The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

A "Paradox" in Confidence Interval Construction Using Sufficient Statistics

Weizhen Wang

To cite this article: Weizhen Wang (2018) A "Paradox" in Confidence Interval Construction Using Sufficient Statistics, The American Statistician, 72:4, 315-320, DOI: 10.1080/00031305.2017.1305292

To link to this article: https://doi.org/10.1080/00031305.2017.1305292

Accepted author version posted online: 30 Mar 2017.
Published online: 10 May 2018.

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 4, 315–320: General
https://doi.org/10.1080/00031305.2017.1305292

A "Paradox" in Confidence Interval Construction Using Sufficient Statistics

Weizhen Wang

College of Applied Sciences, Beijing University of Technology, Beijing, P. R. China; Department of Mathematics and Statistics, Wright State University, Dayton, OH

ABSTRACT

Statistical inference about parameters should depend on raw data only through sufficient statistics—the well known sufficiency principle. In particular, inference should depend on minimal sufficient statistics if these are simpler than the raw data. In this article, we construct one-sided confidence intervals for a proportion which: (i) depend on the raw binary data, and (ii) are uniformly shorter than the smallest intervals based on the binomial random variable—a minimal sufficient statistic. In practice, randomized confidence intervals are seldom used. The proposed intervals violate the aforementioned principle if the search of optimal intervals is restricted within the class of nonrandomized confidence intervals. Similar results occur for other discrete distributions.

ARTICLE HISTORY
Received November 2015
Revised February 2017

KEYWORDS
Admissible confidence interval; Binomial distribution; Nonrandomized inference; One-sided confidence interval; Order

1. Introduction

Sufficiency is a fundamental concept in statistical inference, especially when a parametric model is assumed. For any given sufficient statistic, including any minimal sufficient statistic, statistical inference should depend on the raw data only through the sufficient statistic. This is the well known sufficiency principle (Casella and Berger 2002, p. 272), which is supported by at least two results.

(I) The Rao-Blackwell Theorem: Conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement in variance (Casella and Berger 2002, p. 342).

(II) For any given level-α test with a critical function ϕ, there exists another test ϕ' that depends on a sufficient statistic and has the same power function as the given test. In fact, ϕ' is equal to the conditional expectation of ϕ for the sufficient statistic.

The second result is mathematically correct, but it has a problem in practice. From a mathematical point of view, the critical function ϕ can take any values between 0 and 1. If ϕ only assumes two values: 0 and 1, then it is called crisp in terms of Fuzzy Set Theory (Geyer and Meeden 2005), and a crisp critical function is equivalent to the indicator function of a rejection region. We call ϕ a nonrandomized test; that is, for a given observation $\underline{x}$, either reject or fail to reject the null hypothesis if $\phi(\underline{x}) = 1$ or $\phi(\underline{x}) = 0$, respectively. Hence, the decision by a crisp function ϕ is unique at each $\underline{x}$. For a critical function ϕ that can take values in $[0, 1]$, if $\phi(\underline{x}) = 0.351$ say, then one has to use a uniform random variable U on the interval $[0, 1]$ that is independent of $\underline{X}$ to generate a value u. If u is no larger than 0.351, then reject the null hypothesis, otherwise, fail to reject the null hypothesis. Thus, if u is generated for the second time, the decision may be different from the previous one even for the same critical function ϕ and the same observation $\underline{x}$. This uncertainty makes practitioners reluctant to apply randomized tests. The trouble is, when the observation $\underline{X}$ follows a discrete distribution, ϕ' in result II) is typically a randomized test even for a nonrandomized ϕ. That is, result II) is not necessarily true if only nonrandomized tests are used. In other words, randomization is a prerequisite for the sufficiency principle.

Similarly, there are nonrandomized and randomized confidence intervals. Let $C(\underline{X}) = [L(\underline{X}), U(\underline{X})]$ be a confidence interval for a parameter of interest θ, and let $P(\underline{X}, \theta) \in [0, 1]$ be the associated membership function (Geyer and Meeden 2005). For a given observation $\underline{x}$ and a fixed value θ, the closer $P(\underline{x}, \theta)$ is to 1, the more we think θ is in $C(\underline{x})$; the closer $P(\underline{x}, \theta)$ is to 0, the less we think θ is in $C(\underline{x})$. If $C(\underline{X})$ is nonrandomized, then $P(\underline{X}, \theta)$ assumes two values 0 and 1, and is equal to the indicator function $I_{C(\underline{X})}(\theta)$. For a given observation $\underline{x}$, we think (but we may be wrong) that θ is contained in interval $[L(\underline{x}), U(\underline{x})]$ if and only if $P(\underline{x}, \theta)$ is equal to 1. If $C(\underline{X})$ is randomized, then $P(\underline{X}, \theta)$ may assume any values between 0 and 1. Suppose $P(\underline{x}, \theta) = 0.649$. One has to use a uniform random variable U on the interval $[0, 1]$ that is independent of $\underline{X}$ to generate a value u. If u is less than or equal to 0.649, then we think that θ is contained in $C(\underline{x})$; otherwise θ is contained in the complement of $C(\underline{x})$. Obviously, if one generates u twice, the results may be different. Therefore, similar to randomized tests, a randomized interval $C(\underline{X})$ may not yield a unique decision on a given observation $\underline{x}$, so it is not attractive to practitioners either. In this article, we focus only on nonrandomized confidence intervals. For simplicity, we just call them confidence intervals if it causes no confusion.

Throughout this article, suppose $T(\underline{X})$ is a sufficient statistic and $\sigma(T)$, the $\sigma-$algebra generated by T, is a proper subset of $\sigma(\underline{X})$, the σ-algebra generated by $\underline{X}$. Roughly speaking, $T(\underline{X})$ is simpler than $\underline{X}$.

CONTACT Weizhen Wang *weizhen.wang@wright.edu* College of Applied Sciences, Beijing University of Technology, Beijing 100124, P. R. China.

If the sufficiency principle is still valid for the search of optimal confidence intervals, then the following claim would be true:

For any $1-\alpha$ confidence interval $C(\underline{X})$, there exists a $1-\alpha$ confidence interval $C'(T)$ so that $C'(T)$ is a subset of $C(\underline{X})$. (1.1)

An interval $C(\underline{X})$ is said to be of level $1-\alpha$ if its coverage probability function is no less than $1-\alpha$. We will construct a counter example to (1.1) under a basic model: the binomial experiment. In fact, we establish the opposite of (1.1):

For any $1-\alpha$ confidence interval $C'(T)$, there exists a $1-\alpha$ confidence interval $C(\underline{X})$ so that $C(\underline{X})$ is a proper subset of $C'(T)$. (1.2)

Section 2 provides a complete construction of $1-\alpha$ one-sided confidence intervals for a proportion p based on the raw binary data. Section 3 contains discussion. All proofs are given in the Appendix.

2. One-Sided Confidence Intervals for a Proportion

Let S denote the sample space of n independent and identical Bernoulli trials $X_i \sim \text{Bino}(1, p)$. The goal is to estimate p with a given confidence level by observing successes in these n trials. Then

$$S = \{\underline{x} = (x_1, \ldots, x_n) : x_i = 0 \text{ or } 1\}, \tag{2.1}$$

where $x_i = 1$ means that a success is observed in the ith trial, $x_i = 0$ indicates an observed failure, and the probability mass function is given by

$$p_S(x_1, \ldots, x_n; p) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \tag{2.2}$$

It is well known that $S_n = \sum_{i=1}^n X_i$, the total number of successes in the n trials, follows a binomial distribution $\text{Bino}(n, p)$, is a minimal sufficient statistic for p and is simpler than the raw data $\underline{X} \overset{\text{def}}{=} (X_1, \ldots X_n)$. In fact, S_n contains all information on p since the conditional distribution of $\underline{X}$ for a given S_n is equal to

$$f(x_1, \ldots, x_n | S_n = i) = \frac{1}{\binom{n}{i}},$$

and is independent of p. Thus, an inference based on $\underline{X}$ can be replaced by one based on S_n following the sufficiency principle. Let $S^* = \{s_n : s_n = 0, \ldots, n\}$ be the reduced sample space for S_n with probability mass function

$$p_R(s_n; p) = \binom{n}{s_n} p^{s_n}(1-p)^{n-s_n}. \tag{2.3}$$

Each point $\{s_n : s_n = i\}$ in S^* is a set $\{\underline{x} : s_n(\underline{x}) = i\}$ in S. The information about p is contained in $\{s_n : s_n = i\}$ as a set of s_n, but the different points in $\{\underline{x} : s_n(\underline{x}) = i\}$ as a set of $\underline{x}$ provide the same information about p.

First consider a lower $1-\alpha$ one-sided confidence interval for p. Then, 1) the lower $1-\alpha$ one-sided Clopper-Pearson interval $[L_{\text{CP}}(S_n), 1]$ is the smallest $1-\alpha$ one-sided interval for p based on S_n (see, e.g., Wang 2006); 2) S_n contains the same amount of information about p as $\underline{X}$ does. Therefore, $[L_{\text{CP}}(S_n), 1]$ *would be* the best among all $1-\alpha$ one-sided intervals following (1.1) if that principle is still true. However, we next construct a $1-\alpha$ confidence interval, denoted by $[L_I(\underline{X}), 1]$, that depends on $\underline{X}$ but is a proper subset of $[L_{\text{CP}}(S_n), 1]$.

The sample space S contains 2^n points. An order on S is needed to derive the smallest one-sided interval and can be specified by a rank function R that assigns rank values to points with respect to confidence limits. Here, we say a point with a smaller rank is larger or has a larger confidence limit.

Intuitively, a sample point $\underline{x} = (x_1, \ldots, x_n)$ with a large value of $s_n = \sum_{i=1}^n x_i$ should be a large point. Let

$$A_i = \{\underline{x} \in S : s_n = i\} \tag{2.4}$$

for $i = 0, 1, \ldots, n$. All A_i's form a partition of S.

The set A_n contains a single point $(1, \ldots, 1)$, which should be the largest point, so define $R(1, \ldots, 1) = 1$. Also for $i > i'$ and any $\underline{x} \in A_i$ and $\underline{x}' \in A_{i'}$, we require

$$R(\underline{x}) < R(\underline{x}'), \text{ denoted by } R(A_i) < R(A_i').$$

In general,

$$R(A_n) < R(A_{n-1}) < \cdots < R(A_0). \tag{2.5}$$

Finally, each set A_i contains $\binom{n}{i}$ points. If we use S_n instead of $\underline{X}$ to derive an interval, then these points are tied and have the same confidence limit, and the best associated interval is equal to $[L_{\text{CP}}(S_n), 1]$. Here $L_{\text{CP}}(0) = 0$, and $L_{\text{CP}}(s_n)$ is the solution of

$$1 - \sum_{y=s_n}^{n} p_R(y; p) = 1 - \alpha$$

as a function of p for $s_n > 0$. This result indeed follows Lemma 2.1. However, now we differentiate the points in A_i as follows. For each point $\underline{x} = (x_1, \ldots, x_n)$, introduce $(s_1, \ldots, s_n)$, where each $s_j(\underline{x}) = \sum_{u=1}^j x_u$ is the number of observed successes up to the jth trial for $j = 1, \ldots, n$. Then S in (2.1) can be rewritten as

$$S = \{(s_1, \ldots, s_n) : 0 \le s_1 \le s_2 \le \cdots \le s_n \le n\} \tag{2.6}$$

since each $x_i = s_i - s_{i-1}$ for $i = 1, \ldots, n$ (let $s_0 = 0$). We introduce a rank function $R_I(\underline{X})$ on S as follows:

For any two sample points $\underline{x} = (x_1, \ldots, x_n)$ and $\underline{x}' = (x_1', \ldots, x_n')$ in S,
(1) if $s_n(\underline{x}) > s_n(\underline{x}')$, then define $R_I(\underline{x}) < R_I(\underline{x}')$;
(2) if, for some $i_0 < n$, $(s_{i_0+1}(\underline{x}), \ldots, s_n(\underline{x}))$
$= (s_{i_0+1}(\underline{x}'), \ldots, s_n(\underline{x}')$ but $s_{i_0}(\underline{x})$
$> s_{i_0}(\underline{x}')$, then define $R_I(\underline{x}) < R_I(\underline{x}')$. (2.7)

Therefore, $R_I(x_1, x_2, \ldots, x_n) = 1$ for $(x_1, \ldots, x_n) = (1, 1, \ldots, 1)$, and i_0 is the largest number of trials so that we observe more successes on point $\underline{x}$ than point $\underline{x}'$ up to the i_0th trial. The rank function $R_I(\underline{X})$ satisfies (2.5). More importantly, this function has no ties on S since $R_I(\underline{x}) = R_I(\underline{x}')$ if and only if $\underline{x} = \underline{x}'$. Therefore, the range of R_I is all integers from 1 to 2^n.

On the other hand, the one-sided Clopper-Pearson interval $[L_{\text{CP}}(S_n), 1]$ is obtained by a rank function $R_{\text{CP}}(\underline{X})$ that assumes

a constant value $n - i + 1$ on each set A_i and has a range of 1 through $n + 1$. This function also satisfies (2.5) as $R_I(\underline{X})$ does, but it generates many ties because we apply the sufficiency principle to combine points with the same values of s_n. So $R_I(\underline{X})$ is much finer (assumes more values) than $R_{\text{CP}}(\underline{X})$.

For any given rank function $R(\underline{X})$ defined on S, including $R_I(\underline{X})$ and $R_{\text{CP}}(\underline{X})$, we can derive the smallest confidence interval among those $1 - \alpha$ lower one-sided intervals $[L(\underline{X}), 1]$ for p that satisfy:

a) $L(\underline{x}) = L(\underline{x}')$ if $R(\underline{x}) = R(\underline{x}')$;
b) $L(\underline{x}) \le L(\underline{x}')$ if $R(\underline{x}') \le R(\underline{x})$,

following the lemma below.

Lemma 2.1. Assume $\alpha \in (0, 1)$. For a given rank function $R(\underline{X})$ on S and any $\underline{x} \in S$, consider

$$f_{\underline{x}}(p) = 1 - \sum_{\{\underline{x}' \in S: R(\underline{x}') \le R(\underline{x})\}} p_S(\underline{x}'; p). \tag{2.8}$$

Let

$$G_{\underline{x}} = \{p \in [0, 1] : f_{\underline{x}}(p') \ge 1 - \alpha, \ \forall \ p' < p\}. \tag{2.9}$$

Define

$$L_S(\underline{x}) = \begin{cases} \sup G_{\underline{x}}\ , & \text{if } G_{\underline{x}} \ne \emptyset; \\ 0, & \text{otherwise,} \end{cases} \tag{2.10}$$

then

(i) $[L_S(\underline{X}), 1]$ is a $1 - \alpha$ confidence interval for p and $L_S(\underline{X})$ satisfies a) and b);
(ii) for any $1 - \alpha$ confidence interval $[L(\underline{X}), 1]$ for p, where $L(\underline{X})$ satisfies a) and b), $L(\underline{X}) \le L_S(\underline{X})$.

Interval $[L_S(\underline{X}), 1]$ is the smallest interval under rank function $R(\underline{X})$ because, due to (ii), it is a subset of any intervals satisfying a) and b). Therefore, $[L_S(\underline{X}), 1]$ is the best interval. This lemma is a simple application of Theorem 4 in Wang (2010). The idea of deriving the best one-sided confidence interval under an order (or a rank function) was first proposed by Buehler (1957).

Example 1. Table 1 contains the sample space, the rank functions $R_{\text{CP}}(\underline{X})$ and $R_I(\underline{X})$ and their associated best intervals of level 95%, $[L_{\text{CP}}(\underline{X}), 1]$ and $[L_I(\underline{X}), 1]$, when $n = 4$. To illustrate the computation, pick, for example, $\underline{x} = (1, 1, 0, 1)$. When using $R_{\text{CP}}(\underline{X})$,

$$\{\underline{x}' \in S : R_{\text{CP}}(\underline{x}') \le R_{\text{CP}}(\underline{x})\} = \{(1, 1, 1, 1), (1, 1, 1, 0), (1, 1, 0, 1), (1, 0, 1, 1), (0, 1, 1, 1)\}.$$

Following (2.8),

$$f_{\underline{x}}(p) = 1 - p_R(4; p) - p_R(3; p) = 1 - p^4 - 4p^3(1 - p),$$

where the pmf p_R is given in (2.3). Then $L_{\text{CP}}(1, 1, 0, 1)$ is equal to 0.2486, the smallest solution of $f_{\underline{x}}(p) = 0.95$. On the other hand, when using $R_I(\underline{X})$,

$$\{\underline{x}' \in S : R_I(\underline{x}') \le R_I(\underline{x})\} = \{(1, 1, 1, 1), (1, 1, 1, 0), (1, 1, 0, 1)\}.$$

Following (2.8) again,

$$f_{\underline{x}}(p) = 1 - p_S(1, 1, 1, 1; p) - p_S(1, 1, 1, 0; p) - p_S(1, 1, 0, 1; p) = 1 - p^4 - 2p^3(1 - p),$$

where the pmf p_S is given in (2.2). Then $L_I(1, 1, 0, 1)$ is equal to 0.3092, the smallest solution of $f_{\underline{x}}(p) = 0.95$, and is larger than $L_{\text{CP}}(1, 1, 0, 1) = 0.2486$. To obtain intervals of level 0.95 the confidence limits are always computed by rounding down rather than to the nearest digit. Also we see $L_I(\underline{x}) \ge L_{\text{CP}}(\underline{x})$ for all $\underline{x}$'s in Table 1, and the strict inequality holds on many $\underline{x}$'s. Therefore, this is an example for a uniform improvement over $[L_{\text{CP}}(\underline{X}), 1]$, the best interval based on a sufficient statistic S_n. In fact, the example is a special case of the following result.

Theorem 2.1. Let $[L_{\text{CP}}(\underline{X}), 1]$ and $[L_I(\underline{X}), 1]$ be the intervals following Lemma 2.1 using the rank functions $R_{\text{CP}}(\underline{X})$ and $R_I(\underline{X})$, respectively. Then

$$[L_I(\underline{X}), 1] \subset [L_{\text{CP}}(\underline{X}), 1].$$

Remark 2.1. Theorem 2.1 indicates that to make the sufficiency principle valid we have to apply randomized confidence intervals and tests based on a sufficient statistic T, which, however, does not happen in practice. Therefore, this principle is not valid from the application's point of view. The interval $[L_I(\underline{X}), 1]$ based on the raw data $\underline{X}$ is much better than the interval $[L_{\text{CP}}(\underline{X}), 1] \overset{\text{def}}{=} [L_{\text{CP}}(S_n), 1]$ based on a minimal sufficient statistic S_n as the former is uniformly contained in the latter. It is also interesting to see that a sufficient statistic tries to combine sample points to generate a coarse partition (i.e., the rank function assumes fewer values) on the sample space but a shorter confidence interval is produced by a finer partition (i.e., the rank function assumes more values).

The order of $R_I(\underline{X})$ on sets A_i given by (2.5) is reasonable as it follows common sense, however, the order of $R_I(\underline{X})$ within each A_i is debatable. In Table 1, for example, why on set A_3 do we define $R_I(1, 1, 1, 0) < R_I(1, 1, 0, 1) < R_I(1, 0, 1, 1) < R_I(0, 1, 1, 1)$? This order says a point with earlier successes is larger even though the total number of successes on these four points are identical, that is, we weigh more on early successes. Table 1 gives another order (a rank function $R_A(\underline{X})$ and an interval $[L_A(\underline{X}), 1]$), where we assume, for example, $R_A(1, 1, 1, 0) > R_A(1, 1, 0, 1) > R_A(1, 0, 1, 1) > R_A(0, 1, 1, 1)$ on set A_3, that is, we weigh more on later successes. The new interval is different from but has the same total length over S as $[L_I(\underline{X}), 1]$, and it is also uniformly contained in $[L_{\text{CP}}(\underline{X}), 1]$. Next we provide a class of intervals which are as good as $[L_I(\underline{X}), 1]$.

Each set $A_i = \{s_n = i\}$ contains $k_i = \binom{n}{i}$ points. Let $\tau_i(A_i) = \{\underline{x}_j\}_{j=1}^{k_i}$ be any permutation of A_i that defines a rank function $R(\underline{X})$ on A_i by requiring that $R(\underline{x}_j)$ is strictly increasing for $j = 1, \ldots, k_i$. Also we require that $R(\underline{X})$ satisfies (2.5). Therefore, $R(\underline{X})$ assumes all integer values between 1 and 2^n, $R(1, \ldots, 1) = 1$ and $R(0, \ldots, 0) = 2^n$. See two such examples $R_I(\underline{X})$ and $R_A(\underline{X})$ in Table 1 when $n = 4$.

Theorem 2.2. Let $R(\underline{X})$ be a rank function that satisfies (2.5) and assumes all integer values between 1 and 2^n. For any $\underline{x}_j$ in A_i with $R(\underline{x}_j) < 2^n$, consider an equation

$$f_{\underline{x}_j}(p) \overset{\text{def}}{=} 1 - \sum_{u=i+1}^{n} p_R(u; p) - \frac{j}{k_i} p_R(i; p) = 1 - \alpha. \tag{2.11}$$

Then

Table 1. Three 95% confidence intervals, $[L_{CP}(\underline{X}), 1]$, $[L_I(\underline{X}), 1]$ and $[L_A(\underline{X}), 1]$, on the sample space S when $n = 4$.

$\underline{x} = (x_1, x_2, x_3, x_4)$	$\underline{s} = (s_1, s_2, s_3, s_4)$	R_{CP}	L_{CP}	R_I	L_I	R_A	L_A
(1,1,1,1), A_4	(1,2,3,4)	1	0.4728	1	0.4728	1	0.4728
(1,1,1,0), A_3	(1,2,3,3)	2	0.2486	2	0.3684	5	0.2486
(1,1,0,1), A_3	(1,2,2,3)	2	0.2486	3	0.3092	4	0.2731
(1,0,1,1), A_3	(1,1,2,3)	2	0.2486	4	0.2731	3	0.3092
(0,1,1,1), A_3	(0,1,2,3)	2	0.2486	5	0.2486	2	0.3684
(1,1,0,0), A_2	(1,2,2,2)	3	0.0976	6	0.1950	11	0.0976
(1,0,1,0), A_2	(1,1,2,2)	3	0.0976	7	0.1590	10	0.1068
(0,1,1,0), A_2	(0,1,2,2)	3	0.0976	8	0.1353	9	0.1188
(1,0,0,1), A_2	(1,1,1,2)	3	0.0976	9	0.1188	8	0.1353
(0,1,0,1), A_2	(0,1,1,2)	3	0.0976	10	0.1068	7	0.1590
(0,0,1,1), A_2	(0,0,1,2)	3	0.0976	11	0.0976	6	0.1950
(1,0,0,0), A_1	(1,1,1,1)	4	0.0127	12	0.0445	15	0.0127
(0,1,0,0), A_1	(0,1,1,1)	4	0.0127	13	0.0250	14	0.0169
(0,0,1,0), A_1	(0,0,1,1)	4	0.0127	14	0.0169	13	0.0250
(0,0,0,1), A_1	(0,0,0,1)	4	0.0127	15	0.0127	12	0.0445
(0,0,0,0), A_0	(0,0,0,0)	5	0	16	0	16	0

(i) $f_{\underline{x}_j}(p)$ is a strictly decreasing function and assumes values between 0 and 1.
Let $L_R(\underline{x}_j)$ be the unique solution of (2.11), and let $L_R(0, \ldots, 0) = 0$. Then
(ii) interval $[L_R(\underline{X}), 1]$ is the smallest $1 - \alpha$ one-sided interval for p under the rank function $R(\underline{X})$, and
(iii) $L_{CP}(\underline{X}) \le L_R(\underline{X})$.

Remark 2.2. Lemma 2.1 indeed claims that the smallest solution of $f_{\underline{x}}(p) = 1 - \alpha$ is the lower confidence limit; however, Theorem 2.2 goes one step further and says that the lower confidence limit is equal to the *only* solution of a strictly decreasing function $f_{\underline{x}_j}(p)$. This simplifies interval computation.

Remark 2.3. Example 1 is a special case of Theorem 2.2. One can show that there are in total $\Pi_{i=1}^{n} k_i!$ rank functions on S that satisfy (2.5) and have no ties. Each rank function yields a lower one-sided confidence interval following Theorem 2.2. Let $\mathcal{B}$ be the collection of all these intervals. Then $\mathcal{B}$ is a complete class of admissible $1 - \alpha$ lower one-sided intervals with a nondecreasing confidence limit $L(\underline{X})$ in the A_i's, that is,

$$L(\underline{x}_0) \le L(\underline{x}_1) \le \cdots \le L(\underline{x}_{n-1}) \le L(\underline{x}_n), \quad \forall\ \underline{x}_i \in A_i, i = 0, \ldots, n. \tag{2.12}$$

Here an interval is called admissible if any of its proper subintervals is of level strictly less than $1 - \alpha$; the interval class $\mathcal{B}$ is complete if for any $1 - \alpha$ one-sided interval $[L(\underline{X}), 1]$ for p there exists an interval $[L'(\underline{X}), 1]$ in $\mathcal{B}$ so that the latter is a subset of the former (i.e., $L'(\underline{X}) \ge L(\underline{X})$). All intervals in $\mathcal{B}$ can be easily obtained by permuting the interval $[L_I(\underline{X}), 1]$. One such example is $[L_A(\underline{X}), 1]$ in Table 1. Thus, any of these intervals can represent the entire class $\mathcal{B}$, and all intervals in $\mathcal{B}$ have the same total length and the same coverage probability function. For an interval $C(\underline{X}) = [L(\underline{X}), U(\underline{X})]$, its total length and its coverage probability are given by

$$\text{Tlength}(C) = \sum_{\underline{x} \in S} (U(\underline{x}) - L(\underline{x})) \text{ and } \text{Cover}_C(p) = P(p \in C(\underline{X})), \tag{2.13}$$

respectively. From the mathematical point of view, these intervals are equally good. However, they do make different inferences! Therefore, it is totally up to the practitioner to decide which rank function is used in applications.

For any given rank function $R(\underline{x})$ defined on S, we can also derive the best (smallest) confidence interval among those upper $1 - \alpha$ one-sided intervals $[0, U(\underline{X})]$ for p that, similar to conditions a) and b) before Lemma 2.1, satisfy:
(a′) $U(\underline{x}) = U(\underline{x}')$ if $R(\underline{x}) = R(\underline{x}')$;
(b′) $U(\underline{x}) \le U(\underline{x}')$ if $R(\underline{x}') \le R(\underline{x})$.
We state a theorem below but skip the proof as it is similar to Theorem 2.2.

Theorem 2.3. Let $R(\underline{X})$ be a rank function that satisfies (2.5) and assumes all integer values between 1 and 2^n. For any $\underline{x}_j$ in A_i with $R(\underline{x}_j) > 1$, consider an equation

$$f_{\underline{x}_j}(p) \stackrel{\text{def}}{=} 1 - \sum_{u=0}^{i-1} p_R(u; p) - \frac{k_i - j + 1}{k_i} p_R(i; p) = 1 - \alpha. \tag{2.14}$$

Then
(i) $f_{\underline{x}_j}(p)$ is a strictly increasing function and assumes values between 0 and 1.
Let $U_R(\underline{x}_j)$ be the unique solution of (2.14), and let $U_R(1, \ldots, 1) = 1$. Then
(ii) interval $[0, U_R(\underline{X})]$ is the smallest $1 - \alpha$ interval for p under the rank function R, that is, for any $1 - \alpha$ confidence interval $[0, U(\underline{X})]$ for p, where $U(\underline{X})$ satisfies conditions a′) and b′), $U_R(\underline{X}) \le U(\underline{X})$.
(iii) $U_R(\underline{X}) \le U_{CP}(\underline{X})$, where $[0, U_{CP}(\underline{X})]$ is the upper $1 - \alpha$ one-sided Clopper-Pearson interval for p. That is, $U_{CP}(\underline{x}) = U_{CP}(x_1, \ldots, x_n)$ is the solution of

$$1 - \sum_{y=0}^{s_n} p_R(y; p) = 1 - \alpha \text{ for } s_n = x_1 + \cdots + x_n$$

if $\underline{x} \ne (1, \ldots, 1)$; and is equal to 1 if $\underline{x} = (1, \ldots, 1)$.

Finally, we establish a one-to-one relationship between $L_R(\underline{X})$ and $U_R(\underline{X})$, which is similar to the following invariant property

$$L_{CP}(S_n) = 1 - U_{CP}(n - S_n).$$

Thus $L_R(\underline{X})$ can be determined by $U_R(\underline{X})$ and vice versa.

Theorem 2.4. Let $R(\underline{X})$ be a rank function that satisfies (2.5) and assumes all integer values between 1 and 2^n. Let $[L_R(\underline{X}), 1]$

Table 2. Two 95% confidence intervals $[0, U_I(\underline{X})]$ and $[0, U_A(\underline{X})]$ on the sample space S when $n = 4$.

$\underline{x} = (x_1, x_2, x_3, x_4)$	R_I	U_I	R_A	U_A
(1,1,1,1), A_4	1	1	1	1
(1,1,1,0), A_3	2	0.9873	5	0.9555
(1,1,0,1), A_3	3	0.9831	4	0.9750
(1,0,1,1), A_3	4	0.9750	3	0.9831
(0,1,1,1), A_3	5	0.9555	2	0.9873
(1,1,0,0), A_2	6	0.9024	11	0.8050
(1,0,1,0), A_2	7	0.8932	10	0.8410
(0,1,1,0), A_2	8	0.8812	9	0.8647
(1,0,0,1), A_2	9	0.8647	8	0.8812
(0,1,0,1), A_2	10	0.8410	7	0.8812
(0,0,1,1), A_2	11	0.8050	6	0.8932
(1,0,0,0), A_1	12	0.7514	15	0.6316
(0,1,0,0), A_1	13	0.7269	14	0.6908
(0,0,1,0), A_1	14	0.6908	13	0.7269
(0,0,0,1), A_1	15	0.6316	12	0.7514
(0,0,0,0), A_0	16	0.5272	16	0.5272

and $[0, U_R(\underline{X})]$ be the smallest lower and upper $1-\alpha$ one-sided confidence intervals under the rank function $R(\underline{X})$ following Theorems 2.2 and 2.3, respectively. Let $\underline{x}$ and $\underline{x}'$ satisfy $R(\underline{x}) + R(\underline{x}') = 1 + 2^n$. Then

$$L_R(\underline{x}) = 1 - U_R(\underline{x}'). \tag{2.15}$$

Example 1 (continued). We compute $[0, U_I(\underline{X})]$ under the rank function R_I in two ways: (i) solve (2.14), (ii) solve (2.11) and use (2.15), and obtain the same results, which are reported in Table 2. We also compute $[0, U_A(\underline{X})]$ under the rank function $R_A(\underline{X})$, which is a permutation of $[0, U_I(\underline{X})]$.

3. Discussion

The definition of a sufficient statistic implies that it contains all information about parameters. So, we believe the inference based on it "should be" good enough. This, however, is not true in practice because practitioners prefer a nonrandomized inference. As shown in Theorem 2, the best (smallest) one-sided nonrandomized confidence interval based on the binomial variable, a minimal sufficient statistic, is uniformly dominated by a nonrandomized interval based on the raw binary data. Similarly, any nonrandomized interval based on the binomial variable can be uniformly improved by a nonrandomized interval based on the raw data. Hence, (1.2) is established in the binomial case. Nonrandomized tests based on the raw data can be easily obtained by converting nonrandomized confidence intervals, and these tests are uniformly more powerful than nonrandomized tests based on the binomial variable.

This "paradox" makes us reconsider using a sufficient statistic for data reduction at least for those cases involving discrete distributions. Similar results can be obtained in simple random sampling (sampling without replacement), multinomial sampling, and product multinomial sampling, and for the Poisson distribution, the logistic regression model, and any discrete distribution model when a sufficient statistic $T(\underline{X})$ is simpler than the raw data $\underline{X}$.

Here is another illustration for the Poisson distribution. Suppose a sample $(Y_1, Y_2) = (1, 1)$ of size 2 is observed from a Poisson distribution Poi(μ), where μ is the unknown mean of Poi(μ) to be estimated. Then $Y = Y_1 + Y_2 (= 2) \sim$ Poi(2μ) is a minimal sufficient statistic for μ. Among all 95% confidence intervals $[L(Y), +\infty)$ for μ that depend on Y and have a nondecreasing lower confidence limit, the smallest interval $[L_{S,Y}(y), +\infty)$ by Lemma 2.1 satisfies:

$$0.95 = 1 - P_\mu(Y \geq y) = \sum_{z=0}^{y-1} e^{-2\mu} \frac{(2\mu)^z}{z!} \text{ for any } y > 0,$$
$$\text{and } L_{S,Y}(0) = 0.$$

So, $L_{S,Y}(2) = 0.1776$. If, however, we stay with the raw data (Y_1, Y_2) and define an order for the lower confidence limit $L(Y_1, Y_2)$ on set $\{y = 2\} = \{(y_1, y_2) : (2, 0), (1, 1), (0, 2)\}$ as $L(2, 0) \geq L(1, 1) \geq L(0, 2)$, then the lower limit $L_{S,Y_1,Y_2}(1, 1)$ of the smallest 95% interval based on (Y_1, Y_2) is equal to 0.2042, the smallest solution of

$$0.95 = 1 - P_\mu(Y > 2) - P_\mu((Y_1, Y_2) = \{(1, 1), (2, 0)\}).$$

Similarly, solve $L_{S,Y_1,Y_2}(0, 2)$ and $L_{S,Y_1,Y_2}(2, 0)$ and obtain

$$L_{S,Y_1,Y_2}(2, 0) = 0.3069 > L_{S,Y_1,Y_2}(1, 1) = 0.2042 > L_{S,Y_1,Y_2}(0, 2) = 0.1776 = L_{S,Y}(2). \tag{3.1}$$

Therefore, we obtain smaller 95% confidence intervals than $[L_{S,Y}(2), +\infty)$ by differentiating the three sample points in the set $\{y = 2\}$. This can be generalized to any set $\{y = y_0\}$ for $y_0 > 0$ and to the aforementioned discrete distributions. As noted by a referee, since the three probabilities at $(2, 0)$, $(1, 1)$ and $(0, 2)$ are not equal given that the total is two, which is different from the binomial case, the intervals generated following another order on these three points may have different values from (3.1). For example, if we choose an order of $L(1, 1) \geq L(2, 0) \geq L(0, 2)$, then

$$L(1, 1) = 0.2438 > L(2, 0) = 0.2042 > L(0, 2) = 0.1776 = L_{S,Y}(2).$$

Both $L_{S,Y_1,Y_2}(y_1, y_2)$ and $L(y_1, y_2)$ dominate $L_{S,Y}(y_1 + y_2)$ when $y_1 + y_2 = 2$, but $L(y_1, y_2)$'s cannot be obtained by permuting $L_{S,Y_1,Y_2}(y_1, y_2)$'s as in the binomial case.

When the observation is continuous, the sufficient statistic is also continuous. Then the technique developed in this article cannot be applied to obtain shorter intervals. For example, suppose X_1 through X_n are independently identically distributed $N(\mu, 1)$. Then $\bar{X}$ is a minimal sufficient statistic, and the interval $[\bar{X} - z_\alpha/\sqrt{n}, +\infty)$ is the smallest one-sided interval under the order specified by $\bar{X}$. A shorter lower one-sided interval based on X_1 through X_n cannot be derived as was done in Theorem 2.2 for the binary data simply because $P(\bar{X} = c)$ is 0 for any constant c.

As also noted by the same referee, when the observation is continuous, the rank-based nonparametric inferences like the rank sum test can be improved using the proposed technique. In this case, the continuous raw data $X_1, \ldots, X_n$ are converted to their ranks $R_1, \ldots, R_n$ (note that they typically are not sufficient statistics). Then these ranks have discrete distributions, and so also does the rank-based test statistic. Similar to the case where a sufficient statistic is able to combine sample points together, the rank-based test statistic also combines those ranks $(r_1, \ldots, r_n)$ on which the test statistic assumes the same value. Therefore, the proposed ordering method can differentiate these points and would lead to more powerful tests. This, however, deserves further study.

Finally, the inference about a proportion p is a basic and important statistical problem. We provide in Remark 3 a complete class $\mathcal{B}$ of admissible $1-\alpha$ one-sided confidence intervals with restriction (2.12), and all intervals in this class have the same total length and the same coverage probability. Which interval should be used in practice? The rank function $R_I(\underline{X})$ in (2.7) has a clear interpretation of awarding early successes, so, the smallest interval $[L_I(\underline{X}), 1]$ under $R_I(\underline{X})$ is a reasonable choice for use. Any other interval in the class may be applied instead if the practitioner has a preference on a certain rank function before he/she collects data. A $1-2\alpha$ two-sided confidence interval $[L_I(\underline{X}), U_I(\underline{X})]$ for p can be obtained easily by taking the intersection of two lower and upper $1-\alpha$ one-sided intervals.

Appendix

Proof of Theorem 2.1. Note that $L_{\mathrm{CP}}(\underline{X})$ is a constant on each A_i. $L_{\mathrm{CP}}(\underline{X})$ is increasing in A_i, and $R_I(\underline{X})$ satisfies (2.5). So $L_{\mathrm{CP}}(\underline{X})$ satisfies conditions a) and b) before Lemma 2.1 if choose $R(\underline{X}) = R_I(\underline{X})$. Then the theorem follows Lemma 2.1. □

Proof of Theorem 2.2. For (i), obviously $f_{\underline{x}_j}(0) = 1$ and $f_{\underline{x}_j}(1) = 0$. Let $B_{\underline{x}_j} = \cup_{u=i+1}^{n} A_u \cup \{\underline{x}_v\}_{v=1}^{j}$ be a subset of S. Then $f_{\underline{x}_j}(p) = 1 - P(B_{\underline{x}_j})$. For any $k \in [1, n]$ denote $\hat{\underline{x}}_k = (x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$. For any fixed $\hat{\underline{x}}_k$, let $G_k(\hat{\underline{x}}_k) = \{x_k : (x_1, \ldots, x_k, \ldots, x_n) \in B_{\underline{x}_j}\}$ and let $g_k(\hat{\underline{x}}_k) = \min G_k(\hat{\underline{x}}_k)$ if $G_k(\hat{\underline{x}}_k) \neq \emptyset$. Thus, in this case, $(x_1, \ldots x_{k-1}, g_k(\hat{\underline{x}}_k), x_{k+1}, \ldots, x_n) \in B_{\underline{x}_j}$, and then $(x_1, \ldots x_{k-1}, g_k(\hat{\underline{x}}_k) + 1, x_{k+1}, \ldots, x_n) \in B_{\underline{x}_j}$ if $g_k(\hat{\underline{x}}_k) = 0$. Therefore,

$$B_{\underline{x}_j} = \{(x_1, \ldots, x_n) : x_k \geq g_k(\hat{\underline{x}}_k)\}.$$

Consider independent binary variables $Y_i \sim \mathrm{Bino}(1, p_i)$'s for $i = 1, \ldots, n$. Let

$$h(p_1, \ldots, p_n) = 1 - P(Y_k \geq g_k(Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n)).$$

Thus $h(p_1, \ldots, p_n)$ is strictly decreasing in p_k when the other $p_{k'}$'s are fixed since the distribution of Y_k is monotone likelihood ratio in p_k. Also note $f_{\underline{x}_j}(p) = h(p, p, \ldots, p)$. Then, for $p < p'$,

$$\begin{aligned} f_{\underline{x}_j}(p) = h(p, p, \ldots, p) &> h(p', p, \ldots, p) > h(p', p', p, \ldots, p) \\ &> \cdots > h(p', \ldots, p') = f_{\underline{x}_j}(p'). \end{aligned}$$

For (ii), note that $f_{\underline{x}_j}(p)$ given in (2.11) is equal to $f_{\underline{x}}(p)$ given in (2.8), then the claim (ii) follows (i), (2.9) and (2.10).

The claim (iii) follows the proof of Theorem 2.1. □

Proof of Theorem 2.4. Let $\underline{x}$ and $\underline{x}'$ be two sample points with $R(\underline{x}) + R(\underline{x}') = 1 + 2^n$. If $\underline{x} \in A_0$, then $\underline{x} = (0, \ldots, 0)$, $\underline{x}' = (1, \ldots, 1)$, and $L_R(0, \ldots, 0) = 1 - U_R(1, \ldots, 1)$. Suppose $\underline{x} \in A_i = \{s_n = i\}$ for some $i > 0$. Then $\underline{x}' \in A_{n-i}$. Also if $\underline{x}$ is the jth (for some j) point in A_i in terms of the increasing order of $R(\underline{X})$, then $\underline{x}'$ is the $(k_{n-i} - j + 1)$th point in A_{n-i} in terms of the increasing order of $R(\underline{X})$. From (2.11),

$$1 - \sum_{u=i+1}^{n} p_R(u; L_R(\underline{x})) - \frac{j}{k_i} p_R(i; L_R(\underline{x})) = 1 - \alpha,$$

and $p_R(u; p) = p_R(n - u; 1 - p)$, we have

$$1 - \sum_{u=0}^{n-i-1} p_R(u; 1 - L_R(\underline{x})) - \frac{j}{k_i} p_R(n - i; 1 - L_R(\underline{x})) = 1 - \alpha.$$

Also note $k_i = k_{n-i}$ and $j = k_{n-i} - (k_{n-i} - j + 1) + 1$, we conclude (2.15). □

Acknowledgments

The author thanks two referees, the associate editor, the editor, and Professor Daniel T. Voss for their helpful comments and suggestions that substantially improved the article.

References

Buehler, R. J. (1957), "Confidence Intervals for the Product of Two Binomial Parameters." *Journal of the American Statistical Association*, 52, 482–493. [317]

Casella, G., and Berger, R. L. (2002), *Statistical Inference* (2nd ed.), Pacific Grove, CA: Duxbury Press. [315]

Geyer, C. J., and Meeden, G. D. (2005), "Fuzzy and Randomized Confidence Intervals and p-values," *Statistical Science*, 20, 358–366. [315]

Wang, W. (2006), "Smallest Confidence Intervals for One Binomial Proportion," *Journal of Statistical Planning and Inference*, 136, 4293–4306. [316]

——— (2010), "On Construction of the Smallest One-Sided Confidence Interval for the Difference of Two Proportions," *The Annals of Statistics*, 38, 1227–1243. [317]