A New Confidence Interval for the Mean of a Bounded Random Variable

Erik Learned-Miller and Philip S. Thomas
College of Information and Computer Sciences
University of Massachusetts Amherst
{elm,pthomas}@cs.umass.edu

Abstract

We present a new method for constructing a confidence interval for the mean of a bounded random variable from samples of the random variable. We conjecture that the confidence interval has guaranteed coverage, i.e., that it contains the mean with high probability for all distributions on a bounded interval, for all samples sizes, and for all confidence levels. This new method provides confidence intervals that are competitive with those produced using Student's t-statistic, but does not rely on normality assumptions. In particular, its only requirement is that the distribution be bounded on a known finite interval.

1 Introduction

Consider one of the fundamental problems in statistics: how to use n samples of a real-valued random variable to obtain a confidence interval on its mean. Methods for constructing such confidence intervals are used across all branches of science. In the natural and social sciences, the confidence interval based on Student's t-statistic (Student, 1908) is one of the standard tools for quantifying uncertainty about the results of an empirical study. In theoretical work, concentration inequalities like Hoeffding's inequality (Hoeffding, 1963) are often used to analyze properties of algorithms in machine learning, data science, and other areas. Providing methods for obtaining tighter confidence intervals from fewer samples is critical to scientific advancement, enabling stronger conclusions to be drawn from the same experimental data.

In particular, there is a practical need today for confidence intervals that hold for small sample sizes. Since the confidence interval produced using Student's t-statistic, which we refer to hereafter as the *Student-t interval*, relies on the (near) normality of the sample mean, it is recommended that sample sizes be at least 30 for it to be used, unless there is a specific reason to believe that the population distribution is approximately normal. While other confidence intervals that hold for small sample sizes exist (such as Anderson's (1969)),

they produce intervals that are so wide as to be of little use in practice. This leaves the practitioner with the following choices:

- use methods, such as bootstrap methods, with no performance guarantees;
- use methods with unrealistic assumptions, such as the Student-t interval;
- use valid but weak methods such as Hoeffding or Anderson's inequalities that provide little information about the mean;
- abandon the idea of obtaining useful confidence intervals from the data.

In this paper, we introduce a new confidence interval for bounded distributions that is much tighter than other confidence intervals that come with guarantees. We conjecture that it holds for all bounded distributions, all samples sizes, and all confidence levels. We suggest that for many applications, this is the first practical confidence interval for sample sizes less than 30. We now offer a formal statement of the problem we are addressing.

1.1 Problem Statement

Let $X_1, \ldots, X_n$ be n independent and identically distributed real-valued random variables. Let each X_i take values in the interval $[0, 1]$ and have expected value μ. For now we focus on a high-confidence *upper* bound, i.e., we desire a function m_α such that, for all distributions of X_i, all sample sizes $n \geq 1$, and all confidence levels $1 - \alpha \in [0, 1]$:

$$\Pr\left(m_\alpha(X_1, \ldots, X_n) \geq \mu\right) \geq 1 - \alpha. \tag{1}$$

That is, with probability at least $1-\alpha$, $m_\alpha(X_1, \ldots, X_n)$ should be greater than or equal to the mean. Critically, in this statement the random quantity is the high-confidence upper bound, not the mean. Any definition of m_α that satisfies (1) for all bounded distributions, samples sizes $n \geq 1$, and $1 - \alpha \in [0, 1]$, is said to have *guaranteed coverage*.

In this paper we present a new method for constructing confidence intervals on the mean: a new m_α. We conjecture that our function m_α satisfies (1), i.e., that it has guaranteed coverage. If our conjecture holds, this is the first confidence interval with tightness comparable to the Student-t interval, but with guaranteed coverage in this setting. We prove in Section 7 that it *dominates* several other known confidence intervals with guaranteed coverage. That is, for every possible sample $(x_1, x_2, ..., x_n)$, it produces a confidence interval with width less than or equal to these previous methods, often with a much smaller width. This makes our confidence interval suitable for small sample sizes where other methods are not practical.

After defining m_α in the next section, we present the following results:

- a proof of (1) for a class of distributions that includes Bernoulli distributions, for all samples sizes n, and for all confidence levels $1 - \alpha$;

- the sketch of a proof that our intervals are always at least as tight as those provided by Anderson (1969), which, in turn, are strictly tighter than those of Hoeffding (1963);
- results of extensive simulations on a wide variety of distributions that are consistent with (1) for many sample sizes and confidence intervals;
- empirical comparisons (through Monte Carlo simulations) with previous methods, demonstrating that the confidence intervals produced by m_α are consistently tighter than or as tight as the intervals produced by existing methods.

2 A New Confidence Interval for the Mean

In this section we present our new confidence interval. We also present our conjecture that it holds for all distributions bounded on $[0, 1]$, for all sample sizes, and for all confidence levels.

Let $\mathbf{X} \stackrel{\text{def}}{=} (X_1, X_2, \ldots, X_n)$. Let $\mathbf{Z} \stackrel{\text{def}}{=} (Z_1, Z_2, \ldots, Z_n)$ be the *order statistics* of $\mathbf{X}$, i.e., $\mathbf{Z}$ is a vector containing the sorted values of $\mathbf{X}$ such that $Z_1 \leq Z_2 \leq \cdots \leq Z_n$. Let $\mathbf{z} \stackrel{\text{def}}{=} (z_1, z_2, ..., z_n)$ denote a particular sample of $\mathbf{Z}$ and $\mathbf{x} \stackrel{\text{def}}{=} (x_1, x_2, \ldots, x_n)$ a sample of $\mathbf{X}$. For notational convenience, we alternate between viewing m_α as a function of $\mathbf{z}$ or $\mathbf{x}$. So, when we write $m_\alpha(\mathbf{z})$ subsequently, this corresponds to a definition of $m_\alpha(\mathbf{x})$ where $\mathbf{z}$ are the order statistics of $\mathbf{x}$.

Let $\mathbf{U}$ be the order statistics of a sample of size n from the continuous uniform distribution on $[0, 1]$, with $\mathbf{u} \stackrel{\text{def}}{=} (u_1, u_2, \ldots, u_n)$ being a particular sample of $\mathbf{U}$. Since $\mathbf{u}$ are order statistics, $0 \leq u_1 \leq u_2 \leq ... \leq u_n \leq 1$. We define a function of two ordered vectors:

$$m(\mathbf{z}, \mathbf{u}) \stackrel{\text{def}}{=} 1 - \sum_{i=1}^{n} u_i(z_{i+1} - z_i),$$

where $z_{n+1} \stackrel{\text{def}}{=} 1$. Let $Q(1 - \alpha, Y)$ be the *quantile function* of the scalar random variable Y, i.e.,

$$Q(1 - \alpha, Y) \stackrel{\text{def}}{=} \inf\{y \in \mathbb{R} : F_Y(y) \geq 1 - \alpha\}, \tag{2}$$

where $F_Y(y)$ is the cumulative distribution function (CDF) of Y.

Consider the random quantity $m(\mathbf{z}, \mathbf{U})$, which depends upon a fixed sample $\mathbf{z}$ (non-random) and also on the random variable $\mathbf{U}$. We define $m_\alpha(\mathbf{z})$ to be the $(1 - \alpha)$-quantile of $m(\mathbf{z}, \mathbf{U})$, i.e.,

$$m_\alpha(\mathbf{z}) \stackrel{\text{def}}{=} Q(1 - \alpha, m(\mathbf{z}, \mathbf{U})). \tag{3}$$

We conjecture that this definition of m_α satisfies (1):

Conjecture 1. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be* n *independent and identically distributed random variables bounded in the interval* $[0, 1]$*, each with mean* μ*. Let* $\mathbf{Z} = (Z_1, \ldots, Z_n)$ *be the order statistics of* $\mathbf{X}$*. Then for all* $\alpha \in [0, 1]$*:*

$$\Pr\left(m_\alpha(\mathbf{Z}) \geq \mu\right) \geq 1 - \alpha,$$

where m_α *is defined in* (3)*.*

Several extensions of this conjecture are apparent. First, since each X_i is bounded above by 1, this conjecture implies that $1 - m_\alpha(1 - \mathbf{z})$ is a $(1 - \alpha)$-confidence *lower* bound on μ. Second, if our main conjecture holds, we further conjecture that the assumption that the random variables are in $[0, 1]$ can be extended to $(-\infty, 1]$, or $[0, \infty)$ for the high-confidence lower bound. Furthermore, the deterministic upper bound of 1 can be loosened to only require an almost-sure upper bound of 1. Although these extensions may be important for some applications, hereafter we focus on the basic setting introduced previously.

3 Understanding $m_\alpha(\mathbf{z})$

Our high-confidence bound (for brevity, hereafter we refer to it as simply our bound) is given by the function m_α defined above. In this section we introduce the following concepts, which provide intuition for m_α:

- *ordered CDF pairs*,
- the *conservative completion* of a set of ordered CDF pairs,
- the *induced mean* of a set of ordered CDF pairs, via conservative completion.

3.1 Ordered CDF pairs

For any order statistic vector $\mathbf{z}$, each element of $\mathbf{z}$ can be paired with an element from a non-decreasing sequence of numbers, $u_1, u_2, ..., u_n$, to form n pairs:

$$(z_1, u_1), (z_2, u_2), ..., (z_n, u_n).$$

Assuming the u's are all in the interval $[0, 1]$ (as is the case if $\mathbf{u}$ is a sample of $\mathbf{U}$), these pairs can be viewed as points on a CDF F, i.e., $u_i = F(z_i)$. For this reason, we refer to these n pairs as *ordered CDF pairs*, and write $(\mathbf{z}, \mathbf{u})$ to denote such a set of ordered CDF pairs. We say that a set of ordered CDF pairs is *consistent* with a CDF F if $u_i = F(z_i)$ for all $i \in \{1, 2, \ldots, n\}$. Notice that a set of ordered CDF pairs is consistent with many (usually infinitely many) different CDFs—all non-decreasing functions on the interval $[0, 1]$ that pass through these n points (see Figure 1).

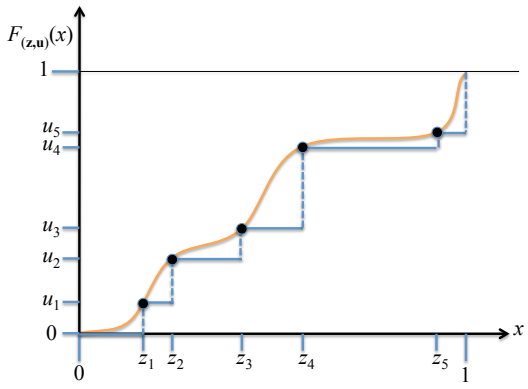

Figure 1: Given a sample $\mathbf{z}$ and a vector $\mathbf{u}$ of sorted uniform samples, the ordered CDF pairs (black points) are compatible with a large family of CDFs. Two of the CDFs compatible with these points are shown, a smooth orange one, and a stairstep blue one. The blue one represents the CDF $F_{(\mathbf{z},\mathbf{u})}(x)$, which has the greatest mean among all such CDFs, since it puts mass "as far right" as possible in a way that is still compatible with the ordered CDF pairs. We refer to this CDF as the *conservative completion* of the ordered CDF pairs $(\mathbf{z}, \mathbf{u})$.

3.2 Conservative Completion of Ordered CDF Pairs

Given a set of ordered CDF pairs, one may ask which of the (usually infinitely many) CDFs that are consistent with the ordered CDF pairs represents the distribution with the greatest mean, and is this CDF unique? This CDF *is* unique, and we refer to it as the *conservative completion* of the ordered CDF pairs. That is, the mean of the distribution characterized by the conservative completion represents an upper bound on the mean of any distribution consistent with the set of ordered CDF pairs.

The conservative completion for a set of ordered CDF pairs, $(\mathbf{z}, \mathbf{u})$, is illustrated in Figure 1. It is given by the CDF:

$$F_{(\mathbf{z},\mathbf{u})}(x) \stackrel{\text{def}}{=} \begin{cases} 0, & \text{for } x < z_1 \\ u_1, & \text{for } z_1 \leq x < z_2 \\ u_2, & \text{for } z_2 \leq x < z_3 \\ \ldots, & \ldots \\ u_n, & \text{for } z_n \leq x < 1 \\ 1, & \text{for } x \geq 1. \end{cases}$$

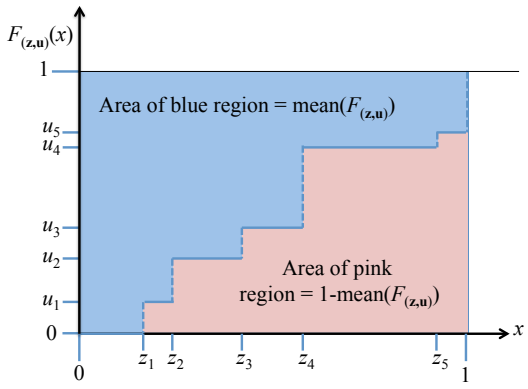

Figure 2: For CDFs defined on the interval $[0, 1]$, the mean of the distribution characterized by $F_{(\mathbf{z},\mathbf{u})}(x)$ is given by the area of the region above the CDF (blue), or one minus the area of the region below the CDF (pink).

3.3 The Induced Mean, $m(\mathbf{z}, \mathbf{u})$

We introduce $m(\mathbf{z}, \mathbf{u})$ to represent the mean of the distribution characterized by $F_{(\mathbf{z},\mathbf{u})}(x)$. This quantity is, for distributions over $[0, 1]$, equivalent to the area of the region above the CDF, as depicted in Figure 2. The geometry of this figure suggests two methods for calculating this mean. The first, derived by decomposing the blue in Figure 2 into a set of horizontal strips, is

$$m(\mathbf{z}, \mathbf{u}) = \sum_{i=1}^{n+1} z_i(u_i - u_{i-1}),$$

where $u_0 \stackrel{\text{def}}{=} 0$, $u_{n+1} \stackrel{\text{def}}{=} 1$, and $z_{n+1} \stackrel{\text{def}}{=} 1$. Another formula is given by dividing the pink region into vertical strips, and is given by

$$m(\mathbf{z}, \mathbf{u}) = 1 - \sum_{i=1}^{n} u_i(z_{i+1} - z_i).$$

Next, we consider the distribution of such means obtained by allowing $\mathbf{u}$ to vary in a particular fashion.

3.4 A Distribution of Induced Means

Recall that $\mathbf{U}$ is a random vector containing n samples from the continuous uniform distribution on $[0, 1]$, sorted such that $0 \leq U_1 \leq U_2 \leq ... \leq U_n \leq 1$. We now consider a distribution of induced means obtained by replacing the fixed $\mathbf{u}$ in $m(\mathbf{z}, \mathbf{u})$ with a random vector $\mathbf{U}$ to form a new scalar random variable $m(\mathbf{z}, \mathbf{U})$.

Recall the definition of the quantile function from (2). We define $m_\alpha(\mathbf{z})$ to be the $(1-\alpha)$-quantile of the random variable $m(\mathbf{z}, \mathbf{U})$, i.e.,

$$m_\alpha(\mathbf{z}) \stackrel{\text{def}}{=} Q(1-\alpha, m(\mathbf{z}, \mathbf{U})).$$

Thus, $m_\alpha(\mathbf{z})$ considers the set of all $\mathbf{u}$'s that can be used to form an ordered CDF pair with a particular sample $\mathbf{z}$ and chooses the $(1-\alpha)$-quantile of the resulting induced means. As we shall see, this turns out to be just "conservative enough" to provide a valid high-confidence bound for Bernoulli-like distributions, and appears to be looser for distributions that are not Bernoulli-like.

4 The Order Statistic Simplex and Feasible Set

In this section, we define the *order statistic simplex* and the notion of *feasible* and *infeasible sets* of samples of order statistics. These definitions will be used in Section 5 to prove that our bound holds for all Bernoulli distributions and also for a more general set of Bernoulli-like distributions.

4.1 A Conditional Analysis

Our general method of proof (in the next section) will rely on a conditional analysis for a specific set of distributions. In particular, we will analyze our bound specifically for

- a fixed sample size n,
- a subset of the distributions with a specific mean, μ,
- a specific confidence level, $1-\alpha$ (or equivalently, failure rate α).

If we can show that the bound, (1), holds for each tuple $(n, \mu, 1-\alpha)$, then we have a complete proof for the set of distributions under consideration.

Before proceeding with this method of proof, we define a few necessary terms.

4.2 The Order Statistic Simplex

Consider the *order statistic simplex* in n dimensions—the set of all possible order statistic vectors, $\mathbf{z}$, which forms a polytope of dimension n with $n+1$ vertices, i.e., a simplex. For distributions on $[0,1]$, we define the order statistic simplex as:

$$\mathcal{Z} \stackrel{\text{def}}{=} \{\mathbf{z} = (z_1, z_2, ..., z_n) : 0 \leq z_1 \leq z_2 ... \leq z_n \leq 1\}.$$

The order statistic simplex for $n = 2$ is depicted by the blue region in Figure 3.

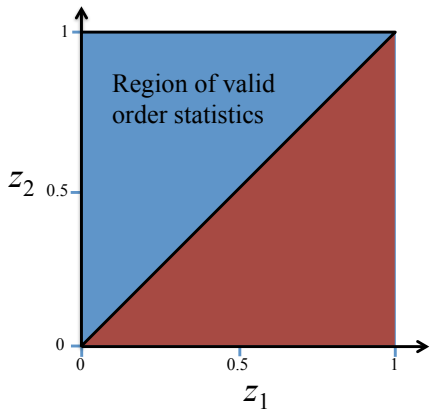

Figure 3: The order statistic simplex in $n = 2$ dimensions. The upper left region (blue) shows the region of possible (valid) order statistics for a sample of size $n = 2$. For other n, this region is defined by a polytope of n dimensions and $n+1$ vertices, i.e., a simplex. We refer to this as the *order statistic simplex* in n dimensions.

4.3 The Infeasible Set of z's

Let the sample size, n, be fixed. For distributions with a specific mean, μ, and for a specific confidence level, $1-\alpha$, let $\mathcal{Z}_\alpha^\mu \subseteq \mathcal{Z}$ denote the set of $\mathbf{z}$'s for which $m_\alpha(\mathbf{z}) \geq \mu$. We refer to $\mathcal{Z}_\alpha^\mu$ as the *feasible set*, and say that $\mathbf{z}$ *satisfies the bound* if $\mathbf{z} \in \mathcal{Z}_\alpha^\mu$. Let $\overline{\mathcal{Z}}_\alpha^\mu$ denote the complement of this set, the set of $\mathbf{z}$'s for which $m_\alpha(\mathbf{z}) < \mu$. We refer to $\overline{\mathcal{Z}}_\alpha^\mu$ as the *infeasible set* of $\mathbf{z}$'s, and say that $\mathbf{z}$ *does not satisfy the bound* if it is in $\overline{\mathcal{Z}}_\alpha^\mu$. Note that, given the mean, μ, of a distribution, the feasible and infeasible sets have no other dependency on the unknown distribution of $\mathbf{X}$.

These ideas are illustrated in Figure 4. For sample size $n = 2$, each plot shows the feasible (blue) and infeasible (green) regions for given confidence levels $1-\alpha$ and means μ. Each row shows results for the same μ and different confidence levels. Note that in some cases, such as $1-\alpha = 0.6$ and $\mu = 0.3$, the entire order statistic simplex is feasible (there are no green pixels).

5 Bernoulli and Half-Bernoulli Distributions

In this section, we present a proof of our conjecture for Bernoulli distributions and for a generalization of Bernoullis that we refer to as *half-Bernoulli* distributions. While Bernoulli distributions have point masses on both 0 and 1, half-

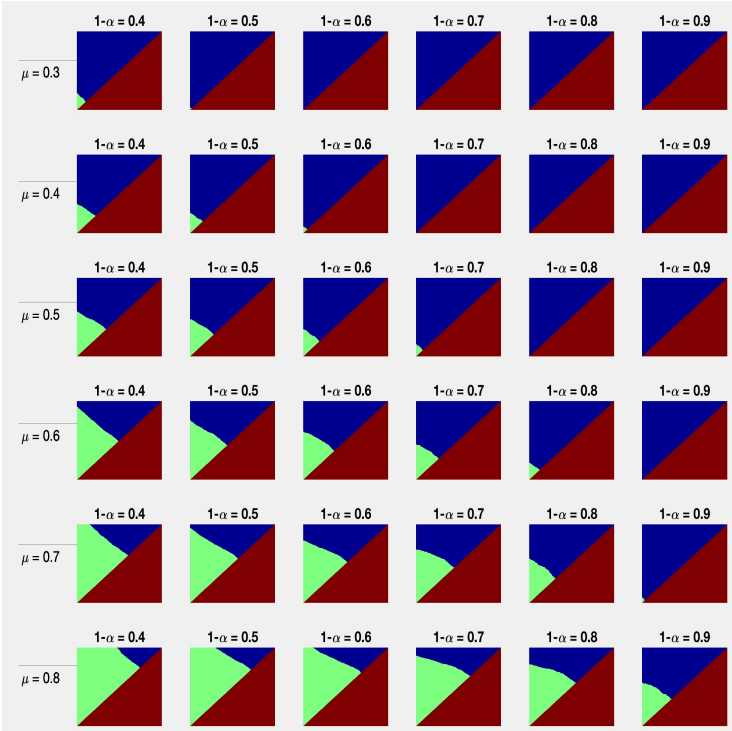

Figure 4: Feasible and infeasible regions for $n = 2$. For various values of the confidence level $1 - \alpha$ and the mean μ, we show the **feasible regions** in blue (for which the bound is greater than or equal to the mean) and the **infeasible regions** in green (for which the bound is less than the mean).

Bernoullis can have point masses at two positions: k and 1, where $0 \leq k < 1$. Thus, they are a generalization of Bernoullis that allow the lower value to be any non-negative value less than 1. Let $H_{k,\mu}$ be the half-Bernoulli distribution where the point masses are at k and 1, and the mean is μ. With k and μ specified by $H_{k,\mu}$, the probability p_k of sampling k is given by

$$p_k = \frac{1 - \mu}{1 - k}.$$

Bernoullis (and half-Bernoullis) are prime candidates for distributions for which the bound will fail (violate (1)), since it is not uncommon to have a sample of all 0's (or all k's), despite having a relatively large mean. For example, the probability of obtaining a sample of $[0, 0, 0, 0]$ from a Bernoulli distribution with parameter $p = 0.5$ is $0.5^4 = 0.0625$. Since the probability of getting this sample is greater than 0.05, the bound must produce a result greater than $\mu = 0.5$ for this sample at the confidence level $1 - \alpha = 0.95$. As we demonstrate below, the bound holds for all Bernoulli and half-Bernoulli distributions.

5.1 Finding Worst Case Half-Bernoullis for $(n, \mu, 1 - \alpha)$

Our approach will be to find "worst case" distributions among the set of half-Bernoulli distributions. By worst case, we mean that the probability of drawing

a sample $\mathbf{z}$ for which the bound fails is as high as possible. We will show that for the worst case half-Bernoulli distributions (there can be more than one of these for each tuple $(n, \mu, 1-\alpha)$, the probability of drawing a sample $\mathbf{z} \in \overline{\mathcal{Z}}_\alpha^\mu$ is no more than α. Since the bound holds for the worst case half-Bernoullis, it must hold for all half-Bernoullis.

5.1.1 Enumerating possible z's

Consider a half-Bernoulli distribution with probability masses at k and 1, and with mean μ. For a given n, there are $n+1$ possible order statistics, $\mathbf{z}$:

$$\begin{gathered}
[k, k, \ldots, k, k] \\
[k, k, \ldots, k, 1] \\
[k, k, \ldots, 1, 1] \\
\vdots \\
[k, 1, \ldots, 1, 1] \\
[1, 1, \ldots, 1, 1].
\end{gathered}$$

Let $\mathbf{z}_{j,n}$ be the sample $\mathbf{z}$ with j out of n values of k, i.e., for $j \in \{0, 1, \ldots, n\}$:

$$\mathbf{z}_{j,n} \stackrel{\text{def}}{=} [\underbrace{k, \ldots, k}_{j}, \underbrace{1, \ldots, 1}_{n-j}].$$

5.1.2 Monotonicity of $m_\alpha(\mathbf{z})$

Notice that m_α is monotonic in the following sense. For two samples $\mathbf{y}$ and $\mathbf{z}$:

$$\texttt{If } \forall i,\, y_i \le z_i \texttt{ then } m_\alpha(\mathbf{y}) \le m_\alpha(\mathbf{z}). \tag{4}$$

It follows from (4) that $m_\alpha(\mathbf{z}_{i,n}) \le m_\alpha(\mathbf{z}_{j,n})$ whenever $i < j$.

5.1.3 An expression for $\Pr(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu)$

For any half-Bernoulli distribution $H_{k,\mu}$, sample size n, and confidence level $1-\alpha$, let

$$j_{\min}(H_{k,\mu}, 1-\alpha, n) = \min\left\{j \in \{0, 1, \ldots, n\} : \mathbf{z}_{j,n} \in \overline{\mathcal{Z}}_\alpha^\mu\right\},$$

where $j_{\min}(H_{k,\mu}, 1-\alpha, n) = n+1$ if $\mathbf{z}_{n,n} \in \mathcal{Z}_\alpha^\mu$.

For example, suppose $n = 5$ and $[k, k, k, 1, 1] \in \overline{\mathcal{Z}}_\alpha^\mu$ but $[k, k, 1, 1, 1] \in \mathcal{Z}_\alpha^\mu$. Then $j_{\min} = 3$, where here and in the following the arguments of $j_{\min}$ are implicit. By the monotonicity of the bound (see (4)), all of the samples with $\texttt{count}(k) \ge j_{\min}$ will be in $\overline{\mathcal{Z}}_\alpha^\mu$.

For a given half-Bernoulli distribution we can now write an expression for the probability that $\mathbf{Z}$ will not satisfy the bound:

$$\begin{aligned}\Pr_{H_{k,\mu}}\left(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu\right) &= \sum_{i=j_{\min}}^{n} \Pr_{H_{k,\mu}}\left(\mathbf{Z} = \mathbf{z}_{i,n}\right) \\ &= \sum_{i=j_{\min}}^{n} \text{Binomial}(i; n, p_k) \\ &= \beta_{\text{cdf}}(p_k; j_{\min}, n - j_{\min} + 1), \end{aligned} \tag{5}$$

where $\beta_{\text{cdf}}(x; a, b)$ is the CDF of a beta distribution with parameters a and b. The above derivation uses the property that each $\mathbf{z}_{i,n}$ can be viewed as a sample from a binomial distribution, and in the last step we use a well-known identity that relates the sum of binomials to the CDF of a beta distribution.

5.1.4 Simplification of $m(\mathbf{z}_{j,n})$ due to the simple structure of $\mathbf{z}_{j,n}$

Before continuing with deriving the p_k that maximizes the failure rate of the bound, we show how $m(\mathbf{z}, \mathbf{U})$ simplifies for samples from half-Bernoulli distributions.

Recall that our bound is a quantile of the function $m(\mathbf{z}, \mathbf{U})$ with respect to the uniform random variable $\mathbf{U}$. For samples of the form $\mathbf{z}_{j,n}$, this function reduces to a particularly simple form:

$$\begin{aligned} m(\mathbf{z}_{j,n}, \mathbf{U}) &= 1 - \sum_{i=1}^{n} U_i((\mathbf{z}_{j,n})_{i+1} - (\mathbf{z}_{j,n})_i) \\ &= 1 - [0, ..., 0, 1-k, 0, ...0] \cdot \mathbf{U} \\ &= 1 - (1-k)U_j. \end{aligned}$$

That is, with the exception of the j^{th} term, all of the successive differences of $\mathbf{z}_{j,n}$ are 0,[1] leaving us with a simple function of the j^{th} order statistic, U_j. Later it will be useful to note that the j^{th} order statistic when taking n samples from a uniform distribution is beta distributed with parameters j and $n-j+1$ (Casella and Berger, 2002, Example 5.4.5).

5.1.5 Choosing p_k to maximize the failure rate

For a fixed $(n, \mu, 1-\alpha)$, consider the set of distributions, $H_{k,\mu}$, with $j_{\min} = j$, for some value j. We are interested in the $H_{k,\mu}$ that maximizes $\Pr(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu)$. Since we are only considering half-Bernoulli distributions with a particular mean, μ, the entire distribution is specified if p_k is specified, and so we solve for the p_k

[1] We can ignore the case where $j = n$, since the bound is trivial for $\mathbf{z}_{n,n} = (1, 1, ..., 1)$.

that maximizes $\Pr(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu)$:

$$\begin{aligned}\underset{p_k : j_{\min}=j}{\arg\max} \Pr_{H_{k,\mu}}\left(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu\right) &\overset{(a)}{=} \underset{p_k : j_{\min}=j}{\arg\max}\, \beta_{\text{cdf}}(p_k; j, n-j+1) \\ &\overset{(b)}{=} \underset{p_k : j_{\min}=j}{\arg\max}\, p_k.\end{aligned}$$

Step (a) follows from (5). Step (b) follows since all beta CDFs are monotonic in their first argument. In other words, within the set of $H_{k,\mu}$ that have the same μ and $j_{\min}$, the failure rate of the bound (the probability that $\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu$) is monotonic in p_k.

Although the failure rate is monotonic in p_k, this does not mean that the worst-case is when $p_k = 1$, since this monotonicity result is restricted to the set of half-Bernoulli distributions with $j_{\min} = j$. We therefore now solve for the maximum p_k such that $j_{\min} = j$ in order to obtain the half-Bernoulli distribution with mean μ that maximizes the failure rate:

$$\begin{aligned}
& \max\left\{p_k \in [0,1] : j_{\min} = j\right\} \\
= \;& \max\Big\{p_k \in [0,1] : \{\mathbf{z}_{1,n}, \mathbf{z}_{2,n}, ..., \mathbf{z}_{j-1,n}\} \subseteq \mathcal{Z}_\alpha^\mu, \\
& \qquad\qquad \{\mathbf{z}_{j,n}, \mathbf{z}_{j+1,n}, ..., \mathbf{z}_{n,n}\} \subseteq \overline{\mathcal{Z}}_\alpha^\mu\Big\} \\
\overset{(a)}{=} \;& \max\left\{p_k \in [0,1] : \mathbf{z}_{j,n} \in \overline{\mathcal{Z}}_\alpha^\mu\right\} \\
= \;& \max\left\{p_k \in [0,1] : \Pr_U\left(m(\mathbf{z}_{j,n}, U) < \mu\right) \geq 1-\alpha\right\} \\
\overset{(b)}{=} \;& \max\left\{p_k \in [0,1] : \Pr_U\left(1-(1-k)U_j < \mu\right) \geq 1-\alpha\right\} \\
= \;& \max\left\{p_k \in [0,1] : \Pr_U\left(U_j > \frac{1-\mu}{1-k}\right) \geq 1-\alpha\right\} \\
= \;& \max\left\{p_k \in [0,1] : \Pr_U\left(U_j > p_k\right) \geq 1-\alpha\right\} \\
= \;& \max\left\{p_k \in [0,1] : 1 - \Pr_U\left(U_j \leq p_k\right) \geq 1-\alpha\right\} \\
\overset{(c)}{=} \;& \max\left\{p_k \in [0,1] : 1 - \beta_{\text{cdf}}\left(p_k; j, n-j+1\right) \geq 1-\alpha\right\} \\
= \;& \max\left\{p_k \in [0,1] : \beta_{\text{cdf}}\left(p_k; j, n-j+1\right) \leq \alpha\right\} \\
= \;& \max\left\{p_k \in [0,1] : \beta_{\text{cdf}}^{-1}(\alpha; j, n-j+1) \geq p_k\right\} \\
= \;& \beta_{\text{cdf}}^{-1}\left(\alpha; j, n-j+1\right).
\end{aligned}$$

Step (a) follows from the monotonicity of the bound and Section 5.1.2. Step (b) uses the result of (8). Step (c) uses the the fact that the j^{th} order statistic of a uniform sample of size n is beta distributed with parameters j and $n-j+1$.

5.1.6 Bringing the pieces together

We have established that, for a given n, of all half-Bernoulli distributions with mean μ and $j_{\min} = j$, the one that maximizes the failure rate of the bound has $p_k = \beta_{\text{cdf}}^{-1}(\alpha; j, n-j+1)$. Plugging this into (5), we have that

$$\begin{aligned}\max_{H_{k,\mu}: j_{\min}=j} \Pr_{H_{k,\mu}}(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu) =& \beta_{\text{cdf}}(\beta_{\text{cdf}}^{-1}(\alpha; j, n-j+1); j, n-j+1) \\ =& \alpha.\end{aligned}$$

Thus, we have seen that by maximizing the probability that $\mathbf{Z}$ causes $m_\alpha(\mathbf{Z}) > \mu$, i.e., maximizing $\Pr(\mathbf{Z} \in \overline{\mathcal{Z}}_\alpha^\mu)$, we can produce a probability of violation of at most α. Thus, the bound holds with probability at least $1-\alpha$. Since this is true for all half-Bernoulli distributions for all values of $j_{\min}$ and arbitrary tuples $(n, \mu, 1-\alpha)$, it is true for all half-Bernoullis, all sample sizes and all confidence levels.

6 Computing m_α

In this section, we discuss two methods for computing our bound, $m_\alpha(\mathbf{z})$, for a particular sample $\mathbf{z}$. The first is based upon a geometric analysis of the bound and the second uses a Monte Carlo sampling technique.

6.1 Geometric Computation of the Bound

Recall that the random variable $\mathbf{U}$ represents the order statistics of a uniform sample, and hence lies in the order statistic simplex defined in Section 4.2. Figure 5 shows the order statistic simplex for $n = 2$. Note that the order statistic simplex of dimension n has volume $\frac{1}{n!}$.

Let $\mathbf{s} \stackrel{\text{def}}{=} [z_2 - z_1, z_3 - z_2, ..., z_{n+1} - z_n]$ be the *spacings* of the sample. Consider an example in which $\mathbf{z} = [0.3, 0.8]$. Then $\mathbf{s} = [0.5, 0.2]$, as shown in the figure.

Starting from the definition of our bound $m_\alpha(\mathbf{z})$ and expanding the definition of the quantile function, we have

$$\begin{aligned}m_\alpha(\mathbf{z}) &= \inf\{\hat{\mu} \in \mathbb{R} : \Pr(m(\mathbf{z}, \mathbf{U}) \leq \hat{\mu}) \geq 1-\alpha\} \\ &= \inf\{\hat{\mu} \in \mathbb{R} : n! \,\text{Volume}(\{\mathbf{u} : m(\mathbf{z}, \mathbf{u}) \leq \hat{\mu}\}) \geq 1-\alpha\} \\ &= \inf\{\hat{\mu} \in \mathbb{R} : n! \,\text{Volume}(\{\mathbf{u} : 1 - \mathbf{u} \cdot \mathbf{s} \leq \hat{\mu}\}) \geq 1-\alpha\} \\ &= \inf\{\hat{\mu} \in \mathbb{R} : n! \,\text{Volume}(\{\mathbf{u} : \mathbf{u} \cdot \mathbf{s} \geq 1-\hat{\mu}\}) \geq 1-\alpha\}.\end{aligned}$$

This final expression has a clear geometric interpretation. The set of points $\mathbf{u}$ such that $\mathbf{u} \cdot \mathbf{s}$ is greater than some value is the upper right region of the order statistic simplex, depicted by the pink region in Figure 5. The bound is defined to be the least value of $\hat{\mu}$ such that the volume of the pink region is a fraction $1-\alpha$ of the simplex volume.

This value of $\hat{\mu}$ can be found by evaluating the volume of the section of the order statistic simplex above the hyperplane l, a hyperplane orthogonal to the

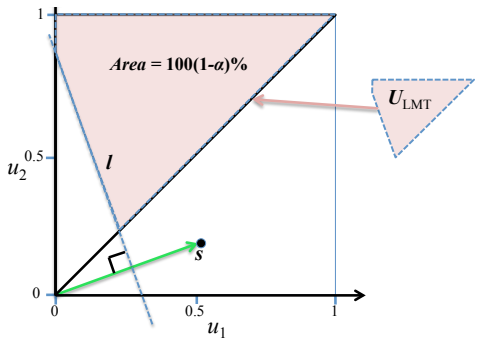

Figure 5: The figure shows several quantities related to the geometric computation of our bound. The upper left triangle represents the order statistic simplex. The point $\mathbf{s}$ represents the spacings of the sample $\mathbf{z}$. The pink region, which we define later to be $\mathcal{U}_{\text{LMT}}$, is a section of the order statistic simplex cut by the hyperplane l, which represents the set of vectors $\mathbf{u}$ for which $\mathbf{u} \cdot \mathbf{s}$ is greater than or equal to some value. The goal is to find the maximum such value, and thereby the minimum $\hat{\mu}$ such that the volume of the pink region is $100(1-\alpha)\%$ of the volume of the order statistic simplex.

spacings vector $\mathbf{s}$. Thus, we seek the smallest value of $\hat{\mu}$ such that the volume of this section is $100(1-\alpha)\%$ of the volume of the order statistic simplex. This value of $\hat{\mu}$ is our bound.

Closed-form expressions for sections of simplexes cut by hyperplanes have been published by several authors, including Lasserre (2015). These expressions lead to efficient calculations of the bound in most cases. However, these formulas have singularities that cause problems for certain samples $\mathbf{z}$, such as samples with repeated values. Thus, we explore a more reliable Monte Carlo approach for computing our bound below.

6.2 Monte Carlo Estimate of the Bound

Since the bound is defined in terms of a quantile of a function that depends upon a uniform random variable, it is simple to develop a Monte Carlo estimate. This

is provided in Algorithm 1.

Algorithm 1: Monte Carlo Estimation of m_α
This pseudocode uses zero-based indexing of arrays.

Input : A sample $\mathbf{x}$, confidence parameter, α, and Monte Carlo sampling parameter l.
Output: An estimate of $m_\alpha(\mathbf{x})$.

$n \leftarrow \text{length}(\mathbf{x})$;
$\mathbf{z} \leftarrow \text{sort}(\mathbf{x}, \text{ascending})$;
Create array **ms** to hold l floating point numbers, and initialize it to zero;
Create arrays **u** and **s**, each to hold n floating point numbers;
for $i \leftarrow 1$ **to** $n - 1$ **do**
  $\mathbf{s}[i] = z[i+1] - z[i]$;
end
$\mathbf{s}[n] \leftarrow 1 - z[n]$;
for $i \leftarrow 1$ **to** l **do**
  for $j \leftarrow 1$ **to** n **do**
    $\mathbf{u}[j] \sim \text{Uniform}(0, 1)$;
  end
  $\text{sort}(\mathbf{u}, \text{ascending})$;
  $\mathbf{ms}[i] \leftarrow 1 - \mathbf{s} \cdot \mathbf{u}$;
end
$\text{sort}(\mathbf{ms}, \text{ascending})$;
return $\mathbf{ms}[\lceil (1-\alpha)(n-1) \rceil]$;

Algorithm 1 can be implemented more efficiently if m_α will be estimated multiple times for the same n, since samples of $\mathbf{U}$ can be computed and sorted a single time. Also, notice that the number of Monte Carlo samples, l, does not scale poorly for distributions with rare values, since we are estimating the $(1-\alpha)$-quantile of $m(\mathbf{z}, \mathbf{U})$, which is robust to outliers (e.g., $\alpha = 0.5$ makes this the median, which is well known to be robust to outliers).

In practice, we find that $l = 10{,}000$ tends to provide a reasonable approximation of $m_\alpha(\mathbf{x})$ for $\alpha = 0.05$. Note that (1) may not hold when using this Monte Carlo estimate of m_α due to error in the estimate. In practice this can be remedied by increasing l, or by incorporating high-probability bounds on the error in the Monte Carlo estimate into the bound. Also, note that as α decreases, l should be increased.

7 Related Work

In this section, we review other methods for computing high-confidence upper bounds on the mean of a random variable from samples (several of which we compare to in the subsequent numerical analysis section). Although some of these methods extend to more general settings (e.g., Hoeffding's inequality does not require identically distributed samples, and Anderson's inequality does not require a lower-bound on the random variable), here we consider only the

standard setting that we have discussed in this paper, wherein the samples are i.i.d. and the random variable always takes values in the interval $[0, 1]$. We divide this section into two parts: prior methods that provide guaranteed coverage, and prior methods that do not provide guaranteed coverage. We present these prior methods as functions, $m_\alpha^{\text{Hoeffding}}$, $m_\alpha^{\text{Maurer\&Pontil}}$, etc., each of which provides an alternative to m_α.

7.1 Prior Methods with Guaranteed Coverage

The methods presented in this subsection have guaranteed coverage in the setting that we have described—they satisfy (1) if used in place of m_α.

Using Hoeffding's inequality (Hoeffding, 1963) to construct a high-confidence upper-bound on μ is perhaps the best known, and simplest, prior method with guaranteed coverage:

$$m_\alpha^{\text{Hoeffding}}(\mathbf{x}) \stackrel{\text{def}}{=} \bar{\mathbf{x}} + \sqrt{\frac{\ln(1/\alpha)}{2n}},$$

where $\bar{\mathbf{x}}$ is the sample mean, i.e., $\bar{\mathbf{x}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i$. In cases where the variance of the random variable is significantly less than one, the upper bounds provided by Maurer and Pontil's empirical Bernstein bound (Maurer and Pontil, 2009) can be tighter than those produced by Hoeffding's inequality. This is achieved by leveraging not just the sample mean, $\bar{\mathbf{x}}$, but also the sample variance, $\widehat{\text{Var}}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2$:

$$m_\alpha^{\text{Maurer\&Pontil}}(\mathbf{x}) \stackrel{\text{def}}{=} \bar{\mathbf{x}} + \sqrt{\frac{2\widehat{\text{Var}}(\mathbf{x})\ln(2/\alpha)}{n}} + \frac{7\ln(2/\alpha)}{3(n-1)}.$$

Going one step further, Anderson's inequality provides high-confidence upper bounds on the mean by using the entire sample CDF (rather than only the sample mean and variance):

$$m_\alpha^{\text{Anderson}}(\mathbf{z}) \stackrel{\text{def}}{=} m(\mathbf{z}, \mathbf{u}^{\text{DKW}}),$$

where for $i \in \{1, 2, \ldots, n\}$,

$$\mathbf{u}_i^{\text{DKW}} \stackrel{\text{def}}{=} \max\left\{0, i/n - \sqrt{\ln(1/\alpha)/2n}\right\} \tag{22}$$

is a vector that Anderson derived from the Dvoretsky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956). Note that the form we present for Anderson's inequality uses the tight constant for the DKW inequality found by Massart (1990), which relies on the assumption that $\alpha \leq 0.5$. This restriction is not restrictive because high-confidence bounds are typically applied with small values of α, e.g., $\alpha = 0.05$.

Following these three methods, several alternatives have been proposed, including other approaches that rely only on the sample mean (Chen, 2008),

methods that extend Maurer and Pontil's empirical Bernstein bound to provide tighter bounds for random variables with long tails (Bubeck et al., 2012; Thomas et al., 2015a), and methods that provide alternatives to Anderson's inequality that use alternative methods of defining inclusion envelopes for a distribution's CDF (Learned-Miller and DeStefano, 2008; Diouf and Dufour, 2005).

7.2 Prior Methods without Guaranteed Coverage

All of the methods presented in this subsection do *not* have guaranteed coverage in the setting that we have described, but are often used to compute high-confidence upper bounds on the mean.

Perhaps the most common method for constructing high-confidence upper bounds is based on Student's t-statistic (Student, 1908):

$$m_\alpha^{\text{Student}}(\mathbf{x}) \stackrel{\text{def}}{=} \bar{\mathbf{x}} + \sqrt{\frac{\widehat{\text{Var}}(\mathbf{x})}{n}} t_{1-\alpha,n-1},$$

where $t_{1-\alpha,\nu}$ denotes the $100(1-\alpha)$ percentile of the Student's t distribution with ν degrees of freedom. We refer to this confidence interval as the Student-t interval. If $\bar{\mathbf{X}}$ is normally distributed, then $m_\alpha^{\text{Student}}$ *does* provide guaranteed coverage. The central limit theorem implies that $\bar{\mathbf{X}}$ tends towards a normal distribution as n increases, and so this method is often applied in scientific research if $n \geq 30$, even though this does not provide coverage guarantees.

Bootstrap methods tend to provide the tightest confidence intervals for the mean. However, this comes at a large cost: they do not have guaranteed coverage, even with normality assumptions. Despite concerns about their reliability, bootstrap methods remain in common use due to their tight confidence intervals and tendency to produce error rates roughly around α for many common distributions (Hanna et al., 2017; Thomas et al., 2015b). Since bootstrap methods are not easily expressed as closed-form alternatives to m_α, we refer the reader to the work of Efron and Tibshirani (1993) for details on these approaches. The two that we focus on in our subsequent experiments are the most common, the percentile bootstrap, and one of the most sophisticated, the *bias corrected and accelerated* (BCa) bootstrap.

One limitation of BCa, the more sophisticated bootstrap method, is that it is not defined in some cases (e.g., if all of the samples take the same value) and can encounter numerical issues in other cases. In our implementations, whenever numerical issues are detected, the method automatically reverts to the percentile bootstrap.

8 Theoretical Analysis

In this section, we provide an analytic comparison of our bound to two prior methods that provide guaranteed coverage: Hoeffding's inequality and Anderson's inequality. We will show that, for any sample $\mathbf{z}$, our high-confidence upper

bound is less than that resulting from Hoeffding's inequality, and never greater than that of Anderson's inequality. We break this result into two components. First we show that for all $\mathbf{z}$, $m_\alpha(\mathbf{z}) \leq m_\alpha^{\text{Anderson}}(\mathbf{z})$. Second, we show that for all $\mathbf{z}$, $m_\alpha^{\text{Anderson}}(\mathbf{z}) \leq m_\alpha^{\text{Hoeffding}}(\mathbf{z})$, which implies that $m_\alpha(\mathbf{z}) \leq m_\alpha^{\text{Hoeffding}}(\mathbf{z})$, where these inequalities are strict if $\alpha \leq 0.5$.

8.1 Theoretical Comparison to Anderson's Inequality

In this section we compare m_α to $m_\alpha^{\text{Anderson}}$.

Theorem 1. *For all possible values* $\mathbf{z}$ *of* $\mathbf{Z}$ *and all* $\alpha \in [0, 0.5]$,

$$m_\alpha(\mathbf{z}) \leq m_\alpha^{Anderson}(\mathbf{z}).$$

Proof. We present a sketch of the proof. Consider the diagram in Figure 6. This figure depicts, for $n = 2$, the space of possible vectors $\mathbf{u}$, which are sorted uniform samples. The point (u_1, u_2) represents $\mathbf{u}^{\text{DKW}}$, defined in (22). $\mathcal{U}_{\text{DKW}}$ denotes the set of vectors that are element-wise greater than $\mathbf{u}^{\text{DKW}}$. It follows from the DKW inequality, with the tight constants found by Massart (1990), that the probability $\mathbf{U}$ is in $\mathcal{U}_{\text{DKW}}$ is at least $1-\alpha$. The region $\mathcal{U}_{\text{LMT}}$ is any set of $\mathbf{u}$'s that result in the lowest induced means, $m(\mathbf{z}, \mathbf{u})$, while ensuring that the probability that $\mathbf{U}$ is in $\mathcal{U}_{\text{LMT}}$ is precisely $1-\alpha$. Note that any point that is not contained within the pink region must represent a vector $\mathbf{u}$ that results in an induced mean, $m(\mathbf{z}, \mathbf{u})$, which is greater than the induced mean of any point in the pink region. Our bound is effectively the maximum over induced means of points in the pink region and Anderson's bound is the maximum over points in the blue region. Since the probability that $\mathbf{U}$ is in $\mathcal{U}_{\text{DKW}}$ cannot be less than the probability that $\mathbf{U}$ is in $\mathcal{U}_{\text{LMT}}$, and $\mathcal{U}_{\text{LMT}}$ contains the $\mathbf{u}$ vectors that minimize the induced mean, our bound cannot be larger than Anderson's. □

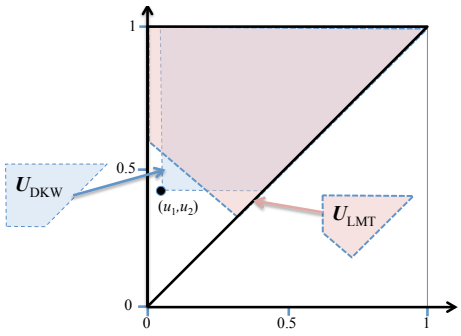

Figure 6: Diagram comparing our bound and Anderson's.

8.2 Analytic Comparison to Hoeffding's Inequality

In this section we prove the following theorem:

Theorem 2. *For all possible values* $\mathbf{z}$ *of* $\mathbf{Z}$ *and all* $\alpha \in [0,1]$,

$$m_\alpha^{Anderson}(\mathbf{z}) \leq m_\alpha^{Hoeffding}(\mathbf{z}),$$

where the inequality is strict if $\alpha \leq 0.5$.

Proof. We begin with $m_\alpha^{\text{Anderson}}(\mathbf{z})$ and present a sequence of inequalities that conclude with $m_\alpha^{\text{Hoeffding}}(\mathbf{z})$, where one inequality is strict if $\alpha \leq 0.5$:

$$\begin{aligned}
m_\alpha^{\text{Anderson}}(\mathbf{z}) =& m\left(\mathbf{z}, \mathbf{u}^{\text{DKW}}\right) \\
=& 1 - \sum_{i=1}^{n}(z_{i+1} - z_i)\mathbf{u}_i^{\text{DKW}} \\
=& 1 - \sum_{i=1}^{n}(z_{i+1} - z_i)\max\left\{0, \frac{i}{n} - \sqrt{\frac{\ln(1/\alpha)}{2n}}\right\} \\
\leq& 1 - \sum_{i=1}^{n}(z_{i+1} - z_i)\left(\frac{i}{n} - \sqrt{\frac{\ln(1/\alpha)}{2n}}\right).
\end{aligned}$$

If $\alpha \leq 0.5$, then this final inequality is strict because, when $i = 1$, we have that for any n, $0 > i/n - \sqrt{\ln(1/\alpha)/2n}$, and so

$$\max\left\{0, \frac{i}{n} - \sqrt{\frac{\ln(1/\alpha)}{2n}}\right\} > \frac{i}{n} - \sqrt{\frac{\ln(1/\alpha)}{2n}}.$$

Continuing, we have:

$$\begin{aligned}
m_\alpha^{\text{Anderson}}(\mathbf{z}) \leq& 1 - \sum_{i=1}^{n}(z_{i+1} - z_i)\frac{i}{n} + \sum_{i=1}^{n}(z_{i+1} - z_i)\sqrt{\frac{\ln(1/\alpha)}{2n}} \\
=& 1 + \frac{1}{n}\left(\sum_{i=1}^{n} i z_i - \sum_{i=1}^{n} i z_{i+1}\right) + (z_{n+1} - z_1)\sqrt{\frac{\ln(1/\alpha)}{2n}} \\
=& 1 + \frac{1}{n}\left(\sum_{i=1}^{n} i z_i - \left(\sum_{i=2}^{n}(i-1)z_i\right) - n\right) + (1 - z_1)\sqrt{\frac{\ln(1/\alpha)}{2n}} \\
=& \frac{1}{n}\sum_{i=1}^{n} z_i + (1 - z_1)\sqrt{\frac{\ln(1/\alpha)}{2n}} \qquad (23) \\
\leq& \frac{1}{n}\sum_{i=1}^{n} z_i + \sqrt{\frac{\ln(1/\alpha)}{2n}} \\
=& m_\alpha^{\text{Hoeffding}}(\mathbf{z}).
\end{aligned}$$

Notice that (23) provides an expression similar to Hoeffding's inequality, but where the lower bound on the random variable (in our case, zero) is replaced by

the smallest observed sample, z_1. This presents a tighter variant of Hoeffding's inequality that holds when $\alpha \leq 0.5$ and the random variables are i.i.d. (the general form of Hoeffding's inequality holds for random variables that are not necessarily identically distributed). □

It then follows from Theorem 2 that our bound is always at least as tight as Hoeffding's inequality, and is strictly tighter if $\alpha \leq 0.5$:

Corollary 1. *For all possible values* $\mathbf{z}$ *of* $\mathbf{Z}$ *and all* $\alpha \in [0, 1]$,

$$m_\alpha(\mathbf{z}) \leq m_\alpha^{Hoeffding}(\mathbf{z}),$$

where the inequality is strict if $\alpha \leq 0.5$.

Proof. This follows immediately from Theorems 1 and 2. □

9 Numerical Analysis

In this section we present results from a numerical analysis of our bound. These empirical results aim to answer the following research questions:

RQ1 For a variety of distributions, confidence levels, and number of samples, are results consistent with (1)?

RQ2 For a variety of distributions that resemble common use-cases, how do the confidence intervals produced by our bound compare to those of previous methods that have guaranteed coverage (i.e., those that satisfy (1))?

RQ3 This question is the same as RQ2, but for methods that do **not** have guaranteed coverage.

RQ4 Can our bound provide confidence intervals that are practical for scientific experiments with fewer than 30 samples?

9.1 Numerical Studies on Guaranteed Coverage

In this subsection we study RQ1 with experiments that are consistent with the conjecture that m_α, as we have defined it, has guaranteed coverage (satisfies (1)). Although they show that (1) appears to hold for a variety of settings, this does not imply that settings do not exist under which (1) does not hold.

To study RQ1, we selected a variety of different distributions (uniform, beta, and Bernoulli, each with various parameters), confidence levels $1 - \alpha$, and number of samples, n. For each such tuple, $(\text{distribution}, 1 - \alpha, n)$, we collected 10,000 samples of $\mathbf{Z}$, computed $m_\alpha(\mathbf{Z})$, and checked whether $m_\alpha(\mathbf{Z}) \geq \mu$. From these 10,000 tests, we estimated the coverage—the probability that the bound holds. That is, we estimated $\Pr(m_\alpha(\mathbf{Z}) \geq \mu)$ by dividing the number of samples of $\mathbf{Z}$ such that $m_\alpha(\mathbf{Z}) \geq \mu$ by 10,000.

Although or goal here is to study RQ1, to facilitate the interpretation of the presented results, we provide results using two prior methods that exhibit

the different types of behavior that our bound might produce. This comparison also provides some insight into RQ2 (via the comparison to Heoffding's inequality) and RQ3 (via comparison to the Student-t interval). However, note that subsequent experiments further study these two research questions.

First consider Figure 7, which presents results using Hoeffding's inequality. The top left plot shows the coverage for a variety of beta distributions, but with n fixed. To interpret this plot, consider the curve $\beta(1,5)$ $n = 10$ at the 0.7 position on the horizontal axis. The position 0.7 on the horizontal axis indicates that we requested an upper bound that holds with probability at least 0.7. Since, at horizontal position 0.7, the curve $\beta(1,5)$ $n = 10$ (red curve) lies above 0.7 (blue curve), the upper bound produced by Hoeffding's inequality held with probability greater than 0.7. Hence, a curve remaining above the blue line indicates that the desired confidence level was achieved. However, notice that the curve is far above the blue line—this indicates that Hoeffding's inequality was overly conservative. It provided a high-confidence upper bound that was greater than or equal to the mean far more often than requested. This means that for this distribution the confidence interval provided by Hoeffding's inequality is not tight, and could be improved.

The three other plots in Figure 7 are similar, but use different parameters. The top row presents results for beta distributions, while the bottom row presents results for Bernoulli distributions. The left column presents results as the parameters of the distributions are varied, while the right column presents results as the number of samples is varied. Overall Figure 7 shows the behavior that we would expect of a bound that has guaranteed coverage, but which provides loose high-confidence upper bounds.

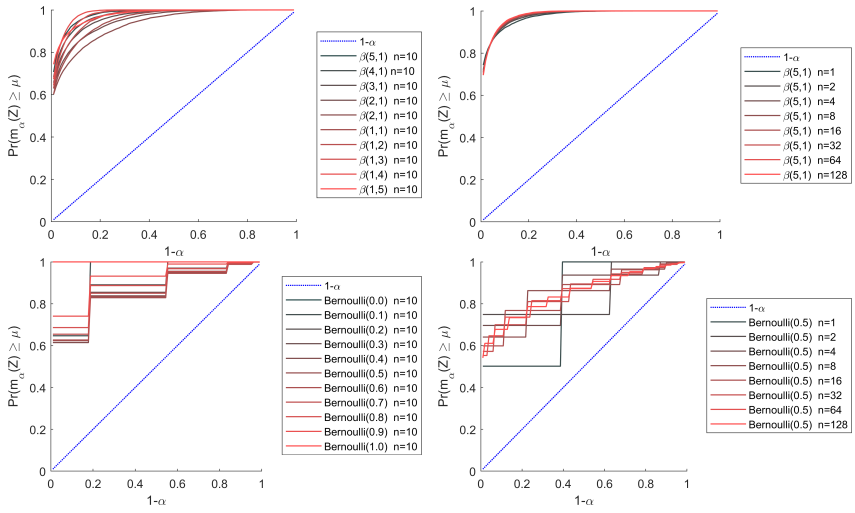

Figure 7: Estimated probability that the high-confidence upper bound produced using Hoeffding's inequality is greater than or equal to the true mean.

Now consider Figure 8, which is identical to Figure 7, except that it uses the Student-t interval instead of Hoeffding's inequality. This plots shows very different behavior: the curves tend to be much lower, indicating tighter confidence intervals around the sample mean. However, the curves often cross the blue line, indicating that in these settings the Student-t interval does *not* provide guaranteed coverage—if you ask for an 0.8-confidence upper bound, you may only get a 0.6-confidence upper bound. Hence Figure 8 shows the behavior that we would expect of a bound that does *not* have guaranteed coverage, but which provides tight "high-confidence" upper bounds.

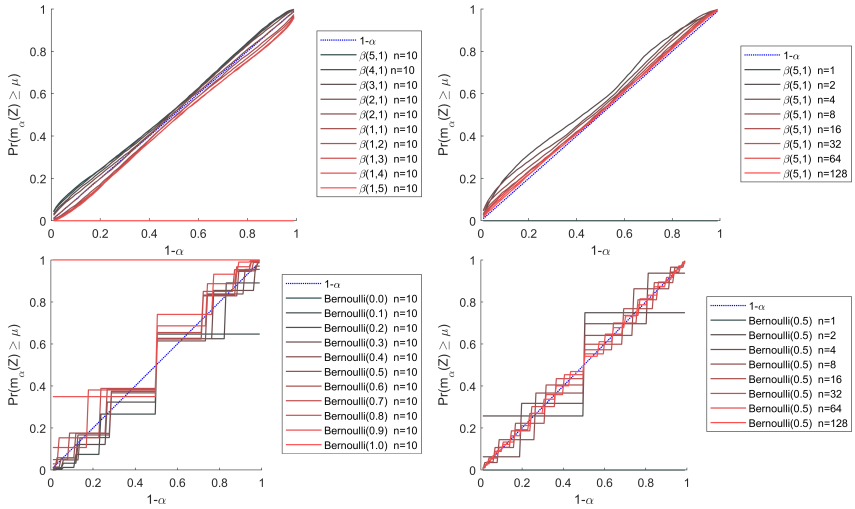

Figure 8: Estimated probability that the high-confidence upper bound produced using the Student-t interval is greater than or equal to the true mean.

The desired behavior of a high-confidence upper bound would blend the desirable properties of Hoeffding's inequality and the Student-t interval. In these plots, this would result in curves that *always* remain above the blue curve (guaranteed coverage), but are otherwise as low as possible (tight). Figure 9 presents the results of this same experiment, conducted using our bound. It achieves this desired behavior—it always remains above the blue line (consistent with guaranteed coverage), but tends to be significantly lower than Hoeffding's inequality.

9.2 Numerical Comparison to Previous Methods

In this subsection we focus on RQ2 and RQ3 with experiments that compare the tightness of our bound to that of previous methods (both with and without guaranteed coverage). A variety of different statistics can be used to capture how tight the high-confidence upper bounds produced by a method are, including the mean upper bound, i.e., $\mathbf{E}[m_\alpha(\mathbf{Z})]$, and the median upper bound. Here we

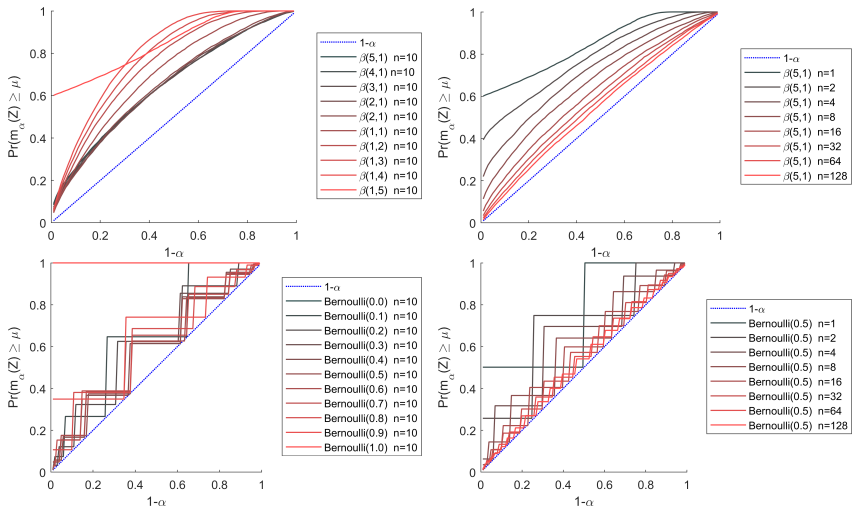

Figure 9: Estimated probability that the high-confidence upper bound produced using our bound is greater than or equal to the true mean.

report the mean upper bound: we gather 1,000 samples of $\mathbf{Z}$ from a distribution and compute the upper bounds produced by our bound and several previous methods and report the sample mean of the upper bounds for each method. For simplicity, here we vary the distribution and n but fix $\alpha = 0.05$ to obtain 95%-confidence upper bounds.

First compare the blue curve (our bound) to the black curves (previous methods with guaranteed coverage), noting that the horizontal axis uses a logarithmic scale. In every case, the blue curve remains strictly below the black curves, indicating that in every setting our bound produces lower values on average. Notice that frequently our bound obtains mean upper bounds that previous methods require an order of magnitude more samples to achieve, indicating that our bound is a drastic improvement in tightness and/or data efficiency.

Next, compare the blue curve (our bound) to the red curves (previous methods that do *not* have guaranteed coverage). The two bootstrap methods do not provide guaranteed coverage, even with normality assumptions. So, although they produce tight confidence intervals (as is evident in these plots), the high-confidence bounds that they produce cannot be relied upon.

Next consider the Student-t interval: for the uniform distribution, it produces high-confidence upper bounds that are similar to those produced by our bound. When the Student-t interval is computed from normally distributed data, it produces a tighter high-confidence upper-bound than our bound. However, when the sampling distribution includes right-skew (e.g., $\beta(1, 10)$), the Student-t interval tends to be overly optimistic—it does not have guaranteed coverage (as is evident in Figure 8 with $\beta(1, 5)$). Hence, although the upper bound of the Student-t interval tends to be lower than those produced by our

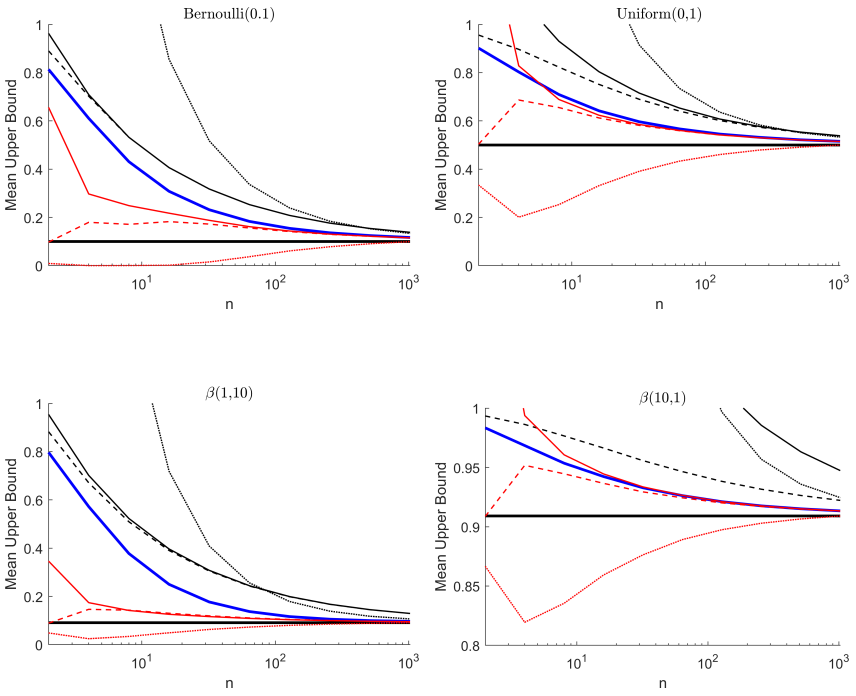

Figure 10: These plots depict the mean upper bounds (over 1,000 trials) for various distributions (the titles on the plots describe the distribution) and using various methods. All figures share the following legend:

method for $\beta(1, 10)$, it does *not* have guaranteed coverage. On the other hand, when the sampling distribution includes left-skew (e.g., $\beta(10, 1)$, the Student-t interval is overly-conservative (like Hoeffding's inequality). Hence, in Figure 10, the $\beta(10, 1)$ plot indicates that our bound is *tighter* than the Student-t interval. This is further evidence that our bound is combining the desirable properties of Hoeffding's inequality and the Student-t interval: it roughly preserves the tightness of the Student-t interval, except in the cases where the Student-t interval is too tight to provide guaranteed coverage (in which case our bound is sufficiently looser to provide guaranteed coverage).

9.3 Numerical Support for Practical use With $n < 30$

Of the many potential uses of our bound, one stands out: it provides a valid method for constructing confidence intervals for scientific studies with fewer than 30 samples. Even though the Student-t interval does *not* have guaranteed coverage when the sampling distribution is not normal, the central limit theorem tells us that the sample mean tends towards a normal distribution as n increases. Hence, the Student-t interval becomes reasonable when n is large.[2] However, without knowing the sampling distribution, it is not clear how large n must be for the Student-t interval to be reasonable. A common rule of thumb used in current scientific research is that n must be at least 30.

This raises the question: what should one do when fewer than 30 samples are available? Our bound provides an answer (assuming our conjecture is true), as it provides confidence intervals of comparable tightness, but with guaranteed support for *any* n and without any normality assumptions. The only requirement is the ability to identify limits on the support of the distribution. To answer RQ3, we present an experiment that shows how our bound can be used to obtain confidence intervals based on fewer than 30 empirical measurements.

Specifically, we used data from the United States Census from the year 2000 to obtain an estimate of the distribution of people's ages, considering only people zero to 84 years old. We then consider the problem of obtaining a tight high-confidence upper bound on the mean age of people ages zero to 84 based on $n < 30$ samples. The results of this experiment are presented in Figure 11, which is a similar form to Figure 10 (but without the logarithmic horizontal axis). The key observations from this plot are: **1)** previous methods with guaranteed coverage are too loose to provide useful high-confidence bounds with so few samples, **2)** the Student-t interval is sufficiently tight, but it cannot be applied responsibly with such a small n, and **3)** our bound produces high-confidence bounds that are comparable to the Student-t interval (while maintaining guaranteed coverage, if our conjecture holds).

[2] Notice that even with arbitrarily large n, the Student-t interval may *not* have guaranteed coverage, so here saying that the Student-t interval is "reasonable" does *not* mean that it has guaranteed coverage.

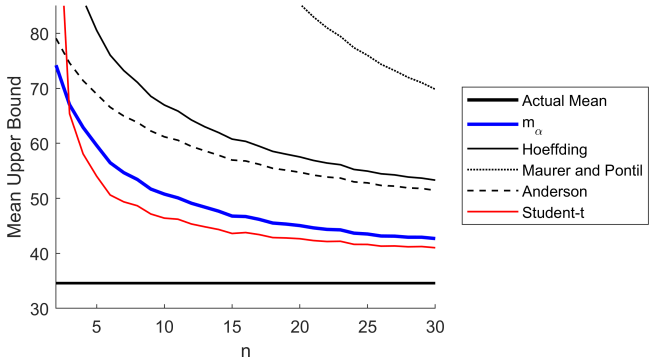

Figure 11: The mean upper bounds (over 1,000 trials) produced by various methods when n is varied and the sampling distribution is an approximation of the distribution of ages (bounded in $[0, 84]$) in the United States in the year 2000.

10 Acknowledgements

This work benefitted significantly from conversations with Vince Lysinski and George Bissias, as well as from discussions with Andrew McGregor, Don Towsley, Berthold Horn, Archan Ray, Justin Domke, Gary Huang, Dan Sheldon, Luc Rey-Bellet, and Markos Katsoulakis. Some of this work grew out of early efforts to improve Anderson's bound by Benjamin Mears while at the University of Massachusetts, Amherst.

References

T. W. Anderson. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of The International and Statistical Institute*, 43:249–251, 1969.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *Arxiv*, arXiv:1209.1727, 2012.

G. Casella and R. L. Berger. *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

X. Chen. Confidence interval for the mean of a bounded random variable and its applications in point estimation. *arXiv preprint arXiv:0802.3458*, 2008.

M. A. Diouf and J. M. Dufour. Improved nonparametric inference for the mean of a bounded random variable with application to poverty measures. 2005. URL `http://web.hec.ca/scse/articles/Diouf.pdf`.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669, 1956.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, London, 1993.

J. P. Hanna, P. Stone, and S. Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 538–546. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. B. Lasserre. Volume of slices and sections of the simplex in closed form. *Optimization Letters*, 9(7):1263–1269, 2015.

E. Learned-Miller and J. DeStefano. A probabilistic upper bound on differential entropy. *IEEE Transactions on Information Theory*, 54(11):5223–5230, 2008.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 1990.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Twenty-Second Annual Conference on Learning Theory*, pages 115–124, 2009.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, 2015a.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015b.