J. Jpn. Soc. Comp. Statist., **15.1**(2002), 19–32

DETERMINATION OF SAMPLE SIZES FOR USE IN CONSTRUCTING CONFIDENCE INTERVALS FOR A BINOMIAL PARAMETER

Manabu Iwasaki* **and Noriko Hidaka†**

ABSTRACT

This article considers exact and approximate confidence intervals for a binomial parameter p. Specifically, we will deal with the problem of determination of sample sizes which guarantee the probability that $100(1-\alpha)\%$ confidence intervals for p do not include pre-specified constants is greater than $1-\beta$. It is shown here that the coverage probability of confidence intervals is not an increasing function of the sample size, which is a consequence of the discreteness of binomial distributions. Illustrative numerical examples and some theoretical consideration for such anomalous behavior are given. We also briefly treat the case that the initial guess of the true binomial parameter is vague.

1. Introduction

We will consider a binomial distribution $B(n,p)$ for n independent Bernoulli trials with proportion p. Since forming a confidence interval for p is one of the most basic analyses in statistical inference, most elementary textbooks deal with this problem in some depth. Although this problem seems to be a simple and fundamental one, there still remain issues to be discussed carefully. In fact several articles have been published in recent statistical journals; see, for example, Agresti and Caffo (2000), Agresti and Coull (1998), Iwasaki and Hidaka (2001), Leemis and Trivedi (1996), Newcombe (1998), Sahai and Khurshid (1996), Vollset (1993), Wang (2000) and Wardell (1997). These articles mainly concern anomalous behavior of actual coverage probabilities of confidence intervals: the coverage probabilities are much greater or smaller than the nominal confidence coefficient and are not monotonic in proportion p.

In this article, we consider confidence intervals from the viewpoint of sample size determination. It is shown here that the larger sample sizes do not necessarily provide better results. Let $(\hat{p}_L, \hat{p}_U)$ be a $100(1-\alpha)\%$ confidence interval for p based on an observed frequency of a random variable X which follows $B(n,p)$. For pre-specified quantities p_0 and p_1 $(0 \leq p_0 < p_1 \leq 1)$, we consider the situation that it is required either the lower confidence limit $\hat{p}_L$ is greater than p_0 or the upper confidence limit $\hat{p}_U$ is less than p_1, or both. For example, if p is the probability of recovery of a patient in a clinical trial, then a lower bound would be placed to specify the minimum required probability of such recovery. On the other

*Department of Industrial Engineering and Information Sciences, Seikei University, Musashino-shi, Tokyo 180-8633 Japan E-mail: iwasaki@is.seikei.ac.jp

†Pharmaceutical Division, Suntory Co. Ltd., Kojimachi, Chiyoda-ku, Tokyo 102-8530 Japan E-maill: Noriko_Hidaka@suntory.co.jp

Key words: Clopper-Pearson interval; Coverage probability; Exact calculus; Wald-type interval

hand, if p is the probability of occurrence of some adverse event, then an upper bound would be more appropriate. Since $n - X \sim B(n, 1-p)$, the upper-bound problem can be easily solved from the corresponding lower-bound counterpart. Therefore, in this article, we consider the lower-bound problem only.

For pre-specified small probabilities a and b, and assuming that an initial guess p_g of p is true, our aim is to determine appropriate sample sizes which ensure

$$\Pr(p_0 \leq \hat{p}_L \,|\, p = p_g) \geq 1 - \beta.$$

It is well known that a confidence interval consists of parameter values which are not rejected by the corresponding statistical test; see, for example, Cox and Hinkley (1974). Specifically, if a random variable X follows $B(n,p)$ and x^* is an observed value, the p-value (actual significance probability) for a one-sided test $H_0 : p = p_0$ vs. $H_1 : p > p_0$ is given by $\Pr(x^* \leq X \,|\, p = p_0)$. A confidence interval for p is the set of p_0 values for which this p-value is greater than a pre-specified significance level. The power under $H_1 : p = p_g > p_0$ is also defined as $\Pr(x^* \leq X \,|\, p = p_g)$. Our objective is to obtain sample sizes such that the power of detecting H_1 is greater than or equal to $1-\beta$ when the true parameter value is p_g. For this reason, the quantity α and $1-\beta$ will be also called "significance probability" and the "power", respectively, in the subsequent sections. For detailed argument of sample size determination for testing, see, for example, Fisher and van Belle (1993) and Fleiss (1981).

In Section 2, various formulae of sample-size determination are given, in which we deal with three procedures: the exact interval of Clopper and Pearson (1934), the score confidence interval ascribed to Wilson (1927), and the Wald-type interval which is frequently referred to in elementary textbooks. Section 3 gives numerical examples and relevant theoretical consideration. In Section 4, we briefly discuss the case that the initial guess p_g of the true p is vague. Finally in Section 5, we conclude the article and give practical recommendations to practitioners.

2. Formulae for sample sizes

The problem of sample size determination can be formulated as follows. First, three quantities are to be specified; the confidence coefficient $1-\alpha$, the power $1-\beta$ and the lower bound p_0. We also need an initial guess p_g of the true binomial parameter, in which the subscript g is for "guess". Then, for a particular construction method of $100(1-\alpha)\%$ confidence interval $(\hat{p}_L,\ \hat{p}_U)$ for p, the required sample size n is determined to meet the following requirements. Let X_n be a random variable which has $B(n,p)$ and $x(n)^*$ be the smallest integer such that $x(n)^* \leq X_n$ implies $p_0 \leq \hat{p}_L$. The parameters n remind us that these quantities are functions of sample size n to be determined. Then, for the value $x(n)^*$ that satisfies

$$\Pr(p_0 \leq \hat{p}_L \,|\, p = p_0) = \Pr(x(n)^* \leq X_n \,|\, p = p_0) \leq \alpha\,/\,2, \tag{1}$$

which ensures the actual significance level is at most $\alpha/2$, the sample size n is to be chosen to satisfy the power requirement

$$\Pr(p_0 \leq \hat{p}_L \,|\, p = p_g) = \Pr(x(n)^* \leq X_n \,|\, p = p_g) \geq 1 - \beta. \tag{2}$$

Different calculation methods of the probabilities (1) and (2) give different results as shown below.

We will consider here the following three methods to construct confidence interval for a binomial parameter p. The first one is an exact interval in the sense that the construction is based on exact calculus of binomial probabilities of (1), whereas the others are approximate in that they utilize normal approximation to calculation of (1). Although some other methods have been proposed in the literature these three methods are most frequently referred to in statistical books and papers. For some other methods, see Leemis and Trivedi (1996), Newcombe (1998), Sahai and Khurshid (1996) and Vollset (1993).

1. Exact interval of Clopper and Pearson (1934): This method makes exact calculation of the probability (1). For an observed frequency x, the lower limit $\hat{p}_L$ and the upper limit $\hat{p}_U$ of the interval are given by the values that satisfy

$$\Pr(x \leq X_n \,|\, \hat{p}_L) = \sum_{k=x}^{n} {}_nC_k \hat{p}_L{}^k (1-\hat{p}_L)^{n-k} = \alpha/2 \tag{3}$$

and

$$\Pr(X_n \leq x \,|\, \hat{p}_U) = \sum_{k=0}^{x} {}_nC_k \hat{p}_U{}^k (1-\hat{p}_U)^{n-k} = \alpha/2, \tag{4}$$

respectively. These confidence limits are obtained from the relationship between binomial distributions and F distributions. If we denote by $F_{a,b}(\alpha/2)$ the upper $100\alpha/2\%$ percentile of an F distribution with degrees of freedom (a, b), then we have

$$\hat{p}_L = \frac{x}{x + (n-x+1)F_{k_1,k_2}(\alpha/2)},$$

where $k_1 = 2(n-x+1), k_2 = 2x$, and

$$\hat{p}_U = \frac{x+1}{x+1+(n-x)/F_{l_1,l_2}(\alpha/2)},$$

where $l_1 = 2(x+1), l_2 = 2(n-x)$. In this construction method, we have to make clear the choice of the confidence coefficient when the observed frequency x is 0 or n. Wang (2000) suggests using α instead of $\alpha/2$ in (3) and (4) from the viewpoint of fiducial inference. On the other hand, other authors, Agresti and Coull (1998), Newcombe (1998), Vollset (1993) and also the original Clopper and Pearson (1934) use $\alpha/2$. In this article, even when $x = 0$ or n, we will use $\alpha/2$. See Iwasaki and Hidaka (2001) for relevant discussion.

2. Score-type approximate interval: This method uses a normal approximation to (1) and is obtained from the score test for proportion p. Since this interval first appeared in Wilson (1927), it is sometimes called the Wilson method (Agresti and Coull, 1998). Since the sample proportion $\hat{p} = X_n/n$ approximately distributes as a normal distribution $N(p, p(1-p)/n)$ when n is sufficiently large, we have

$$\Pr\left(-z(\alpha/2) \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq z(\alpha/2)\right) = \alpha,$$

where $z(\alpha/2)$ is the upper $100\alpha/2\%$ percentile of $N(0,1)$. The lower and the upper confidence limits are given by solving the equation

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

with respect to p. After some calculation we obtain

$$\hat{p}_L = \frac{\hat{p} + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{1}{2} - z(\alpha/2)\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{1}{4n}}}{1 + \dfrac{z(\alpha/2)^2}{n}}$$

and

$$\hat{p}_U = \frac{\hat{p} + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{1}{2} + z(\alpha/2)\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n} + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{1}{4n}}}{1 + \dfrac{z(\alpha/2)^2}{n}}$$

respectively. Although these formulae seem complicated at a first glance, Agresti and Coull (1998) pointed out that they can be interpreted as a weighted mixture of $p = \hat{p}$ and $p = 1/2$ with weights 1 and $z(\alpha/2)^2/n$. See also Agresti and Caffo (2000).

3. Wald-type approximate interval: This interval is based on the asymptotic normality of the Wald-type statistic $(\hat{p} - p)/\sqrt{\hat{p}(1-\hat{p})/n}$. Solving the equation

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

with respect to p, we obtain the Wald-type interval

$$(\,\hat{p}_L = \hat{p} - z(\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p}_U = \hat{p} + z(\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}\,), \tag{5}$$

which frequently appears in elementary textbooks.

Before we obtain required sample sizes for the construction methods above, we first give a conventional textbook formula to determine the sample size to meet the required conditions. This is based on the Wald-type approximate confidence interval (5). It also uses normal approximation in calculating the power (2). The sample-size formula is given by

$$n \geq \frac{\left\{z(\alpha/2)\sqrt{p_g(1-p_g)} + z(\beta)\sqrt{p_0(1-p_0)}\right\}^2}{(p_0 - p_g)^2}, \tag{6}$$

cf. Miyahara and Tango (1995), p. 265, (11.4). We shall denote by n_0 the smallest integer that satisfies (6). For example, when $\alpha = 0.05$, $1 - \beta = 0.8$, $p_0 = 0.6$ and $p_g = 0.8$, we have $n_0 = 43$. In this case, the power $1 - \beta$ under each sample size n greater than n_0 is expected to be greater than the power of n_0. This is true for continuous distributions, and it is the reason why we calculate the smallest value n_0 only. For discrete cases, however, this is not true. Hence, we have to report not only the smallest value but also some other values as shown below.

The sample size based on the exact interval is obtained as follows. We will use the relationship between confidence interval and corresponding hypothesis testing, which was briefly discussed in Section 1. Let $x_e(n)^*$ be the smallest integer that satisfies

$$\Pr(x_e(n)^* \leq X_n \,|\, p_0) = \sum_{x_e(n)^* \leq k} {}_nC_k {p_0}^k (1-p_0)^{n-k} \leq \alpha/2, \tag{7}$$

in which the subscript e stands for "exact". Required sample sizes n should satisfy

$$\Pr(x_e(n)^* \leq X_n \,|\, p_g) = \sum_{x_e(n)^* \leq k} {}_nC_k {p_g}^k (1-p_g)^{n-k} \geq 1-\beta. \tag{8}$$

As will be shown numerically in the next section, even if an n satisfies both (7) and (8), $n+1$ does not always satisfy (8). For example, when $\alpha = 0.05$, $1-\beta = 0.8$, $p_0 = 0.6$ and $p_g = 0.8$ as before, the sample sizes $n = 45$ and 46 satisfy the required conditions, while $n = 47$ does not meet the requirement. In fact, the corresponding power becomes $1-\beta = 0.783$ when $n = 47$. This seems problematic.

The sample size for the score-type interval becomes as follows. The critical value $x_s(n)^*$ is the smallest integer that satisfies

$$\frac{x/n - p_0}{\sqrt{p_0(1-p_0)/n}} \geq z(\alpha/2)$$

in which the subscript s stands for "score." Although we used a normal approximation to evaluate (1), we will make exact calculation of (2). In this case, the required sample size n is the integer such that

$$\Pr(x_s(n)^* \leq X \,|\, p_g) = \sum_{x_s(n)^* \leq k} {}_nC_k {p_g}^k (1-p_g)^{n-k} \geq 1-\beta.$$

The sample size of this method has almost the same property as that of the exact method. For example, under the same setting of $\alpha = 0.05$, $1-\beta = 0.8$, $p_0 = 0.6$ and $p_g = 0.8$, we see that $n = 41$ satisfies the required conditions, while $n = 42$ and 43 do not meet the requirement, whose powers are 0.795 and 0.771 respectively.

The sample size for the Wald-type interval becomes as follows. In order that $p_0 \leq \hat{p}_L$, we should have

$$\frac{x/n - p_0}{\sqrt{(x/n)(1-x/n)/n}} \geq z(\alpha/2),$$

from which we have

$$(\hat{p} - p_0)^2 \geq (z(\alpha/2))^2 \{\hat{p}(1-\hat{p})/n\}.$$

Hence we get a critical value $x_W(n)^*$, where the subscript W is for "Wald", which is the smallest integer that satisfies

$$x_W(n)^* \geq \frac{np_0 + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{n}{2} + z(\alpha/2)\sqrt{np_0(1-p_0) + \dfrac{z(\alpha/2)^2}{n} \cdot \dfrac{n}{4}}}{1 + \dfrac{z(\alpha/2)^2}{n}}. \tag{9}$$

The required sample size is obtained to satisfy

$$\Pr(x_W(n)^* \leq X \,|\, p_g) = \sum_{x_W(n)^* \leq k} {}_nC_k {p_g}^k (1-p_g)^{n-k} \geq 1-\beta.$$

This easy-to-use method also suffers from the same difficulty as before. Actually when $\alpha = 0.05$, $1-\beta = 0.8$, $p_0 = 0.6$ and $p_g = 0.8$, we see that $n = 36$ and 37 satisfy the required conditions, while $n = 38$ does not meet the requirement, in fact $1-\beta = 0.784$. The situation is much worse, because many authors have pointed out that the true coverage probability of this method can be considerably lower than expected. Therefore, the Wald-type interval cannot be recommended to use in practice.

3. Numerical illustrations and some relevant theory

In this section, we will make numerical comparison of the sample sizes calculated by the methods of the previous section. The calculation methods as well as the derived sample sizes are referred to as "textbook", "exact", "score" and "Wald" for short. We let $\alpha = 0.05$ and $1 - \beta = 0.8$, which is one of the most frequent settings in practice. For illustration we consider the cases $p_0 = 0.6$ and 0.9 in detail. The choice $p_0 = 0.6$ can be a minimum requirement in a clinical trial for a new anti-hypertension drug. For sensitivity and specificity of diagnostic or screening tests the probability to obtain a correct result should be large, and hence $p_0 = 0.9$ might be required.

Table 1 shows the actual significance probabilities α^* and the actual power $1 - \beta^*$ of exact confidence intervals for various values of n and p_g when $\alpha = 0.05$ and $p_0 = 0.6$. We can draw two types of figures from Table 1; one shows the behavior of $1 - \beta^*$ in the rows and the other shows that in the columns of the table. They are given in Figure 1 (the row graph) and Figure 2 (the column graph) for some selected p_g and n.

We observe in Figure 1 that the power is not an increasing function of n, which means that the larger sample size does not always provide higher power. The reason for this phenomenon will be given below. In Figure 2 we see that, for any fixed n, the power $1 - \beta^*$ is an increasing function of the initial guess p_g. This fact can be easily shown from the relationship between the binomial distribution and the F distribution. It should be noted here that the power function of a larger sample does not necessarily lie above the functions of smaller sample sizes. In fact, we see in Figure 2 that the power functions lie in the order of $n = 48, 45, 46, 47$ and 44 from the above.

Figure 3 shows the same graph as Figure 1 for $p_0 = 0.9$. In Figure 3 we see much worse behavior of the power function. The gaps of the zigzags are large. Actually the gap is almost 0.35 between $n = 53$ and 54. It is frequently recommended to use exact methods when the binomial parameter is near 0 or 1, since in such cases the normal approximation would not be appropriate. In such cases we have to be aware of this unexpected behavior of the power function.

Required sample sizes to meet the conditions for each calculation method are summa-

Table 1: Critical values x^*, actual significance probability α^* and the power $1 - \beta$ of "exact" intervals for various n and p_g when $\alpha = 0.05$ and $p_0 = 0.6$

								p_g					
n	x^*	α^*	0.75	0.76	0.77	0.78	0.79	0.80	0.81	0.82	0.83	0.84	0.85
40	31	0.016	0.440	0.498	0.558	0.618	0.676	0.732	0.783	0.830	0.870	0.905	0.933
41	32	0.012	0.405	0.463	0.524	0.585	0.646	0.704	0.759	0.809	0.853	0.891	0.923
42	32	0.021	0.512	0.572	0.632	0.690	0.745	0.795	0.840	0.879	0.911	0.938	0.958
43	33	0.016	0.477	0.538	0.599	0.660	0.717	0.771	0.820	0.862	0.899	0.928	0.951
44	34	0.012	0.442	0.504	0.567	0.629	0.689	0.746	0.798	0.845	0.884	0.917	0.943
45	34	0.022	0.546	0.607	0.668	0.725	0.778	0.826	0.868	0.903	0.931	0.953	0.970
46	35	0.017	0.511	0.575	0.637	0.697	0.753	0.805	0.850	0.889	0.921	0.946	0.965
47	36	0.013	0.478	0.542	0.606	0.668	0.728	0.783	0.832	0.874	0.909	0.937	0.959
48	36	0.022	0.577	0.639	0.700	0.756	0.807	0.852	0.890	0.922	0.946	0.965	0.978
49	37	0.017	0.544	0.608	0.671	0.730	0.785	0.834	0.875	0.910	0.938	0.959	0.974
50	38	0.013	0.511	0.577	0.642	0.704	0.762	0.814	0.860	0.898	0.929	0.952	0.970

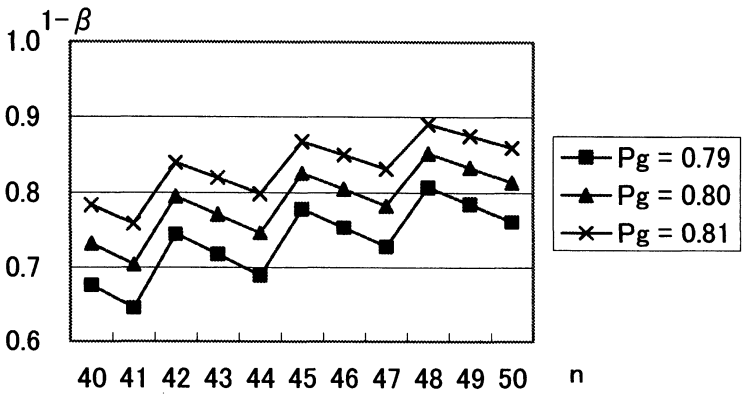

Fig. 1: The power $1-\beta$ of "exact" intervals for various n's and three selected p_g's when $\alpha = 0.05$ and $p_0 = 0.6$

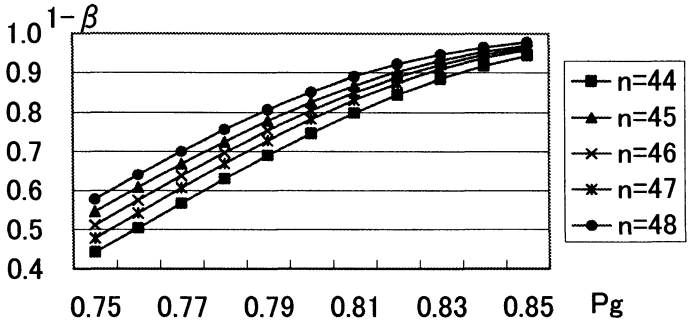

Fig. 2: The power $1-\beta$ of "exact" intervals as the functions of p_g for five selected n's when $\alpha = 0.05$ and $p_0 = 0.6$

rized in Table 2 ($p_0 = 0.6$) and Table 3 ($p_0 = 0.9$). In each table, the first column shows the values of p_g. The second column gives the sample sizes calculated by the textbook formula (7). The third and the fourth columns show the sample sizes for the exact interval, in which the third column "Exact(L)" gives the minimum sample sizes which meet the required conditions, and the forth column "Exact(H)" shows the minimum sample size such that all the sample sizes greater than or equal to that value meet the requirement (L stands for "lowest", while H stands for "highest"). Then we see that each of the sample sizes $n = \text{Exact(H)} - 1$ does not meet the requirement unless it happens to be Exact(L). For example, in the row "$p_g = 0.8$" in Table 2, we read Exact(L) = 45 and Exact(H) = 48. This means that the minimum required sample size is 45 and that $n = 46$ and 47 do not meet the requirement. We also understand that all the sample sizes $n \geq 48$ meet the requirement. This fact can be also confirmed from Table 1. The meanings of subsequent columns are the same.

It is worth noting that the sample sizes of the Wald interval for the case that p_0 is very close to 1 would give misleading results, since the normal approximation no longer works. If $n = 1$ and $x = 1$ then the right-hand side of (9) becomes less than 1, and hence the formula

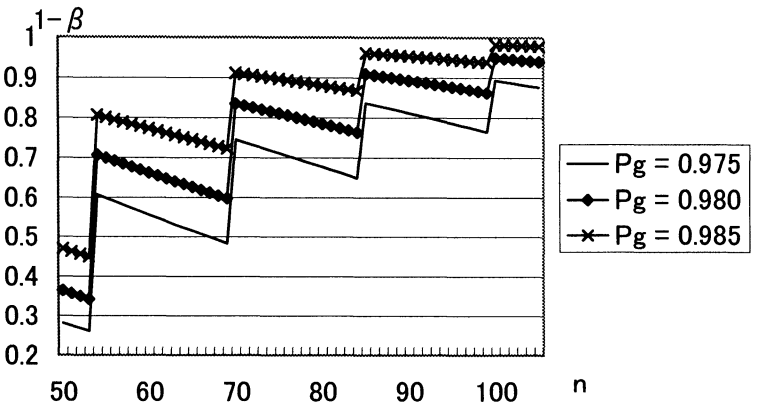

Fig. 3: The power $1-\beta$ of "exact" intervals for various n's and three selected p_g's when $\alpha = 0.05$ and $p_0 = 0.9$

Table 2: Sample sizes for various initial guesses p_g when $\alpha = 0.05$, $1-\beta = 0.8$ and $p_0 = 0.6$

p_g	Textbook	Exact(L)	Exact(H)	Score(L)	Score(H)	Wald(L)	Wald(H)
0.70	182	181	195	177	186	162	179
0.71	149	151	162	144	153	134	146
0.72	125	126	134	119	128	109	118
0.73	106	106	115	102	111	92	104
0.74	91	92	98	85	97	81	83
0.75	78	80	86	76	82	66	75
0.76	69	72	78	65	74	57	63
0.77	60	63	69	59	65	48	57
0.78	53	54	60	50	56	42	48
0.79	48	48	54	44	50	39	42
0.80	43	45	48	41	44	36	39
0.81	38	39	45	34	41	29	36
0.82	35	36	39	31	34	26	33
0.83	31	32	36	28	31	23	26
0.84	28	29	32	28	28	23	26
0.85	26	26	29	25	28	19	23

concludes $x^* = 1$. If $\Pr(X = 1 \,|\, p = p_g) = p_g > 1-\beta$, $n = 1$ is automatically identified to be a required sample size. This is of course unrealistic and erroneous. Hence, we have to put a restriction for Wald(L). In Table 3, since the minimum sample size which ensures $\alpha^* \leq 0.025$ is 36, we set Wald(L) $= 36$ as a minimum requirement.

Let us consider the reason why the power is not an increasing function of n. We denote the probability function of $B(n,p)$ by

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

where ${}_nC_x = n!/\{x!(n-x)!\}$ is the binomial coefficient. The lower and upper cumulative

Table 3: Sample sizes for various initial guesses p_g when $\alpha = 0.05$, $1 - \beta = 0.8$ and $p_0 = 0.9$

p_g	Textbook	Exact(L)	Exact(H)	Score(L)	Score(H)	Wald(L)	Wald(H)
0.950	239	231	255	204	254	161	186
0.955	193	180	206	179	204	121	147
0.960	158	157	164	140	166	93	121
0.965	131	127	141	127	140	64	93
0.970	110	100	114	99	113	48	79
0.975	93	85	100	84	99	48	64
0.980	78	70	85	69	84	36	48
0.985	66	54	70	53	69	36	36
0.990	56	54	54	53	53	36	36
0.995	47	36	54	35	53	36	36

probabilities are to be denoted by

$$p(x;n,p) = \Pr(X \leq x) = \sum_{k=0}^{x} p(k;n,p) \tag{10}$$

and

$$Q(x;n,p) = \Pr(X \geq x) = \sum_{k=x}^{n} p(k;n,p), \tag{11}$$

respectively. Note that $P(x;n,p) + Q(x;n,p)$ is not equal to 1 because $p(x;n,p)$ is involved in both cumulative probabilities. First we will show a lemma.

Lemma 1 *For the probabilities of $B(n,p)$, it holds*

$$x \leq (n+1)p \Leftrightarrow p(x;n+1,p) \leq p(x;n,p) \tag{12}$$

and

$$x \leq (n+1)p - 1 \Leftrightarrow p(x+1;n+1,p) \geq p(x;n,p). \tag{13}$$

Proof. Since

$$\frac{p(x;n+1,p)}{p(x;n,p)} = \frac{{}_{n+1}C_x p^x (1-p)^{n+1-x}}{{}_nC_x p^x (1-p)^{n-x}}$$

$$= \frac{\dfrac{(n+1)!}{x!(n+1-x)!} p^x (1-p)^{n+1-x}}{\dfrac{n!}{x!(n-x)!} p^x (1-p)^{n-x}} = \frac{n+1}{n+1-x}(1-p),$$

the inequality

$$\frac{n+1}{n+1-x}(1-p) \leq 1$$

follows if and only if $x \leq (n+1)p$, and hence (12) holds. Similarly for (13), since we have

$$\frac{p(x+1;n+1,p)}{p(x;n,p)} = \frac{{}_{n+1}C_{x+1}p^{x+1}(1-p)^{n+1-(x+1)}}{{}_nC_xp^x(1-p)^{n-x}}$$

$$= \frac{\dfrac{(n+1)!}{(x+1)!(n-x)!}p^{x+1}(1-p)^{n-x}}{\dfrac{n!}{x!(n-x)!}p^x(1-p)^{n-x}} = \frac{n+1}{x+1}p,$$

the inequality

$$\frac{n+1}{x+1}p \geq 1,$$

follows if and only if $x \leq (n+1)p - 1$, which completes the proof. □

The zigzag behavior of the power function shown in Figures 1 and 3 can be deduced from the following theorem.

Theorem 1 *For the lower cumulative probability* (10) *we obtain*

$$x \leq (n+1)p \Rightarrow P(x;n+1,p) < P(x;n,p) \tag{14}$$

and

$$x \leq (n+1)p - 1 \Rightarrow P(x+1;n+1,p) > P(x;n,p). \tag{15}$$

For the upper cumulative probability (11) *we have*

$$x \geq (n+1)p \Rightarrow Q(x;n+1,p) > Q(x;n,p) \tag{16}$$

and

$$x \geq (n+1)p - 1 \Rightarrow Q(x+1;n+1,p) < Q(x;n,p). \tag{17}$$

Proof. Since

$$P(x;n,p) - P(x;n+1,p) = \sum_{k=0}^{n}\{p(k;n,p) - p(k;n+1,p)\}$$

and

$$Q(x;n,p) - Q(x+1;n+1,p) = \sum_{k=x}^{n}\{p(k;n,p) - p(k+1;n+1,p)\}$$

the inequalities (14) and (17) obviously hold because the inequalities (12) and (13) hold for each k and noting that at least one strict inequality hold in the summation. For (15), it follows that

$$\begin{aligned} &P(x+1;n+1,p) - P(x;n,p) \\ &= \sum_{k=0}^{x+1}p(k;n+1,p) - \sum_{k=0}^{x}p(k;n,p) \\ &= (1-p)^{n+1} + \sum_{k=0}^{x}\{p(k+1;n+1,p) - p(k;n,p)\}, \end{aligned}$$

which is positive from (13). Also for (16), we have

$$\begin{aligned} & Q(x; n+1, p) - Q(x; n, p) \\ &= \sum_{k=x}^{n+1} p(k; n+1, p) - \sum_{k=x}^{n} p(k; n, p) \\ &= p^{n+1} + \sum_{k=x}^{n} \{p(k; n+1, p) - p(k; n, p)\}, \end{aligned}$$

which is also shown to be positive from (12), as is required. □

For the exact interval, let X_n have $B(n, p)$ and if the sample size is n and the critical value is x^*, then the actual significance probability ${\alpha_n}^*$ and the actual power $1 - {\beta_n}^*$, in which the subscript n denotes the sample size being considered, can be expressed as

$${\alpha_n}^* = \Pr(x^* \leq X_n \mid p_0) = Q(x^*; n, p_0)$$

and

$$1 - {\beta_n}^* = \Pr(x^* \leq X_n \mid p_g) = 1 - P(x^* - 1; n, p_g)$$

from which we have

$${\beta_n}^* = P(x^* - 1; n, p_g)$$

Since x^* is obviously greater than $(n+1)p_0$, from Theorem 1 we have

$$Q(x^*; n+1, p_0) \geq Q(x^*; n, p_0)$$

and

$$Q(x^* + 1; n+1, p_0) \leq Q(x^*; n, p_0).$$

If the required power is greater than 0.5, which is a condition which should be achieved in practice, x^* is less than $(n+1)p_g$. Hence, we have

$$P(x^* - 1; n+1, p_g) \leq P(x^* - 1; n, p_g)$$

and

$$P(x^*; n+1, p_g) \geq P(x^* - 1; n, p_g).$$

The above results provide a proof of the findings in our numerical illustration. If the critical value becomes $x^* + 1$ under $n+1$ then both the actual significance probability α^* and the power $1 - \beta^*$ decrease, that is

$${\alpha_{n+1}}^* = Q(x^* + 1; n+1, p_0) \leq Q(x^*; n, p_0) = {\alpha_n}^*$$

and

$$1 - {\beta_{n+1}}^* = 1 - P(x^*; n+1, p_g) \leq 1 - P(x^* - 1; n, p_g) = 1 - {\beta_n}^*.$$

However, if the critical value still remains x^* even under $n+1$ then α^* and $1-\beta^*$ will both increase, i.e.

$$\alpha_{n+1}{}^* = Q(x^*; n+1, p_0) \geq Q(x^*; n, p_0) = \alpha_n{}^*$$

and

$$1-\beta_{n+1}{}^* = 1 - P(x^*-1; n+1, p_g) \geq 1 - P(x^*-1; n, p_g) = 1-\beta_n{}^*.$$

We actually observe in Table 1 that when the sample size n increases the actual significance probability α^* and the power $1-\beta^*$ behave similarly, that is, when α^* increases the power $1-\beta^*$ also increases, and vice versa. The same behavior of the power in "score" and "Wald" can be exemplified in the same way.

4. Vague initial guess

In the previous sections we have developed our argument under the assumption that we have a concrete initial guess p_g. This might be, however, unrealistic in practice in that our prior knowledge for p_g is sometimes vague. Then we have to take into account such vagueness in our sample size determination. An attempt for dealing with such vagueness is given in this section. We will focus on the exact interval only in this section. The other types of intervals can be treated similarly.

We first specify by ourselves or ask the researcher to specify an interval (c, d) in which the true p probably lies from present subject-matter knowledge. Then the most obvious method to give appropriate sample sizes is as follows. For a specified interval (c, d), we calculate Exact(L) under c and also give Exact(H) under d for recommended sample sizes. These sample sizes usually form an interval. From this interval we will choose an appropriate sample size. For example, when $\alpha = 0.05$, $1-\beta = 0.8$ and $p_0 = 0.6$, if the specified interval is (0.78, 0.8) then the recommended sample sizes are between 48 (Exact(H) under 0.8) and 54 (Exact(L) under 0.78).

We see that the interval (48, 54) above seems wider than expected even under such precise specification of (0.78, 0.8). This is a consequence of the fact that the required sample size is very sensitive to the specification of p_g unless p_g is far from p_0.

5. Conclusion

For the problem of sample size determination, we have seen that the discreteness of binomial distributions causes a serious difficulty. It is necessary to report not only the minimum sample size but also all the sample sizes that satisfy the required conditions. Since the sample size $n+1$ does not always provide better result than n, the term "minimum" should be used with some caution.

It is worth noting that employment of the minimum required sample size, n^* say, would be dangerous because the gap of the power between n^*-1 and n^* is larger than expected as is observed in Table 1 ($p_0 = 0.6$) and Table 3 ($p_0 = 0.9$). Table 3 also indicates that the situation becomes worse if the binomial probability is near 1 (or also near 0). We observe in these figures that if one observation happens to be lost for some reason, then the power may considerably decrease. It is also observed that even if one more observation happens to be obtained, the power does not necessarily increase. In fact, the sample size of n^*+1 is not a clever choice because the power is mostly less than n^*. Then, how should we

choose an appropriate sample size? What information should be provided to practitioners in determining the correct sample size?

One possible and realistic solution is to show the graphs such as Figures 1 and 2. These graphs contain all the relevant information to be provided in sample size determination. Nowadays, such figures can be easily drawn by using a PC. In order to determine an appropriate sample size we should consider many things, some of which are non-statistical issue such as cost, deadline and feasibility of the study. For determining an appropriate sample size, we have to take into account such non-statistical things and also the statistical concepts such as the probable value of true p and the power to exclude the required lower bound as well.

Added to the proof

During the reviewing process of the present paper, three articles, Brown, Cai and DasGupta (2001), Cesana, Reina and Marubini (2001) and Henderson and Meyer (2001) came to the authors' attention. These articles deal with issues closely related to our paper. In particular, a review article Brown *et al.* (2001) and subsequent discussions are very helpful to understand the background of our paper.

Acknowledgements

The authors are very grateful to the editor and the referees for their careful reading of the earlier drafts of the present paper. Their comments and suggestions were very helpful to make our manuscript much more readable.

REFERENCES

Agresti, A. and Caffo, B. (2000). Simple and effective confidence 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.

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

Cesana, B.M., Reina, G. and Marubini, E. (2001). Sample size for testing a proportion in clinical trials: A "two-step" procedure combining power and confidence interval expected width. *American Statistician* **55**, 288–292.

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.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman & Hall.

Fisher, L.D. and van Belle, G. (1993). *Biostatistics. A Methodology for the Health Sciences.* New York: John Wiley & Sons.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, Second Edition. New York: John Wiley & Sons.

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.

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.

Leemis, L.M. and Trivedi, K.S. (1996). A comparison of approximate interval estimators for the Bernoulli parameter. *American Statistician* **50**, 63–68.

Miyahara, H. and Tango, T. (Eds.) (1995). *Handbook of Medical Statistics.* Tokyo: Asakura Shoten. (in Japanese)

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

Sahai, H. and Khurshid, A. (1996). Confidence intervals for the probability of success in the binomial distribution: a review. *Metron* **54**, 153–180.

Vollset, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine* **12**, 809–824.

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

Wang, Y.H. (2000). Fiducial intervals: What are they? *American Statistician* **54**, 105–111.

Wardell, D.G. (1997) Small-sample interval estimation of Bernoulli and Poisson parameters. *American Statistician* **51**, 321–325.

(Received June 2001, Accepted April 2002)