Sequential Fixed Width Confidence Intervals

Thesis by

Tihomir Zlatev Asparouhov

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2000

(Submitted March 13, 2000)

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Gary Lorden, whose support and supervision made this research possible. It has been a great pleasure to work under his guidance.

The Mathematics Department of the California Institute of Technology was a great place for conducting the research, and for that I would like to thank all the members and staff of the department.

I would like to thank my wife Elena and my son Delian for their constant support and encouragement.

Abstract

We consider the problem of constructing confidence intervals of fixed width d and confidence level γ for the success probability p in Bernoulli trials. Algorithms are given for calculating numerical lower bounds on the average expected sample size required and an asymptotic lower bound is obtained as $d \to 0$. Sequential and two-stage procedures are proposed that attain the asymptotic lower bounds and nearly attain the numerical lower bounds. Asymptotically optimal sequential and two-stage confidence intervals of fixed width and confidence level are proposed for the mean in a general (non-Bernoulli) context.

Contents

Chapter 1 Introduction

Let $X_1, X_2, ...$ be independent Bernoulli random variables with unknown parameter $0 \le p \le 1$, i.e., for each $i = 1, 2...$, $P(X_i = 1) = p = 1 - P(X_i = 0)$. Let $[L, R]$ be a confidence interval for the parameter p, where L and R are functions of the observed X_i's. The confidence interval $[L, R]$ has confidence level γ if its coverage probability is at least γ, i.e., $\inf_p P_p(L \le p \le R) \ge \gamma$. It has width d if $P_p(R - L \le d) = 1$ for all p. We are interested in finding level γ width d confidence intervals that require a minimal amount of sampling. There are two ways to approach this problem: by a fixed sample size method, or by a sequential method. A fixed sample size method consists of (n, L, R), where n is the number of observations needed, and $[L, R]$ is the confidence interval based on the observations. In contrast, a sequential method consists of (N, L, R) where N is a stopping time (also called a stopping rule) denoting the number of observations taken, which is a random variable based on the observations. Our results concern sequential methods, which require less sampling than fixed sample size methods. Throughout this work γ and d are given numbers, $0 < \gamma < 1$, $d > 0$ and we denote by $h = d/2$ the length of the half interval.

There are two classical fixed sample size methods to obtain confidence intervals for p.

SAM (Standard Approximate Method) By the Central Limit Theorem and Slutsky's ([10]) theorem the confidence interval

$$[L,R] = \left[\bar{X}_n - c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

has coverage probability approximately equal to γ, where c is the $(1+\gamma)/2$ quantile of the standard normal distribution, i.e., $c = \Phi^{-1}((1+\gamma)/2) > 0$ where Φ is the standard normal distribution function. However, for p close to 0 and 1 this approximation breaks down and the coverage probability approaches 0.

SAM (Standard Exact Method) This method is based on constructing the two uniformly most accurate $(1+\gamma)/2$ one-sided confidence intervals, described in Chapter 3.5 of [11]. Given that $S_n = \sum_{i=1}^{n} X_i = k$ the confidence interval $[L,R]$ is obtained by solving these two equations for L and R :

$$P_{p=L}\Big(S_n \geq k\Big) = \frac{1-\gamma}{2}$$

$$P_{p=R}\Big(S_n \leq k\Big) = \frac{1-\gamma}{2}$$

Neither of these methods yields fixed width confidence intervals. There are two improvements of the classical methods up to now, both designed to give confidence intervals of fixed width.

Classical Sequential Method (Chow-Robbins [4]) The idea of this method is to use the confidence interval based on the Central Limit Theorem approximation

$$[L,R] = \left[\bar{X}_n - c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

but to sample until the width of this interval becomes less than d. Hence the method uses the following stopping rule: N is the smallest n for which $2c\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} < d$. This method also has only approximate coverage probability γ and the approximation breaks down when p is close to 0 and 1.

Pushed Confidence Intervals (Lorden [12]) Lorden has obtained the best fixed sample size method for giving confidence intervals of exact coverage probability γ and fixed width d. It is the best in the sense that it uses the smallest possible sample size n. Chapter 4 includes a discussion of this method and our application of it in the sequential case.

Let $\lambda(p)$ be a probability density function defined on $[0,1]$. Define

$$B(\gamma, d, \lambda) = \inf \int_0^1 E_p N \, \lambda(p) \, dp,$$

where the infimum is taken over all confidence intervals (N, L, R) of level γ and width d. The quantity $B(\gamma, d, \lambda)$ represents the minimal amount of sampling needed to obtain a level γ width d confidence interval, in the sense of minimizing the average over p (using λ as a weight function) of the expectation, when p is true, of the sample size.

With few exceptions, problems in sequential analysis about best possible performance, such as the one defined by $B(\gamma, d, \lambda)$, lead to theorems proved only in asymptotic form, as the sample size becomes large. In the present problem, asymptotic theory is most naturally developed by letting $d \to 0$ with γ held fixed. In Chapter 2 we use a Bayes technique to obtain an asymptotic lower bound on $B(\gamma, d, \lambda)$ as $d \to 0$. In Chapter 3 we modify an idea of Chow and Robbins [4] to construct level γ width d confidence intervals that achieve the asymptotic lower bound. Combining these two results yields the following theorem.

Theorem 1. *For any positive continuous function λ defined on $[0, 1]$*

$$\lim_{d \to 0} d^2 B(\gamma, d, \lambda) = 4c^2 \int_0^1 p(1-p)\lambda(p)\, dp$$

The method proposed in Chapter 3 is developed as a solution to the problem of constructing a confidence interval for the unknown mean of a distribution belonging to a general class of distributions. Under some mild moment conditions, we construct fixed width d level γ confidence intervals, that are asymptotically efficient in a certain sense. We also consider an appealing special class of sequential methods called two-stage procedures. These rely on a preliminary (first-stage) sample of a fixed size, m, followed if necessary by a second stage of variable size. We construct two-stage confidence interval procedures that are asymptotically efficient and in the Bernoulli case attain $B(\gamma, d, \lambda)$ in the limit as $d \to 0$.

Another goal of this investigation is to determine algorithms for constructing "nearly optimal" sequential confidence intervals, along with methods for calculating how close they come to attaining $B(\gamma, d, \lambda)$. Such constructions have two parts: the stopping rule N and the so called *terminal decision rule* that calculates the interval $[L, R]$ based on N and $\bar{X}_N$. It turns out that rather than center the interval at $\bar{X}_N$ and use probability estimates to guarantee coverage probability γ, which suffices for the asymptotic theory of Chapter 3, it is much more effective for small sample sizes to apply Lorden's "push" algorithm [12] to optimize the construction of $[L, R]$ for a given stopping rule N. These considerations reduce the problem to one of finding the best (or nearly best) stopping rules N. In Chapter 4 we give a method for computing explicit lower bounds on $B(\gamma, d, \lambda)$ and describe two new methods for constructing nearly optimal stopping rules. The first method is based on the auxiliary Bayes problem of Chapter 2 and uses a Backward Induction scheme to compute the stopping rule. The second method is based on selecting a member of the asymptotically optimal family of solutions proposed in Chapter 3. For both methods we use as a terminal decision the confidence intervals obtained by Lorden's push algorithm. We also give a method for obtaining nearly optimal two-stage confidence interval procedures. As the numerical results in Table 1 of Chapter 4 illustrate, the algorithms for both the sequential and two-stage rules come very close to solving the problem of attaining $B(\gamma, d, \lambda)$ in practice.

Chapter 2 Lower Bound on Average Expected Sample Size

2.1 Bayes Auxiliary Problem

Given a probability density function $\pi(p)$ defined on the interval $[0,1]$ and $c > 0$, we consider the problem of minimizing

$$\int_0^1 \left\{ \frac{h^2 \phi(c)}{c} E_p N - p(1-p) P_p(L \leq p \leq R) \right\} \pi(p)\, dp \tag{2.1}$$

over all width $d = 2h$ confidence intervals (N, L, R) for p, assuming without loss of generality that $R = L + d$. There is no restriction on the confidence level of the intervals. Our goal is to find a lower bound on the quantity (2.1), useful for constructing lower bounds on $B(\gamma, d, \lambda)$. A standard way to solve this problem is to use the following Bayes technique. Let p be a random variable, $p \in [0,1]$ with density function π, also called the *prior* density of p. Define the loss function for (N, L, R) as a function of p by

$$\mathcal{L}(N, L, R) = \frac{h^2 \phi(c)}{c} N - p(1-p)\, 1_{\{L \leq p \leq R\}}$$

Then (2.1) can be interpreted in the form

$$E\left(\mathcal{L}\left(N,L,R\right)\right)=\int_{0}^{1}\left\{\frac{h^{2}\phi\left(c\right)}{c}E_{p}N-p\left(1-p\right)P_{p}\left(L\leq p\leq R\right)\right\}\pi\left(p\right)\,dp,$$

where the expectation on the left-hand side is taken with respect to the sequence of Bernoulli random variables $\{X_i\}$ with probability of success p, where p is a random variable having probability density function π. In this form the problem is known as a Bayes problem and $E\left(\mathcal{L}\left(N,L,R\right)\right)$ is called the *integrated risk* of the procedure (N,L,R). We are interested in a lower bound on the integrated risk in terms of h,γ and π. A standard way to study the integrated risk is to express it in terms of a conditional expectation, i.e.,

$$E\left(\mathcal{L}\left(N,L,R\right)\right)=E\left\{E\left(\mathcal{L}\left(N,L,R\right)|N,S_{N}\right)\right\}$$

and to obtain bounds on $E\left(\mathcal{L}\left(N,L,R\right)|N,S_{N}\right)$, which is called the *posterior expected loss.* Define $s=S_N$ as the number of "successes" at termination, i.e., the number of observed 1's, and $f=N-s$ as the number of "failures" at termination, i.e., the number of observed 0's. We use the following notation for the conditional expectation given N and $S_N: E\left(\cdot|s,f\right)=E\left(\cdot|N=s+f,S_N=s\right)$. Then the posterior expected loss is

$$E\left(\mathcal{L}|s,f\right)=\frac{h^{2}\phi\left(c\right)}{c}\left(s+f\right)-\int_{L}^{L+d}p\left(1-p\right)\pi\left(p|s,f\right)\,dp \tag{2.2}$$

where $\pi(p|s,f)$ is the conditional density of p given $N = s+f$ and $S_N = s$, which is also known as the *posterior* density of p

$$\pi(p|s,f) = \frac{p^s(1-p)^f \pi(p)}{\int_0^1 p^s(1-p)^f \pi(p)}.$$

The posterior expected loss will be minimal when L is chosen to maximize the integral on the right-hand side of equation (2.2), and since the integrand depends only on s and f, the optimal value for L will depend only on s and f. Such an optimal value is called *a Bayes terminal decision*, which we denote by $L(s,f)$. Without loss of generality we assume that $L = L(s,f)$, which reduces the problem to choosing N optimally.

We depict the random walk generated by $X_1, X_2, \ldots$ in the following way: it starts at the origin of the two-dimensional coordinate system, where the s axis is horizontal and the f axis is vertical. For each success it moves to the right, i.e., it increases the s coordinate by 1 and for each failure it increases the f coordinate by 1. Thus (s,f) is the position of the random walk after observing $s+f$ data points. A stopping rule N determines a partition of the integer lattice points – each (s,f) is either a stopping point or a continuation point.

First we will find a lower bound on the integrated risk in the case when the prior distribution of p is Beta(a,b), $a,b > 0$, i.e., the prior density function of p is the

Beta(a,b) density function

$$f_{a,b}(p) = \frac{p^{a-1}(1-p)^{b-1}}{\mathcal{B}(a,b)},$$

where $\mathcal{B}(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}\,dp$, is the Beta function.

2.2 Lower Bound on the Posterior Expected Loss with Uniform Prior

In this section we assume that p has a uniform distribution on $[0,1]$, i.e., a Beta$(1,1)$ distribution.

Lemma 1. *There exists a constant C such that*

$$E(\mathcal{L}|s,f) \geq (c\phi(c) - \gamma)\,E(p(1-p)\,|s,f) - Ch. \tag{2.3}$$

Proof. If p has uniform prior density, then the posterior density of p at (s,f) is the Beta$(s+1,f+1)$ density (see page 193 of [16]). If the confidence interval associated with the stopping point (s,f) is $[z, z+2h]$, then the posterior expected loss is

$$E(\mathcal{L}|s,f) = \frac{h^2\phi(c)}{c}(s+f) - \frac{\int_z^{z+2h} p(1-p)\,p^s(1-p)^f\,dp}{\mathcal{B}(s+1,f+1)}$$

$$= \frac{h^2\phi(c)}{c}(s+f) - \frac{\mathcal{B}(s+2,f+2)}{\mathcal{B}(s+1,f+1)}\,\frac{\int_z^{z+2h} p^{s+1}(1-p)^{f+1}\,dp}{\mathcal{B}(s+2,f+2)}$$

$$= \frac{h^2\phi(c)}{c}(s+f) - \frac{(s+1)(f+1)}{(s+f+2)(s+f+3)} \frac{\int_z^{z+2h} p^{s+1}(1-p)^{f+1}\,dp}{\mathcal{B}(s+2,f+2)}, \tag{2.4}$$

where we used Euler's formula for the Beta function, $\mathcal{B}(s,f) = \frac{\Gamma(s)\Gamma(f)}{\Gamma(s+f)}$, to evaluate $\mathcal{B}(s+2,f+2)/\mathcal{B}(s+1,f+1)$. Also, since the posterior distribution of p is Beta$(s+1,f+1)$, the posterior expectation of $p(1-p)$ is

$$E(p(1-p)|s,f) = \frac{\mathcal{B}(s+2,f+2)}{\mathcal{B}(s+1,f+1)} = \frac{(s+1)(f+1)}{(s+f+2)(s+f+3)}. \tag{2.5}$$

From now on, we will assume that $s \leq f$. By symmetry all the claims need only be proved in the case $s \geq f$.

Case 1. Suppose that $s \neq f$ and

$$\frac{(s+1)(f+1)}{2(f-s)(s+f+2)} \leq h$$

Then (2.3) is satisfied.

Proof of Case 1. We have

$$E(\mathcal{L}|s,f) = \frac{h^2\phi(c)}{c}(s+f) - \frac{(s+1)(f+1)}{(s+f+2)(s+f+3)} \frac{\int_z^{z+2h} p^{s+1}(1-p)^{f+1}\,dp}{\int_0^1 p^{s+1}(1-p)^{f+1}\,dp}$$

$$\geq \frac{h^2\phi(c)}{c}(s+f) - \frac{(s+1)(f+1)}{(s+f+2)(s+f+3)} \geq -E(p(1-p)|s,f)$$

Therefore

$$E(\mathcal{L}|s,f) - (c\phi(c) - \gamma) E(p(1-p)|s,f) \geq -(1-\gamma+c\phi(c)) \frac{(s+1)(f+1)}{(s+f+2)(s+f+3)}$$

$$\geq -(1-\gamma+c\phi(c)) \frac{2h(f-s)}{s+f+3} \geq -(1-\gamma+c\phi(c))\, 2h,$$

which proves (2.3)

Case 2. Suppose that

$$\frac{(s+1)(f+1)}{2(s+f+2)} \geq h(f-s) \tag{2.6}$$

Then (2.3) is satisfied.

Proof of Case 2. Let $a = s+1$ and $b = f+1$. Our goal is to find an upper bound on

$$g(z) = \int_{z}^{z+2h} p^a (1-p)^b \, dp.$$

Define $g_1(p) = p^a (1-p)^b$. Suppose g achieves its maximum at z_0, then

$$0 = g'(z_0) = g_1(z_0 + 2h) - g_1(z_0).$$

However, $g_1(p)$ is an increasing function in the interval $[0, a/(a+b)]$ and decreasing on $[a/(a+b), 1]$, which implies $z_0 < a/(a+b) < z_0 + 2h$. For a change of variable in $g(z)$ set

$$p = \frac{a}{a+b} + x.$$

The above inequalities imply that for all $p \in [z_0, z_0 + 2h]$,

$$|x| < 2h.$$

The inequality

$$log\,(1+x) \leq x - \frac{x^2}{2} + \frac{x^3}{3} \tag{2.7}$$

is valid for all $x > -1$. Now

$$\log\left(p^a (1-p)^b\right) = a\log\left(\frac{a}{a+b} + x\right) + b\log\left(\frac{b}{a+b} - x\right)$$

$$= a\log\left(\frac{a}{a+b}\right) + b\log\left(\frac{b}{a+b}\right) + a\log\left(1 + \frac{x\,(a+b)}{a}\right) + b\log\left(1 - \frac{x\,(a+b)}{b}\right)$$

$$= \log\left(\frac{a^a b^b}{(a+b)^{a+b}}\right) + \left(a\log\left(1 + \frac{x\,(a+b)}{a}\right) + b\log\left(1 - \frac{x\,(a+b)}{b}\right)\right)$$

By assumption $b \geq a$. We use $|x| < 2h$ and (2.7) to get

$$a\log\left(1 + \frac{x\,(a+b)}{a}\right) + b\log\left(1 - \frac{x\,(a+b)}{b}\right) \leq -\frac{x^2\,(a+b)^3}{2ab} + \frac{x^3\,(a+b)^4\,(b-a)}{3a^2b^2}$$

$$\leq -\frac{x^2\,(a+b)^3}{2ab} + \frac{x^2 2h\,(a+b)^4\,(b-a)}{3a^2b^2} = -\frac{x^2\,(a+b)^3}{2ab}\left(1 - \frac{4h\,(b^2-a^2)}{3ab}\right) = -\frac{x^2}{2\sigma^2},$$

where σ is defined by the last equation, i.e.,

$$\sigma = \sqrt{\frac{ab}{(a+b)^3\left(1 - \frac{4h(b^2-a^2)}{3ab}\right)}}.$$

Notice now that (2.6) can be restated as

$$\frac{h\left(b^2-a^2\right)}{ab}\leq\frac{1}{2}.$$

Now we combine the above inequalities

$$\frac{\int\limits_{z}^{z+2h} p^a\left(1-p\right)^b\,dp}{\int\limits_{0}^{1} p^a\left(1-p\right)^b\,dp}\leq\frac{(a+b+1)!}{a!b!}\int\limits_{z_0}^{z_0+2h} p^a\left(1-p\right)^b\,dp$$

$$=\frac{(a+b+1)!}{a!b!}\frac{a^ab^b}{(a+b)^{a+b}}\int\limits_{z_0-a/(a+b)}^{z_0+2h-a/(a+b)} e^{a\log\left(1+\frac{x(a+b)}{a}\right)+b\log\left(1-\frac{x(a+b)}{b}\right)}\,dx$$

$$\leq\frac{(a+b+1)!}{a!b!}\frac{a^ab^b}{(a+b)^{a+b}}\int\limits_{z_0-a/(a+b)}^{z_0+2h-a/(a+b)} e^{-\frac{x^2}{2\sigma^2}}\,dx$$

$$\leq\frac{(a+b+1)!}{a!b!}\frac{a^ab^b}{(a+b)^{a+b}}\sqrt{2\pi}\sigma\left(2\Phi\left(\frac{h}{\sigma}\right)-1\right),$$

since the highest probability density interval for the normal distribution is the symmetric interval around zero. Using Stirling's formula in the form found in [6] page 52,

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n+1}}<n!<\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n+\frac{1}{12n}},$$

and we obtain

$$\frac{a^ab^b}{(a+b)^{a+b}}\frac{(a+b+1)!}{a!b!}\leq\frac{a^ab^b}{(a+b)^{a+b}}(a+b+1)\frac{\sqrt{2\pi}\,(a+b)^{a+b+\frac{1}{2}}\,e^{-a-b+\frac{1}{12(a+b)}}}{\sqrt{2\pi}a^{a+\frac{1}{2}}e^{-a+\frac{1}{12a+1}}\sqrt{2\pi}b^{b+\frac{1}{2}}e^{-b+\frac{1}{12b+1}}}$$

$$= \sqrt{\frac{(a+b+1)^2(a+b)}{2\pi ab}} e^{\frac{1}{12(a+b)} - \frac{1}{12a+1} - \frac{1}{12b+1}} < \sqrt{\frac{(a+b+1)^2(a+b)}{2\pi ab}}.$$

Finally we obtain

$$\frac{\int_z^{z+2h} p^a (1-p)^b \, dp}{\int_0^1 p^a (1-p)^b \, dp} \leq \frac{a+b+1}{a+b} \frac{1}{\sqrt{1 - \frac{4h(b^2-a^2)}{3ab}}} \left(2\Phi\left(\frac{h}{\sigma}\right) - 1\right)$$

$$= \frac{1}{\sqrt{1 - \frac{4h(b^2-a^2)}{3ab}}} \left(2\Phi\left(\frac{h}{\sigma}\right) - 1\right) + \frac{1}{a+b} \frac{1}{\sqrt{1 - \frac{4h(b^2-a^2)}{3ab}}} \left(2\Phi\left(\frac{h}{\sigma}\right) - 1\right). \quad (2.8)$$

By the mean value theorem

$$\frac{\Phi(x) - 1/2}{x} \leq \max \phi(x) = \frac{1}{\sqrt{2\pi}},$$

so for all positive x,

$$2\Phi(x) - 1 \leq \frac{2x}{\sqrt{2\pi}}.$$

An upper bound for the last term in (2.8) can be obtained from the inequality

$$\frac{1}{a+b} \frac{1}{\sqrt{1 - \frac{4h(b^2-a^2)}{3ab}}} \left(2\Phi\left(\frac{h}{\sigma}\right) - 1\right) \leq \sqrt{\frac{2}{\pi}} \sqrt{\frac{a+b}{ab}} h \leq \frac{2h}{\sqrt{\pi}}.$$

By (2.6)

$$\frac{4h(b^2-a^2)}{3ab} \leq \frac{2}{3}.$$

For any $0 < x \le 2/3$

$$\frac{1}{\sqrt{1-x}} = 1 + \frac{1-\sqrt{1-x}}{\sqrt{1-x}} \le 1 + \frac{x}{\sqrt{1-x}} \le 1 + \sqrt{3}x.$$

Now we deal with the first term in (2.8)

$$\frac{1}{\sqrt{1-\frac{4h(b^2-a^2)}{3ab}}}\left(2\Phi\left(\frac{h}{\sigma}\right)-1\right) \le 2\Phi\left(\frac{h}{\sigma}\right) - 1 + \sqrt{3}\frac{4h\left(b^2-a^2\right)}{3ab}\left(2\Phi\left(\frac{h}{\sigma}\right)-1\right)$$

$$\le 2\Phi\left(h\sqrt{\frac{(a+b)^3}{ab}}\right) - 1 + \sqrt{3}\frac{4h\left(b^2-a^2\right)}{3ab}.$$

Finally

$$\frac{\int\limits_z^{z+2h} p^a (1-p)^b \, dp}{\int\limits_0^1 p^a (1-p)^b \, dp} \le 2\Phi\left(h\sqrt{\frac{(a+b)^3}{ab}}\right) - 1 + \sqrt{3}\frac{4h\left(b^2-a^2\right)}{3ab} + \frac{2h}{\sqrt{\pi}}.$$

Now we substitute this inequality in (2.4)

$$E\left(\mathcal{L}|s,f\right) \ge \frac{h^2\phi(c)}{c}(a+b-2) - \frac{ab}{(a+b)(a+b+1)}\left(2\Phi\left(h\sqrt{\frac{(a+b)^3}{ab}}\right)-1\right) -$$

$$-\frac{ab}{(a+b)(a+b+1)}\sqrt{3}\frac{4h\left(b^2-a^2\right)}{3ab} - \frac{ab}{(a+b)(a+b+1)}\frac{2h}{\sqrt{\pi}}$$

$$\ge \frac{h^2\phi(c)}{c}(a+b-2) - \frac{ab}{(a+b)(a+b+1)}\left(2\Phi\left(h\sqrt{\frac{(a+b)^3}{ab}}\right)-1\right) - \frac{4\sqrt{3}h}{3} - \frac{h}{2\sqrt{\pi}}$$

For all $x \geq 0$ define the function

$$K(x) = \frac{\phi(c)}{c}x - \left(2\Phi\left(\sqrt{x}\right) - 1\right).$$

Its derivative

$$K'(x) = \frac{\phi(c)}{c} - \frac{\phi\left(\sqrt{x}\right)}{\sqrt{x}}$$

is an increasing function and has a unique root $x = c^2$. In fact, $K'(x) < 0$ when $0 < x < c^2$ and $K'(x) > 0$ for $x > c^2$. Therefore,

$$K(x) \geq K\left(c^2\right) = c\phi(c) - \gamma.$$

Using this inequality we complete the proof.

$$E(\mathcal{L}|s, f) \geq \frac{h^2\phi(c)}{c}(a+b) - \frac{ab}{(a+b)(a+b+1)}\left(2\Phi\left(h\sqrt{\frac{(a+b)^3}{ab}}\right) - 1\right) -$$

$$-2\frac{h^2\phi(c)}{c} - \frac{4\sqrt{3}h}{3} - \frac{h}{2\sqrt{\pi}} \geq \frac{ab}{(a+b)(a+b+1)}\cdot$$

$$\cdot\left(\frac{h^2\phi(c)}{c}\frac{(a+b)^2(a+b+1)}{ab} - \left(2\Phi\left(h\sqrt{\frac{(a+b)^2(a+b+1)}{ab}}\right) - 1\right)\right) - C_3h$$

$$\geq \frac{ab}{(a+b)(a+b+1)}K\left(\frac{h^2(a+b)^2(a+b+1)}{ab}\right) - C_3h$$

$$\geq E(p(1-p)|s, f)(c\phi(c) - \gamma) - C_3h,$$

where C_3 depends only on γ. This completes the proof of Lemma 1.

The above argument suggests as a solution of the Bayes problem the stopping boundary

$$\frac{(s+f+2)^2\,(s+f+3)}{(s+1)\,(f+1)}=\frac{c^2}{h^2}.$$

This stopping boundary is a special case of the Chow-Robbins stopping rules.

2.3 Lower Bound on the Integrated Risk with Beta Prior

In this section we assume that p is a Beta(a,b) random variable, where a and b are positive numbers. Denote by $E_{a,b}$ the expectation, when the prior distribution of p is Beta(a,b).

Lemma 2. *There is a constant C_1 depending only on a and b such that*

$$E_{a,b}\left(\mathcal{L}|s,f\right)\geq\left(c\phi\left(c\right)-\gamma\right)E_{a,b}\left(p\left(1-p\right)|s,f\right)-C_1h$$

Proof. By (2.2)

$$E_{a,b}\left(\mathcal{L}|s,f\right)=\frac{h^2\phi\left(c\right)}{c}\left(s+f\right)-E_{a,b}\left(p\left(1-p\right)1_{L\leq p\leq R}|s,f\right)$$

where the second expectation depends only on the posterior distribution of p, which is Beta$(a+s, b+f)$. Therefore,

$$E_{a,b}(\mathcal{L}|s,f) = E_{1,1}(\mathcal{L}|s+(a-1), f+(b-1)) - (a+b-2)\, h^2 \frac{\phi(c)}{c} \tag{2.9}$$

By Lemma 1

$$E_{1,1}(\mathcal{L}|s+(a-1), f+(b-1)) \geq E_{1,1}(p(1-p)\,|s+(a-1), f+(b-1)) - Ch$$

$$= E_{a,b}(p(1-p)\,|s,f) - Ch,$$

again because the conditional distribution of p is the same. Combine these two results and the fact that $h < 1$ to complete the proof of the lemma.

The optimal boundary suggested by the above argument when the prior is Beta(a, b) is given by

$$\frac{(s+f+a+b)^2(s+f+a+b+1)}{(s+a)(f+b)} = \frac{c^2}{h^2}.$$

Lemma 3. *There is a constant C_1 depending only on a and b such that*

$$E_{a,b}(\mathcal{L}) \geq (c\phi(c) - \gamma)\, E_{a,b}(p(1-p)) - C_1 h.$$

Proof. Integrate the inequality in Lemma 2 with respect to the distribution of (s, f).

2.4 Asymptotic Lower Bound on the Sample Size of a Fixed Width Confidence Interval

Now we apply the result from the previous section to obtain a lower bound on the average expected sample size of any sequential confidence interval (N, L, R) of width $2h$ and confidence level γ. We can of course assume that the width is exactly $2h$. Recall that $f_{a,b}(p)$ is the Beta(a, b) density function.

Lemma 4. *For any positive numbers a and b and any confidence interval (N, L, R) of width $2h$ and confidence level γ, a lower bound on the expected sample size averaged with respect to $f_{a,b}(p)$ as $h \to 0$ is given by*

$$h^2 \int_0^1 E_p(N) f_{a,b}(p)\, dp \geq c^2 \int_0^1 p(1-p) f_{a,b}(p)\, dp + O(h) = c^2 \frac{ab}{(a+b)(a+b+1)} + O(h).$$

Proof. Fix p and integrate

$$\mathcal{L} = \frac{h^2 \phi(c)}{c} N - p(1-p) 1_{\{L \leq p \leq R\}}$$

with respect to $X_1, X_2, \ldots$ to get

$$E(\mathcal{L}|p) = \frac{h^2 \phi(c)}{c} E_p N - p(1-p) P_p(L \leq p \leq R) \leq \frac{h^2 \phi(c)}{c} E_p N - p(1-p)\gamma.$$

Now we integrate with respect to p and get

$$E_{a,b}(\mathcal{L}) \leq \frac{h^2\phi(c)}{c} \int_0^1 E_p(N) f_{a,b}(p)\, dp - \gamma E_{a,b} p(1-p).$$

By Lemma 3, therefore,

$$c\phi(c) E_{a,b}(p(1-p)) + O(h) \leq \frac{h^2\phi(c)}{c} \int_0^1 E_p(N) f_{a,b}(p)\, dp$$

and the conclusion of the lemma follows, using the fact that

$$E_{a,b}(p(1-p)) = \frac{ab}{(a+b)(a+b+1)}$$

by (2.5).

Now we extend this result to a general class of priors.

Theorem 2. *Let $f(p)$ be a positive continuous function defined on the interval $[0,1]$. If $\{(N(h), L(h), R(h))\}$ is a family of confidence intervals of width $2h$ and confidence level γ, then*

$$\liminf_{h\to 0} h^2 \int_0^1 E_p(N) f(p)\, dp \geq c^2 \int_0^1 p(1-p) f(p)\, dp.$$

Proof. Choose $\delta > 0$ such that $f(p) \geq \delta$ for all $p \in [0,1]$. It is well known that the Bernstein polynomials

$$(B_n f)(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}$$

converge uniformly to any continuous function. Therefore, for a fixed ε there exist n and some positive numbers l_k such that

$$|f(p) - \sum_{k=0}^{n} l_k f_{k,n-k}(p)| < \varepsilon.$$

Now

$$\liminf_{h\to 0} h^2 \int_0^1 E_p(N)(f(p)+\varepsilon)\,\mathrm{d}p \geq \sum_{k=0}^{n} l_k \liminf_{h\to 0} h^2 \int_0^1 E_p(N) f_{k,n-k}(p)\,\mathrm{d}p$$

$$\geq \sum_{k=0}^{n} l_k c^2 \int_0^1 p(1-p) f_{k,n-k}(p)\,\mathrm{d}p \geq c^2 \int_0^1 p(1-p)(f(p)-\varepsilon)\,\mathrm{d}p$$

$$= c^2 \int_0^1 p(1-p) f(p)\,\mathrm{d}p - \frac{c^2}{6}\varepsilon.$$

On the other hand

$$\liminf_{h\to 0} h^2 \int_0^1 E_p(N)(f(p)+\varepsilon)\,\mathrm{d}p$$

$$\leq \liminf_{h\to 0} \left(h^2 \int_0^1 E_p(N) f(p)\,\mathrm{d}p + \varepsilon h^2 \int_0^1 E_p(N) \frac{f(p)}{\delta}\mathrm{d}p \right)$$

$$\leq \left(1+\frac{\varepsilon}{\delta}\right) \liminf_{h\to 0} h^2 \int_0^1 E_p(N) f(p)\,\mathrm{d}p.$$

Therefore,

$$\left(1+\frac{\varepsilon}{\delta}\right)\liminf_{h\to 0} h^2 \int\limits_0^1 E_p(N)\, f(p)\, \mathrm{d}p \geq c^2 \int\limits_0^1 p(1-p)\, f(p)\, \mathrm{d}p - \frac{c^2}{6}\varepsilon.$$

Letting ε go to zero, we obtain the desired result.

Remark. If we consider only stopping rules $N(h)$ for which $E(N) = O(h^{-2})$ in the above theorem, we can drop the assumption $f(p) > 0$. Note that all reasonable stopping rules are of this type. In fact the optimal fixed size stopping rule needs no more than $O(h^{-2})$ observations.

Chapter 3 Asymptotically Efficient Fixed Width Confidence Intervals of Exact Coverage for the Mean of a Population

A method for constructing width $2h$ confidence intervals (N, L, R) for the mean μ of independent identically distributed random variables $X_1, X_2, \ldots$ with finite variance is described by Chow and Robbins in [4]. The stopping rule considered is

$$N(h) = \min\left\{n \geq 1 : V_n \leq \frac{h^2 n}{c_n^2}\right\},$$

where $\lim_{n\to\infty} c_n = c = \Phi^{-1}((1+\gamma)/2)$, and γ is the desired coverage level. Here

$$V_n = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + 1}{n},$$

defining a sequence of estimators of the unknown variance. Upon stopping the confidence interval is

$$[L, R] = [\bar{X}_{N(h)} - h, \bar{X}_{N(h)} + h].$$

Theorem 3 (Chow-Robbins). *Let σ^2 be the variance of the X_i's. If $0 < \sigma < \infty$,* *then*

$$\lim_{h \to 0} h^2 N(h) = c^2 \sigma^2 \quad a.s.$$

$$\lim_{h \to 0} h^2 E(N(h)) = c^2 \sigma^2$$

$$\lim_{h \to 0} P(\bar{X}_{N(h)} - h \leq \mu \leq \bar{X}_{N(h)} + h) = \gamma. \tag{3.1}$$

The coverage probability of these confidence intervals is asymptotically γ, i.e., the probability of their containing the true μ will be arbitrarily close to γ when h is sufficiently small. However, in a parametric problem like the Bernoulli case, the definition of a confidence interval requires more than just asymptotic coverage. At a minimum what is needed is the so-called *strong asymptotic coverage*, which specifies that the convergence of the coverage probability in (3.1) is uniform in p. Then for h sufficiently small the coverage probability will be arbitrarily close to γ simultaneously for all p. This is critical since p is unknown. Theorem 4 below resolves these issues; in fact, our procedures guarantee coverage probability γ, once h is suffuciently small. Moreover, that theorem and its Corollary at the end of Section 3.2 are used along with the lower bound proved in Theorem 2 in Chapter 2 to prove Theorem 1.

We consider the general problem of constructing width $2h$ level γ confidence intervals for the mean of a population in a parametric context. Let $X_1, X_2, ...$ be independent and identically distributed random variables with unknown distribution F which belongs to a known class of distributions $\mathcal{F}$. Let μ_F and σ_F^2 be the mean and the variance of F. A level γ confidence interval (N, L, R) for the mean $\mu = \mu_F$

is a procedure satisfying

$$P_F(L \leq \mu_F \leq R) \geq \gamma \quad \text{for all} \quad F \in \mathcal{F}.$$

Fix $0 < \gamma < 1$ and $c = \Phi^{-1}((1+\gamma)/2)$.

In this chapter, but only in this chapter, we will be interested in width $2h$ confidence intervals for μ that are centered at the point estimate $\bar{X}_{N(h)}$, i.e.,

$$[\bar{X}_{N(h)} - h, \bar{X}_{N(h)} + h].$$

Therefore, we identify the confidence interval $(N(h), L = \bar{X}_{N(h)} - h, R = \bar{X}_{N(h)} + h)$ with the stopping time $N(h)$, and will refer to this confidence interval by referring only to the stopping time $N(h)$.

The stopping time we propose is defined by

$$N(h) = \min\left\{n \geq K \,:\, n - K \geq V_n \frac{c_h^2}{h^2}\right\}$$

where $K = K(h)$ and c_h are parameters to be chosen. We will find conditions on K and c_h sufficient for coverage probability γ, which also yield asymptotically efficient $N(h)$ in the sense of Theorem 4. Throughout this chapter we assume that:

1. $K = o(h^{-2})$.

2. $\lim_{h \to 0} c_h = c$.

Also, without loss of generality assume that c_h is bounded between fixed numbers c'

and c'', i.e., $0 < c' < c_h < c''$, for all h.

Theorem 4 describes a method for constructing width $2h$ level γ confidence intervals for the mean. Assume that all $F \in \mathcal{F}$ have finite fourth moment. Let

$$\rho_F = E_F|X_i - \mu_F|^3 \quad \text{and} \quad w_F^2 = E_F(X_i - \mu_F)^4.$$

We will omit the subscript F when there is no ambiguity.

Theorem 4. *Suppose that for all $F \in \mathcal{F}$, $\min\{w_F, w_F/\sigma_F^2\} < B$ for some positive constant B. Then for $K = h^{-1.85}$ and $c_h = c + h^{0.13}$ the confidence intervals defined by $N(h)$ achieve confidence level γ, i.e.,*

$$\inf_{F \in \mathcal{F}} P_F(\bar{X}_{N(h)} - h \leq \mu_F \leq \bar{X}_{N(h)} + h) \geq \gamma. \tag{3.2}$$

In addition $N(h)$ is asymptotically efficient for all $F \in \mathcal{F}$, in the sense that

$$\lim_{h \to 0} h^2 E_F N(h) = c^2 \sigma_F^2. \tag{3.3}$$

Moreover, N is asymptotically efficient with respect to any probability measure π defined on $\mathcal{F}$ for which $\int_0^1 \sigma_F^2 \, d\pi < \infty$, i.e.,

$$\lim_{h \to 0} h^2 \int E_F N \, d\pi = c^2 \int \sigma_F^2 \, d\pi. \tag{3.4}$$

The meaning of "asymptotically efficient" in the above theorem, also sometimes called *Chow-Robbins asymptotic efficiency*, is based on the following observation. If

σ were known, then by the Central Limit Theorem we would need to take approximately $c^2\sigma^2/h^2$ observations to achieve coverage probability γ using fixed sample size symmetric confidence intervals. Thus, (3.3) shows that for unknown σ a sequential stopping rule can use the sequence $\{V_n\}$ estimating σ^2 to achieve on the average the same sample size as would be required if σ^2 were known. Note that this definition of asymptotic efficiency does not necessarily guarantee best possible performance among sequential rules. However, Theorem 2 shows that in the case when $\mathcal{F}$ is the family of Bernoulli distributions and π is a probability distribution with positive continuous density, asymptotic efficiency with respect to π in the sense of (3.4) does guarantee best possible performance among sequential procedures. A discussion of related papers in the literature is given at the end of this chapter.

Remark 1. The choice of K and c_h is not the optimal choice we can make. To make such a choice one needs to look at the eight terms that provide a lower bound on the coverage probability in Lemma 9, for each term deduce a linear inequality that needs to be satisfied by $\log_h(K)$, $\log_h(c_h)$ and the other functions involved, and solve these strict linear inequalities. The actual optimal choice is close to the one we made.

3.1 Asymptotic Efficiency

This is Lemma 2 from [15].

Lemma 5. *For any $a > 0$, and $m \geq 1$,*

$$P\left(\sup_{n\geq m} |\bar{X}_n - \mu| \geq a\right) \leq \frac{8\sigma^2}{ma^2}$$

The following direct consequence of Lemma 5 is of use later.

Lemma 6.

$$E \sup_n |\bar{X}_n| < |\mu| + 16\sigma^2 + 1.$$

Proof.

$$E \sup_n |\bar{X}_n| \leq E \sup_n |\bar{X}_n - \mu| + |\mu|.$$

By Lemma 5

$$E \sup_n |\bar{X}_n - \mu| \leq 1 + \sum_{i=0}^{\infty} 2^{i+1} P(\sup_n |\bar{X}_n - \mu| \geq 2^i) \leq 1 + 8\sigma^2 \sum_{i=0}^{\infty} \frac{2^{i+1}}{2^{2i}} = 1 + 16\sigma^2$$

and the result follows. Define

$$s_n^2 = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2 = V_n - \frac{1}{n}.$$

Next we derive a lemma similar to Lemma 5.

Lemma 7. *For any $a > 0$, and $m \geq 1$,*

$$P\left(\sup_{n \geq m} |s_n^2 - \sigma^2| \geq a \right) \leq \frac{16}{m} \left(2\frac{w^2 - \sigma^4}{a^2} + \frac{\sigma^2}{a} \right),$$

where $w^2 = w_F^2 = E_F(X_i - \mu_F)^4$.

Proof.

$$
\begin{aligned}
&P\left(\sup_{n\geq m}\left|\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{n}-\sigma^2\right|\geq a\right)\\
&=P\left(\sup_{n\geq m}\left|\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}-\sigma^2-(\bar{X}_n-\mu)^2\right|\geq a\right)\\
&\leq P\left(\sup_{n\geq m}\left|\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}-\sigma^2\right|\geq \frac{a}{2}\right)+P\left(\sup_{n\geq m}\left|(\bar{X}_n-\mu)^2\right|\geq\frac{a}{2}\right)\\
&\leq \frac{32(w^2-\sigma^4)}{ma^2}+\frac{16\sigma^2}{ma}
\end{aligned}
$$

by the preceding lemma.

Lemma 8 (Asymptotically Efficient Sample Size). *Suppose that*

$$\lim_{h\to 0} h^2K(h)=0 \quad \textit{and} \quad \lim_{h\to 0} c_h = c.$$

Then

(i) for every $F\in\mathcal{F}$, (3.3) holds,

(ii) for any probability measure π on $\mathcal{F}$ satisfying $\int \sigma_F^2\,d\pi<\infty$, (3.4) holds.

Proof. (i) We use N as a shorthand for $N(h)$. First we establish a lower bound on N. Since $V_n\geq n^{-1}$ and

$$N\geq N-K\geq V_N(c_h/h)^2\geq N^{-1}(c_h/h)^2,$$

taking square roots shows that

$$N \geq c_h/h > c'/h. \tag{3.5}$$

Thus $N \to \infty$ as $h \to 0$. Since also $\lim_{n\to\infty} V_n = \sigma^2$ a.s. ,

$$\lim_{h\to 0} V_N = \sigma^2 \text{ a.s.}$$

The stopping rule also satisfies

$$(K+1)h^2 + V_{N-1}c_h^2 \geq Nh^2 \geq Kh^2 + V_N c_h^2. \tag{3.6}$$

As $h \to 0$ therefore

$$\lim_{h\to 0} Nh^2 = (\sigma c)^2 \text{ a.s.}$$

To prove (3.3), i.e., that

$$\lim_{h\to 0} E(N)h^2 = (\sigma c)^2$$

it suffices to show that $\{Nh^2\}_{h>0}$ are bounded by an integrable function for h in a neighborhood of 0. From inequality (3.6) we deduce that for h sufficiently small

$$Nh^2 \leq (c'')^2 \sup_n V_n + 1$$

It suffices to show that $E \sup_n V_n < \infty$, which follows from

$$V_n \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{n} + 1$$

by Lemma 6 and the fact that X_i has a finite fourth moment.

(ii) In order to establish (3.4), we need to show that $\{h^2 E(N)\}$ are uniformly integrable with respect to π for all h in a neighborhood of 0. From (3.6) we get

$$N(N-1)h^2 \leq (c'')^2 + (N-1)(K+1)h^2 + (c'')^2 \sum_{i=1}^{N-1} (X_i - \bar{X}_{N-1})^2$$

$$\leq (c'')^2 + N(K+1)h^2 + (c'')^2 \sum_{i=1}^{N-1} (X_i - \mu)^2$$

$$\leq (c'')^2 + N(K+1)h^2 + (c'')^2 \sum_{i=1}^{N} (X_i - \mu)^2.$$

Therefore,

$$N^2 h^2 \leq (c'')^2 + N(K+2)h^2 + (c'')^2 \sum_{i=1}^{N} (X_i - \mu)^2.$$

Using Wald's equation ([7])

$$E\left(\sum_{i=1}^{N} (X_i - \mu)^2\right) = \sigma^2 E(N),$$

and hence

$$h^2 (EN)^2 \leq h^2 E(N^2) \leq E(N)(K+2)h^2 + (c'')^2 (\sigma^2 E(N) + 1).$$

Therefore,

$$h^2E(N) \le (K+2)h^2 + (c'')^2(\sigma^2+1)$$

Since σ^2 is integrable with respect to π, we get the uniform integrability of $h^2E(N)$ with respect to π, which completes the proof of Lemma 8.

3.2 Exact Coverage Probability

Our choice for the parameters c_h and K will be based on the following technical lemma. Define $n_{min} = \max\{K, c'/h\} - 1$. Since $N(h) \ge K$, (3.5) ensures that $N(h) - 1 \ge n_{min}$. Use N as a shorthand for $N(h)$.

Lemma 9. *Suppose that ε and ε_1 are such that*

$$\frac{1}{2} > \varepsilon > \frac{2}{n_{min}\sigma^2} \tag{3.7}$$

$$\frac{c_h - c}{2} + \frac{1}{2}\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2 + Kh^2/\sigma^2} + c''} - c\varepsilon - \varepsilon_1 = \zeta > 0 \tag{3.8}$$

Then for h sufficiently small and for all $F \in \mathcal{F}$, a lower bound on the coverage *probability of the confidence intervals defined by $N(h)$ is given by*

$$P(\bar{X}_{N(h)} - h \le p \le \bar{X}_{N(h)} + h) \ge \gamma - \frac{32}{n_{min}}\left(2\frac{w^2-\sigma^4}{\left(\varepsilon\sigma^2+\frac{2}{n_{min}}\right)^2} - \frac{\sigma^2}{\varepsilon\sigma^2+\frac{2}{n_{min}}}\right) - \frac{\varepsilon}{\varepsilon_1^2}$$
$$- \frac{6\rho}{\sigma^3\sqrt{K+\sigma^2(c')^2/h^2}} - C_1\varepsilon - C_2\varepsilon_1 + C_3(c_h - c) + C_4\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2+Kh^2/\sigma^2}+c''}, \tag{3.9}$$

where the C_i's are positive and depend on γ, but not on h.

Proof. Since $N-1 \ge n_{min}$, $N-K-1 < V_{N-1}(c_h/h)^2$ and $s_n^2 = V_n - \frac{1}{n}$, we get

$$\frac{c_h^2 s_N^2}{h^2} + K \le \frac{V_N c_h^2}{h^2} + K \le N \le \frac{c_h^2 s_{N-1}^2}{h^2} + \frac{c_h^2}{(N-1)h^2} + K + 1$$
$$\le \frac{c_h^2 s_{N-1}^2}{h^2} + \frac{c_h^2}{n_{min}h^2} + K + 1.$$

Let

$$N_0 = \left\lceil \frac{c_h^2\sigma^2}{h^2} + K \right\rceil.$$

Now

$$P\left(\left|\frac{N}{N_0} - 1\right| > \varepsilon\right) = P\left(|N - N_0| > \varepsilon N_0\right)$$
$$\le P\left(\left|\frac{c_h^2 s_N^2}{h^2} - \frac{c_h^2\sigma^2}{h^2}\right| > \varepsilon N_0 - 1\right) + P\left(\left|\frac{c_h^2 s_{N-1}^2}{h^2} - \frac{c_h^2\sigma^2}{h^2}\right| > \varepsilon N_0 - 1 - \frac{c_h^2}{n_{min}h^2}\right)$$
$$\le 2P\left(\sup_{n\ge n_{min}}\left|\frac{c_h^2 s_n^2}{h^2} - \frac{c_h^2\sigma^2}{h^2}\right| > \varepsilon N_0 - 1 - \frac{c_h^2}{n_{min}h^2}\right).$$

Since $N_0 \geq c_h^2\sigma^2/h^2$ and $c_h > c'$ we get

$$P\left(\left|\frac{N}{N_0}-1\right|>\varepsilon\right) \leq 2P\left(\sup_{n\geq n_{min}}\left|\frac{c_h^2 s_n^2}{h^2}-\frac{c_h^2\sigma^2}{h^2}\right|>\varepsilon\frac{c_h^2\sigma^2}{h^2}-1-\frac{c_h^2}{n_{min}h^2}\right)$$

$$\leq 2P\left(\sup_{n\geq n_{min}}|s_n^2-\sigma^2|>\varepsilon\sigma^2-\frac{h^2}{(c')^2}-\frac{1}{n_{min}}\right)$$

Notice that for h sufficiently small $n_{min} \leq (c')^2/h^2$, since $\lim_{h\to 0} Kh^2 = 0$. So for h sufficiently small $1/n_{min} \geq h^2/(c')^2$ and we get that

$$P\left(\left|\frac{N}{N_0}-1\right|>\varepsilon\right) \leq 2P\left(\sup_{n\geq n_{min}}|s_n^2-\sigma^2|>\varepsilon\sigma^2-\frac{2}{n_{min}}\right). \tag{3.10}$$

By (3.7) we can apply Lemma 7 to the right-hand side of (3.10)

$$P\left(\left|\frac{N}{N_0}-1\right|>\varepsilon\right) \leq \frac{32}{n_{min}}\left(2\frac{w^2-\sigma^4}{\left(\varepsilon\sigma^2-\frac{2}{n_{min}}\right)^2}+\frac{\sigma^2}{\varepsilon\sigma^2-\frac{2}{n_{min}}}\right). \tag{3.11}$$

Let $S_n = \sum_1^n X_i$. Now estimate the probability that μ is not covered.

$$P(\mu \text{ is not covered}) = P\left(\left|\frac{S_N}{N}-\mu\right|>h\right) = P(|S_N - N\mu| > hN)$$

$$\leq P\left(\left|\frac{N}{N_0}-1\right|>\varepsilon\right) + P\left(\left|\frac{N}{N_0}-1\right|\leq\varepsilon \quad \text{and} \quad |S_N-N\mu|>hN\right)$$

$$\leq P\left(\left|\frac{N}{N_0}-1\right|>\varepsilon\right) + P\left(|S_{N_0}-N_0\mu|>hN_0-\varepsilon hN_0-\varepsilon_1\sigma\sqrt{N_0}\right)$$

$$+P\left(\left|\frac{N}{N_0}-1\right|\leq\varepsilon \quad \text{and} \quad |S_N-N\mu-S_{N_0}+N_0\mu|>\varepsilon_1\sigma\sqrt{N_0}\right) \tag{3.12}$$

At this point we need to estimate the last two terms in (3.12). We begin with the last term. Let $n' = \min\{N, N_0\}$. By Kolmogorov's ([6]) inequality we get

$$P\left(\left|\frac{N}{N_0} - 1\right| \le \varepsilon \quad \text{and } |S_N - N\mu - S_{N_0} + N_0\mu| > \varepsilon_1 \sigma\sqrt{N_0}\right)$$

$$\le P\left(\max_{0\le j\le N_0\varepsilon}\left|\sum_{i=n'}^{n'+j}(X_i - \mu)\right| > \varepsilon_1\sigma\sqrt{N_0}\right) \le \frac{N_0\varepsilon\sigma^2}{\varepsilon_1^2 N_0\sigma^2} = \frac{\varepsilon}{\varepsilon_1^2}.$$

The middle term in (3.12) will be estimated by normal approximation. Let $F_n(x)$ be the distribution function of

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

From the Berry-Esseen ([6]) theorem we know that

$$|F_n(x) - \Phi(x)| \le \frac{3\rho}{\sigma^3\sqrt{n}}$$

for all x. Now

$$P\left(|S_{N_0} - N_0\mu| > hN_0 - \varepsilon h N_0 - \varepsilon_1\sigma\sqrt{N_0}\right)$$

$$= P\left(\frac{|S_{N_0} - N_0\mu|}{\sigma\sqrt{N_0}} > h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right)$$

$$\le 2\left[1 - \Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right)\right] + \frac{6\rho}{\sigma^3\sqrt{N_0}}$$

$$\le 2\left[1 - \Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right)\right] + \frac{6\rho}{\sigma^3\sqrt{K + \sigma^2(c')^2/h^2}}. \tag{3.13}$$

Finally we need a lower bound for this term

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma}-\varepsilon_1\right)\geq\Phi\left(h(1-\varepsilon)\sqrt{\frac{\frac{\sigma^2c_h^2}{h^2}+K}{\sigma^2}}-\varepsilon_1\right)$$

$$=\Phi\left(h(1-\varepsilon)\sqrt{\frac{\frac{\sigma^2c_h^2}{h^2}+K}{\sigma^2}}-\varepsilon_1\right)=\Phi\left((1-\varepsilon)\sqrt{c_h^2+\frac{Kh^2}{\sigma^2}}-\varepsilon_1\right). \qquad (3.14)$$

By (3.7)

$$(1-\varepsilon)\sqrt{c_h^2+Kh^2/\sigma^2}=(1-\varepsilon)(c_h+\sqrt{c_h^2+Kh^2/\sigma^2}-c_h)$$

$$=(1-\varepsilon)(c_h+\frac{Kh^2/\sigma^2}{\sqrt{c_h^2+Kh^2/\sigma^2}+c_h})\geq(1-\varepsilon)\left(c+(c_h-c)+\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2+Kh^2/\sigma^2}+c''}\right)$$

$$\geq(1-\varepsilon)c+\frac{c_h-c}{2}+\frac{1}{2}\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2+Kh^2/\sigma^2}+c''}=c-\zeta+\varepsilon_1.$$

Therefore,

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma}-\varepsilon_1\right)\geq\Phi(c-\zeta).$$

Notice that for h sufficiently small ζ is less than 1. Also by (3.8) $\zeta>0$. Denote the minimum of ϕ on the interval $(c,c+1)$ by M. By the mean value theorem, for some $\xi\in(c,c+\zeta)$

$$\Phi(c+\zeta)=\Phi(c)+\zeta\phi(\xi)\geq\Phi(c)+\zeta M.$$

Finally

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma}-\varepsilon_1\right)\geq\Phi(c)+\left(-c\varepsilon+\frac{c_h-c}{2}+\frac{1}{2}\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2+Kh^2/\sigma^2}+c''}-\varepsilon_1\right)M$$

We now combine this inequality with inequality (3.13) to get the desired estimate of the middle term in (3.12), and the proof of the lemma is complete.

Lemma 10. *Suppose that there are positive σ_0 and b such that for all $F\in\mathcal{F}$, $\sigma_0^2\leq\sigma_F^2$ and $w_F/\sigma_F^2<b$. If $c_h\geq c+h^{0.13}$ and $K\geq 0$, then for h sufficiently small the confidence intervals defined by $N(h)$ achieve confidence level γ.*

Proof. Choose $\varepsilon=h^{0.42}$ and $\varepsilon_1=h^{0.14}$. It is easy to see that for h sufficiently small (3.7) and (3.8) are satisfied. We also see that

$$\varepsilon_1,\varepsilon,\frac{\varepsilon}{\varepsilon_1^2}=o(c_h-c).$$

Also

$$\frac{6\rho}{\sigma^3\sqrt{K+\sigma^2(c')^2/h^2}}<\frac{6h\rho}{\sigma^4c'}<\frac{6hw^{3/2}}{\sigma^4c'}<\frac{6hw^2}{\sigma^4c'\sigma_0}=o(c_h-c).$$

Finally for h sufficiently small

$$\frac{32}{n_{min}}\left(2\frac{w^2-\sigma^4}{\left(\varepsilon\sigma^2-\frac{2}{n_{min}}\right)^2}+\frac{\sigma^2}{\varepsilon\sigma^2-\frac{2}{n_{min}}}\right)\leq\frac{32}{n_{min}}\left(2\frac{w^2/\sigma^4-1}{\left(\varepsilon-\frac{2}{n_{min}\sigma_0^2}\right)^2}+\frac{1}{\varepsilon-\frac{2}{n_{min}\sigma_0^2}}\right)$$

$$\leq\frac{1}{n_{min}\varepsilon^2}\leq\frac{h}{h^{0.84}}=o(c_h-c).$$

By Lemma 9 for h sufficiently small the dominating nonconstant term in (3.9) is $c_h - c$ and therefore

$$P(p \text{ is not covered}) < 1 - \gamma.$$

Lemma 11. *Suppose that for some positive constant w_0, $w_F^2 \leq w_0^2$, for all $F \in \mathcal{F}$. Then if $K \geq h^{-1.85}$ and $c_h \geq c$, for h sufficiently small the confidence intervals defined by $N(h)$ achieve confidence level γ.*

Proof. From Lemma 5 we see that if the variance is small (depending on h), the desired coverage is easy to establish:

$$P\left(|\bar{X_N} - \mu| > h\right) \leq P\left(\sup_{n \geq K} |\bar{X}_n - \mu| > h\right) \leq \frac{8\sigma^2}{Kh^2}.$$

So if $\sigma^2 < Kh^2(1-\gamma)/8$ there is nothing to prove. Assume that $\sigma^2 \geq Kh^2(1-\gamma)/8 \geq (1-\gamma)h^{0.15}/8$. Again in (3.9) choose $\varepsilon = h^{0.48}$ and $\varepsilon_1 = h^{0.16}$. It's easy to see that for h sufficiently small, (3.7) and (3.8) are satisfied. Also

$$\varepsilon, \varepsilon_1, \frac{\varepsilon}{\varepsilon_1^2} = o(Kh^2).$$

Notice that $n_{min} = K$, for h sufficiently small. Since w is uniformly bounded, so are σ and ρ, i.e., $\sigma \leq \sigma_0$ and $\rho \leq \rho_0$. For h sufficiently small

$$\frac{32}{n_{min}}\left(2\frac{w^2 - \sigma^4}{\left(\varepsilon\sigma^2 - \frac{2}{n_{min}}\right)^2} + \frac{\sigma^2}{\varepsilon\sigma^2 - \frac{2}{n_{min}}}\right) < \frac{1}{K\varepsilon^2\sigma^4} < \frac{1}{h^4K^3\varepsilon^2} \leq h^{0.59} = o(Kh^2).$$

Also for h sufficiently small

$$\frac{6\rho}{\sigma^3\sqrt{K+\sigma^2(c')^2/h^2}} < \frac{6\rho_0}{\sigma^3 K^{0.5}} < \frac{1}{h^3K^2} \leq h^{0.7} = o(Kh^2).$$

Finally, for h sufficiently small

$$\frac{Kh^2/\sigma^2}{\sqrt{(c'')^2+Kh^2/\sigma^2}+c''} \geq \frac{Kh^2/\sigma_0^2}{\sqrt{(c'')^2+Kh^2/\sigma_0^2}+c''} \geq \frac{Kh^2}{c''\sigma_0^2}.$$

From (3.9) we get that for h sufficiently small

$$P(p \text{ is not covered}) < 1-\gamma.$$

Remark 2. The minimal possible order of magnitude for K in the above proof is actually $h^{-1.8}$ in the following sense. There is a constant, say C, depending only on the confidence level γ, such that if $K = Ch^{-1.8}$, the symmetric confidence intervals will achieve precise coverage for h sufficiently small. To see this we need to repeat the above proof with $\varepsilon = h^{0.6}$, $\varepsilon_1 = h^{0.2}$. The magnitudes of the terms in (3.9) are as follows:

$$\frac{1}{K^3h^4\varepsilon^2} \sim Kh^2 \sim \varepsilon_1 \sim \frac{\varepsilon}{\varepsilon_1^2} \sim h^{0.2}$$

$$\frac{1}{h^3K^2} \sim h^{0.6}$$

The role of the constant C is to ensure that the coefficient in front of Kh^2 is larger than all the coefficients in front of the terms of the same magnitude $h^{0.2}$, which will make Kh^2 again the dominating term in (3.9).

Proof of Theorem 4. If $w_F < B$ then Lemma 11 applies. If $w_F > B$, then $B/\sigma_F^2 < w_F/\sigma_F^2 < B$, which implies that $\sigma_F > 1$ and Lemma 10 applies.

Remark 3. The condition that $\min\{w, w/\sigma^2\}$ is uniformly bounded can be restated as

$$\lim_{w \to \infty} \frac{w}{\sigma^2} < \infty$$

and it is satisfied for most common family of distributions. Examples are the Normal, Exponential, Chi-Square, Extreme Value, Poisson and Geometric families of distributions. This condition will fail only for classes of distributions with very heavy tails.

The next corollary says that if the random variables X_i are bounded, we can make a simple choice for c_h, namely c. In particular this applies to the family of Bernoulli distributions.

Corollary 1. *Let X_i be bounded iid random variables, i.e., $|X_i| < B$ for some constant B independent of F. Let N be the stopping rule*

$$N = \min\left\{n \geq K \ : \ n - K \geq V_n \frac{c^2}{h^2}\right\}.$$

For $K = h^{-1.85}$ this stopping rule is asymptotically efficient for any F in the sense of (3.3) and for any probability measure π on $\mathcal{F}$ in the sense of (3.4), and the confidence intervals defined by N achieve confidence level γ.

Proof. By Lemma 11 confidence level γ is achieved, and the asymptotic efficiency follows from Lemma 8. Notice that since the X_i's are bounded, so is σ and therefore

$\int \sigma^2 \, d\pi < \infty$, for any π.

Application of this corollary together with Theorem 2 completes the proof of Theorem 1.

3.3 Two Stage Confidence Intervals

Here we discuss the two stage procedures associated with our stopping rules. Theorem 5 shows that the advantage of taking the observations fully sequentially will not appear in the first order term of $E(N)$.

Theorem 5. *If* $\min\{w, w/\sigma^2\}$ *is uniformly bounded for* $F \in \mathcal{F}$, *then for* $K = \lceil h^{-1.85} \rceil$ *and* $c_h = c + h^{0.13}$ *the two stage procedure*

1. *Take* K *observations.*

2. *If* $K < \left\lceil V_K \frac{c_h^2}{h^2} \right\rceil$ *then take* $\left\lceil V_K \frac{c_h^2}{h^2} \right\rceil - K$ *more observations; otherwise stop.*

3. $[L, R] = [\bar{X}_{N(h)} - h, \bar{X}_{N(h)} + h]$.

achieves confidence level γ *and is asymptotically efficient for any* $F \in \mathcal{F}$ *in the sense of (3.3). In addition if* $\pi(F)$ *is a probability measure on* $\mathcal{F}$ *such that* $\int \sigma^2 \; d\,\pi(F) < \infty$, *this two stage procedure is asymptotically efficient with respect to* π *in the sense of (3.4).*

Proof: The proof is based on the same ideas we used in the proof of Theorem 4. Take

$$N = \left\lceil \max\left\{K, V_K \frac{c_h^2}{h^2}\right\} \right\rceil$$

$$N_0 = \left\lceil \max\left\{K, \sigma^2 \frac{c_h^2}{h^2}\right\} \right\rceil.$$

It is easy to see that

$$|N - N_0| \le \frac{c_h^2}{h^2}|\sigma^2 - V_K| + 1.$$

Define $n_{min} = K - 1$. Inequalities (3.10), (3.11) and (3.12) are obtained the same way. The analogues of (3.13) and (3.14) are

$$P\left(|S_{N_0} - N_0\mu| > hN_0 - \varepsilon h N_0 - \varepsilon_1 \sigma\sqrt{N_0}\right)$$

$$\le 2\left[1 - \Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right)\right] + \frac{6\rho}{\sigma^3\sqrt{K}}$$

and

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right) \ge \Phi\left((1-\varepsilon)c_h - \varepsilon_1\right).$$

As before we obtain

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right) \ge \Phi\left(c - c\varepsilon + \frac{c_h - c}{2} - \varepsilon_1\right).$$

$$\Phi\left(h(1-\varepsilon)\frac{\sqrt{N_0}}{\sigma} - \varepsilon_1\right) \ge \Phi(c) + \left(-c\varepsilon + \frac{c_h - c}{2} - \varepsilon_1\right)\Phi'(c+1),$$

and as in (3.9) we get

$$P(p \text{ is not covered}) \le \frac{32}{n_{min}}\left(2\frac{w^2 - \sigma^4}{\left(\varepsilon\sigma^2 - \frac{2}{n_{min}}\right)^2} + \frac{\sigma^2}{\varepsilon\sigma^2 - \frac{2}{n_{min}}}\right) + \frac{\varepsilon}{\varepsilon_1^2} + \frac{6\rho}{\sigma^3\sqrt{K}} +$$

$$+C_1\varepsilon + C_2\varepsilon_1 - C_3(c_h - c) + 2(1 - \Phi(c)).$$

Now we take $K = h^{-1.85}$, $\varepsilon = h^{0.42}$, $\varepsilon_1 = h^{0.14}$ and $c_h = c + h^{0.13}$. As in Lemma 11 we can deduce that $\sigma^2 \geq C_4 h^{0.15}$ where C_4 is independent of h. The only new term we have in the non-coverage probability is

$$\frac{6\rho}{\sigma^3\sqrt{K}} \leq \frac{6w^{3/2}}{\sigma^3\sqrt{K}}.$$

Suppose that $\min\{w, w/\sigma^2\} \leq B$. Then for all $w > B$, $w/\sigma^2 \leq B$ and

$$\frac{6\rho}{\sigma^3\sqrt{K}} \leq \frac{6B^{3/2}}{\sqrt{K}} = o(c_h - c)$$

If $w \leq B$

$$\frac{6\rho}{\sigma^3\sqrt{K}} \leq \frac{6B^{3/2}}{\sigma^3\sqrt{K}} \leq \frac{6B^{3/2}}{(C_4 h^{0.15})^{3/2}\sqrt{K}} = o(c_h - c).$$

This establishes $P(p \text{ is not covered}) \leq 1 - \gamma$ for h sufficiently small. The remaining claims are proved as in Theorem 4.

Remark 4. When the random variables X_i are bounded, then as in the sequential case we can choose $c_h = c$. However, to guarantee coverage probability γ the second stage has to be modified to "take $\left\lceil V_K \frac{c^2}{h^2} \right\rceil$ more observations."

3.4 Related Literature

A parametric problem was first considered by Anscombe in [1] in the case of the normal distribution. This problem is quite different from the Bernoulli problem considered in this thesis, since the sample mean and the sample variance are independent in the case of the normal distribution. Other notable results for normal populations include those of Simons [17], who showed that a modified Chow-Robbins procedure achieves exact coverage, Starr [18], who investigates numerically the errors in the coverage probability approximation for a Chow-Robbins type boundary, and Woodroofe [20], who finds an asymptotic lower bound on expected sample size under the condition of uniform convergence of the coverage probability. Also in the normal population case, second order results are obtained by Simons [17] and Woodroofe [20]. There are, however, no results in the literature establishing for non-normal distributions asymptotic optimality in the sense of Theorem 1, where confidence level γ is guaranteed. While there is considerable research which studies specific sampling plans and compares them to the hypothetical fixed sample size plan if the variance were known, there is no research which seeks to find optimal sequential procedures and lower bounds on their expected sample sizes, except [20] in the normal case. We conjecture that a lower bound on the average expected sample size, similar to Theorem 2 in the Bernoulli case, is true also for a general class of distributions, under suitable conditions.

The first important theorem about a two-stage interval estimation procedure was established by Stein [19] in the case of the normal distribution. There have been other modifications and improvements of this procedure, for example [9] and [13]. However, the literature contains no optimality results, like those in Theorem 5, for non-normal distributions, except in [8], where the exponential distribution is considered.

For a comprehensive and up-to-date review in the area of fixed-width sequential and multistage confidence intervals, one can refer to [7] and [14].

Chapter 4 Exact Lower Bounds and Nearly Optimal Procedures

4.1 Backward Induction

Here we describe a method for computing explicit lower bounds on $B(\gamma, d, \lambda)$. We define a Bayes problem similar to the problem defined in Chapter 2. Let p be a random variable with density function λ and let $X_1, X_2, \ldots$ be Bernoulli random variables with parameter p. Define the loss function of a sequential procedure (N, L, R), $R = L + d$ to be

$$\mathcal{L} = eN - (p(1-p))^l 1_{\{L \leq p \leq R\}}$$

where e and l are parameters to be specified later. We are interested in finding procedures that minimize the integrated risk $I = E(\mathcal{L})$.

There is a nice interpretation of this problem. An unknown p is selected from a known distribution λ. A statistician can sample from Bernoulli(p) and each observation costs e. The statistician decides how much to sample and after stopping announces an interval of width d. The statistician receives $(p(1-p))^l$ if p belongs to that interval and otherwise receives 0. We are interested in the strategy that maximizes the earnings. The solution of this problem is given by Optimal Stopping Theory in the form of an algorithm called Backward Induction ([3], [5]). By Theorem

3, Chapter 7, in [3] there exists a t such that for the optimal strategy $P_p(N \leq t) = 1$ for all p. Therefore, we can restrict the search for the optimal strategy to the class of strategies with uniformly bounded sample size. The smallest possible posterior loss for (N, L, R), after observing s successes and f failures and stopping is

$$E(\mathcal{L}|f,s) = E(\mathcal{L}|N = s+f, S_N = s) = e(s+f) - \max_L \int_L^{L+d} (p(1-p))^l \lambda(p|s,f)\,dp.$$

Let $I(s,f)$ denote the smallest possible posterior loss of an optimal strategy, given that the point (s,f) is reached. Then

$$I(s,f) = \min\{E(\mathcal{L}|s,f), P(X_{s+f+1} = 1|s,f)I(s+1,f) + P(X_{s+f+1} = 0|s,f)I(s,f+1)\} \tag{4.1}$$

This equation expresses the fact that the smallest possible conditional expected loss, given that the point (s,f) is reached, is the minimum of the smallest loss if one stops and the smallest loss if one continues. The latter is evaluated by considering the probabilities of the cases $X_{s+f+1} = 1$ and $X_{s+f+1} = 0$ and multiplying by the best possible conditional expected losses in those cases. Now we wish to compute $I = I(0,0)$. To begin, we compute $I(s,f)$ for $(s,f) = (0,t), (1,t-1), ..., (t,0)$ using

$$I(s,f) = E(\mathcal{L}|f,s).$$

Then for each $k = t-1, t-2, ..., 0$ in succession, the "backward induction" proceeds by computing $I(s,f)$ for $(s,f) = (0,k), (1,k-1), ..., (k,0)$ using (4.1). The optimal stopping time is defined by: stop at the first n such that

$$I(S_n, n - S_n) = E(\mathcal{L}|S_n, n - S_n),$$

i.e., the stopping points are those (s, f) for which $I(s, f) = E(\mathcal{L}|s, f)$, and each point which is not a stopping point is a continuation point.

Denote I by $I_{e,l}$ and the optimal stopping time by $N^*(e, l)$ to indicate the dependence on the parameters e and l. The stopping rule N^*, with parameters chosen properly, is the stopping rule which we want to calculate.

Also, the value of $I_{e,l}$ leads to a lower bound on $B(\gamma, d, \lambda)$ in the following way. If (N, L, R) is a width d level γ confidence interval, then

$$I_{e,l} \leq E(\mathcal{L}(N, L, R)) = e \int_0^1 E_p N \lambda(p)\, dp - \int_0^1 P_p(L \leq p \leq R)(p(1-p))^l \lambda(p)\, dp$$

$$\leq e \int_0^1 E_p N \lambda(p)\, dp - \gamma \int_0^1 (p(1-p))^l \lambda(p)\, dp,$$

which implies the lower bound

$$B(\gamma, d, \lambda) \geq \frac{I_{e,l} + \gamma \int_0^1 (p(1-p))^l \lambda(p)\, dp}{e}. \tag{4.2}$$

Relation (4.2) is used to obtain a numerical lower bounds on $B(\gamma, d, \lambda)$ by choosing pairs of values of e and l, computing $I_{e,l}$ by backward induction, and using the largest value of the right-hand side of (4.2) obtained in this way.

Remark 5. For other families of distributions the factor $(p(1-p))^l$ should be replaced by $(\sigma_F)^{2l}$.

Remark 6. We can obtain analytically a value of t that can be used in the algorithm but such a t would be unnecessarily large. Instead we guess a value of t, start the algorithm, and if there are continuation points on the line $s + f = t$, we increase t, repeating the procedure until there are no continuation points on that line. As an initial guess of t we take c^2/d^2, which is the asymptotic result for $p = 0.5$.

4.2 Asymptotically Efficient Stopping Times

Here we describe explicitly the stopping rules suggested by the asymptotic theory. Suppose that we are interested in finding a width $2h$ level γ confidence interval (N, L, R) for p that minimizes $\int_0^1 E_p(N) f_{a,b}(p)\, dp$, where $f_{a,b}(p)$ is the Beta(a, b) density function. The stopping rule $N'(K)$ suggested by the asymptotic theory is

$$N'(K) = \min\left\{n \;:\; n - K \geq \frac{c^2(s+a)(f+b)}{h^2n^2}\right\},$$

where $c = \Phi^{-1}\left((1+\gamma)/2\right)$ and $s = S_n, f = n - S_n$. The stopping rule depends on a parameter K which is restricted to integer values, for simplicity. The two stage procedure, $N''(c_h, K)$, depending on the parameters c_h and K, is defined by: the sample size for the first stage is K and the sample size for the second stage is

$$\max\left\{0, \left\lceil \frac{c_h^2(s+a)(f+b)}{h^2K^2} \right\rceil - K\right\},$$

where $s = S_K$ is the number of observed successes during the first stage and $f = K - S_K$ is the number of observed failures during the first stage. When the second stage sample size is 0, the sampling terminates after the first stage. The parameters c_h and K will be determined numerically as described in Section 4.4.

Remark 7. It was suggested by the asymptotic theory that for the fully sequential procedures only the parameter K needs to be selected and c_h can be taken to be c. For the two-stage procedure both parameters c_h and K need to be selected numerically. Our numerical investigation shows that if we restrict c_h to c in the two-stage procedure, the performance suffers. This was also suggested by the asymptotic theory.

4.3 Lorden's Push Algorithm

In this section we will discuss Lorden's method for constructing confidence intervals that improve upon the performance of $[\bar{X}_n - h, \bar{X}_n + h]$. Lorden shows ([12]) that in the fixed sample size case one can improve the minimum coverage probability obtained by fixed width d confidence intervals by letting $[L, R]$ increase as a function of $\bar{X}_n$ as rapidly as possible, i.e., the intervals $[L, R]$ are *pushed* to the right as much as possible. This *push* algorithm can also be applied in the sequential setting for any given stopping rule, N (see Theorem 7), and the algorithm constructs the so-called *pushed confidence intervals.* The pushed confidence intervals applied to the stopping rules $N^*(e, l), N'(K)$ and $N''(c_h, K)$ described in the previous sections form the confidence interval procedures that we propose for use.

Fix a stopping rule N. In order to apply the push algorithm we need to define an ordering of the stopping points $(s, f) = (S_N, N - S_N)$. The push algorithm uses this ordering to set confidence intervals covering smaller p values for smaller points, i.e., if (s, f) precedes (s', f') then $L(s, f) \leq L(s', f')$ and $R(s, f) \leq R(s', f')$. In the fixed sample size case the optimal (see [12]) ordering of the stopping points is defined by their maximum likelihood estimate of p, i.e., by the number of successes. In the sequential case such an optimality property is not known, i.e., there is no reasonable conjecture that determines the optimal ordering of the stopping points. In the hope of achieving near-optimality, we propose to use the maximum likelihood estimate of p, i.e., if (s, f) and (s', f') are two stopping points, we say that (s, f) precedes (s', f') if $s/(s+f) < s'/(s'+f')$. If $s/(s+f) = s'/(s'+f')$, we say that (s, f) precedes (s', f') if $s < s'$. Index the stopping points as (s_i, f_i) for $i = 1, ..., \bar{j}$. Let J be the index of the actual stopping point $(S_N, N - S_N)$. Now we have to define the confidence interval $[L, R]$ for all the possible values of J, $1, ..., \bar{j}$, so that L and R increase as a function of J as rapidly as possible.

There is another consideration that plays an important role. The sample space of J is finite, and if the confidence intervals are based only on J, then the coverage probability as a function of p will have relatively large jumps at the endpoints of the confidence intervals. Greater efficiency, i.e., larger γ, is achievable by reducing the size of the jumps through a technique called *randomization*. Let U be a uniformly distributed random variable on the interval $[-0.5, 0.5]$, independent of the X_i's. Let $Y = S_N + U$. The randomized pushed confidence intervals for p are based on the value of Y, i.e., $L = L(Y)$ and $R = L + d$ increase in Y as rapidly as possible.

The optimality property that determines the construction of Lorden's pushed intervals, even in the fixed sample size case, requires that the endpoints L and R of the confidence intervals be restricted to a finite grid $\{0, \frac{1}{m}, \frac{2}{m}, ..., \frac{m-1}{m}, 1\}$ in order to make a numerical algorithm possible. As explained in [12], the construction of optimal $[L, R]$ that are not limited to a grid requires taking monotonic majorants, which are not computable. In our numerical investigations we use $m = 10^5$, so that little is lost by this restriction. Theorem 8 shows that $\frac{2}{m}$ is an upper bound on the difference between the unrestricted shortest-width intervals and shortest-width intervals that are grid limited, i.e., that have end points in $\{0, \frac{1}{m}, ..., 1\}$. Let $p_k = \frac{k}{m}$ for $k = 0, ..., m$.

Definition 1. A confidence interval $[L, R]$ has grid coverage level γ if for $k = 1, ..., m$

$$P_p(L \leq p_{k-1} < p_k \leq R) \geq \gamma \tag{4.3}$$

for $p = p_{k-1}$ and $p = p_k$.

Choose m and r so that $\frac{r}{m} = d$.

Theorem 6 (Lorden). *Grid-limited level γ confidence intervals have grid coverage level γ.*

Proof. Let $[L, R]$ be a grid-limited level γ confidence interval. Let $y_k = \inf\{y | L(y) \geq p_k\}$ for $k = 0, ..., m$ and $y_k = -0.5$ for $k = -r, ..., -1$. Note that

$\{y_k\}_{k=-r}^{m}$ is a nondecreasing sequence. Suppose $p_{k-1} < p < p_k$. Then

$$\gamma \leq P_p(L(Y) \leq p \leq R(Y)) = P_p(L(Y) \leq p_{k-1} < p_k \leq R(Y)) \tag{4.4}$$

Now $L(y) \leq p_{k-1}$ if and only if $y < y_k$ or $y = y_k$ and $L(y_k) \leq p_{k-1}$. Also $R(y) \geq p_k$ if and only if $L(y) \geq p_{k-r}$ if and only if $y > y_{k-r}$ or $y = y_{k-r}$ and $L(y_{k-r}) = p_{k-r}$. Since Y is a continuous random variable, for all p

$$P_p(L(Y) \leq p_{k-1} < p_k \leq R(Y)) = P_p(y_{k-r} \leq Y \leq y_k).$$

Since $P_p(y_{k-r} \leq Y \leq y_k)$ is a continuous function of p, we obtain that $P_p(L(Y) \leq p_{k-1} < p_k \leq R(Y))$ is a continuous function of p, and since inequality (4.4) is valid for $p \in (p_{k-1}, p_k)$ we conclude that

$$\gamma \leq P_p(L(Y) \leq p_{k-1} < p_k \leq R(Y))$$

for $p = p_{k-1}$ and $p = p_k$.

In the fixed sample size case, the converse of the above theorem is also true, as shown in [12]. However, in the sequential case it is not true in general for all stopping rules, but it is true for some classes of stopping rules. Discussion of this will appear in [2]. When m is large the coverage level of a confidence interval is well approximated by its grid-coverage level and the procedures that we propose achieve grid-coverage level γ.

Denote by F_k the distribution function of Y when p_k is true, i.e., $F_k(y) = P_{p=p_k}(Y \le y)$. Since Y is a continuous random variable, the inverse functions $F_k^{-1}(\alpha)$, $k = 0, ..., m$, defining the α quantiles are well-defined and are strictly increasing for $0 < \alpha < 1$. It is convenient to define for all k

$$F_k^{-1}(\alpha) = \begin{cases} -.5 & if\ \alpha = 0 \\ \bar{j} + .5, & if\ \alpha = 1 \\ +\infty, & if\ \alpha > 1, \end{cases}$$

which preserves the strictly increasing property. We are interested in finding the smallest integer r such that there exist grid-limited width r/m level γ confidence intervals for p that are nondecreasing functions of Y.

Theorem 7 (Lorden). *Suppose a stopping time N is given and also an ordering $\{(s_i, f_i), i = 1, ...\bar{j}\}$ of the stopping points. Fix m and require that $L, R \in \{0, 1/m, 2/m, ..., 1\}$. Suppose that there exists a width r/m confidence interval $[L, R]$ with grid coverage level γ such that $L(Y)$ and $R(Y)$ are nondecreasing in Y. Then there exists a nondecreasing sequence $\{y_k\}_{k=-r}^{m}$ such that*

$$if \quad y_k < y < y_{k+1} \quad then \quad [L(y), R(y)] = [p_k, p_{k+r}]. \tag{4.5}$$

The sequence $\{y_k\}_{k=-r}^{m}$ satisfies

$$y_k \ge y_{k-1} \vee F_{k-1}^{-1}(\gamma + F_{k-1}(y_{k-r})) \vee F_k^{-1}(\gamma + F_k(y_{k-r})). \tag{4.6}$$

As a consequence, $y_k \geq x_k$ *for all* k*, where* $\{x_k\}_{k=-r}^{m}$ *is defined by* $x_k = -0.5$ *if* $k < 0$ *and if* $k \geq 0$

$$x_k = x_{k-1} \vee F_{k-1}^{-1}(\gamma + F_{k-1}(x_{k-r})) \vee F_k^{-1}(\gamma + F_k(x_{k-r})). \tag{4.7}$$

Then

$$[L^*(y), R^*(y)] = [p_k, p_{k+r}] \quad if \quad x_k \leq y < x_{k+1} \tag{4.8}$$

defines grid-limited width r/m *confidence intervals with grid coverage level* γ*, and* $L^*(y) \geq L(y)$ *and* $R^*(y) \geq R(y)$ *for all* y*.*

Proof. Let $y_k = \inf\{y | L(y) \geq p_k\}$ Then (4.5) is clearly satisfied. We have

$$\gamma \leq P_{p=p_{k-1}}(L(Y) \leq p_{k-1} < p_k \leq R(Y)) = P_{p=p_{k-1}}(y_{k-r} \leq Y \leq y_k) \tag{4.9}$$

and

$$\gamma \leq P_{p=p_k}(L(Y) \leq p_{k-1} < p_k \leq R(Y)) = P_{p=p_k}(y_{k-r} \leq Y \leq y_k). \tag{4.10}$$

The two inequalities yield

$$y_k \geq F_{k-1}^{-1}(\gamma + F_{k-1}(y_{k-r}))$$

and

$$y_k \geq F_k^{-1}(\gamma + F_k(y_{k-r})),$$

which proves (4.6), since $\{y_k\}$ is nondecreasing. Now it is straightforward to prove by induction on k that $y_k \geq x_k$. The induction step uses the fact that $y_{k-r} \geq x_{k-r}$ implies $F_j(y_{k-r}) \geq F_j(x_{k-r})$ for $j = k-1, k$ and the fact that F_j^{-1} is nondecreasing. The confidence interval $[L^*, R^*]$ has grid confidence level γ since (4.7) implies that (4.9) and (4.10) hold with x_k in place of y_k.

Note that the sequence $\{x_k\}$ is well-defined whether or not width r/m confidence intervals of grid coverage level γ exist. The terms x_k are all finite if and ony if $F_{k-1}(x_{k-r}) \leq 1 - \gamma$ for all k. This is how we numerically check whether grid level γ width d confidence intervals exist: we compute the sequence $\{x_k\}$ using equation (4.7), and if x_m is finite the algorithm succeeds and equation (4.8) defines the pushed confidence intervals. Otherwise, Theorem 7 says that width r/m intervals with grid coverage level γ do not exist. Thus, given N and m, we can use the push algorithm to find the smallest r (by trial and error) for which width r/m confidence intervals of grid coverage level γ exist as a nondecreasing function of Y.

The following theorem shows that grid-limited confidence intervals are not much longer than non-grid-limited ones.

Theorem 8 (Lorden). *Suppose that r is the smallest integer such that grid limited width r/m level γ confidence intervals nondecreasing in Y exist. If there exists a level γ width d confidence interval procedure that is nondecreasing in Y, then $d > (r-2)/m$.*

Proof. Suppose that $d \leq (r-2)/m$. If $p_j \leq L(y) < p_{j+1}$ then $R(y) < p_{j+r-1}$.

Define $[L^*(y), R^*(y)] = [p_j, p_{j+r-1}]$ for such y. Then clearly $[L^*(y), R^*(y)]$ defines a nondecreasing grid-limited width $(r-1)/m$ confidence interval with coverage probability γ, contrary to the assumption that r/m is the smallest possible level γ monotonic grid limited confidence interval.

Remark 8. The pushed confidence intervals are not symmetric with respect to the symmetry on the interval $[0,1] : p \to 1-p$. However, Lorden has shown in [12] that the pushed confidence intervals, once they are constructed, can be easily modified to be symmetric and all the optimalities continue to hold.

4.4 Numerical Examples

In this section we report numerical results obtained for specific sets of parameters. We use inequality (4.2) to produce explicit lower bounds on $B(\gamma, d, \lambda)$. The maximization with respect to the parameters l and e is done on a grid. Given γ, d, and λ we wish to determine the set of parameters defining N^*, N' and N'' for which the push algorithm succeeds, i.e., such that the push algorithm constructs width d level γ confidence intervals for p and

$$\int_0^1 E_p(N)\lambda(p)\,dp \tag{4.11}$$

is minimal.

For N^* this is done in the following way. For a fixed l, $E_p(N^*(e,l))$ is a decreasing function of e. Therefore, for l on a grid we determine the largest e for which the push algorithm succeeds. Then we minimize (4.11) with respect to l. The values of l we used are all in the interval $[0, 2]$.

The value of K that specifies N' is selected to be the smallest K for which the push confidence intervals succeed, the reason being again that $E_p(N'(K))$ in increasing in K. Since K is an integer this computation is very fast. In fact the biggest advantage of N' is that it requires much less computation when compared to N^* and N''. The search for K is initiated at $K = 0$.

The search for the optimal parameters c_h and K for the two stage procedure $N''(c_h, K)$ is done in the following way. Since $E_p(N''(c_h, K))$ is increasing in c_h, for a fixed K we determine the smallest c_h for which the push algorithm succeeds and then we minimize (4.11) with respect to K.

Tables 1 and 2 show that all three procedures that we propose come very close to being optimal. Another conclusion that we can draw (something we already suspect from the asymptotic theory) is that the advantage of a fully sequential procedure when compared to a two-stage procedure is relatively small, and the two-stage procedures are very effective in the Bernoulli problem. Table 2 also shows that occasionally the two-stage procedure N'' performs better than the sequential, but this is due to the fact that the two stage procedure is optimized with respect to two parameters and the sequential is optimized only with respect to one parameter, which we artificially restricted to the integer values for simplicity. Also in Table 2 in one instance the Bayesian procedure performed worse than the sequential and the two-stage procedure, which we think is due to the discrete nature of $E(N)$. Table 3 shows $\max_p E_p(N^*)$

and max N^*, which occurs for $p = 0.5$. The results show that our procedures always perform better than the Classical Extreme Tails method. However, they perform worse than Lorden's best fixed sample size method for some values of p. Table 4 shows the optimal choice for the parameters defining the stopping rules. Figures 4.1 and 4.2 show the stopping boundaries N^* and N', which are very close. Figures 4.3 and 4.4 show the stopping boundaries for the two-stage rule. The meaning of the graph is the following: the straight line represents the first stage sample size and to get the second stage sample size one needs to find the intersection point of the second boundary and the line connecting the origin of the coordinate system and the position of the random walk after the first stage. The sum of the two coordinates of that intersection point represents the total number of observations. Figures 4.5 and 4.6 show $E_p(N^*)$ as a function of p when the prior function is Beta$(1, 1)$ and Beta$(3, 3)$ compared to the classical fixed sample method and Lorden's best fixed sample method. Figures 4.7 and 4.8 show the curves defining the pushed confidence intervals and, for comparison, the straight lines defining the centered confidence intervals.

Table 1. Average Expected Sample Size with Uniform Prior, γ=0.95

h	0.1	0.065	0.05	0.03
Classical Extreme Tails	104	240	402	1096
Lorden Fixed-Sample Size	78	199	347	1004
Two-Stage Procedure(N'')	65.1	154.3	261.4	722.2
Sequential Procedure(N')	63.1	152.5	256.9	714.2
Bayesian Procedure(N^*)	62.6	151.3	256.0	712.3
Lower Bound	61.1	149.0	253.7	709.3

Table 2. Average Expected Sample Size with Beta(3,3) Prior, γ=0.95

h	0.1	0.065	0.05	0.03
Classical Extreme Tails	104	240	402	1096
Lorden Fixed-Sample Size	78	199	347	1004
Two-Stage Procedure(N'')	76.3	188.9	323.3	909.4
Sequential Procedure(N')	76.7	189.2	323.3	908.5
Bayesian Procedure(N^*)	76.1	188.2	323.1	909.7
Lower Bound	74.6	187.0	321.3	906.6

Table 3. $\max_p E_p(N^*)$ **and** $\max N^*$ **with Uniform Prior, γ=0.95**

h	0.1	0.065	0.05	0.03
$\max_p E_p(N^*)$	90.9	222.1	376.8	1065.8
$\max N^*$	92	225	378	1068

Table 4. Parameter Specification with Uniform Prior, γ=0.95

h	0.1	0.065	0.05	0.03
(c_h, K) for N''	(1.8,36)	(1.87,52)	(1.908,95)	(1.938,190)
(K) for N'	-7	-5	-5	-3
$(e10^{-3}, l)$ for N^*	(1.69,0.81)	(0.7283,0.87)	(0.4368,0.87)	(0.1594,1)
Lower Bound	(1.8,1.2)	(0.76,1.17)	(0.45,1.1)	(0.16,1.05)

Figure 4.1: N^* and N'(dotted line) , Uniform Prior, $h = 0.1$, $\gamma = 0.95$

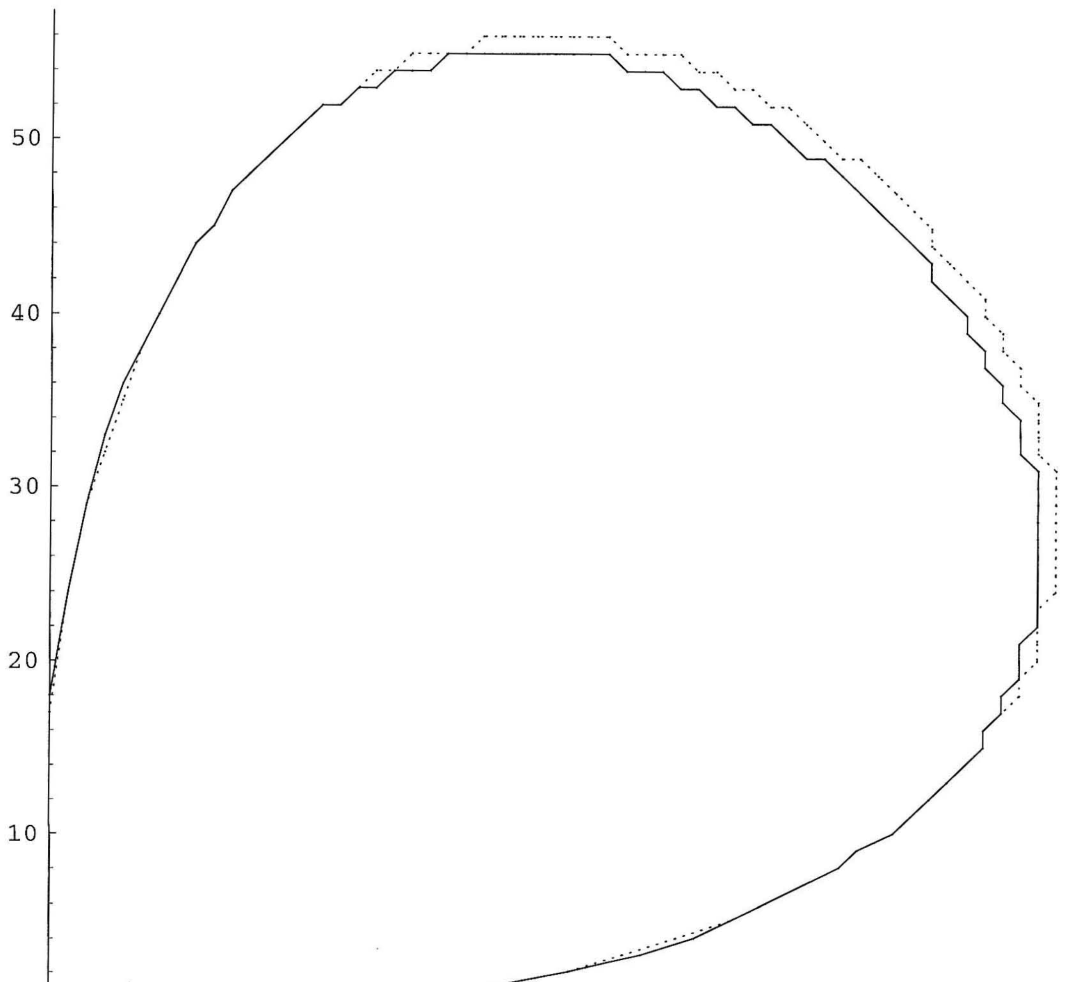


Figure 4.2: N^* and N'(dotted line), Uniform Prior, $h = 0.03$, $\gamma = 0.95$

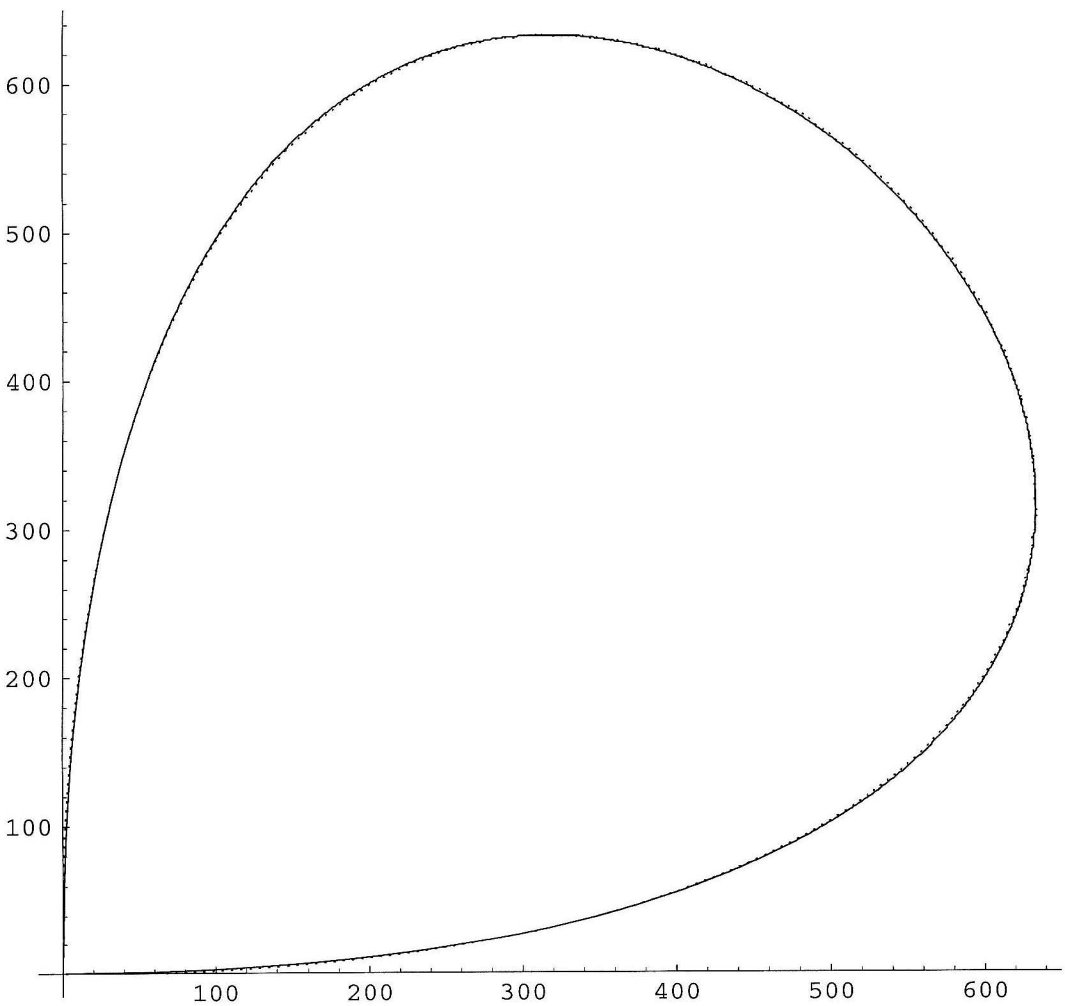


Figure 4.3: N'', Uniform Prior, $h = 0.1$, $\gamma = 0.95$

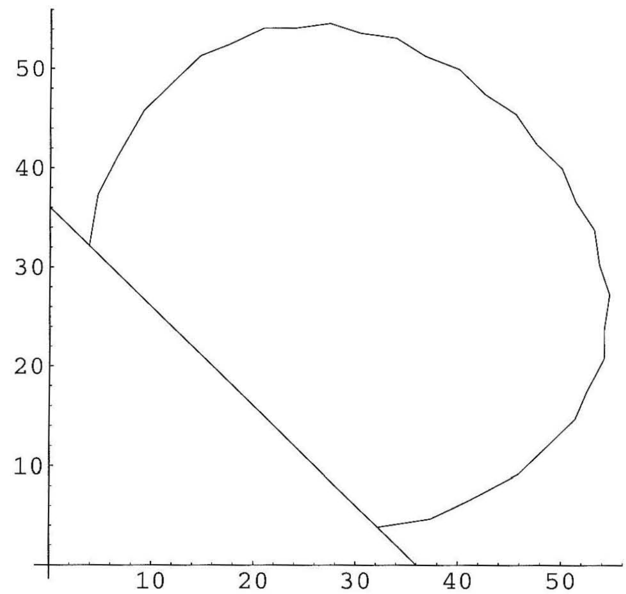


Figure 4.4: N'', Uniform Prior, $h = 0.03$, $\gamma = 0.95$

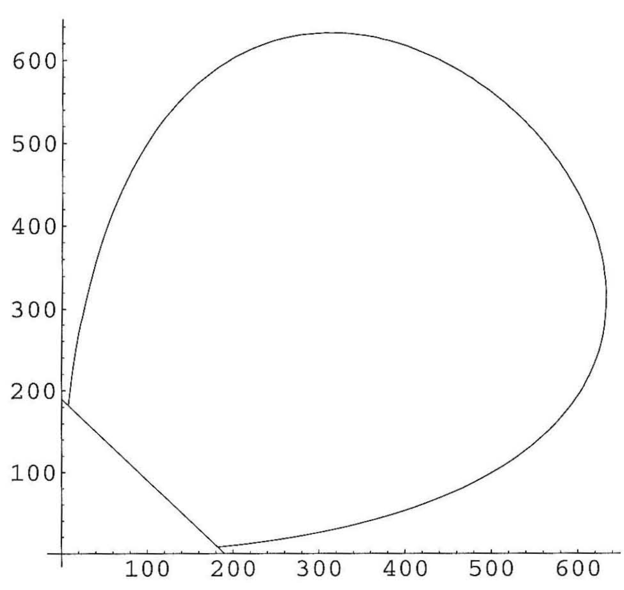


Figure 4.5: $E_p(N^*)$, Uniform Prior (the more peaked graph) and Beta(3,3) Prior, Classical Fixed Sample and Best Fixed Sample, $h = 0.1$, $\gamma = 0.95$

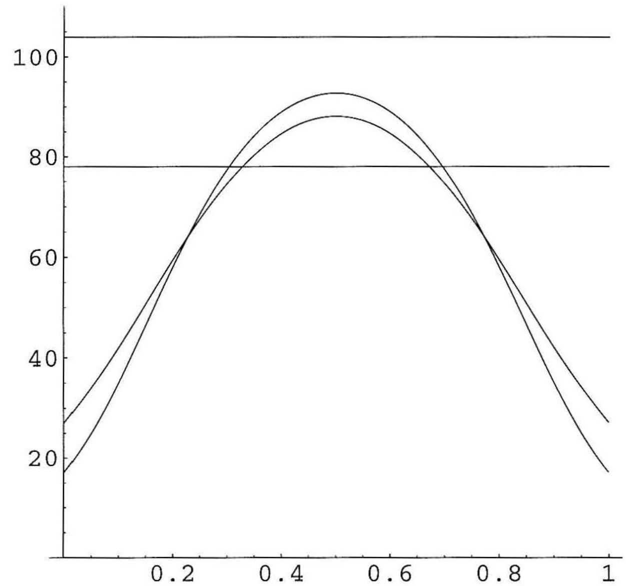


Figure 4.6: $E_p(N^*)$, Uniform Prior (the more peaked graph) and Beta(3,3) Prior, Classical Fixed Sample and Best Fixed Sample, $h = 0.03$, $\gamma = 0.95$

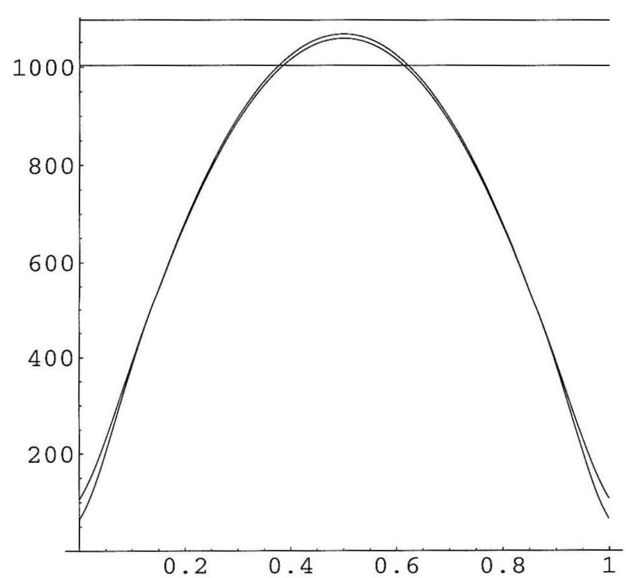


Figure 4.7: Pushed Confidence Intervals for N^* and centered intervals, Uniform Prior, $h = 0.1$, $\gamma = 0.95$

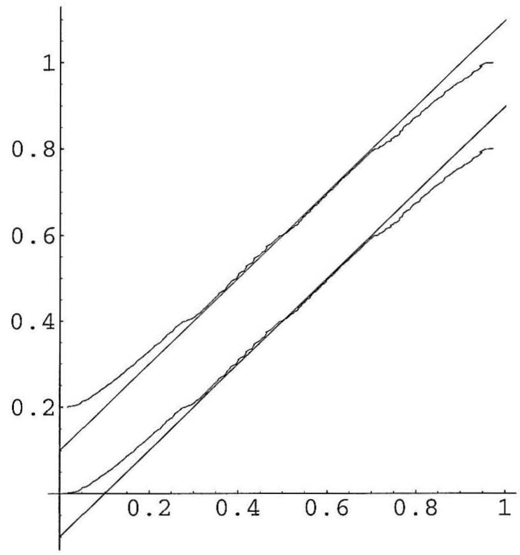
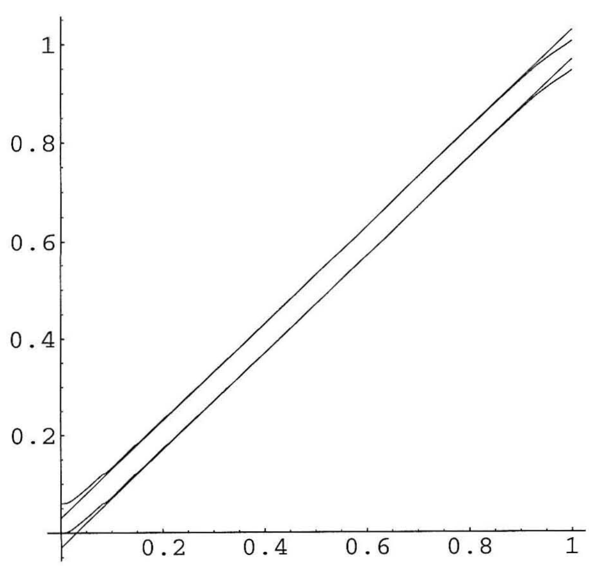


Figure 4.8: Pushed Confidence Intervals for N^* and centered intervals, Uniform Prior, $h = 0.03$, $\gamma = 0.95$



Bibliography

[1] Anscombe, F.J. Sequential Estimation, J.R. Statist. Soc., Ser. B, 15(1953)1–21

[2] Asparouhov, T. On the Unimodality of some Binomial Probabilities (in preparation)

[3] Berger, J. O. Statistical decision theory and Bayesian analysis. Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1985

[4] Chow, Y. S.; Robbins, H. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. Ann. Math. Statist. 36 1965 457–462

[5] Chow, Y. S.; Robbins, H.; Siegmund, D. Great expectations: the theory of optimal stopping. Houghton Mifflin Co., Boston, Mass., 1971

[6] Feller, W. An Introduction to Probability Theory and its Applications, Vol. I, 1957, second edition. John Wiley & Sons, New York

[7] Ghosh, M.; Mukhopadhyay, Nitis; Sen, Pranab K. Sequential estimation. Wiley Series in Probability and Statistics: Probability and

Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1997

[8] Ghurye, S. G. Note on sufficient statistics and two-stage procedures. Ann. Math. Statist.(1958) 29 155-166

[9] Hall, P. Asymptotic theory of triple sampling for sequential estimation of the mean, Ann. Statist. 9(1981) 1229–1238

[10] Karr, A. Probability, Springer-Verlag New York, Inc., 1993

[11] Lehmann, E. L. Testing statistical hypotheses. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London, 1959

[12] Lorden, G. Shortest Fixed Width Confidence Intervals for the Bernoulli parameter p (to appear)

[13] Mukhopadhyay, N. A consistent and asymptotically efficient two-stage procedure to construct fixed-width confidence intervals for the mean. Metrika, 27, 281–284

[14] Mukhopadhyay, N.; Solanky, T. K. S. Multistage selection and ranking procedures. Second-order asymptotics. Statistics: Textbooks and Monographs, 142. Marcel Dekker, Inc., New York, 1994

[15] Robbins, H.; Siegmund, D.; Wendel, J. The limiting distribution of the last time $s_n \geq n\varepsilon$. Proc. Nat. Acad. Sci. U.S.A. 61, 1968 1228–1230

[16] Ross, Sheldon M. Introduction to Probability and Statistics for Engineers and Scientists. Wiley Series in Probability and Mathematical Statistics 1987 John Wiley & Sons, Inc., New York

[17] Simons, G. On the cost of not knowing the variance when making a fixed width confidence interval for the mean, Ann. Math. Statist. 39 (1968), 1946–1952

[18] Starr, N. The performance of a sequential procedure for fixed width confidence interval estimation of a mean, Ann. Math. Statist. 37(1966), 1215–1218

[19] Stein, C. A two-sample test for a linear hypothesis whose power is independent of the variance, Ann. Math. Statist. 5(1977), 984–995

[20] Woodroofe, M. Asymptotic optimality in sequential interval estimation. Adv. in Appl. Math. 7 (1986), no. 1, 70–79