j+1}]$ , then  $H(\theta) = \sum p\{x: x/n \in I_j; \theta\}$ , and therefore,  $H(\theta) = \sum_{x=j}^n p\{x; \theta\} - \sum_{x=i+j+1}^n p\{x; \theta\}$ . On the other hand (see Rohatgi, 1984),

$$H(\theta) = \int_0^\theta n \binom{n-1}{j-1} t^{j-1} (1-t)^{n-j} - n \binom{n-1}{i+j} t^{i+j} (1-t)^{n-i-j-1} dt.$$

Hence

$$\begin{aligned} H'(\theta) &= n \binom{n-1}{j-1} \theta^{j-1} (1-\theta)^{n-j} - n \binom{n-1}{i+j} \theta^{i+j} (1-\theta)^{n-i-j-1} \\ &= n \binom{n-1}{j-1} \theta^{j-1} (1-\theta)^{n-j} \left(1 - \frac{\binom{n-1}{i+j}}{\binom{n-1}{j-1}} \left(\frac{\theta}{1-\theta}\right)^{i+1}\right). \end{aligned}$$

Let  $g(\theta) = 1 - \frac{\binom{n-1}{i+j}}{\binom{n-1}{j-1}} (\theta/(1-\theta))^{i+1}$ , then  $g'(\theta) = \frac{\binom{n-1}{i+j}}{\binom{n-1}{j-1}} (\theta/(1-\theta))^i (i+1) (1/(1-\theta))^2 > 0$ , so that  $g(\theta)$  is increasing. Furthermore,

$$g(\theta_j) = 1 - \frac{\binom{n-1}{i+j}}{\binom{n-1}{j-1}} \frac{\binom{n}{j-1}}{\binom{n}{i+j}} = 1 - \frac{n-i-j}{n-j+1} < 0$$

and

$$g(\theta_{j+1}) = 1 - \frac{\binom{n-1}{i+j}}{\binom{n-1}{j-1}} \frac{\binom{n}{j}}{\binom{n}{i+j+1}} = 1 - \frac{i+j+1}{j} > 0,$$

so that  $g(\theta)$  has only one zero in  $I_j$ . Hence,  $H(\theta)$  has a maximum in  $I_j$ , therefore, in this interval  $p(\theta)$  is minimum at one of the end points  $\theta_j$  or  $\theta_{j+1}$ .  $\square$

## References

- Desu, M.M., 1988. Sample Size Methodology, Academic Press, New York.  
Joseph, L., Wolfson, D., de Berger, R., 1995. Sample size calculations for binomial proportions via highest posterior density intervals. The Statist. 44, 143-154.

- Joseph, L., du Berger, R., Bélisle, P., 1997. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist. Med.* 16, 769–781.
- Kupper, L.L., Hafner, K.B., 1989. How appropriate are popular sample size formulas? *The Amer. Statist.* 43, 101–105.
- Lemeshow, S., Hosmer Jr., D.W., Klar, J., Lwanga, S.K., 1990. *Adequacy of Sample Size in Health Studies*. Wiley, New York.
- Rohatgi, V.K., 1984. *Statistical Inference*. Wiley, New York.