A SEQUENTIAL CONFIDENCE INTERVAL
FOR THE ODDS RATIO

BY

D. SIEGMUND

TECHNICAL REPORT NO. 7
FEBRUARY 2, 1980

PREPARED UNDER CONTRACT
N00014-77-C-0306 (NR-042-373)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A SEQUENTIAL CONFIDENCE INTERVAL
FOR THE ODDS RATIO

by

D. Siegmund

Technical Report No. 7

February 2, 1980

Prepared under Contract

N00014-77-C-0306 (NR-042-373)

for the Office of Naval Research

D. Siegmund, Project Director

Reproduction in whole or in part is permitted for
any purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also issued as Technical Report No. 141 under National Science
Foundation Grant MCS77-16974 - Department of Statistics,
Stanford University.
ABSTRACT

A sequential fixed width confidence interval is proposed for the log odds ratio of a $2 \times 2$ table. It is shown that the proposed interval has asymptotically the correct coverage probability and is asymptotically efficient uniformly in the unknown parameters.

AMS 1970 Subject Classification. Primary 62L10

Key Words and Phrases. Fixed width confidence interval, log odds ratio, stopping rule, uniformly asymptotically normal.
A SEQUENTIAL CONFIDENCE INTERVAL FOR THE ODDS RATIO

1. Introduction

For $i = 1, 2$ let $s_{in}^1$ and $f_{in}^1 = n_i - s_{in}^1$ be respectively the numbers of successes and failures in $n_i$ independent Bernoulli trials with constant success probability $p_i$ on each trial. A simple large sample approximate confidence interval for the log odds ratio, $\log \left( \frac{p_1 q_2}{p_2 q_1} \right)$, is

$$\log \left( s_{2n_2}^1 f_{2n_2}^1 / s_{2n_2} f_{2n_2}^1 \right) \pm z_\alpha \left[ \frac{n_1}{s_{2n_1} f_{2n_1}} + \frac{n_2}{s_{2n_2} f_{2n_2}} \right]^{1/2},$$

where $\int_{-\infty}^{\infty} (2\pi)^{-1/2} \exp(-x^2/2)dx = \alpha/2$ (Cox, 1970, p. 35). The confidence coefficient, $1 - \alpha$, is asymptotically correct for fixed $p_1, p_2$ as $\min(n_1, n_2) \to \infty$.

These intervals have two defects when $p_1$ and $p_2$ may be near 0 or 1. On the one hand the rate of approach to normality can be very slow, so that use of asymptotic theory is questionable. More importantly, however, even with exact calculations, no fixed sample size design will permit one to estimate the log odds ratio by an interval of preassigned width in these boundary cases.

For one binomial population with success probability $p$, Robbins and Siegmund (1974) proposed a sequential scheme for obtaining approximately a confidence interval of preassigned width for $\log(p/q)$. However, they do not consider the question of the uniformity of their procedure for $p$ near 0 or 1, when a sequential procedure would presumably be of greatest value.
The purpose of this paper is to consider the two population analogue of the procedure of Robbins and Siegmund. The procedure will be seen to attain asymptotically the required coverage probability and to be asymptotically efficient uniformly in $0 < p_1, p_2 < 1$.

In Section 2 the one-population case is reviewed, and the results of Robbins and Siegmund are appropriately strengthened to provide the tools for the two-population problem. It is also shown that Robbins and Siegmund's uncritical acceptance of Haldane's (1955) modification of the empirical log odds ratio is inappropriate in the sequential case.

Section 3 is concerned with the case of two populations. Remarks about further extensions are collected in Section 4.
2. One Population

Let \( x_j \) be independent with \( P\{x_j = 1\} = p \), \( P\{x_j = 0\} = q = 1 - p \) \( (j = 1, 2, \ldots) \). Let \( s_n = x_1 + \ldots + x_n \) and \( f_n = n - s_n \). For large \( n \), \( \log(s_n/f_n) \) is approximately normally distributed with mean \( \log(p/q) \) and variance \( 1/(npq) \). Hence to find a confidence interval for \( \log(p/q) \) of preassigned width, or equivalently in large samples to estimate \( \log(p/q) \) by an estimator with preassigned variance \( 1/c \), Robbins and Siegmund (1974) define

\[
T = \inf\{ n : s_n/f_n > nc \}.
\]

They propose estimating \( \log(p/q) \) by

\[
\log\left[ \frac{(s_n + \frac{1}{2})/(f_n + \frac{1}{2})} \right],
\]

which they show is asymptotically normally distributed with mean \( \log(p/q) \) and variance \( 1/c \) as \( c \to \infty \). The modification of the empirical log odds by adding \( 1/2 \) to numerator and denominator was originally suggested by Haldane (1955) as a bias reducing device in the fixed sample case. Robbins and Siegmund also show that \( ET \sim c/(pq) \) as \( c \to \infty \). This may be interpreted as showing that their procedure is asymptotically efficient in the sense of requiring asymptotically about the same number of observations as a fixed sample procedure chosen to be appropriate for a value \( p_0 \) which happens to be the actual value of \( p \).

In this section it is shown that the asymptotic normality of (3) holds uniformly over \( 0 < p < 1 \). This is in marked contrast to the fixed sample case, as was noted in the Introduction. It will
also be shown that the analogue of Haldane's bias reducing device in this sequential context is to subtract \( \frac{1}{2} \) from numerator and denominator of the empirical odds ratio. However, for simplicity and because the appropriate modification for the two sample case is unknown, in most of what follows only the unmodified empirical odds ratio is considered.

The main result of this section is Theorem 1. Lemma 1, which was obtained by Robbins and Siegmund (1974), is of interest in its own right. It says that the asymptotic efficiency of (2) is uniform in \( 0 < p < 1 \). Repeated use will be made of the algebraic identity

\[
s_n f_n / n = (q - p)(s_n - np) + npq - (s_n - np)^2 / n .
\]

Theorem 1. For the stopping rule \( T \) defined in (2), uniformly in \( 0 < p < 1 \)

\[
\lim_{c \to \infty} P\{c^{\frac{1}{2}}[\log(s_T / f_T) - \log(p/q)] \leq x\} = \Phi(x) ,
\]

where

\[
\Phi(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} \exp(-u^2/2)du .
\]

The proof utilizes the following lemmas. For the simple proof of Lemma 1 based on (4), see Robbins and Siegmund (1974).

Lemma 1. \( c < pq ET < (c + 1)/[1 - (4c)^{-1}] \).
Lemma 2. There exists a $c_0$ such that for all $c \geq c_0$ and all $0 < p < 1$

$$(pq)^2 E(T- c/pq)^2 \leq 4c .$$

Lemma 3. For each $0 < \epsilon < 1$ and $c \geq c_0$, where $c_0$ is defined as in Lemma 2,

$$P\{|s_T- pT| \geq \epsilon pq T\} < \kappa/\epsilon^2 c ,$$

where $\kappa$ does not depend on $\epsilon$ or $c$.

Proof of Lemma 2. Squaring (4) gives

$$\left( s_{n-f-n} \right)^2 = (q- p)^2 (s-n)^2 + (pq)^2 (n-c/pq)^2 + (s_n-np)^4/n^2$$


By the Schwarz inequality and Wald's second moment identity

$$|E(s_T- pT)(T- c/pq)| \leq \{pq ET E(T- c/pq)^2\}^{1/2} .$$

Hence, since $(s_T^2 f_T/T- c)^2 < 1$, Wald's second moment identity yields

$$1 > (q- p)^2 pq E(T)+ (pq)^2 E(T- c/pq)^2 - 2pq|q- p|\{pq E(T)E(T- c/pq)^2\}^{1/2}$$

$$- 2|q- p|pq ET - 2(pq)^2 ET ,$$

or

$$(pq)^2 E(T- c/pq)^2 - 2pq|q- p|\{pq ET E(T- c/pq)^2\}^{1/2} + pq(q- p)^2 ET $$

$$< 1 + 2pq ET .$$
Taking square roots in this expression, then rearranging terms and squaring yields

\[(pq)^2 E(T - c/pq)^2 \leq \{(pq ET)^{1/2} + (1 + 2pq ET)^{1/2}\}^2\]

\[\leq 1 + 3pq ET \leq 1 + 3(c+1)/(1-1/4c),\]

where the last inequality follows from Lemma 1. This completes the proof.

**Proof of Lemma 3.** Let \(0 < \delta < 1\) and \(n_0 = c/pq\). By Lemma 2

\[P\{|T - n_0| > \delta c/pq\} \leq (\delta c)^{-2} \ (pq)^2 E(T - n_0)^2 \leq 4/\delta^2 c\]

Hence, by Wald’s lemma for the second moment and Lemma 1

\[P\{|s_T - pT| > \varepsilon pq T\} \leq 4/\delta^2 c + P\{|s_T - pT| > \varepsilon pq T, |T - n_0| < \delta c/pq\}\]

\[\leq 4/\delta^2 c + P\{|s_T - pT| > \varepsilon (1 - \delta) c\} \leq 4/\delta^2 c + E(s_T - pT)^2/\varepsilon^2 (1 - \delta)^2 c^2\]

\[\leq 4/\delta^2 c + 2/[\varepsilon^2 (1 - \delta)^2 c]\]

**Proof of Theorem 1.** From the mean value theorem follows

\[c^{1/2}[\log(s_T - n_T) - \log(p/q)] = c^{1/2}(s_T - pT)/(pq)\frac{1}{T} + c^{1/2}[((s_T - pT)/(pq))^T \frac{1}{n_T(1 - n_T)} - 1]\]

\[= n_0^{1/2}(s_T - pT)/(pq) + n_0^{1/2}[(s_T - pT)/(pq)]^T \frac{1}{n_T(1 - n_T)} - 1\]

where \(|n_T - p| \leq |T^{-1} s_T - p|\), and as before \(n_0 = c/pq\). Hence it suffices to show that uniformly in \(0 < p < 1\)
\[
\lim_{c \to \infty} \Pr\{n_0^{\frac{1}{2}}(s_T - pT)/(pq)^{\frac{1}{2}} T \leq x\} = \Phi(x)
\]

and

\[
pq/n_T(1-n_T) \xrightarrow{P} 1.
\]

The second statement follows easily from Lemma 3, and the first may be obtained by minor modifications in the standard proof of Anscombe's theorem (e.g., Rényi, 1966, p. 390).

An asymptotically more precise approximation to \(E(T)\) than that provided by Lemma 1, although one which is decidedly not uniform in \(p\), is

\[
(5) \quad pq \ E_T = c + \frac{1}{2}(p-q)^2 + \frac{1}{2} pq + o(1) \quad (c \to \infty),
\]

which is valid for all \(p\) for which \((p/q)^2\) is irrational. This result follows easily from (4) and Theorem 2 of Lai and Siegmund (1979).

As an estimator of \(\log(p/q)\), Haldane (1955) considered

\[
\log\{(s_n + a)/(\ell_n + a)\}
\]

and showed by a Taylor series expansion that the choice of a minimizing the asymptotic bias of this estimator is

\(a = \frac{1}{2}\). The following heuristic calculation shows that \(a = - \frac{1}{2}\) is appropriate in the present context. The machinery for justifying this calculation may be found in Pollak and Siegmund (1975). It should be noted that this result is appropriate for the stopping rule \(T\) defined by (2). It does not carry over to the two-population case discussed in Section 3.
A two term Taylor series expansion gives

$$\log\{\left(\frac{s_T + a}{f_T + a}\right)\} - \log\left(\frac{p}{q}\right) = \left(\frac{a}{s_T - pT + a}\right)/pT - \left(\frac{a}{f_T - qT + a}\right)/qT$$

$$-(s_T - pT)^2/2(pT)^2 + (f_T - qT)^2/2(qT)^2 + o_p(c^{-1})$$.

Since $T \overset{p}{\sim} c/pq$ and hence $E\{\left(\frac{s_T - pT}{T}\right)^2\}$

$$\sim (pqc^{-1})^2 E\{\left(\frac{s_T - pT}{pqT}\right)^2\} = (pqc^{-1})^2 pq ET \sim (pq)^2/c$$,

one obtains

$$E[\log\{\left(\frac{s_T + a}{f_T + a}\right)\}] - \log\left(\frac{p}{q}\right)$$

(6)

$$\sim E\{\left(\frac{s_T - pT}{pqT}\right)\} + c^{-1}(q - p)(a - \frac{1}{2})$$.

It is shown below that

$$E\{\frac{s_T - pT}{T}\} \sim c^{-1} pq(q - p) \quad (c \to \infty)$$,

which shows that the right hand side of (6) is $\sim c^{-1}(q - p)(a + \frac{1}{2})$,

leading to the optimal choice $a = -\frac{1}{2}$.

Let $\xi_T = s_T f_T/T - c$. By (4) and Taylor expansions, one obtains

$$(s_T - pT)/T = (q - p)^{-1}(c + \xi_T)(pqc^{-1} - (pqc^{-1})^2(T - c/pq))$$

$$+ (pqc^{-1})^3(T - c/pq)^2 + \ldots) - (q - p)^{-1} pq + (q - p)^{-1} (s_T - pT)^2/T^2$$.

It is easy to see from (4) that $c + E_{\xi_T} = pq ET - pq + o(1)$; and
Robbins and Siegmund (1974) have obtained $E(T - c/pq)^2 = (q - p)^2 c/(pq)^2 + O(1)$. Hence by the asymptotic independence of $\xi_T$ and $c^{-1/2}(T - c/pq)$ (Lai and Siegmund, 1977),

$$E\{(s_T - pT)/T\} = (q-p)^{-1}(c + E\xi_T)\{pqc^{-1} - (pqc^{-1})^2[E\xi_T/pq + 1]

+ (pqc^{-1})^3(q-p)^2 c/(pq)^2\} - (q-p)^{-1} pq + (pq)^2/(q-p)c + o(c^{-1})

\sim c^{-1} pq(q-p),$$

as claimed.
3. Two Populations

Consider again the two population case described in the introduction and suppose that observations are taken in pairs, one from each population, so \( n_1 = n_2 = n \), say. This restriction is stronger than necessary, but it simplifies the subsequent analysis. It is easy to modify the results to accommodate the case in which observations are taken from the two populations in an arbitrary fixed ratio. It seems possible to achieve a slight reduction in the total expected sample size by choosing the sampling rates adaptively, but the fairly small improvement seems not to be worth the considerable complication in analysis.

The obvious analogue of the stopping rule (2) is (cf. (1))

\[
T = \inf\{n : n\left(\frac{1}{s_1f_1} + \frac{1}{s_2f_2}\right) \leq \frac{1}{c}\}.
\]

The main results of this section are Theorems 2 and 3, which correspond respectively to Lemma 1 and Theorem 1 in the single population case. Theorem 2 shows that \( T \) defined by (7) is uniformly asymptotically efficient and Theorem 3 shows that it asymptotically provides the correct coverage probability uniformly in \( p_1, p_2 \).

**Theorem 2.** Uniformly in \( 0 < p_1, p_2 < 1 \),

\[
ET \sim c\left(\left(p_1q_1\right)^{-1} + \left(p_2q_2\right)^{-1}\right) \quad (c \to \infty).
\]

The inequality in one direction is a consequence of the following trivial lemma.
Lemma 4. For all $0 < p_1, p_2 < 1$ and all $c$

$$ET \geq c \{(p_1q_1)^{-1} + (p_2q_2)^{-1} \}.$$

**Proof.** From (4), Wald's identity and Jensen's inequality one obtains

$$c^{-1} \geq \frac{ET}{E[1/s_{1T}f_{1T} + 1/s_{2T}f_{2T}]} \geq \{E(s_{1T}f_{1T}/T)\}^{-1} + \{E(s_{2T}f_{2T}/T)\}^{-1} = \{p_1q_1 ET - E((s_{1T} - p_1 T)^2)/T\}^{-1} + \{p_2q_2 ET - E((s_{2T} - p_2 T)^2)/T\}^{-1} \geq (ET)^{-1} \{(p_1q_1)^{-1} + (p_2q_2)^{-1} \}.$$ 

To obtain asymptotic upper bounds on $E(T)$ it is useful to define (cf. (2))

$$T_1(c) = \inf\{n : n/s_{\text{in}}f_{\text{in}} \leq 1/c \}.$$ 

Since $s_{\text{in}}f_{\text{in}}/n$ increases with $n$, for all $\alpha > 1$ and $\beta > 1$ with $1/\alpha + 1/\beta = 1$,

$$T \leq \max(T_1(\alpha c), T_2(\beta c)).$$

(8)

In what follows $\alpha = (p_1q_1 + p_2q_2)/p_2q_2$ and $\beta = (p_1q_1 + p_2q_2)/p_1q_1$, so

(9) $\alpha/p_1q_1 = \beta/p_2q_2 = (p_1q_1 + p_2q_2)/(p_1q_1p_2q_2) = (p_1q_1)^{-1} + (p_2q_2)^{-1}$.

With these fixed values of $\alpha$ and $\beta$ there is no ambiguity in writing $T_1$ for $T_1(\alpha c)$ and $T_2$ for $T_2(\beta c)$.

It is now possible to complete the proof of Theorem 2. Obviously from (8)
\[(10) \quad E(T) \leq E(\max(T_1, T_2)) = \int_{\{T_1 < T_2\}} T_2 dP + \int_{\{T_2 < T_1\}} T_1 dP.\]

Let $\varepsilon > 0$ be arbitrary. Then

\[\int_{\{T_1 < T_2\}} T_2 dP \leq \int_{\{T_1 < T_2, T_2 < \beta c(1+\varepsilon)/p_2 q_2\}} T_2 dP + \int_{\{T_2 > \beta c(1+\varepsilon)/p_2 q_2\}} T_2 dP\]

\[\leq (p_2 q_2)^{-1} \beta c(1+\varepsilon) P\{T_1 < T_2\} + (p_2 q_2)^{-1} \beta c \ P\{T_2 > (p_2 q_2)^{-1} \beta c(1+\varepsilon)\}\]

\[+ \int_{\{T_2 > (p_2 q_2)^{-1} \beta c(1+\varepsilon)\}} |T_2 - (p_2 q_2)^{-1} \beta c| dP.\]

By Lemma 2

\[(p_2 q_2)^{-1} \beta c \ P\{T_2 > (p_2 q_2)^{-1} \beta c(1+\varepsilon)\} \leq 4(p_2 q_2)^{-1} \varepsilon^{-2};\]

and by the Schwarz inequality and Lemma 2 again

\[\int_{\{T_2 > (p_2 q_2)^{-1} \beta c(1+\varepsilon)\}} |T_2 - (p_2 q_2)^{-1} \beta c| dP \leq \left[(p_2 q_2)^{-2} E[p_2 q_2 T_2 - \beta c]^2 P\{T_2 > (p_2 q_2)^{-1} \beta c(1+\varepsilon)\}\right]^{1/2}

\[\leq 4(p_2 q_2)^{-1} \varepsilon^{-1}.\]

Putting these inequalities together with (9), (10), and (11) yields

\[ET \leq c((p_1 q_1)^{-1} + (p_2 q_2)^{-1})(1 + \varepsilon + 8/\varepsilon^2 c),\]

which completes the proof, as $\varepsilon$ is arbitrarily small.

**Theorem 3.** For $T$ defined by (7), uniformly in $0 < p_1, p_2 < 1$

\[\lim_{c \to \infty} P\{\sqrt{c} [\log(s_{1T} f_{2T}/s_{2T} f_{1T}) - \log(p_1 q_2/p_2 q_1)] < x}\} = \Phi(x) .\]
With the help of Lemma 5 below, the proof of Theorem 3 may be carried out along the same lines as the proof of Theorem 1.

**Lemma 5.** Let \( \mu = \{(p_1 q_1)^{-1} + (p_2 q_2)^{-1}\}^{-1} \). For all \( \varepsilon > 0 \) and all large \( c \) (not depending on \( \varepsilon \))

\[
P\{|\mu T - c| > c \varepsilon\} \leq \frac{8}{c \varepsilon^2}.
\]

**Proof.** The proof of Theorem 2 shows that

\[
P\{T > c(1+\varepsilon)\mu^{-1}\} = P\{T_1 < T_2, \ T_2 > (p_2 q_2)^{-1} \beta c (1+\varepsilon)\}
\]

\[+ P\{T_2 < T_1, \ T_1 > (p_1 q_1)^{-1} ac(1+\varepsilon)\} \leq \frac{4}{c \varepsilon^2}.
\]

The same upper bound for \( P\{T < c(1-\varepsilon)\mu^{-1}\} \) follows by a similar calculation and the observation that \( T \geq \min(T_1(ac), T_2(\beta c)) \).
4. Remarks

(a) Unpublished numerical computations of H. Levene in the one-sample case show that the asymptotic theory of Section 2 provides good approximations for \( c \geq 10 \) and reasonable ones for \( c \) as small as 3. It seems likely that similar results hold for two populations.

(b) The heuristic principle which suggests the stopping rules (2) and (7) is quite common in the literature of fixed precision estimation (e.g., Anscombe, (1953)), and it leads to reasonable stopping rules for more complicated log linear models. However, the uniform asymptotic theory developed here seems to require new ideas for very simple extensions.

One important generalization is a set of \( 2 \times 2 \) tables with equal odds ratios. Appropriate asymptotic theory might involve a large number of observations from each of a small number of tables or a large number of tables.

Another interesting variation is log linear regression. In this case one might also wish to consider sequential design in selecting values of the independent variable.
References


A sequential confidence interval for the odds ratio

Also issued as Technical Report No. 141 under National Science Foundation Grant MCS77-16974 - Department of Statistics, Stanford University.

A sequential fixed width confidence interval is proposed for the log odds ratio of a 2 x 2 table. It is shown that the proposed interval has asymptotically the correct coverage probability and is asymptotically efficient uniformly in the unknown parameters.