Communications in Statistics—Simulation and Computation®, 36: 1171–1186, 2007

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610910701569218

Inference

Some Properties of the Exact and Score Methods for Binomial Proportion and Sample Size Calculation

K. KRISHNAMOORTHY AND JIE PENG

Department of Mathematics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA

In this article, we point out some interesting relations between the exact test and the score test for a binomial proportion p. Based on the properties of the tests, we propose some approximate as well as exact methods of computing sample sizes required for the tests to attain a specified power. Sample sizes required for the tests are tabulated for various values of p to attain a power of 0.80 at level 0.05. We also propose approximate and exact methods of computing sample sizes needed to construct confidence intervals with a given precision. Using the proposed exact methods, sample sizes required to construct 95% confidence intervals with various precisions are tabulated for $p = .05(.05).5$. *The approximate methods for computing sample sizes for score confidence intervals are very satisfactory and the results coincide with those of the exact methods for many cases.*

Keywords Clopper–Pearson interval; Coverage probability; Expected length; One-sided limits; Sizes; Wilson interval.

Mathematics Subject Classification Primary 62H15; Secondary 62H17.

1. Introduction

The binomial distribution is the oldest subject in statistical sciences which has been receiving continuous interest among researchers and practitioners. The binomial model is commonly postulated for making inference about the proportion p of individuals in a population with a particular attribute of interest. Even though there are several inferential methods are proposed in the literature, all of them are subject to some criticisms. For example, the exact confidence intervals due to Clopper and Pearson (1934) for the binomial proportion are too conservative,

Received May 19, 2006; Accepted March 23, 2007
Address correspondence to K. Krishnamoorthy, Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA; E-mail: krishna@louisiana.edu

yielding confidence intervals that are unnecessarily wide. There are articles that recommend the approximate score confidence intervals due to Wilson (1927) for the binomial proportion (e.g., Agresti and Coull, 1998). The review article by Brown et al. (2001) evaluates the merits of several approximate as well as some exact confidence intervals for the binomial proportion. These authors recommend the score interval, among others, for practical use because of its simplicity and satisfactory accuracy. Even though score intervals are, in general, shorter than the exact confidence intervals, their coverage probabilities are very unstable, and may go well below the nominal level for some parameter and sample size combinations (see Brown et al., 2001; Casella, 2001; Geyer and Meeden, 2005, for two-sided intervals, and Cai, 2005, for one-sided intervals).

There are other alternative approaches which are less conservative and produce shorter intervals than the Clopper–Pearson exact intervals. Blyth and Still (1983) and Cai and Krishnamoorthy (2005) proposed shorter intervals for binomial proportions which control the coverage probabilities very close to the nominal level. These intervals are shorter than the classical exact intervals, but they are computationally intensive, and are not so simple as the Clopper–Pearson exact method or the score interval. Randomized intervals (Blyth and Hutchinson, 1960) and, recently, fuzzy intervals (Geyer and Meeden, 2005) are also proposed for estimating binomial proportion, but these intervals are not commonly used in applications. Even though other approaches produce shorter intervals with some desirable properties, the score intervals and the exact intervals are still popular and they appear in common textbooks. Softwares such as Minitab and S-Plus, and online calculators (e.g., http://statpages.org/confint.html) use the exact method to compute confidence intervals. The Department of Health of the Washington state government (http://www.doh.wa.gov/), recommends the score interval, Draft Guidance for Industry and FDA Staff (http://www.fda.gov/cdrh/oivd/guidance/1171.pdf) suggests using either of the intervals, and the National Institute of Standard and Technology (NIST) describes the score and the exact intervals.

As mentioned above, there are many articles that compare the score intervals, Wald intervals, and the exact intervals with respect to coverage probabilities and expected lengths. However, not many comparison studies were made for hypothesis testing, especially for one-sided hypothesis testing. At first glance, one may think that the interval estimation problem is dual to hypothesis testing, and so the properties of hypothesis tests can be easily deduced from those of interval estimation procedures. However, as will be seen later in the sequel, this is not always the case. Another important problem that has not been well addressed in the literature is the sample size calculation for computing confidence intervals with a specified precision (margin of error) or for hypothesis testing with a specified power. For example, as the score interval is recommended for applications, one may want to know the reduction in sample size by using the score interval instead of the exact interval.

The primary goal of this article is to outline sample size calculation methods for Wilson's approach and the exact methods, and provide a useful reference for practitioners of statistics. Keeping these objectives in mind, this article is organized as follows. In the following section, we outline the score test, exact test, and exact method of computing their powers. Exact size and power properties of the tests are presented. Sample size computation for a given power and nominal level is outlined. Exact sample sizes required to attain a power of 0.80 at level 0.05 are given in

Table 1(a)

Sample sizes for the score and exact tests for testing $H_0 : p \leq p_0$ vs. $H_a : p > p_0$ when $\alpha = 0.05$ and power is 0.80

	p									
p_0	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50
.90	179(.049)									
.85	59(.047)	281(.050)								
.80	30(.044)	82(.046)	365(.048)							
.75	23(.049)	45(.045)	103(.046)	437(.050)						
.70	14(.047)	28(.047)	49(.048)	119(.048)	501(.050)					
.65	12(.042)	20(.044)	31(.046)	55(.049)	133(.048)	545(.050)				
.60	10(.046)	14(.040)	21(.037)	36(.045)	62(.049)	143(.048)	585(.049)			
.55	9(.039)	12(.042)	15(.042)	24(.036)	37(.043)	70(.045)	150(.050)	613(.049)		
.50	8(.035)	8(.035)	13(.046)	18(.048)	23(.047)	37(.049)	69(.046)	158(.047)	620(.050)	
.45	4(.041)	7(.036)	9(.050)	14(.043)	16(.049)	25(.044)	42(.042)	70(.047)	154(.050)	618(.050)
.40	4(.026)	6(.041)	8(.050)	11(.029)	15(.034)	19(.035)	28(.050)	42(.038)	71(.044)	158(.048)
.35	3(.043)	6(.022)	8(.025)	13(.046)	13(.046)	13(.046)	19(.035)	26(.038)	41(.048)	68(.046)
.30	3(.027)	5(.031)	5(.031)	7(.029)	10(.047)	10(.047)	14(.031)	17(.040)	25(.044)	39(.050)
.25	6(.038)	6(.038)	6(.038)	6(.038)	6(.038)	8(.027)	9(.049)	14(.038)	17(.040)	26(.040)
	.45	.40	.35	.30	.25	.20	.15	.10	.05	
.40	604(.050)									
.35	148(.049)	584(.049)								
.30	67(.047)	144(.047)	549(.050)							
.25	36(.046)	62(.043)	129(.049)	494(.050)						
.20	21(.043)	35(.034)	56(.043)	116(.049)	433(.049)					
.15	14(.047)	22(.037)	28(.049)	48(.048)	101(.043)	360(.048)				
.10	13(.034)	13(.034)	19(.035)	26(.040)	40(.042)	78(.045)	270(.047)			
.05	6(.033)	7(.044)	14(.030)	14(.030)	16(.043)	27(.044)	52(.045)	169(.045)		
.01	3(.030)	4(.039)	4(.039)	5(.049)	23(.022)	23(.022)	23(.022)	29(.034)	121(.034)	

Table 1(b)

Sample size for testing $H_0 : p = p_0$ vs. $H_a : p \neq p_0$ when $\alpha = 0.05$ and power is 0.80; S–score test, E–exact test

		p									
p_0	Test	.95	.90	.85	.80	.75	.70	.65	.60	.55	.50
.90	S	231(.048)									
	E	231(.048)									
.85	S	74(.049)	365(.047)								
	E	75(.037)	356(.047)								
.80	S	41(.049)	114(.046)	466(.049)							
	E	41(.049)	107(.040)	466(.049)							
.75	S	27(.042)	55(.041)	133(.045)	553(.049)						
	E	27(.042)	54(.041)	132(.044)	553(.049)						
.70	S	23(.037)	31(.048)	61(.049)	156(.044)	641(.047)					
	E	22(.035)	31(.048)	61(.049)	155(.044)	633(.046)					
.65	S	14(.045)	22(.042)	41(.048)	72(.048)	171(.045)	701(.048)				
	E	14(.045)	22(.042)	41(.048)	73(.037)	167(.043)	701(.048)				
.60	S	12(.035)	20(.037)	26(.043)	45(.047)	77(.047)	181(.048)	733(.050)			
	E	12(.035)	19(.035)	26(.043)	45(.034)	80(.040)	181(.048)	744(.047)			
.55	S	11(.029)	13(.047)	20(.040)	31(.045)	44(.048)	83(.046)	191(.049)	765(.050)		
	E	10(.028)	14(.028)	20(.040)	28(.037)	47(.040)	83(.046)	191(.049)	775(.047)		
.50	S	9(.039)	12(.039)	15(.035)	20(.041)	30(.043)	49(.044)	90(.045)	199(.047)	786(.050)	
	E	9(.039)	12(.039)	15(.035)	20(.041)	30(.043)	49(.044)	90(.045)	199(.047)	786(.050)	
.45	S	4(.041)	8(.026)	13(.047)	15(.036)	23(.034)	31(.045)	50(.045)	84(.048)	199(.046)	785(.048)
	E	8(.026)	8(.026)	13(.025)	18(.030)	22(.033)	31(.045)	50(.045)	89(.042)	199(.046)	789(.049)
.40	S	4(.026)	7(.047)	9(.035)	13(.045)	16(.037)	21(.046)	32(.046)	49(.040)	87(.049)	193(.047)
	E	7(.019)	7(.019)	10(.018)	14(.026)	16(.037)	22(.029)	32(.046)	49(.040)	90(.040)	195(.048)
.35	S	3(.043)	6(.022)	10(.039)	10(.039)	12(.031)	16(.033)	23(.046)	28(.047)	48(.048)	81(.048)
	E	4(.015)	6(.022)	9(.032)	11(.021)	14(.045)	16(.033)	24(.030)	34(.046)	48(.048)	85(.040)
.30	S	3(.027)	5(.031)	5(.031)	7(.029)	11(.041)	13(.028)	14(.038)	19(.043)	29(.041)	45(.049)
	E	4(.008)	6(.011)	8(.011)	10(.011)	11(.041)	13(.028)	18(.035)	23(.037)	31(.048)	47(.037)
.25	S	3(.016)	5(.016)	5(.016)	6(.038)	6(.038)	8(.027)	9(.049)	16(.037)	17(.048)	28(.046)
	E	3(.016)	5(.016)	5(.016)	7(.013)	9(.010)	10(.020)	13(.048)	17(.020)	21(.040)	30(.032)

(continued)

Table 1(b)
Continued

		p								
p_0	Test	.45	.40	.35	.30	.25	.20	.15	.10	.05
.40	S	756(.049)								
	E	767(.047)								
.35	S	178(.049)	714(.050)							
	E	192(.041)	737(.049)							
.30	S	76(.045)	175(.047)	680(.049)						
	E	83(.041)	183(.043)	689(.046)						
.25	S	41(.046)	73(.042)	153(.050)	607(.049)					
	E	46(.039)	78(.035)	165(.047)	624(.047)					
.20	S	24(.041)	38(.040)	66(.045)	141(.045)	524(.049)				
	E	26(.046)	41(.049)	72(.038)	144(.047)	540(.046)				
.15	S	14(.047)	24(.040)	31(.041)	51(.047)	114(.048)	428(.049)			
	E	16(.024)	24(.040)	34(.028)	62(.030)	127(.046)	449(.047)			
.10	S	11(.019)	13(.034)	18(.028)	25(.033)	44(.038)	83(.044)	313(.048)		
	E	11(.019)	16(.017)	22(.018)	29(.022)	49(.027)	94(.036)	341(.047)		
.05	S	6(.033)	7(.044)	12(.020)	14(.030)	16(.043)	27(.044)	52(.045)	191(.044)	
	E	9(.008)	10(.012)	12(.020)	18(.011)	21(.019)	33(.023)	67(.018)	213(.039)	
.01	S	3(.030)	4(.039)	4(.039)	5(.049)	18(.014)	18(.014)	19(.015)	29(.034)	110(.025)
	E	6(.001)	7(.002)	8(.003)	9(.003)	11(.005)	14(.008)	19(.015)	42(.009)	134(.012)

Table 2
Sample size for constructing 95% confidence intervals with a given precision;
S–score method, E–exact method

		Precision d							
p	Method	.20	.15	.10	.05	.04	.03	.02	.01
.05	S					125(.964)[1]	215(.961)	474(.956)[6]	1837(.953)
	E					136(.973)	233(.963)[1]	508(.959)	1927(.959)
.10	S				141(.952)[1]	223(.957)[5]	394(.957)[7]	867(.953)	3464(.953)[4]
	E				155(.956)	238(.961)	414(.960)	914(.959)	3557(.953)
.15	S			48(.960)	195(.956)	306(.955)[1]	545(.953)[1]	1224(.950)	4897(.950)
	E			56(.961)	213(.957)	328(.956)	576(.959)	1273(.955)	4997(.952)
.20	S		26(.954)[1]	60(.966)[1]	244(.955)	384(.952)[1]	681(.951)	1537(.952)[3]	6144(.950)
	E		31(.979)	68(.967)	263(.963)	406(.953)	715(.956)	1585(.952)	6245(.952)
.25	S	15(.969)	30(.968)[1]	72(.960)[3]	285(.953)	447(.951)	797(.950)	1798(.950)	7200(.952)
	E	20(.962)	35(.971)	79(.963)	305(.953)	472(.957)	832(.955)	1849(.953)	7301(.952)
.30	S	18(.965)[1]	33(.966)[1]	78(.953)[1]	320(.956)[1]	503(.954)[3]	894(.951)[1]	2013(.951)	8065(.951)[2]
	E	22(.965)	39(.965)	88(.965)	340(.956)	527(.954)	928(.955)	2065(.951)	8165(.951)
.35	S	18(.955)	36(.965)[1]	83(.950)	346(.952)	542(.953)	974(.953)[7]	2181(.952)	8735(.950)
	E	24(.970)	42(.966)	94(.961)	366(.958)	569(.957)	1002(.953)	2233(.952)	8837(.952)
.40	S	*20(.963)[1]	37(.956)	88(.950)	365(.952)	572(.950)	1021(.952)[1]	2305(.952)[4]	9215(.951)
	E	25(.977)	44(.970)	99(.960)	386(.952)	599(.955)	1056(.956)	2353(.952)	9317(.952)
.45	S	20(.960)[1]	40(.963)[2]	91(.955)	376(.951)	590(.953)	1052(.953)	2373(.950)	9509(.951)[6]
	E	26(.971)	46(.963)	102(.964)	397(.956)	617(.957)	1088(.956)	2425(.952)	9605(.951)
.50	S	20(.959)	40(.966)	92(.953)	380(.955)	597(.951)[1]	1063(.950)	2397(.950)	9600(.951)[1]
	E	26(.971)	46(.974)	103(.952)	401(.954)	623(.955)	1098(.950)	2449(.952)	9701(.951)

*The approximate sample size in (18) is the reported exact sample size for the score interval minus the number in the superscript.

Tables 1(a) and 1(b). Also, convenient simple approximations to the sample sizes are given for one-tail and two-tail tests. In Sec. 3, we first outline the score and the exact confidence interval procedures. Expressions for computing exact expected lengths are given. Exact sample sizes required to compute 95% confidence intervals with various precisions are evaluated and presented in Table 2. We also provide simple approximations to compute the sample sizes. Comparison of the exact sample sizes with those based on the approximations indicate that the approximations are remarkably accurate for the score intervals. Some illustrations for using table values are given in Sec. 4, and some concluding remarks are given in Sec. 5.

2. The Tests

Let X be a binomial (n, p) random variable with probability mass function (pmf)

$$f(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{1}$$

Suppose we want to test

$$H_0 : p \le p_0 \quad \text{vs.} \quad H_a : p > p_0, \tag{2}$$

where p_0 is a specified value of p, based on an observed value k of X.

2.1. *The Score Test*

The score test is based on the z-score statistic

$$Z(X, n, p_0) = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}, \tag{3}$$

where the sample proportion $\hat{p} = X/n$. This test rejects the null hypothesis in (2) when $Z(k, n, p_0) \ge z_\alpha$, where z_α denotes the upper αth quantile of the standard normal distribution.

For testing two-sided alternative hypothesis, that is, when

$$H_0 : p = p_0 \quad \text{vs.} \quad H_a : p \ne p_0, \tag{4}$$

the score test rejects the null hypothesis in (4) when $|Z(k, n, p_0)| \ge z_{\alpha/2}$.

2.2. *The Exact Test*

The exact test is based on the exact p-value that can be computed using the binomial pmf. In particular, the p-value for testing (2) is given by

$$P(X \ge k \mid n, p_0) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}. \tag{5}$$

This exact test that rejects H_0 in (2) when the p-value in (5) is less than or equal to α. The p-value for a two-sided alternative hypothesis, that is, for testing (4), is given by $2\min\{P(X \le k \mid n, p_0), P(X \ge k \mid n, p_0)\}$.

These exact tests are level α tests in the sense that the Type I error rates never exceed the nominal level α for all (n, p) configurations.

2.3. *Size and Power Properties of the Tests*

The exact power of the score test for testing one-sided hypotheses in (2) is given by

$$P\big(Z(X, n, p_0) \geq z_\alpha \mid n, p) = P\big(X \geq [np_0 + z_\alpha\sqrt{np_0(1-p_0)}]_+ \mid n, p\big), \tag{6}$$

where $X \sim$ binomial(n, p) and $[x]_+$ denotes the smallest integer greater than or equal to x. Note that, when $p = p_0$, the above expression gives the size (Type-I error rate) of the score test. The power function for testing (4) is given by

$$P\big(X \geq [np_0 + z_{\alpha/2}\sqrt{np_0(1-p_0)}]_+ \mid n, p\big) + P\big(X \leq [np_0 - z_{\alpha/2}\sqrt{np_0(1-p_0)}]_- \mid n, p\big), \tag{7}$$

where $[x]_-$ denotes the largest integer less than or equal to x. For a given $p \neq p_0$, power β and level α, the sample size required for a two-tail score test is the smallest value of n for which the above power in (2) is at least β, and the Type-I error rate is at most α.

The power of the exact test for testing (2) is given by

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} I(P(X \geq k \mid n, p_0) \leq \alpha), \tag{8}$$

where $I(.)$ is the indicator function. We note that the expression in (8) with $p = p_0$ gives the size of the exact test. A power expression for a two-tail test can be obtained using (8) with the indicator function replaced by

$$I(2\min\{P(X \leq k \mid n, p_0), P(X \geq k \mid n, p_0)\} \leq \alpha).$$

The exact sizes and powers of the score test and the exact test are computed using (6) and (8), respectively. The sizes and powers of both tests are plotted as a function of n in Fig. 1(a) for testing $H_0 : p \leq .3$ vs. $H_a : p > .3$; powers are computed at $p = .35$. It is clear from these graphs that the sizes of the score test are always greater than or equal to those of the exact test. In particular, we note from Fig. 1(a) that whenever the size of the score test is below the nominal level it coincides with that of the exact test. The fluctuation in powers reflects the size behaviors for all n. Specifically, whenever the size of the score test is above the nominal level .05, it offers more power than the exact test; otherwise the powers of the tests are the same. These findings indicate that the sample sizes needed for both tests to attain a given power are the same if the score test is required control the Type-I error rate within the nominal level. We also computed the sizes and powers of the tests as a function of n for other values of p. As the plots exhibited similar patterns as those in Fig. 1(a), they are not presented here.

In Fig. 1(b), we plotted the sizes and powers of the tests as a function of n for a two-tail test. We first observe that the size behaviors are different from those for the right-tail test given in Fig. 1(a). We also see from Fig. 1(b) that the sizes of the score test are always larger than that of the exact test; however, there are many sample

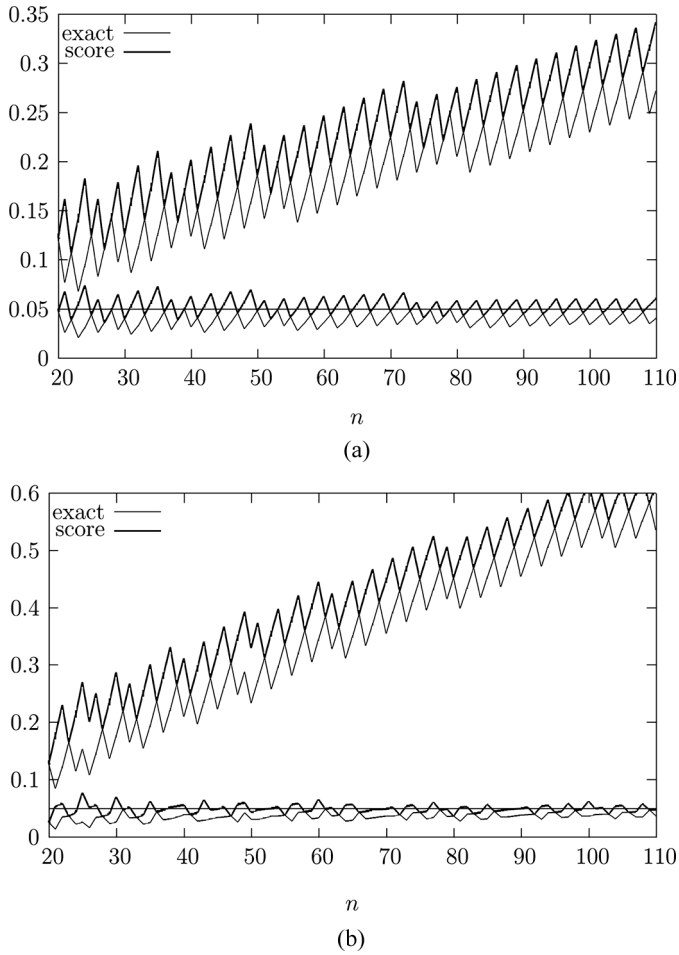

Figure 1. Sizes and powers of the tests for testing (a) $H_0 : p \leq .3$ vs. $H_a : p > .3$ at $\alpha = .05$; (b) $H_0 : p = .3$ vs. $H_a : p \neq .3$ at $\alpha = .05$ powers are computed at $p = .35$.

sizes for which the Type-I error rates of the score test are below the nominal level and greater than those of the exact test. For these sample sizes, the score test not only controls the Type-I error rates, but also provides large power than the exact test.

2.4. *Sample Size Calculation for Hypothesis Tests*

For a given p, p_0, and α, the sample size required for the score test (for one-sided hypothesis) to attain a power of β is the smallest value of n for which the power in (6) is at least β. That is, the smallest sample size for which

$$P\big(X \geq [np_0 + z_\alpha\sqrt{np_0(1-p_0)}]_+ \mid n, p\big) \geq \beta. \tag{9}$$

Notice that the the power functions of both tests are oscillating with respect to the sample size n (see Fig. 1(a)); the power does increase for a large increase in sample size, but it may decrease for small changes in sample size. Therefore, to compute

an accurate sample size, a forward search method, starting from a small value of n, can be used to find the smallest value of n that satisfies (9). Any other search method may produce a sample size that is unnecessarily larger than the one required to attain a specified power. Using (7), the sample size for a two-tail score test can be computed similarly.

An approximation to the sample size can be obtained using the normal approximation to the binomial(n, p) distribution. Specifically, an approximation to the power in (9) can be expressed as

$$1 - \Phi\left(\frac{np_0 + z_\alpha\sqrt{np_0q_0} - np}{\sqrt{npq}}\right), \tag{10}$$

where $q = 1 - p$, $q_0 = 1 - p_0$, and Φ is the standard normal distribution function. Equating the above power to β and solving for n, we get

$$n \simeq \frac{(z_\beta\sqrt{pq} - z_\alpha\sqrt{p_0q_0})^2}{(p - p_0)^2}. \tag{11}$$

It should be noted that z_c, for $0 < c < 1$, is the upper cth quantile of the standard normal distribution. If the hypotheses are two-sided, then it is not feasible to obtain a simple approximation to the sample size using the approach given for one-sided hypotheses. An intuitive approximation is (11) with z_α replaced by $z_{\alpha/2}$. That is,

$$n \simeq \frac{(z_\beta\sqrt{pq} - z_{\alpha/2}\sqrt{p_0q_0})^2}{(p - p_0)^2}. \tag{12}$$

The sample size for the one-sided exact test is the smallest value of n for which the power in (8) is at least β. That is, the least value of n for which

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}I(P(X \geq k \mid n, p_0) \leq \alpha) \geq \beta. \tag{13}$$

The sample size for a two-tail test can be expressed similarly.

For any given p, p_0, α, and power β, we first determine the sample size n_s for the score test to attain the power of at least β and the size at most α, and then we determine the sample size for the exact test using forward search starting from n_s. Using this approach, we computed the sample sizes required for right-tail tests for various values of p and p_0, $\alpha = 0.05$ and power 0.80. As the size and power studies in Sec. 2.3 indicated, the attained powers of both tests are the same for all the cases considered in Table 1(a). For example, when $n = 49$, the power of the score test is 0.809 at $(p, p_0) = (.85, .70)$ and the attained level is 0.048. At this sample size–parameter configuration, the exact test also attained the same level and power. This is true for all the cases reported in Table 1(a).

Sample sizes are presented in Table 1(a) for right-tail tests; the sample size required for a left-tail test can be obtained from Table 1(a) (see Sec. 4). If the hypotheses are two-sided, then there are many situations where the score and the exact tests require different sample sizes, often the score test requires smaller samples than the exact test (see Table 1(b)). Thus, for two-sided hypotheses, the score test may be preferable to the exact test.

Table 3
Comparison of the exact and approximate sample sizes for the score test to attain a power of at least 0.80 at level 0.05

	Right-tail test		Two-tail test	
(p, p_0)	Approx.	Exact	Approx.	Exact
(.95, .90)	183	179	233	231
(.85, .80)	367	365	475	466
(.90, .85)	282	281	362	365
(.75, .50)	22	23	28	30
(.65, .60)	582	585	741	733
(.85, .60)	19	21	25	26
(.75, .60)	60	62	77	77
(.45, .30)	61	67	77	76
(.40, .35)	572	584	725	714

We also checked the approximations in (11) and (12) for their accuracies. Sample sizes based on these approximations and the corresponding exact ones are given in Table 3 for some values of (p, p_0). Comparison of the sample sizes shows that the approximation is not very satisfactory in some situations. Nevertheless, the approximation can provide a reasonable estimate of the sample size, and it can be used as an initial value for determining the exact sample size using (9).

3. Confidence Intervals

We shall now present the score intervals, exact intervals, and expressions for computing their expected lengths.

3.1. *Score Intervals and Expected Lengths*

The score confidence interval is obtained by inverting the score test statistic. Specifically, solving the inequality

$$\left| \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \right| \leq z_{\alpha/2}$$

for p, we get the score interval as

$$\left(\frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} \right) \pm \frac{\frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\hat{p}(1-\hat{p}) + z_{\alpha/2}^2/(4n)}}{1 + \frac{z_{\alpha/2}^2}{n}}. \tag{14}$$

One-sided limits can be obtained by replacing $z_{\alpha/2}$ by z_α.

One half of the expected length of the score interval in (14) is given by

$$\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \left(\frac{\frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{(x/n)(1-x/n) + z_{\alpha/2}^2/(4n)}}{1 + \frac{z_{\alpha/2}^2}{n}} \right). \tag{15}$$

An approximation to the above expression can be obtained using the following lemma.

Lemma 3.1. *Let $f(\hat{p})$ be a real valued function. Then,*

$$Ef(\hat{p}) = f(p) + O(n^{-1}). \tag{16}$$

Proof. Using a Taylor series expansion around p, we have

$$f(\hat{p}) = f(p) + (\hat{p} - p)f'(\hat{p})\,|_{\hat{p}=p} + \frac{(\hat{p} - p)^2}{2!} f''(\hat{p})|_{\hat{p}=p} + \cdots .$$

Now, taking expectation on both sides, we get (16).

Applying the above lemma, we see that (15) is approximately equal to

$$\left(\frac{\frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{p(1-p) + z_{\alpha/2}^2/(4n)}}{1 + \frac{z_{\alpha/2}^2}{n}} \right). \tag{17}$$

An approximation to the sample size n that is required to construct a confidence interval with a given precision d can be calculated as follows. Setting (17) equal to d, and solving the resulting equation for n, we get

$$n \simeq \frac{z_{\alpha/2}^2 \left[(pq - 2d^2) + \sqrt{(pq - 2d^2)^2 - d^2(4d^2 - 1)} \right]}{2d^2}, \tag{18}$$

where $q = 1 - p$. Similarly, an approximate sample size to find a $1 - \alpha$ upper limit within a precision of d from the true p can be computed as the solution of the equation

$$\left(\frac{\frac{z_{\alpha}}{\sqrt{n}} \sqrt{pq + z_{\alpha}^2/(4n)}}{1 + \frac{z_{\alpha}^2}{n}} \right) - p = d. \tag{19}$$

Solving (19) for n, we get

$$n \simeq \frac{z_{\alpha}^2 \left[pq + d - 2d(p+d) + \sqrt{(pq + d - 2d(p+d))^2 - d^2((2(p+d) - 1)^2 - 1)} \right]}{2d^2}. \tag{20}$$

These above approximations are very satisfactory as will be shown later in Sec. 3.3.

3.2. *Exact Confidence Intervals and Expected Lengths*

The endpoints of the Clopper and Pearson (1934) confidence interval for a binomial proportion p can be obtained as solutions of the following two equations. In particular, for a given sample size n and an observed number of successes k, the lower limit p_L for p is the solution of the equation

$$\sum_{i=k}^{n} \binom{n}{i} p_L^i (1 - p_L)^{n-i} = \frac{\alpha}{2},$$

and the upper limit p_U is the solution of the equation

$$\sum_{i=0}^{k} \binom{n}{i} p_U^i (1-p_U)^{n-i} = \frac{\alpha}{2}.$$

Using a relation between the binomial and beta distributions, the endpoints can be expressed as

$$p_L = B^{-1}(\alpha/2; k, n-k+1) \quad \text{and} \quad p_U = B^{-1}(1-\alpha/2; k+1, n-k),$$

where $B^{-1}(c; a, b)$ denotes the cth quantile of a beta distribution with the shape parameters a and b. The interval (p_L, p_U) is an exact $1-\alpha$ confidence interval for p, in the sense that the coverage probability is always greater than or equal the specified confidence level $1-\alpha$. One-sided $1-\alpha$ lower limit for p is $B^{-1}(\alpha; k, n-k+1)$ and one-sided $1-\alpha$ upper limit for p is $B^{-1}(1-\alpha; k+1, n-k)$. When $k=n$, the lower limit is $\alpha^{\frac{1}{n}}$ and the upper limit is set to be 1; when $k=0$, the upper limit is $1-\alpha^{\frac{1}{n}}$ and the lower limit is set to be 0.

For a given confidence level $1-\alpha$ and p, the expected length of (p_L, p_U) is given by

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (p_U - p_L), \tag{21}$$

where p_L and p_U are as defined above. Suppose that one wants to compute the sample size required to have a $1-\alpha$ confidence interval with a specified precision d then the sample size can be computed as the solution of the equation

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} (p_U - p_L) = 2d. \tag{22}$$

For a given p and d, the required sample size for estimating the proportion within the precision d is the smallest value of n for which the above expected length is less than or equal to $2d$. An approximation to the expected length in (21) can be obtained using Lemma 3.1, and is given by

$$B^{-1}(1-\alpha/2; np+1, n-np) - B^{-1}(\alpha/2; np, n-np+1). \tag{23}$$

Our numerical comparison (not reported here) of (23) and the exact expected lengths in (21) showed that the approximation is satisfactory for $n \geq 40$, and is accurate up to two decimal places for $n \geq 100$ regardless of the values of p.

3.3. *Sample Size Calculation for a Given Precision*

We plotted the exact expected lengths of the score and exact intervals as a function of p in Fig. 2(a) and as a function of n in Fig. 2(b). All the plots clearly indicate that the score intervals are in general shorter (not withstanding coverage probabilities) than the exact intervals. The difference between the expected lengths appears to diminish as the sample size increases. Furthermore, we see in Fig. 2(a) that the

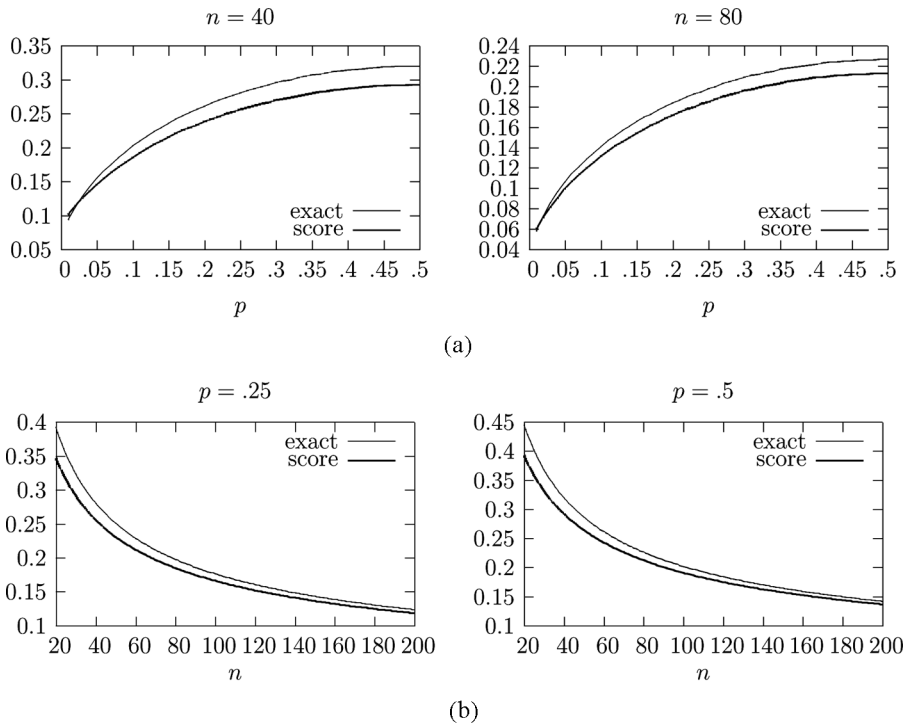

Figure 2. Expected lengths of exact and score confidence intervals (a) as a function of p; (b) as a function of n.

expected lengths of both confidence intervals are increasing with increasing p in (0, .5] and in Fig. 2(b) that they are decreasing with increasing n.

For a given precision d, value of p and confidence level $1 - \alpha$, the exact sample size required to compute the score interval is the smallest value of n for which (15) is less than or equal to d and the coverage probability is at least $1 - \alpha$. We first computed the sample size n_s using the approximation in (18), and then using the n_s as an initial value in (15) we carried out forward/backward search to find the exact sample size required to construct the score interval with a specified precision and coverage probability at least $1 - \alpha$. The sample sizes for the exact intervals are computed similarly using n_s as an initial value in (22), and a forward search method. The sample sizes for 95% confidence intervals are given in Table 2.

We also computed the approximate sample sizes for the score intervals using (18), and compared them with the exact ones based on (15). Our comparison indicted that the approximate sample sizes are identical to those based on the exact approach in many cases. In general, the approximation is very satisfactory. The cases where the approximate sample sizes are not equal to the exact ones are indicated in Table 2. For example, when $p = .10$, $d = .01$ and the confidence level is 0.95, the approximate sample size based on (18) is 3,460 and the exact one is 3,464, which is four units greater than the exact one; this is indicated by placing number 4 at the superscript (see Table 2). As shown in Table 2, in many cases the approximate sample sizes coincide with those based on the exact methods. We

observe from Table 2 that the sample sizes needed for the score intervals are smaller than those for the exact intervals for all the cases.

4. Some Illustrations for Using Tables

Sample size for a given power: Suppose that one wants to determine the sample size for testing $H_0 : p \leq 0.6$ vs. $H_a : p > 0.6$ at level 0.05 when the true value of p is 0.7. Then the required sample size for either of the tests to get the power of 0.80 can be obtained by referring to the value $(p, p_0) = (.7, .6)$ in Table 1(a), and is 143; the actual Type-I error rate is 0.048. If it is a two-tail test, that is, $H_0 : p = 0.6$ vs. $H_a : p \neq 0.6$, then the sample size can be found by referring to $(p, p_0) = (.7, .6)$ in Table 1(b), and is 181; the actual size is 0.048.

Sample size for a left-tail test can be obtained from Table 1a as follows. We first note that, testing $H_0 : p \geq p_0$ vs. $H_a : p < p_0$ based on X is equivalent to testing $H_0 : q \leq q_0$ vs. $H_a : q > q_0$ based on $n - X$, because the number of failures $n - X$ also follows the binomial(n, q) distribution. For example, the problem of finding the sample size for testing $H_0 : p \geq .3$ vs. $H_a : p < .3$ when the true value of $p = 0.25$ is equivalent to the one for testing $H_0 : q \leq .7$ vs. $H_a : q > .7$ when the true value of $q = 1 - p = .75$. Thus, treating (q, q_0) as (p, p_0), and referring to $(p, p_0) = (.75, .7)$ in Table 1(a), we get the sample size for the exact test (or for the score test) as 501.

Sample size for confidence intervals: Suppose that a researcher believes that the true proportion p of individuals with an attribute of interest in a population is 0.20, and he wants to construct a 95% confidence interval for p within a margin of error $\pm 1\%$. Then, by referring to $p = .2$ and precision 0.01 in Table 2, we get 6,144 for the score interval and 6,245 for the exact interval. If $p > .5$, then referring to $1 - p$ and the specified precision, we can get the sample size from Table 2. For example, if $p = .6$, then the required sample size to estimate p within $\pm 5\%$ can be found by referring to the value $(1 - p, d) = (.4, .05)$ in Table 2, and is 365 for the score interval and 386 for the exact interval.

5. Concluding Remarks

In this article, we have shown that, if the sample size for a test has to be determined so that it should control both Type-I and Type-II error rates, then for one-sided hypotheses, the score and exact tests require the same sample size. This finding indicates the sample size for the exact test (if one wants to use the exact test) can be computed using the power expression in (6) because it is easier to compute than the power of the exact test in (8). However, for two-sided hypotheses, the score test can be recommended for applications as it requires smaller samples in many situations. The score method is also preferable to the exact method for constructing confidence intervals as the former requires considerably smaller samples than the latter to attain the same power. We hope that our findings are useful for researcher and practitioners to choose between the score and exact methods for applications.

References

Agresti, A., Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportion. *Amer. Statistician* 52:119–125.

Blyth, C. R., Hutchinson, D. W. (1960). Tables of Neyman-shortest confidence intervals for the binomial parameter. *Biometrika* 47:381–391.

Blyth, C. R., Still, H. A. (1983). Binomial confidence intervals. *J. Amer. Statist. Assoc.* 78:108–116.

Brown, L. D., Cai, T., Das Gupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* 16:101–133.

Cai, T. (2005). One-sided confidence intervals in discrete distributions. *J. Statist. Plan. Infer.* 131:63–88.

Cai, Y., Krishnamoorthy, K. (2005). A simple improved inferential method for some discrete distributions. *Computat. Statist. Data Anal.* 48:605–621.

Casella, G. (2001). Comment on Brown, Cai, and DasGupta A. *Statist. Sci.* 16:120–122.

Clopper, C. J., Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of binomial. *Biometrika* 26:404–413.

Geyer, C. J., Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and p-values. *Statist. Sci.* 20:358–366.

Wilson, E. B. (1927). Probable inference, the law of successions and statistical inference. *J. Amer. Statist. Assoc.* 22:209–212.