Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

ON THE ERROR OF NAIVE RARE-EVENT MONTE CARLO ESTIMATOR

Yuanlu Bai
Henry Lam

Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027, USA

ABSTRACT

We consider the estimation of rare-event probabilities using sample proportions output by naive Monte Carlo. Unlike using variance reduction techniques, this naive estimator does not have a priori relative efficiency guarantee. On the other hand, due to the recent surge of sophisticated rare-event problems arising in safety evaluations of intelligent systems, efficiency-guaranteed variance reduction may face implementation challenges, which motivate one to look at naive estimators. In this paper we investigate this naive rare-event estimator, particularly its conservativeness level and the guarantees in using it to construct confidence bounds for the target probability. We show that the half-width of a valid confidence interval is typically scaled proportional to the magnitude of the target probability and inverse square-root with the number of positive outcomes in the Monte Carlo. We also derive and compare several valid confidence bounds constructed from various techniques.

1 INTRODUCTION

We consider the problem of estimating a minuscule probability, denoted $p = P(A)$, for some rare event A, using data or Monte Carlo samples. This problem, known as rare-event estimation, is of wide interest to communities ranging from financial and insurance risk management (McNeil et al. 2015; Glasserman 2003; Glasserman and Li 2005), engineering reliability (Nicola et al. 1993; Heidelberger 1995) to climatology and hydrology (Gumbel 2012), where it is crucial to estimate the likelihood of events which, though unlikely, can cause catastrophic impacts. There are multiple prominent lines of work addressing this estimation problem, depending on how information is collected. In settings where real-world data are collected, methods based on extreme value theory (Embrechts et al. 1997) are often used to extrapolate distributional tails to assist such estimation. In settings where A is an event described by a simulable model, Monte Carlo methods can be used, and to speed up computation one often harnesses variance reduction tools such as importance sampling.

In this paper, we focus on a more basic setting than some of the above literature, but in a sense fundamental. More precisely, we focus on the situation where all we have to estimate p is a set of i.i.d. Bernoulli observations $I(A)$. A natural point estimate of p is obviously the sample proportion $\hat{p}$, i.e., given a set of Bernoulli data $I_1, \ldots, I_n$ of size n, we output $\hat{p} = (1/n)\sum_{i=1}^n I_i$. We are interested in understanding the statistical error in using $\hat{p}$, in the situation where p could be very small, importantly with no lower bound on how small it could be.

The reason why we focus on the above, apparently simple, setup is related to the limitation of variance reduction techniques in unstructured problems. While variance reduction is greatly beneficial in reducing the number of Monte Carlo samples needed to estimate rare events (Bucklew 2004; Asmussen and Glynn 2007; Rubinstein and Kroese 2016), it is also widely known that they rely heavily on model assumptions

(Juneja and Shahabuddin 2006; Blanchet and Lam 2012). That is, to guarantee the successful performances of these techniques, one typically needs to analyze the underlying model dynamics carefully to design the Monte Carlo scheme. However, recent applications, such as autonomous vehicle safety evaluation (Zhao et al. 2016; Zhao et al. 2017; Huang et al. 2017; O'Kelly et al. 2018) and robustness evaluation of machine learning predictors (Webb et al. 2018; Weng et al. 2018), lead to rare-event estimation problems with extremely sophisticated structures that hinder the design of efficiency-guaranteed variance reduction schemes. On the other hand, with the remarkable recent surge of computational infrastructure, in some situations one could afford to run gigantic amount of simulation trials. Thus, one may consider resorting to the use of naive sample proportion like described above. However, unlike the estimates given by efficiency-guaranteed variance reduction techniques, it is open, at least to our best knowledge, whether using simple sample proportion can give meaningful guarantee to estimating rare-event probabilities, in relation to the sample size n and the (unknown) magnitude of p.

2 PROBLEM SETTING AND MOTIVATING CHALLENGES

To be more concrete, in this paper we would focus on the construction of confidence bounds, i.e., using information from the Bernoulli data, or (equivalently) $\hat{p}$, we would like to construct simple confidence interval for p that has justifiable statistical guarantees. In answering this, we would also quantify the error between the point estimate $\hat{p}$ with p (probabilistically). We will focus on the upper confidence bound, since typically the lower confidence bound can be argued analogously (though not always) and also the upper direction is the more important one due to safety-related applications.

To understand the challenges, we begin by first examining the use of a standard "textbook" confidence interval. This means we use the following as the $(1-\alpha)$-level upper confidence bound

$$\hat{p}^{\text{naive}} = \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{1}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of a standard normal variable. The typical way to justify (1) is a Gaussian approximation using the central limit theorem (CLT), which entails that

$$P(p \le \hat{p}^{\text{naive}}) \approx \bar{\Phi}(-z_{1-\alpha}) = 1-\alpha \tag{2}$$

where we denote $\bar{\Phi}$ (and Φ) as the tail (and cumulative) distribution function of standard normal.

To delve a little further, note that the approximation error in (2) is controlled by the Berry-Essen Theorem. To simplify the discussion, suppose we are in a more idealized (but unrealistic) case that we know the precise value of the standard deviation of the Bernoulli trial, i.e., $\sigma^2 = p(1-p)$, so that we use $\hat{p} + z_{1-\alpha}\sigma/\sqrt{n}$. Then the Berry-Essen theorem stipulates that

$$|P(p \le \hat{p}^{\text{naive}}) - \Phi(z_{1-\alpha})| \le \frac{C\rho}{\sigma^3\sqrt{n}} \tag{3}$$

where $\rho = E|I_i - p|^3 = p(1-p)(1-2p+2p^2)$, and C is a universal constant (≈ 0.4748). Thus, the error in (3) is bounded by

$$\frac{Cp(1-p)(1-2p+2p^2)}{p^{3/2}(1-p)^{3/2}\sqrt{n}} \le \frac{C}{\sqrt{np(1-p)}}. \tag{4}$$

The issue is that when p is tiny, np can also be tiny unless n is sufficiently big, but a priori we would not know what n is "sufficient" (Note that if we have used the "self-normalized" bound given by (1), a similar Berry-Esseen bound would ultimately conclude the same issue as revealed by (4); Shao and Wang 2013). These lead to the following exemplified questions:

1. Suppose we have, say, 30 "success" outcomes among n trials, then we may think that $np \approx 30$, so that from the bound (4) the error of $\hat{p}^{naive}$ appears controlled. As another, more extreme, case, suppose we only have only 1 success, then we may be led to believe $np \approx 1$, so that $\hat{p}^{naive}$ is not "too bad" but its coverage is likely way off from $1-\alpha$. Are these conclusions on $\hat{p}^{naive}$ correct? Note that the guess that $np \approx 30$ or $np \approx 1$ is itself based on some central limit or concentration argument, which apparently leads to a circular reasoning.
2. If we have 1 success among n trials, how do we construct a CI that is guaranteed the correct coverage? Correspondingly, would constructing such a CI be easier if we have 30 successes?
3. For the valid CIs constructed in Question 2, what is the typical "half-width", relative to the point estimate $\hat{p}$? If the relative half-width is too big, then the provided upper bound may not be meaningful. For instance, if n is 0, then (1) is clearly meaningless. What about if n is 1 when we use a valid CI? Note that the use of relative half-width is important since p (and correspondingly $\hat{p}$) could be tiny, and a meaningful upper bound should have a similar magnitude as p (and $\hat{p}$).
4. Following up Question 3, we would also like to understand the relative error of the point estimate, i.e., $(\hat{p}-p)/p$, and the relative error of the confidence bound, e.g., $(\hat{p}^{naive}-p)/p$, the latter related to Question 3.
5. Do all the above answers hold if we stop whenever we observe enough successes (e.g., when the number of successes is 30, or 1) in our simulation experiment?

Note that a quick and implementable approach to Question 2 is to utilize the fact that $n\hat{p}$ follows a binomial distribution and extract a finite-sample confidence region using duality. Though this is computationally easy, we are interested in simpler mathematical forms that allow us to answer Questions 3 and 4 above. In this regard, Wilson's interval (Agresti and Coull 1998) has been studied and shown to give superior empirical performances, even in the case that p is tiny, but we are not aware of any rigorous proof on its validity. In Sections 3 and 4 below, we will offer two different ways of constructing upper confidence bounds for p, one using a concentration inequality, while another one using the Berry-Esseen bound. They both offer answers to Question 2 on top of the binomial confidence region and Wilson's interval. Compared to the binomial confidence region, these two bounds are in a similar form as $\hat{p}^{naive}$, which allows us to answer Question 1. They also have the explicit forms that allow us to investigate their half-widths, thereby answering Question 3.

In fact, we will show that the half-widths of our constructed confidence bounds scale in the same magnitude as $\hat{p}$, and inverse square-root of the number of positive outcomes $n\hat{p}$. Thus, even if we only have one positive outcome in our sample, the resulting valid confidence bound is still not "too loose", and the bounds get better at the canonical rate in the number of positive outcomes. Compared to $\hat{p}^{naive}$, our bound has a longer half-width that is of higher-order of $n\hat{p}$, and this can be viewed as a price of "validity" paid to make $\hat{p}^{naive}$ correct. Lastly, we also compare our bounds to Wilson's bound and see that ours is of a similar magnitude.

We will briefly conclude that, together with an analogous valid lower confidence bound (not shown in this paper), we can answer Question 4. Lastly, we use similar methods to derive confidence bounds for p in the situation asked by Question 5, which is presented in Section 5.

3 CONFIDENCE BOUND VIA INVERTING A CONCENTRATION BOUND

We present our first approach to construct an upper confidence bound for p by utilizing the Chernoff inequality. We can either directly invert this inequality, or consider the following set that resembles the common "p-value" method:

$$\{0 \le p \le 1 : F(\hat{p}) \ge \alpha\}$$

where $F(\cdot)$ is the distribution function of $\hat{p}$ (which depends on p). This region is a $1-\alpha$ confidence region for p, i.e., $P(F(\hat{p}) \ge \alpha) \ge 1-\alpha$. Note that if F was continuous, then this inequality becomes an equality since in this case $F(\hat{p}) \overset{d}{=} Unif[0,1]$. We argue that the inequality holds in the discrete case of $\hat{p}$. Indeed, for any $\alpha \in (0,1]$, there exists $0 \le k \le n$ such that $F((k-1)/n) < \alpha \le F(k/n)$. Then

$$P(F(\hat{p}) \ge \alpha) = P(\hat{p} \ge k/n) = 1 - P(\hat{p} < k/n) = 1 - F((k-1)/n) > 1 - \alpha.$$

By Chernoff's inequality, we have

$$P(\hat{p} \le (1-\delta)p) \le e^{-\delta^2 np/2}$$

for any $0 \le \delta \le 1$. Replacing $(1-\delta)p$ by x, we have

$$F(x) \le \exp\left(-\left(1-\frac{x}{p}\right)^2 \frac{np}{2}\right)$$

for $x \le p$. Hence

$$P\left(\exp\left(-\left(1-\frac{\hat{p}}{p}\right)^2 \frac{np}{2}\right) \vee 1 \ge \alpha\right) \ge P(F(\hat{p}) \ge \alpha) = 1 - \alpha.$$

Therefore,

$$\left\{0 \le p \le 1 : \exp\left(-\left(1-\frac{\hat{p}}{p}\right)^2 \frac{np}{2}\right) \vee 1 \ge \alpha\right\}$$

is a confidence region for $\hat{p}$ with confidence level at least $1-\alpha$. Simplifying the expression above, we get, either $\hat{p} \ge p$, or

$$p - \left(2\hat{p} + \frac{2\log(1/\alpha)}{n}\right) + \frac{\hat{p}^2}{p} \le 0$$

giving

$$\begin{aligned} p &\le \frac{1}{2}\left(2\hat{p} + \frac{2\log(1/\alpha)}{n} + \sqrt{\left(2\hat{p} + \frac{2\log(1/\alpha)}{n}\right)^2 - 4\hat{p}^2}\right) \\ &= \frac{1}{2}\left(2\hat{p} + \frac{2\log(1/\alpha)}{n} + \sqrt{\frac{8\log(1/\alpha)\hat{p}}{n} + \frac{4(\log(1/\alpha))^2}{n^2}}\right) \\ &= \hat{p} + \sqrt{\frac{2\log(1/\alpha)\hat{p}}{n} + \frac{(\log(1/\alpha))^2}{n^2}} + \frac{\log(1/\alpha)}{n}. \end{aligned}$$

Hence, taking union with $\hat{p} \ge p$, we have

$$\hat{p} + \sqrt{\frac{2\log(1/\alpha)\hat{p}}{n} + \frac{(\log(1/\alpha))^2}{n^2}} + \frac{\log(1/\alpha)}{n}$$

being a $(1-\alpha)$-level upper confidence bound for p, for any finite sample n. This can be summarized as:

Theorem 1 The bound

$$\hat{p}^{con} = \hat{p} + \sqrt{\frac{2\log(1/\alpha)\hat{p}}{n} + \frac{(\log(1/\alpha))^2}{n^2}} + \frac{\log(1/\alpha)}{n}$$

is a valid $(1-\alpha)$-level upper confidence bound for p, for any finite sample n. That is, $P(p \le \hat{p}^{con}) \ge 1-\alpha$ for any n.

Note that the bound is trivially ∞ when $n=0$. When $\hat{p}=0$, the bound reduces to $2\log(1/\alpha)/n$ (and in fact we can construct even tighter bounds by using the binomial distribution of $n\hat{p}$ directly in this case). On the other hand, when $\hat{p}>0$, we can re-express using $\hat{n}=n\hat{p}$, the number of 1's among the n trials, to get

$$\hat{p}^{con}=\hat{p}+\hat{p}\sqrt{\frac{2\log(1/\alpha)}{\hat{n}}+\frac{(\log(1/\alpha))^2}{\hat{n}^2}}+\hat{p}\frac{\log(1/\alpha)}{\hat{n}}.$$

We highlight that in this case, the confidence bound has the same magnitude as $\hat{p}$, and the relative discrepancy with $\hat{p}$ shrinks as $\hat{n}$ increases in a square-root fashion.

Finally, if we check the difference between $\hat{p}^{con}$ and the naive bound $\hat{p}^{naive}$, then we will find that it is of the same order as $\hat{p}^{naive}-\hat{p}$. The following theorem presents the details of this claim. We will contrast this result with our bound presented in the next section momentarily.

Theorem 2 Assume that $0<\alpha<1/2$. Then

$$\hat{p}^{con}-\hat{p}^{naive}\geq(\sqrt{2\log(1/\alpha)}-z_{1-\alpha})\sqrt{\frac{\hat{p}}{n}}+\frac{\log(1/\alpha)}{n}.$$

Note that $\sqrt{2\log(1/\alpha)}-z_{1-\alpha}>0$ for $0<\alpha<1/2$.

Proof. We have that

$$\begin{aligned}\hat{p}^{con}-\hat{p}^{naive}&=\sqrt{\frac{2\log(1/\alpha)\hat{p}}{n}+\frac{(\log(1/\alpha))^2}{n^2}}+\frac{\log(1/\alpha)}{n}-z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\\&\geq(\sqrt{2\log(1/\alpha)}-z_{1-\alpha})\sqrt{\frac{\hat{p}}{n}}+\frac{\log(1/\alpha)}{n}.\end{aligned}$$

One may note that as long as $\hat{n}\geq 1$, that is, we have at least one positive observation, then $\sqrt{\hat{p}/n}=\sqrt{\hat{n}}/n\geq 1/n$. Provided that $\sqrt{2\log(1/\alpha)}-z_{1-\alpha}>0$, $\hat{p}^{con}-\hat{p}^{naive}$ is of no higher order than $\hat{p}^{naive}-\hat{p}$. □

4 CONFIDENCE BOUND VIA INVERTING THE BERRY-ESSEEN THEOREM

We develop another upper confidence bound for p by inverting the Berry-Esseen theorem. Here, we assume that p is known to satisfy $p<\frac{1}{2}$ a priori (which is reasonable if we consider rare event). We also assume that $\alpha<\frac{1}{2}$ (typically, we would set α to be a small number like 0.05). In this paper, we use the standard version of the Berry-Esseen theorem, and a potential future investigation is to consider a B-E bound for the studentized statistic (Wang and Jing 1999; Wang and Hall 2009).

By the Berry-Esseen theorem, we have that

$$\left|P\left((\hat{p}-p)\sqrt{\frac{n}{p(1-p)}}\leq x\right)-\Phi(x)\right|\leq\frac{C(1-2p+2p^2)}{\sqrt{np(1-p)}}\leq\frac{C}{\sqrt{np(1-p)}}$$

where C is a universal constant. We use $F()$ to denote the distribution function of $\hat{p}$ and replace x by $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$, and then we get that

$$\left|F(\hat{p})-\Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right)\right|\leq\frac{C}{\sqrt{np(1-p)}}.$$

Then we have that

$$P\left(\Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right)+\frac{C}{\sqrt{np(1-p)}}\geq\alpha\right)\geq P(F(\hat{p})\geq\alpha)\geq 1-\alpha.$$

Hence,

$$\left\{0 < p < 1 : \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right) + \frac{C}{\sqrt{np(1-p)}} \geq \alpha\right\}$$

is a valid $(1-\alpha)$-level confidence region for p. Since we have assumed that $p < 1/2$, the above confidence region can be further shrunk. To summarize, we have the following theorem:

Theorem 3 Assume that $p < 1/2$. Then the set

$$\left\{0 < p < 1/2 : \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right) + \frac{C}{\sqrt{np(1-p)}} \geq \alpha\right\} \tag{5}$$

is a valid $(1-\alpha)$-level confidence region for p, for any finite sample n.

Now we may develop a more explicit $(1-\alpha)$-level upper confidence bound starting from (5). In fact, for any $0 \leq \lambda \leq 1 - \frac{2C}{\sqrt{n}\alpha}$, we have that

$$0 < p < 1/2, \frac{C}{\sqrt{np(1-p)}} \geq (1-\lambda)\alpha \Rightarrow 0 < p \leq \frac{1-\sqrt{1-\frac{4C^2}{n(1-\lambda)^2\alpha^2}}}{2}$$

and

$$0 < p < 1/2, \Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right) \geq \lambda\alpha \Rightarrow 0 < p \leq \frac{1+\frac{2n\hat{p}}{z_{\lambda\alpha}^2}+\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}}{2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)}.$$

Therefore, we get that

$$\left(\frac{1-\sqrt{1-\frac{4C^2}{n(1-\lambda)^2\alpha^2}}}{2}\right) \vee \left(\frac{1+\frac{2n\hat{p}}{z_{\lambda\alpha}^2}+\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}}{2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)}\right)$$

is a $(1-\alpha)$-level upper confidence bound. For simplicity, we denote the two parts as U_1 and U_2 respectively. One may note that λ is not necessarily deterministic. Instead, it can be dependent on the data as long as it stays within the interval $[0, 1-\frac{2C}{\sqrt{n}\alpha}]$. In fact, we may choose λ carefully such that $U_1 \leq U_2$ is guaranteed for sufficiently large n. Specifically, the following theorem proposes another valid upper confidence bound.

Theorem 4 Assume that $p < 1/2$ and $\alpha < 1/2$. Let

$$\lambda = 1 - \frac{\tilde{C}}{\sqrt{n}\alpha}$$

where

$$\tilde{C} = \left(\frac{C}{\sqrt{\hat{p}(1-\hat{p})}}\right) \wedge \left(u\sqrt{n}\alpha\right).$$

Here, $u < 1$ is any constant such that $\frac{C^2}{u^2\alpha^2} < z_{(1-u)\alpha}^2$. In the case that $\hat{p} = 0$ or 1, naturally we set $\tilde{C} = u\sqrt{n}\alpha$. Then there exists $N_0 \in \mathbb{N}$, which does not depend on p and $\hat{p}$, such that for any $n > N_0$,

$$\hat{p}^{BE} = \frac{1+\frac{2n\hat{p}}{z_{\lambda\alpha}^2}+\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}}{2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)}$$

is a valid $(1-\alpha)$-level upper confidence bound for p. In particular, N_0 can be chosen as

$$\left(\frac{2C}{u\alpha}\right)^2 \vee \frac{3z_{(1-u)\alpha}^2 C^2/(u^2\alpha^2)}{z_{(1-u)\alpha}^2 - C^2/(u^2\alpha^2)}.$$

Proof. Following our derivations, it suffices to show that there exists N_0 such that for any $n > N_0$, we have that $0 \le \lambda \le 1 - \frac{2C}{\sqrt{n}\alpha}$ and $U_1 \le U_2$. Obviously, $\frac{\tilde{C}}{\sqrt{n}\alpha} \le u < 1$, so $\lambda > 0$. On the other side, $\lambda \le 1 - \frac{2C}{\sqrt{n}\alpha}$ if and only if $\tilde{C} \ge 2C$, which holds as long as $n \ge \left(\frac{2C}{u\alpha}\right)^2$. Now we prove that $U_1 \le U_2$ for sufficiently large n. Indeed, if $\tilde{C} = C/\sqrt{\hat{p}(1-\hat{p})}$, then $U_1 = \hat{p} \wedge (1-\hat{p}) \le \hat{p}$ and we know that $U_2 \ge \hat{p}$, so $U_1 \le U_2$. In the other case that $\tilde{C} = u\sqrt{n}\alpha$, we have that

$$U_1 = \frac{1 - \sqrt{1 - \frac{4C^2}{nu^2\alpha^2}}}{2} = \frac{C^2}{u^2\alpha^2}\frac{1}{n} + o\left(\frac{1}{n}\right),$$

$$U_2 \ge \frac{1}{1 + \frac{n}{z_{(1-u)\alpha}^2}} = \frac{z_{(1-u)\alpha}^2}{n} + o\left(\frac{1}{n}\right).$$

Since u is chosen such that $\frac{C^2}{u^2\alpha^2} < z_{(1-u)\alpha}^2$, we also have $U_1 \le U_2$ for sufficiently large n. Note that as $u \uparrow 1$, $\frac{C^2}{u^2\alpha^2} \to \frac{C^2}{\alpha^2}$ while $z_{(1-u)\alpha}^2 \to \infty$, and thus such u exists. Therefore, we can find such N_0 which does not depend on the value of p or $\hat{p}$. □

Next we will show that $\hat{p}^{BE} - \hat{p}^{naive}$ is bounded by order $1/n$. In other words, though $\hat{p}^{naive}$ has undesirable coverage probability in the rare-event setting, it is not "too far" from a valid confidence bound. The following theorem states this result.

Theorem 5 Assume that $p < 1/2$ and $\alpha < 1/2$. $\hat{p}^{BE}$ is as defined in Theorem 4. Then there is a constant C_0 which only depends on α and u such that

$$|\hat{p}^{BE} - \hat{p}^{naive}| \le C_0/n.$$

Proof. We have that

$$\hat{p}^{BE} - \hat{p}^{naive} = \frac{1 - 2\hat{p} + \sqrt{1 + \frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}} - 2\left(1 + \frac{n}{z_{\lambda\alpha}^2}\right) z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{2\left(1 + \frac{n}{z_{\lambda\alpha}^2}\right)}.$$

We first deal with

$$\sqrt{1 + \frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}} - 2\left(1 + \frac{n}{z_{\lambda\alpha}^2}\right) z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \frac{\left(1 - \frac{8z_{1-\alpha}^2\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}\right) - \frac{4z_{1-\alpha}^2\hat{p}(1-\hat{p})}{n} + \frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}\left(1 - \frac{z_{1-\alpha}^2}{z_{\lambda\alpha}^2}\right)}{\sqrt{1 + \frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}} + 2\left(1 + \frac{n}{z_{\lambda\alpha}^2}\right) z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}.$$

The denominator satisfies that

$$\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}+2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\geq\left(\frac{2\sqrt{n\hat{p}(1-\hat{p})}(z_{1-\lambda\alpha}+z_{1-\alpha})}{z_{\lambda\alpha}^2}\right)\vee 1.$$

Note that $(z_{1-\lambda\alpha}+z_{1-\alpha})/z_{\lambda\alpha}^2$ increases with the value of λ. Since $\lambda\geq 1-u>0$, we can find a constant C_1 such that

$$\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}+2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\geq\left(C_1\sqrt{n\hat{p}(1-\hat{p})}\right)\vee 1.$$

Then we deal with the numerator. We know that $z_\alpha=\Phi^{-1}(\alpha)$ and $z_{\lambda\alpha}=\Phi^{-1}(\lambda\alpha)$. By Taylor expansion, we have that

$$\frac{1}{z_{\lambda\alpha}^2}=\frac{1}{z_\alpha^2}-\frac{2\sqrt{2\pi}}{z_\alpha^3}e^{\frac{z_\alpha^2}{2}}(\lambda-1)\alpha+r(\lambda).$$

Here, $r(\lambda)$ is continuous in λ and $r(\lambda)/(1-\lambda)\to 0$ as $\lambda\uparrow 1$. We also know that $1-\lambda\leq u$, and thus $|r(\lambda)/(1-\lambda)|=|(\sqrt{n}\alpha r(\lambda))/\tilde{C}|$ is bounded by a constant. Hence $|\sqrt{n\hat{p}(1-\hat{p})}r(\lambda)|$ is bounded by a constant. We have that

$$1-\frac{z_{1-\alpha}^2}{z_{\lambda\alpha}^2}=\frac{2\sqrt{2\pi}}{z_{1-\alpha}}e^{\frac{z_{1-\alpha}^2}{2}}\frac{\tilde{C}}{\sqrt{n}}-z_{1-\alpha}^2 r(\lambda).$$

Thus, the numerator satisfies that

$$\begin{aligned}&\left|\left(1-\frac{8z_{1-\alpha}^2\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}\right)-\frac{4z_{1-\alpha}^2\hat{p}(1-\hat{p})}{n}+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}\left(1-\frac{z_{1-\alpha}^2}{z_{\lambda\alpha}^2}\right)\right|\\ \leq&1+8\hat{p}(1-\hat{p})+\frac{4z_{1-\alpha}^2\hat{p}(1-\hat{p})}{n}+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}\left(\frac{2\sqrt{2\pi}}{z_{1-\alpha}}e^{\frac{z_{1-\alpha}^2}{2}}\frac{\tilde{C}}{\sqrt{n}}-z_{1-\alpha}^2 r(\lambda)\right).\end{aligned}$$

Clearly, the first three terms divided by the denominator are bounded by some constants. Now we consider the fourth term. Since $|\sqrt{n\hat{p}(1-\hat{p})}r(\lambda)|$ is bounded, we can also get that the fourth term divided by the denominator is bounded by some universal constant.

Therefore, combining the above results, we know that

$$\left|1-2\hat{p}+\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}-2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right|\leq C_2$$

where C_2 is a positive constant. We also have that

$$2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)\geq\frac{2n}{z_{(1-u)\alpha}^2}.$$

Hence the error term satisfies that

$$\left|\frac{1-2\hat{p}+\sqrt{1+\frac{4n\hat{p}(1-\hat{p})}{z_{\lambda\alpha}^2}}-2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{2\left(1+\frac{n}{z_{\lambda\alpha}^2}\right)}\right|\leq\frac{C_0}{n}$$

for some constant C_0. From the above derivations, we find that C_0 only depends on α and the choice of u. □

Note that the bound in Theorem 5 can be rephrased as $|\hat{p}^{BE} - \hat{p}^{naive}| \leq C_0 \hat{p}/\hat{n}$. In other words, $\hat{p}^{BE}$ differ from $\hat{p}^{naive}$ by a magnitude that is of higher-order than the half-width of $\hat{p}^{naive}$ (i.e., $\hat{p}^{naive} - \hat{p}$) in terms of $\hat{n}$, while all quantities scale with $\hat{p}$ in a similar manner. Compared to Theorem 2, we see in Theorem 5 that $\hat{p}^{BE}$ is substantially tighter than $\hat{p}^{con}$ when $\hat{n}$ increases, although due to the implicit constant C_0 it may not be the case for small $\hat{n}$.

Compared to our confidence region proposed in (5) which takes care of the Berry-Esseen error term, Wilson's confidence bound is obtained by directly inverting

$$\Phi\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}\right) \geq \alpha.$$

That is,

$$\hat{p}^{Wilson} = \frac{1 + \frac{2n\hat{p}}{z_{1-\alpha}^2} + \sqrt{1 + \frac{4n\hat{p}(1-\hat{p})}{z_{1-\alpha}^2}}}{2\left(1 + \frac{n}{z_{1-\alpha}^2}\right)}$$

works as an approximate $(1-\alpha)$-level upper confidence bound for p. Note that $\hat{p}^{Wilson}$ has a similar form to $\hat{p}^{BE}$. Indeed, we may derive similar results to Theorem 5 (the proof is simple and thus skipped):

Theorem 6 Assume that $\alpha < 1/2$. Then

$$|\hat{p}^{Wilson} - \hat{p}^{naive}| \leq \frac{z_{1-\alpha}^2}{n} + \frac{z_{1-\alpha}^3}{2n^{3/2}}.$$

In practice, the Wilson's confidence bound has a satisfactory performance, in the sense that it is relatively tight while the coverage probability is usually close to the nominal confidence level. The Berry-Esseen confidence region in (5) is conservative by comparison, yet guaranteed to achieve the correct confidence level. Nevertheless, Theorem 5 and 6 show that for both of them, the difference from $\hat{p}^{naive}$ is $O(1/n)$.

Finally, we note that, by deriving results for an analogous lower bound similar to those in Theorems 1 and 2 (or Theorems 4 and 5), we can conclude that $P(|\hat{p}-p| \leq \varepsilon\hat{p}|\hat{n} \geq c) \geq 1-\alpha$ for some small ε depending on the constant c, which is chosen large enough independently. This in turn concludes that $P(|\hat{p}-p| \leq (\varepsilon/(1-\varepsilon))p|\hat{n} \geq c) \geq 1-\alpha$. In other words, given we observe that $\hat{n}$ is large enough, $\hat{p}$ can be viewed as a conditionally relatively efficient estimator. The caution, however, is that this is under the precondition that we can observe enough positive outcomes.

5 CONFIDENCE BOUND UNDER TARGETED STOPPING

Now we consider experiments where we keep sampling until we get n_0 successes. Our goal is to construct a $(1-\alpha)$-level upper confidence bound for p using a similar method to Section 4 and then carry out the analysis on the difference from the naive bound. Like in Section 4, we assume here that $p < 1/2$ and $\alpha < 1/2$. The sample size N is a random variable. More specifically, $N = N_1 + \cdots + N_{n_0}$ where $N_1, \cdots, N_{n_0} \overset{i.i.d.}{\sim} Geometric(p)$. Note that $N \geq n_0$.

We may again apply Berry-Esseen Theorem. Let $X_i = \frac{1}{p} - N_i$. Then $E(X_i) = 0$. Set $\sigma_N^2 = E(X_i^2) = \frac{1-p}{p^2}$ and $\rho_N = E|X_i|^3$. By the theorem, we get that

$$\left|P\left(N \geq \frac{n_0}{p} - x\sqrt{\frac{n_0(1-p)}{p^2}}\right) - \Phi(x)\right| \leq \frac{C\rho_N}{\sigma_N^3\sqrt{n_0}}. \tag{6}$$

We need to deal with ρ_N first. In fact, we know that

$$p^3\rho_N = p^3 E\left|\frac{1}{p} - N_i\right|^3 = E|1-pN_i|^3 \leq 1 + 3pE(N_i) + 3p^2E(N_i^2) + p^3E(N_i^3).$$

Since $N_i \sim Geometric(p)$, we know that

$$E(N_i) = \frac{1}{p}, E(N_i^2) = \frac{2-p}{p^2}, E(N_i^3) = \frac{p^2-6p+6}{p^3},$$

and thus

$$p^3\rho_N \leq p^2 - 6p + 6 + 3(2-p) + 3 + 1 = p^2 - 12p + 16 \leq 16.$$

Hence,

$$\frac{C\rho_N}{\sigma_N^3\sqrt{n_0}} = \frac{Cp^3\rho_N}{(1-p)^{\frac{3}{2}}\sqrt{n_0}} \leq \frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}}$$

where $\tilde{C} = 16C$ is an absolute constant.

We use $F_N()$ to denote the CDF of N. Also, we set $x = \sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)$ in (6). Then we get that

$$\left|1 - F_{N-}(N) - \Phi\left(\sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)\right)\right| \leq \frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}},$$

which gives that

$$F_{N-}(N) \geq 1 - \Phi\left(\sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)\right) - \frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}}. \tag{7}$$

Now we show that $P(F_{N-}(N) \leq 1-\alpha) \geq 1-\alpha$. We know that $F_{N-}(n_0) = 0$ and $F_{N-}(n) \to 1$ as $n \to \infty$. Then we can find $\tilde{n} \geq n_0$ such that $F_{N-}(\tilde{n}) \leq 1-\alpha < F_{N-}(\tilde{n}+1)$. Thus we have that

$$P(F_{N-}(N) \leq 1-\alpha) = P(N \leq \tilde{n}) = F_{N-}(\tilde{n}+1) > 1-\alpha.$$

Also, from (7), we get that

$$F_{N-}(N) \leq 1-\alpha \Rightarrow \Phi\left(\sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)\right) + \frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}} \geq \alpha.$$

Finally, we develop a valid confidence region under this particular setting, which is similar to the one in Section 4:

Theorem 7 Suppose that we keep sampling from *Bernoulli*(p) until we get n_0 successes and the sample size is denoted by N. Assume that $p < 1/2$ and $\alpha < 1/2$. Then

$$\left\{0 < p < 1/2 : \Phi\left(\sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)\right) + \frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}} \geq \alpha\right\}$$

is a valid $(1-\alpha)$-level confidence region for p. Here, $\tilde{C}$ is a universal constant. In particular, one may pick $\tilde{C} = 16C$ where C is the constant in the Berry-Esseen theorem.

For $0 < \lambda < 1$, we have that

$$\Phi\left(\sqrt{\frac{p^2}{n_0(1-p)}}\left(\frac{n_0}{p} - N\right)\right) \geq \lambda\alpha \Leftrightarrow p \leq \frac{(2Nn_0 - z_{\lambda\alpha}^2 n_0) + \sqrt{4z_{\lambda\alpha}^2 Nn_0(N-n_0) + z_{\lambda\alpha}^4 n_0^2}}{2N^2},$$

$$\frac{\tilde{C}}{(1-p)^{\frac{3}{2}}\sqrt{n_0}} \geq (1-\lambda)\alpha \Leftrightarrow p \geq 1 - \left(\frac{\tilde{C}}{(1-\lambda)\alpha\sqrt{n_0}}\right)^{\frac{2}{3}}.$$

Note that if we fix $0<\lambda<1$, then $\left(\frac{\tilde{C}}{(1-\lambda)\alpha\sqrt{n_0}}\right)^{\frac{2}{3}} \to 0$ as $n_0 \to \infty$. Therefore, there exists $\tilde{N}_0 \in \mathbb{N}$ such that $1-\left(\frac{\tilde{C}}{(1-\lambda)\alpha\sqrt{n_0}}\right)^{\frac{2}{3}} \geq \frac{1}{2}$ for any $n_0 > \tilde{N}_0$. In this case,

$$\hat{p}_{n_0}^{BE} = \left((2Nn_0 - z_{\lambda\alpha}^2 n_0) + \sqrt{4z_{\lambda\alpha}^2 N n_0 (N-n_0) + z_{\lambda\alpha}^4 n_0^2}\right)/(2N^2) \tag{8}$$

is a valid $(1-\alpha)$-level confidence upper bound for p.

Naturally, we still denote $\hat{p}=\frac{n_0}{N}$. Moreover, we consider the naive bound $\hat{p}^{naive} = \hat{p} + z_{1-\alpha}\sqrt{\hat{p}(1-\hat{p})/n}$. We may also develop an upper bound for the difference between $\hat{p}_{n_0}^{BE}$ and $\hat{p}^{naive}$:

Theorem 8 Assume that $p<1/2$ and $\alpha<1/2$. $\hat{p}_{n_0}^{BE}$ is as defined in (8). Then

$$|\hat{p}_{n_0}^{BE} - \hat{p}^{naive}| \leq z_{1-\lambda\alpha}\sqrt{n_0}/N.$$

We skip the proof of Theorem 8, which can be handled by similar techniques as for Theorem 5.

Therefore, under the setting of target stopping instead of fixing the sample size, we may also develop a valid confidence region for p by applying the Berry-Esseen Theorem. However, $\hat{p}^{naive} - \hat{p}$ is approximately $z_{1-\alpha}\sqrt{n_0}/N$, where N is a random sample size, so we cannot conclude that $\hat{p}_{n_0}^{BE} - \hat{p}^{naive}$ elicits a higher-order magnitude in $\hat{n}$ as in Theorem 5, which appears to require more delicate analysis.

6 NUMERICAL EXPERIMENTS

To visualize the differences among the confidence bounds, we perform some numerical experiments. The true value is chosen as $p = 1e-6$. For $n = 1/p, 5/p, 10/p, 20/p, 30/p, 50/p$, we respectively do 1000 simulations and calculate the confidence bounds with $\alpha = 0.05$. Table 1 presents the mean values and the coverage probabilities of the four confidence bounds covered in this paper.

Though the CLT-based bound is closest to the true value, it is not reliable especially when the sample size is relatively small. The Chernoff bound and the B-E bound are both very conservative, but as analyzed above, the error of the B-E bound decreases in higher order as the sample size grows. Wilson's bound has a satisfactory performance overall, though theoretically it does not always reach the nominal confidence level. Therefore, the Chernoff bound and the B-E bound proposed in this paper are useful in situations where absolute safety is critical, since they are rigorously justified, but they are conservative. In other general applications, Wilson's bound would be an ideal choice.

Table 1: Results of the numerical experiments. "CLT", "Cher", "BE", "Wil" respectively stand for the CLT-based bound, the Chernoff bound, the B-E bound and Wilson's bound. "mean" stands for the mean value of the confidence bounds and "prob" stands for the coverage probability.

np	CLT-mean	Cher-mean	BE-mean	Wil-mean	CLT-prob	Cher-prob	BE-prob	Wil-prob
1	2.21E-06	7.74E-06	9.02E-05	4.35E-06	0.615	1	1	1
5	1.73E-06	2.85E-06	1.80E-05	2.06E-06	0.869	1	1	0.957
10	1.51E-06	2.12E-06	9.02E-06	1.67E-06	0.923	0.995	1	0.966
20	1.36E-06	1.70E-06	4.51E-06	1.43E-06	0.935	0.998	1	0.961
30	1.29E-06	1.55E-06	3.01E-06	1.34E-06	0.921	0.995	1	0.972
50	1.23E-06	1.41E-06	1.81E-06	1.26E-06	0.935	0.998	1	0.948

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710 and IIS-1849280.

REFERENCES

Agresti, A., and B. A. Coull. 1998. "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions". *The American Statistician* 52(2):119–126.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York, New York: Springer Science & Business Media.

Blanchet, J., and H. Lam. 2012. "State-Dependent Importance Sampling for Rare-Event Simulation: An Overview and Recent Advances". *Surveys in Operations Research and Management Science* 17(1):38–59.

Bucklew, J. 2004. *Introduction to Rare Event Simulation*. New York, New York: Springer Science & Business Media.

Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. *Modelling Extremal Events: for Insurance and Finance*. Germany: Springer Science & Business Media.

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering*. New York, New York: Springer Science & Business Media.

Glasserman, P., and J. Li. 2005. "Importance Sampling for Portfolio Credit Risk". *Management science* 51(11):1643–1656.

Gumbel, E. J. 2012. *Statistics of Extremes*. North Chelmsford, Massachusetts: Courier Corporation.

Heidelberger, P. 1995. "Fast Simulation of Rare Events in Queueing and Reliability Models". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 5(1):43–85.

Huang, Z., H. Lam, D. J. LeBlanc, and D. Zhao. 2017. "Accelerated Evaluation of Automated Vehicles using Piecewise Mixture Models". *IEEE Transactions on Intelligent Transportation Systems* 19(9):2845–2855.

Juneja, S., and P. Shahabuddin. 2006. "Rare-Event Simulation Techniques: an Introduction and Recent Advances". *Handbooks in operations research and management science* 13:291–350.

McNeil, A. J., R. Frey, and P. Embrechts. 2015. *Quantitative Risk Management: Concepts, Techniques and Tools-Revised Edition*. Princeton, New Jersey: Princeton university press.

Nicola, V. F., M. K. Nakayama, P. Heidelberger, and A. Goyal. 1993. "Fast Simulation of Highly Dependable Systems with General Failure and Repair Processes". *IEEE Transactions on Computers* 42(12):1440–1452.

O'Kelly, M., A. Sinha, H. Namkoong, R. Tedrake, and J. C. Duchi. 2018. "Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation". In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, and N. Cesa-Bianchi, 9849–9860. Red Hook, New York: Curran Associates, Inc.

Rubinstein, R. Y., and D. P. Kroese. 2016. *Simulation and the Monte Carlo Method*. Hoboken, New Jersey: John Wiley & Sons.

Shao, Q.-M., and Q. Wang. 2013. "Self-Normalized Limit Theorems: A Survey". *Probab. Surveys* 10:69–93.

Wang, Q., and P. Hall. 2009. "Relative Errors in Central Limit Theorems for Student's t Statistic, with Applications". *Statistica Sinica*:343–354.

Wang, Q., and B.-Y. Jing. 1999. "An Exponential Nonuniform Berry-Esseen Bound for Self-Normalized Sums". *The Annals of Probability* 27(4):2068–2088.

Webb, S., T. Rainforth, Y. W. Teh, and M. P. Kumar. 2018. "A Statistical Approach to Assessing Neural Network Robustness". *arXiv preprint arXiv:1811.07209*.

Weng, T.-W., H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel. 2018. "Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach". *arXiv preprint arXiv:1801.10578*.

Zhao, D., X. Huang, H. Peng, H. Lam, and D. J. LeBlanc. 2017. "Accelerated Evaluation of Automated Vehicles in Car-Following Maneuvers". *IEEE Transactions on Intelligent Transportation Systems* 19(3):733–744.

Zhao, D., H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan. 2016. "Accelerated Evaluation of Automated Vehicles Safety in Lane-Change Scenarios Based on Importance Sampling techniques". *IEEE transactions on intelligent transportation systems* 18(3):595–607.

AUTHOR BIOGRAPHIES

YUANLU BAI is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. She received an M.S. in Operations Research from Columbia University, and a B.S. in Statistics as well as a B. Ec in Economics from Peking University. Her research interest lies in stochastic simulation, such as uncertainty quantification and extreme risk analysis. Her email address is yb2436@columbia.edu.

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics from Harvard University in 2011, and was on the faculty of Boston University and the University of Michigan before joining Columbia in 2017. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is henry.lam@columbia.edu.