Towards Practical Mean Bounds for Small Samples

My Phan [1] Philip S. Thomas [1] Erik Learned-Miller [1]

Abstract

Historically, to bound the mean for small sample sizes, practitioners have had to choose between using methods with unrealistic assumptions about the unknown distribution (e.g., Gaussianity) and methods like Hoeffding's inequality that use weaker assumptions but produce much looser (wider) intervals. In 1969, Anderson (1969a) proposed a mean confidence interval strictly better than or equal to Hoeffding's whose only assumption is that the distribution's support is contained in an interval $[a, b]$. For the first time since then, we present a new family of bounds that compares favorably to Anderson's. We prove that each bound in the family has *guaranteed coverage*, i.e., it holds with probability at least $1 - \alpha$ for all distributions on an interval $[a, b]$. Furthermore, one of the bounds is tighter than or equal to Anderson's for all samples. In simulations, we show that for many distributions, the gain over Anderson's bound is substantial.

1. Introduction

In this work, we revisit the classic statistical problem of defining a confidence interval on the mean μ of an unknown distribution with CDF F from an i.i.d. sample $\mathbf{X} = X_1, X_2, \dots, X_n$, and the closely related problems of producing upper or lower confidence bounds on the mean. For simplicity, we focus on upper confidence bounds (UCBs), but the development for lower confidence bounds and confidence intervals is similar.

To produce a non-trivial UCB, one must make assumptions about F, such as finite variance, sub-Gaussianity, or that its support is contained on a known interval $[a, b]$. We adopt this last assumption, working with distributions whose support is known to fall in an interval $[a, b]$. For UCBs, we refer to two separate settings, the *one-ended support* setting, in which the distribution is known to fall in the interval $[-\infty, b]$, and the *two-ended support* setting, in which the distribution is known to fall in an interval $[a, b]$, where $a > -\infty$ and $b < \infty$.

A UCB has *guaranteed coverage* for a set of distributions $\mathcal{F}$ if, for all sample sizes $1 \leq n \leq \infty$, for all confidence levels $1 - \alpha \in (0, 1)$, and for all distributions $F \in \mathcal{F}$, the bound $\mu^{1-\alpha}_{\texttt{upper}}$ satisfies

$$Prob_F[\mu \leq \mu^{1-\alpha}_{\texttt{upper}}(X_1, X_2, ..., X_n)] \geq 1 - \alpha, \quad (1)$$

where μ is the mean of the unknown distribution F.

Among bounds with guaranteed coverage for distributions on an interval $[a, b]$, our interest is in bounds with good performance on *small sample sizes*. The reason is that, for 'large enough' sample sizes, excellent bounds and confidence intervals already exist. In particular, the confidence intervals based on Student's $t-$statistic (Student, 1908) are satisfactory in terms of coverage and accuracy for most practitioners, given that the sample size is greater than some threshold.[1]

The validity of the Student's t method depends upon the Gaussianity of the sample mean, which, strictly speaking does not hold for any finite sample size unless the original distribution itself is Gaussian. However, for many applications, the sample mean becomes close enough to Gaussian as the sample size grows (due to the effects described by the central limit theorem), that the resulting bounds hold with probabilities close to the confidence level. Such results vary depending upon the unknown distribution, but it is generally accepted that a large enough sample size can be defined to cover any distributions that might occur in a given situation.[2] The question is what to do when the sample size is smaller than such a threshold.

Establishing good confidence intervals on the mean for small samples is an important but often overlooked problem. The t-test is widely used in medical and social sciences.

[1] College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, USA. Correspondence to: My Phan <myphan@cs.umass.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] An adequate sample size for the Student's t method depends upon the setting, but a common rule is $n > 30$.

[2] An example in which the sample mean is still visibly skewed (and hence inappropriate for use with Student's t) even after $n = 80$ samples is given for log-normal distributions in the supplementary material.

Small clinical trials (such as Phase 1 trials), where such tests could potentially be applied, occur frequently in practice (Institute of Medicine, 2001). In addition, there are several machine learning applications. The sample mean distribution of an importance-weighted estimator is skewed even when the sample size is much larger than 30, so tighter bounds with guarantees may be beneficial. Algorithms in Safe Reinforcement Learning (Thomas et al., 2015) use importance weights to estimate the return of a policy and use confidence bounds to estimate the range of the mean. The UCB multi-armed bandit algorithm is designed using the Hoeffding bound - a tighter bound may lead to better performance with guarantees.

In the two-ended support setting, our bounds provide a new and better option for guaranteed coverage with small sample sizes.[3] At least one version of our bound is tighter (or as tight) for *every possible sample* than the bound by Anderson (Anderson, 1969a), which is arguably the best existing bound with guaranteed coverage for small sample sizes. In the limit as $a \to -\infty$, i.e., the one-ended support setting, this version of our bound is equivalent to Anderson.[4]

It can be shown from Learned-Miller & Thomas (2019) that Anderson's UCB is less than or equal to Hoeffding's for *any* sample when $\alpha \leq 0.5$, and is strictly less than Hoeffding's when $\alpha \leq 0.5$ and $n \geq 3$. Therefore our bound is also less than or equal to Hoeffding's for *any* sample when $\alpha \leq 0.5$, and is strictly better than Hoeffding's inequality when $\alpha \leq 0.5$ and $n \geq 3$.

Below we review bounds with coverage guarantees, those that do *not* exhibit guaranteed coverage, and those for which the result is unknown.

1.1. Distribution free bounds with guaranteed coverage

Several bounds exist that have guaranteed coverage. These include Hoeffding's inequality (Hoeffding, 1963), Anderson's bound (Anderson, 1969a), and the bound due to Maurer & Pontil (2009).

Hoeffding's inequality. For a distribution F on $[a, b]$, Hoeffding's inequality (Hoeffding, 1963) provides a bound on the probability that the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, will deviate from the mean by more than some amount $t \geq 0$:

$$\Pr\left(\mu - \bar{X}_n \leq t\right) \leq e^{-\frac{2nt^2}{(b-a)^2}}. \quad (2)$$

Defining α to be the right hand side of this inequality, solving for t as a function of α, and rewriting in terms of α rather than t, one obtains a $1-\alpha$ UCB on the mean of

$$b^{\alpha,\text{Hoeffding}}(\mathbf{X}) \overset{\text{def}}{=} \bar{X}_n + (b-a)\sqrt{\frac{\ln(1/\alpha)}{2n}}. \quad (3)$$

Maurer and Pontil. One limitation of Hoeffding's inequality is that the amount added to the sample mean to obtain the UCB scales with the range of the random variable over $\sqrt{n}$, which shrinks slowly as n increases.

Bennett's inequality (Bennett, 1962) considers both the sample mean and the sample variance and obtains a better dependence on the range of the random variable *when the variance is known*. Maurer & Pontil (2009) derived a UCB for the variance of a random variable, and suggest combining this with Bennet's inequality (via the union bound) to obtain the following $1-\alpha$ UCB on the mean:

$$b^{\alpha,\text{M\&P}}(\mathbf{X}) \overset{\text{def}}{=} \bar{X}_n + \frac{7(b-a)\ln(2/\alpha)}{3(n-1)} + \sqrt{\frac{2\hat{\sigma}^2\ln(2/\alpha)}{n}}.$$

Notice that Maurer and Pontil's UCB scales with the range $(b-a)$, divided by n (as opposed to the $\sqrt{n}$ of Hoeffding's). However, the $\sqrt{n}$ dependence is unavoidable to some extent: Maurer and Pontil's UCB scales with the sample standard deviation $\hat{\sigma}$ divided by $\sqrt{n}$. As a result, Maurer and Pontil's bound tends to be tighter than Hoeffding's when both n is large and the range of the random variable is large relative to the variance. Lastly, notice that Maurer and Pontil's bound requires $n{\geq}2$ for the sample standard deviation to be defined.

Anderson's bound. Anderson (1969a)[5] introduces a bound by defining an 'envelope' of equal width that, with high probability, contains the true CDF. The upper and lower extremes of such an envelope define the CDFs with the minimum and maximum attainable means for distributions that fit within the envelope, and thus bound the mean with high probability.[6]

In practice, Anderson's bound tends to be significantly tighter than Maurer and Pontil's inequality unless the variance of the random variable is miniscule in comparison to the range of the random variable (and n is sufficiently large). However, neither Anderson's inequality nor Maurer and Pontil's inequality strictly dominates the other. That is, neither upper bound is strictly less than or equal to the other

[3] Code accompanying this paper is available at `https://github.com/myphan9/small_sample_mean_bounds`.

[4] At the time of submission, we had established that a particular version of our bound was tighter than or equal to Anderson's for both the one-ended and the two-ended settings. Subsequently, Phan et al. (2021) established that this version of our bound is in fact equivalent to Anderson's for the one-ended setting, but superior for many cases in the two-ended setting. We made minor revisions to the text to incorporate this new information.

[5] An easier to access and virtually equivalent version of Anderson's work can be found in (Anderson, 1969b).

[6] In his original paper, Anderson also suggests a large family of envelopes, each of which produces a distinct bound. Our simulation results in Section 5 are based on the equal-width envelope, but our theoretical results in Section 4 hold for all possible envelopes.

in all cases. However, Anderson's bound *does* dominate Hoeffding's inequality (Learned-Miller & Thomas, 2019).

Some authors have proposed specific envelopes for use with Anderson's technique (Diouf & Dufour, 2005; Learned-Miller & DeStefano, 2008; Romano & Wolf, 2000). However, none of these variations are shown to dominate Anderson's original bound. That is, while they give tighter intervals for some samples, they are looser for others.

Finally we mention a bound due to Fienberg et al. (1977). This bound applies to distributions on a discrete set of support points, but nothing prevents it, in theory, from being applied to an arbitrarily dense set of points on an interval such as $[0, 1]$. This bound has a number of appealing properties, and comes with a proof of guaranteed coverage. However, the main drawback is that it is currently computationally intractable, with a computation time that depends exponentially on the number of points in the support set, precluding many (if not most) practical applications.

1.2. Bounds that do not exhibit guaranteed coverage

Many bounds that are used in practice are known to violate Eq. (1) for certain distributions. These include the aforementioned Student's t method, and various bootstrap procedures, such as the bias-corrected and accelerated (BCa) bootstrap and the percentile bootstrap. See Efron & Tibshirani (1993) for details of these methods. A simple explanation of the failure of bootstrap methods for certain distributions is given by Romano & Wolf (2000, pages 757–758). Presumably if one wants guarantees of Eq. (1), one cannot use these methods (unless one has extra information about the unknown distribution).

1.3. Bounds conjectured to have guaranteed coverage

There are at least two known bounds that perform well in practice but for which no proofs of coverage are known. One of these, used in accounting procedures, is the so-called Stringer bound (Stringer, 1963). It is known to violate Eq. (1) for confidence levels $\alpha > 0.5$ (Pap & van Zuijlen, 1995), but its coverage for $\alpha < 0.5$ is unknown.

A little known bound by Gaffke (2005) gives remarkably tight bounds on the mean, but has eluded a proof of guaranteed coverage. This bound was recently rediscovered by Learned-Miller & Thomas (2019), who do an empirical study of its performance and provide a method for computing it efficiently.

We demonstrate in Section 4 that our bound dominates those of both Hoeffding and Anderson. **To our knowledge, this is the first bound that has been shown to dominate Anderson's bound.**

2. A Family of Confidence Bounds

In this section we define our new upper confidence bound. Let n be the sample size. We use bold-faced letters to denote a vector of size n and normal letters to denote a scalar. Uppercase letters denote random variables and lowercase letters denote values taken by them. For example, $X_i \in \mathcal{R}$ and $\mathbf{X} = (X_1, ..., X_n) \in \mathcal{R}^n$ are random variables. $x_i \in \mathcal{R}$ is a value of X_i, and $\mathbf{x} = (x_1, ..., x_n) \in \mathcal{R}^n$ is a value of $\mathbf{X}$. For a sample $\mathbf{x}$, we let $F(\mathbf{x}) \overset{\text{def}}{=} (F(x_1), \cdots, F(x_n)) \in [0, 1]^n$.

Order statistics play a central role in our work. We denote random variable order statistics $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ and of a specific sample as $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

Given a sample $\mathbf{X} = \mathbf{x}$ of size n and a confidence level $1 - \alpha$, we would like to calculate a UCB for the mean. Let F be the CDF of X_i, i.e., the true distribution and $D \subset \mathcal{R}$ be the support of F. We assume that D has a finite upper bound. Given D and any function $T : D^n \to \mathcal{R}$ we will calculate an upper confidence bound $b^\alpha_{D,T}(\mathbf{x})$ for the mean of F.

We show in Lemma 2.1 that if D^+ is a superset of D with finite upper bound, then $b^\alpha_{D^+,T}(\mathbf{x}) \geq b^\alpha_{D,T}(\mathbf{x})$. Therefore we only need to know a superset of the support with finite upper bound to obtain a guaranteed bound.

Let $s_D \overset{\text{def}}{=} \sup\{x : x \in D\}$. We next describe a method for pairing the sample $\mathbf{x}$ with another vector $\boldsymbol{\ell} \in [0, 1]^n$ to produce a stairstep CDF function $G_{\mathbf{x},\boldsymbol{\ell}}$. Let $x_{(n+1)} \overset{\text{def}}{=} s_D$. Consider the step function $G_{\mathbf{x},\boldsymbol{\ell}} : \mathcal{R} \to [0, 1]$ defined from $\boldsymbol{\ell}$ and $\mathbf{x}$ as follows (see Figure 1):

$$G_{\mathbf{x},\boldsymbol{\ell}}(y) = \begin{cases} 0, & \text{if } x < x_{(1)} \\ \ell_{(i)}, & \text{if } x_{(i)} \leq x < x_{(i+1)} \\ 1, & \text{if } x \geq s_D. \end{cases} \tag{4}$$

In particular, when $\boldsymbol{\ell} = (1/n, \dots, n/n)$, $G_{\mathbf{x},\boldsymbol{\ell}}$ becomes the empirical CDF. Also note that when $\boldsymbol{\ell} = F(\mathbf{x})$, $\forall x, G_{\mathbf{x},\boldsymbol{\ell}}(x) \leq F(x)$, as illustrated in Figure 2.

Following Learned-Miller & Thomas (2019), if we consider $G_{\mathbf{x},\boldsymbol{\ell}}$ to be a CDF, we can compute the mean of the resulting distribution as a function of two vectors $\mathbf{x}$ and $\boldsymbol{\ell}$ as

$$m_D(\mathbf{x}, \boldsymbol{\ell}) \overset{\text{def}}{=} \sum_{i=1}^{n+1} x_{(i)}(\ell_{(i)} - \ell_{(i-1)}) \tag{5}$$

$$= s_D - \sum_{i=1}^{n} \ell_{(i)}(x_{(i+1)} - x_{(i)}), \tag{6}$$

where $\ell_{(0)} \overset{\text{def}}{=} 0$, $\ell_{(n+1)} \overset{\text{def}}{=} 1$ and $x_{(n+1)} \overset{\text{def}}{=} s_D$. When s_D is finite, this is well-defined. Notice that this function is defined in terms of the *order statistics* of $\mathbf{x}$ and $\boldsymbol{\ell}$. Learned-Miller & Thomas (2019) refer to this as the *induced mean*

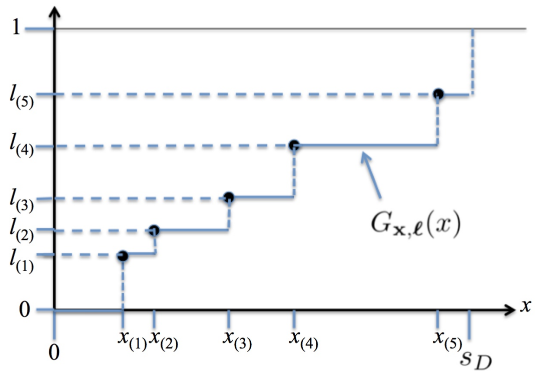

Figure 1. The stairstep function $G_{\mathbf{x},\boldsymbol{\ell}}$, which is a function of the sample $\mathbf{x}$ and a vector $\boldsymbol{\ell}$ of values between 0 and 1. When $\boldsymbol{\ell} = (1/n, \ldots, n/n)$, $G_{\mathbf{x},\boldsymbol{\ell}}$ becomes the empirical CDF.

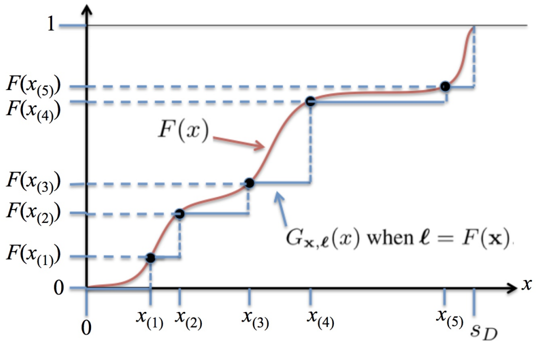

Figure 2. The CDF of a distribution F in red, with a random sample of five order statistics on the x-axis. The blue stairstep function shows the function $G_{\mathbf{x},\boldsymbol{\ell}}(x)$ when $\boldsymbol{\ell} = F(\mathbf{x})$. Notice that for all x, $G_{\mathbf{x},\boldsymbol{\ell}}(x) \leq F(x)$.

for the sample $\mathbf{x}$ by the vector $\boldsymbol{\ell}$. Although we borrow the above terms from Learned-Miller & Thomas (2019), the bound we introduce below is a new class of bounds, and differs from the bounds discussed in their work.

An ordering on D^n. Next, we introduce a scalar-valued function T which we will use to define a total order on samples in D^n, and define a set of samples less than or equal to another sample. In particular, for any function $T : \mathcal{R}^n \to \mathcal{R}$, let $\mathbb{S}_{D,T}(\mathbf{x}) = \{\mathbf{y} \in D^n | T(\mathbf{y}) \leq T(\mathbf{x})\}$.

The greatest induced mean for a given U. Let $\mathbf{U} = U_1, ..., U_n$ be a sample of size n from the continuous uniform distribution on $[0, 1]$, with $\mathbf{u} \overset{\text{def}}{=} (u_1, \cdots, u_n)$ being a particular sample of $\mathbf{U}$.

Now consider the random quantity

$$b_{D,T}(\mathbf{x}, \mathbf{U}) \overset{\text{def}}{=} \sup_{\mathbf{z} \in \mathbb{S}_{D,T}(\mathbf{x})} m_D(\mathbf{z}, \mathbf{U}), \tag{7}$$

which depends upon a fixed sample $\mathbf{x}$ (non-random) and also on the random variable $\mathbf{U}$.

Our upper confidence bound. Let $0 < p < 1$. Let $Q(p, Y)$ be the *quantile function* of the scalar random variable Y, i.e.,

$$Q(p, Y) \overset{\text{def}}{=} \inf\{y \in \mathbb{R} : F_Y(y) \geq p\}, \tag{8}$$

where $F_Y(y)$ is the CDF of Y. We define $b^\alpha_{D,T}(\mathbf{x})$ to be the $(1 - \alpha)$-quantile of the random quantity $b_{D,T}(\mathbf{x}, \mathbf{U})$.

Definition 2.1 (Upper confidence bound on the mean)**.** *Given a sample* $\mathbf{x}$ *and a confidence level* $1 - \alpha$*:*

$$b^\alpha_{D,T}(\mathbf{x}) \overset{\text{def}}{=} Q(1 - \alpha, b_{D,T}(\mathbf{x}, \mathbf{U})), \tag{9}$$

where $b_{D,T}(\mathbf{x}, \mathbf{U})$ *is defined in Eq. 7.*

To simplify notation, we drop the superscript and subscripts whenever clear. We show in Section 2.1 that this UCB has guaranteed coverage for all sample sizes n, for all confidence levels $0 < 1 - \alpha < 1$ and for all distributions F and support D where s_D is finite.

We show below that a bound computed from a superset $D^+ \supseteq D$ will be looser than or equal to a bound computed from the support D. Therefore it is enough to know a superset of the support D to obtain a bound with guaranteed coverage.

Lemma 2.1. *Let* $D^+ \supseteq D$ *where* s_{D^+} *is finite. For any sample* $\mathbf{x}$*:*

$$b^\alpha_D(\mathbf{x}) \leq b^\alpha_{D^+}(\mathbf{x}). \tag{10}$$

Proof. Since s_{D^+} is finite, $m_{D^+}(\mathbf{y}, \mathbf{u})$ is well-defined. Since $D \subseteq D^+$, for any $\mathbf{y}$ and $\mathbf{u}$, $m_D(\mathbf{y}, \mathbf{u}) \leq m_{D^+}(\mathbf{y}, \mathbf{u})$. Then

$$\sup_{\mathbf{y} \in \mathbb{S}_D(\mathbf{x})} m_D(\mathbf{y}, \mathbf{u}) \leq \sup_{\mathbf{y} \in \mathbb{S}_D(\mathbf{x})} m_{D^+}(\mathbf{y}, \mathbf{u}) \tag{11}$$

$$\leq \sup_{\mathbf{y} \in \mathbb{S}_{D^+}(\mathbf{x})} m_{D^+}(\mathbf{y}, \mathbf{u}), \tag{12}$$

where the last inequality is because $\mathbb{S}_D(\mathbf{x}) \subseteq \mathbb{S}_{D^+}(\mathbf{x})$. Let $b_D(\mathbf{x}, \mathbf{U}) = \sup_{\mathbf{z} \in \mathbb{S}_D(\mathbf{x})} m_D(\mathbf{z}, \mathbf{U})$ and $b_{D^+}(\mathbf{x}, \mathbf{U}) = \sup_{\mathbf{z} \in \mathbb{S}_{D^+}(\mathbf{x})} m_{D^+}(\mathbf{z}, \mathbf{U})$. Then $b^\alpha_D(\mathbf{x})$ and $b^\alpha_{D^+}(\mathbf{x})$ are the $(1 - \alpha)$-quantiles of $b_D(\mathbf{x}, \mathbf{U})$ and $b_{D^+}(\mathbf{x}, \mathbf{U})$. Since $b_D(\mathbf{x}, \mathbf{u}) \leq b_{D^+}(\mathbf{x}, \mathbf{u})$ for any $\mathbf{u}$, $b^\alpha_D(\mathbf{x}) \leq b^\alpha_{D^+}(\mathbf{x})$. □

In Section 2.1 we show that the bound has guaranteed coverage. In Section 3 we discuss how to efficiently compute the bound. In Section 4 we show that when T is a certain linear function, the bound is equal to or tighter than Anderson's for any sample. In addition, we show that when the support is known to be $\{0, 1\}$, our bound recovers the well-known Clopper-Pearson confidence bound for binomial distributions (Clopper & Pearson, 1934). In Section 5, we present simulations that show the consistent superiority of our bounds over previous bounds.

2.1. Guaranteed Coverage

In this section we show that our bound has guaranteed coverage in Theorem 2.7. We omit superscripts and subscripts if they are clear from context.

2.1.1. Preview of proof

We explain the idea behind our bound at a high level using a special case. Note that our proof is more general than our special case, which makes assumptions such as the continuity of F to simplify the intuition.

Suppose that F is continuous. Then the *probability integral transform* $F_X(X)$ of X is uniformly distributed on $[0,1]$ (Angus, 1994). Suppose there exists a sample $\mathbf{x}_\mu$ such that $b^\alpha(\mathbf{x}_\mu) = \mu$. Then the probability that a sample $\mathbf{Z}$ outputs $b^\alpha(\mathbf{Z}) < \mu$ is equal to the probability $\mathbf{Z}$ outputs $b^\alpha(\mathbf{Z}) < b^\alpha(\mathbf{x}_\mu)$ (the yellow region on the left of Fig. 3). This is the region where the bound fails, and we would like to show that the probability of this region is at most α.

Let $\mathbf{U} \overset{\text{def}}{=} F(\mathbf{Z})$ and $\mathbf{u} \overset{\text{def}}{=} F(\mathbf{z})$. Then U_i is uniformly distributed on $[0,1]$. If F is invertible, we can transform the region $\{\mathbf{z} : b^\alpha(\mathbf{z}) < b^\alpha(\mathbf{x}_\mu)\}$ to $\{\mathbf{u} : b^\alpha(F^{-1}(\mathbf{u})) < b^\alpha(\mathbf{x}_\mu)\}$ where $F^{-1}(\mathbf{u}) \overset{\text{def}}{=} (F^{-1}(u_1), \ldots, F^{-1}(u_n))$ (the yellow region on the right of Fig. 3).

Through some calculations using the definition of function b, we can show that the yellow region $\{\mathbf{u} : b^\alpha(F^{-1}(\mathbf{u})) < b^\alpha(\mathbf{x}_\mu)\}$ is a subset of the striped region $\{\mathbf{u} : b(\mathbf{x}_\mu, \mathbf{u}) \geq \mu\}$.

Note that since $b^\alpha(\mathbf{x}_\mu) = \mu$, μ is equal to the $1-\alpha$ quantile of $b(\mathbf{x}_\mu, \mathbf{U})$. Therefore, by the definition of quantile, the probability of the striped region is at most α:

$$\mathbb{P}_{\mathbf{U}}(b(\mathbf{x}_\mu, \mathbf{U}) \geq \mu) \leq \alpha, \tag{13}$$

and thus the probability of the yellow region is at most α.

2.1.2. Main Result

In this section, we present some supporting lemmas and then the main result in Theorem 2.7. The proofs of the simpler lemmas have been deferred to the supplementary material.

Lemma 2.2. *Let X be a random variable with CDF F and $Y \overset{\text{def}}{=} F(X)$, known as the probability integral transform of X. Let U be a uniform random variable on $[0,1]$. Then for any $0 \leq y \leq 1$,*

$$\mathbb{P}(Y \leq y) \leq \mathbb{P}(U \leq y). \tag{14}$$

If F is continuous, then Y is uniformly distributed on $[0,1]$.

The next lemma is illustrated by Fig. 2. It shows that by building a 'stairstep CDF' using a random sample and points on the true CDF, the resulting distribution has a mean greater than or equal to the original distribution's mean.

Lemma 2.3. *For any $\mathbf{x} \in D^n$,*[7]

$$m_D(\mathbf{x}, F(\mathbf{x})) \geq \mu. \tag{15}$$

For use in the next lemma, we define a partial order for the samples on D^n. Note that it is defined with respect to the *order statistics* of the sample, not the original components.

Definition 2.2 (Partial Order)**.** *For any two samples $\mathbf{z}$ and $\mathbf{y}$, we define $\mathbf{z} \preceq \mathbf{y}$ to indicate that $z_{(i)} \leq y_{(i)}, 1 \leq i \leq n$.*

Lemma 2.4. *Let $\mathbf{Z}$ be a random sample of size n from F. Let $\mathbf{U} = U_1, ..., U_n$ be a sample of size n from the continuous uniform distribution on $[0,1]$. For any function $T : D^n \to R$ and any $\mathbf{x} \in D^n$:*

$$\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{x})) \leq \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \geq \mu). \tag{16}$$

Proof sketch. Let $\cup$ denote the union of events and $\{\}$ denote an event. Then for any $\mathbf{x} \in D^n$:

$$\begin{aligned}
& \mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{x})) && (17)\\
= \; & \mathbb{P}_{\mathbf{Z}}(\mathbf{Z} \in \mathbb{S}(\mathbf{x})) && (18)\\
= \; & \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{\mathbf{Z} = \mathbf{y}\}) && (19)\\
\leq \; & \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{\mathbf{Z} \preceq \mathbf{y}\}) && (20)\\
\leq \; & \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{F(\mathbf{Z}) \preceq F(\mathbf{y})\}) \text{ by monotone } F && (21)\\
\leq \; & \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{\mathbf{U} \preceq F(\mathbf{y})\}). && (22)
\end{aligned}$$

The last step is by an extension of Lemma 2.2. Recall that $m_D(\mathbf{y}, \mathbf{u}) = s_D - \sum_{i=1}^n u_{(i)}(y_{(i+1)} - y_{(i)})$ where $\forall i, y_{(i+1)} - y_{(i)} \geq 0$. Therefore if $\mathbf{u} \preceq F(\mathbf{y})$ then $m_D(\mathbf{y}, \mathbf{u}) \geq m_D(\mathbf{y}, F(\mathbf{y}))$:

$$\begin{aligned}
& \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{\mathbf{U} \preceq F(\mathbf{y})\}) && (23)\\
\leq \; & \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{m_D(\mathbf{y}, \mathbf{U}) \geq m_D(\mathbf{y}, F(\mathbf{y}))\}) && (24)\\
\leq \; & \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} \{m_D(\mathbf{y}, \mathbf{U}) \geq \mu\}), \text{by Lem. 2.3} && (25)\\
\leq \; & \mathbb{P}_{\mathbf{U}}(\sup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})} m_D(\mathbf{y}, \mathbf{U}) \geq \mu) && (26)\\
= \; & \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \geq \mu). && (27)
\end{aligned}$$

□

We include a more detailed version of the proof for the above lemma in the supplementary material.

Lemma 2.5. *Let $\mathbf{U} = U_1, ..., U_n$ be a sample of size n from the continuous uniform distribution on $[0,1]$. Let $\mathbf{X}$ and $\mathbf{Z}$ denote i.i.d. samples of size n from F. For any function $T : D^n \to \mathcal{R}$ and any $\alpha \in (0,1)$,*

$$\begin{aligned}
& \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{U}}(b_{D,T}(\mathbf{X}, \mathbf{U}) \geq \mu) \leq \alpha\right) \\
& \leq \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{X})) \leq \alpha\right). \qquad (28)
\end{aligned}$$

[7] Phan et al. (2021) show a more general property that is also satisfied by the quantile. Thus this method could also be used to give confidence intervals for the quantile.

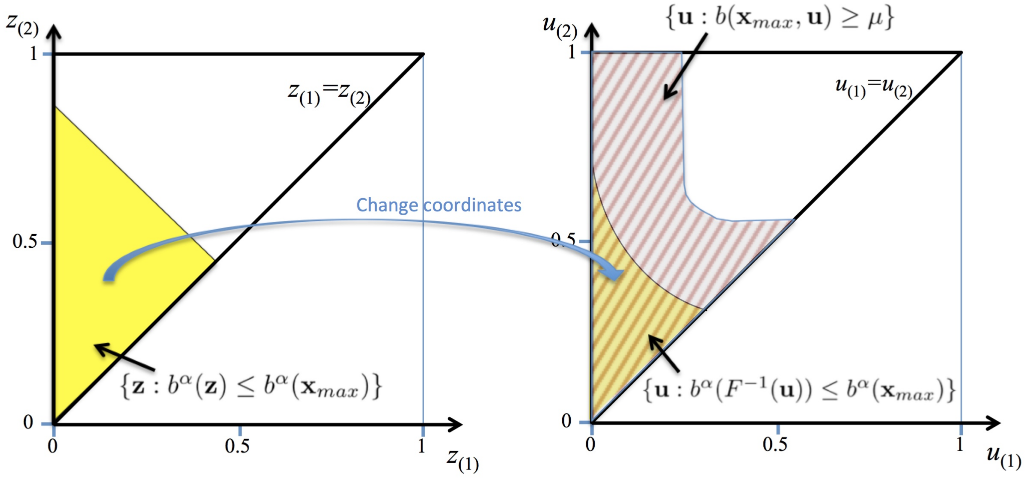
$z_{(2)}$
1
0.5
0
0
0.5
1
$z_{(1)}$
$z_{(1)}=z_{(2)}$
$\{\mathbf{z} : b^\alpha(\mathbf{z}) \le b^\alpha(\mathbf{x}_{max})\}$
Change coordinates
$u_{(2)}$
1
0.5
0
0
0.5
1
$u_{(1)}$
$u_{(1)}=u_{(2)}$
$\{\mathbf{u} : b(\mathbf{x}_{max}, \mathbf{u}) \ge \mu\}$
$\{\mathbf{u} : b^\alpha(F^{-1}(\mathbf{u})) \le b^\alpha(\mathbf{x}_{max})\}$

Figure 3. Illustrations of Section 2.1.1. **Left.** The yellow region shows samples of $\mathbf{z} = [z_{(1)}, z_{(2)}]$ such that $b^\alpha(\mathbf{z}) \le b^\alpha(\mathbf{x}_{max})$. **Right.** The same yellow region, but in the coordinates $\mathbf{u} = F^{-1}(\mathbf{z})$. We will show that the yellow region is a subset of the striped, which contains $\mathbf{u}$ such that $b(\mathbf{x}_{max}, \mathbf{u}) \ge \mu$.

Proof. From Lemma 2.4 for any sample $\mathbf{x}$,

$$\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \le T(\mathbf{x})) \le \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \ge \mu). \tag{29}$$

Therefore,

$$\mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \le T(\mathbf{X})) \le \alpha\right) \tag{30}$$
$$\ge \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{U}}(b(\mathbf{X}, \mathbf{U}) \ge \mu) \le \alpha\right). \tag{31}$$

□

Lemma 2.6. *Let $\mathbf{U} = U_1, ..., U_n$ be a sample of size n from the continuous uniform distribution on $[0, 1]$. Let $\mathbf{X}$ be a random sample of size n from F. For any function $T : D^n \to \mathcal{R}$ and any $\alpha \in (0, 1)$,*

$$\mathbb{P}_{\mathbf{X}}(b^\alpha_{D,T}(\mathbf{X}) < \mu) \tag{32}$$
$$\le \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{U}}(b_{D,T}(\mathbf{X}, \mathbf{U}) \ge \mu) \le \alpha\right). \tag{33}$$

Proof. Because $b^\alpha(\mathbf{x})$ is the $1 - \alpha$ quantile of $b(\mathbf{x}, \mathbf{U})$, by the definition of quantile: $\mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \le b^\alpha(\mathbf{x})) \ge 1 - \alpha$. Therefore $\mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \ge b^\alpha(\mathbf{x})) \le \alpha$. If $b^\alpha(\mathbf{x}) < \mu$ then $\mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \ge \mu) \le \alpha$. Since $b^\alpha(\mathbf{x}) < \mu$ implies $\mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \ge \mu) \le \alpha$, we have

$$\mathbb{P}_{\mathbf{X}}(b^\alpha(\mathbf{X}) < \mu) \tag{34}$$
$$\le \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{U}}(b(\mathbf{X}, \mathbf{U}) \ge \mu) \le \alpha\right). \tag{35}$$

□

We now show that the bound has guaranteed coverage.

Theorem 2.7. *Let $\mathbf{X}$ be a random sample of size n from F. For any function $T : D^n \to R$ and for any $\alpha \in (0, 1)$:*

$$\mathbb{P}_{\mathbf{X}}(b^\alpha_{D,T}(\mathbf{X}) < \mu) \le \alpha. \tag{36}$$

Proof. Let $\mathbf{Z}$ be a random sample of size n from F.

$$\mathbb{P}_{\mathbf{X}}(b^\alpha(\mathbf{X}) < \mu) \tag{37}$$
$$\le \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{U}}(b(\mathbf{X}, \mathbf{U}) \ge \mu) \le \alpha\right) \text{ by Lemma 2.6} \tag{38}$$
$$\le \mathbb{P}_{\mathbf{X}}\left(\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \le T(\mathbf{X})) \le \alpha\right) \text{ by Lemma 2.5} \tag{39}$$
$$= \mathbb{P}\left(W \le \alpha\right) \text{ where } W \overset{\text{def}}{=} \mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \le T(\mathbf{X})) \tag{40}$$
$$\le \alpha \text{ by Lemma 2.2.} \tag{41}$$

□

3. Computation

In this section we present a Monte Carlo algorithm to compute the bound. First we note that since the bound only depends on $\mathbf{x}$ via the function $T(\mathbf{x})$, we can precompute a table of the bounds for each value of $T(\mathbf{x})$. We discuss how to adjust for the uncertainty in the Monte Carlo result in Appendix D.

Let the superset of the support D^+ be a closed interval with a finite upper bound. If m is a continuous function,

$$\sup_{\mathbf{y} \in \mathbf{S}_{D+}(\mathbf{x})} m(\mathbf{y}, \mathbf{u}) = \max_{\mathbf{y} \in \mathbf{S}_{D+}(\mathbf{x})} m(\mathbf{y}, \mathbf{u}). \tag{42}$$

Therefore $b_{D+}(\mathbf{x}, \mathbf{u})$ is the solution to

$$\max_{y_{(1)}, \ldots, y_{(n)}} m(\mathbf{y}, \mathbf{u}) \tag{43}$$

subject to:

1. $T(\mathbf{y}) \le T(\mathbf{x})$,
2. $\forall i \in \{1, \ldots, n\}, y_{(i)} \in D^+$,
3. $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$.

When D^+ is an interval and T is linear, this is a linear programming problem and can be solved efficiently.

Algorithm 1 Monte Carlo estimation of $m^{\alpha}_{D^+,T}(\mathbf{x})$ where $D^+ = [0,1]$. This pseudocode uses 1-based array indexing.

```
**Input:** A sample $\mathbf{x} \in D^n$, confidence parameter $1-\alpha < 1$, a function $T : [0,1]^n \to \mathcal{R}$ and Monte Carlo sampling parameter $l$.
**Output:** An estimation of $m^{\alpha}_{D^+,T}(\mathbf{x})$
$n \leftarrow length(\mathbf{x})$.
Create array **ms** to hold $l$ floating point numbers, and initialize it to zero.
Create array **u** to hold $n$ floating point numbers.
for $i \leftarrow 1$ to $l$ do
  for $j \leftarrow 1$ to $n$ do
    **u**[j] $\sim$ Uniform(0,1).
  end for
  Sort(**u**, ascending).
  Solve: $M = \max_{y_{(1)},\cdots,y_{(n)}} m(\mathbf{y}, \mathbf{u})$ subject to:
    1) $T(\mathbf{y}) \le T(\mathbf{x})$.
    2) $\forall i : 1 \le i \le n, 0 \le y_{(i)} \le 1$.
    3) $y_{(1)} \le y_{(2)} \le ... \le y_{(n)}$.
  $\mathbf{ms}[i] = M$.
end for
Sort(**ms**, ascending).
Return $\mathbf{ms}[\lceil (1-\alpha) l \rceil]$.
```

We can compute the $1-\alpha$ quantile of a random variable M using Monte Carlo simulation, sampling M l times. Letting $m_{(1)} \le ... \le m_{(l)}$ be the sorted values, we output $m_{(\lceil (1-\alpha) l \rceil)}$ as an approximation of the $1-\alpha$ quantile.

Running time. When T is linear, the algorithm needs to solve a linear programming problem with n variables and $2n$ constraints l times. For sample size $n = 50$, computing the bound for each sample $\mathbf{x} \in D^n$ takes just a few seconds using $l = 10{,}000$ Monte Carlo samples.

4. Relationships with Existing Bounds

In this section, we compare our bound to previous bounds including those of Clopper and Pearson, Hoeffding, and Anderson. Proofs omitted in this section can be found in the supplementary material.

4.1. Special Case: Bernoulli Distribution

When we know that $D = \{0, 1\}$, the distribution is Bernoulli. If we choose T to be the sample mean, our bound becomes the same as the Clopper-Pearson confidence bound for binomial distributions (Clopper & Pearson, 1934). See the supplementary material for details.

4.2. Comparisons with Anderson and Hoeffding

In this section we show that for any sample size n, any confidence level α and for any sample $\mathbf{x}$, our method produces a bound no larger than Anderson's bound (Theorem 4.3) and Hoeffding's bound (Theorem 4.4).

Note that if we only know an upper bound b of the support (1-ended support setting), we can set $D^+ = (-\infty, b]$ and our method is equal to Anderson's (Phan et al., 2021) and dominates Hoeffding's. As the lower support bound increases (2-ended setting), our bound becomes tighter or remains constant, whereas Anderson's remains constant, as it does not incorporate information about a lower support. Thus, in cases where our bound can benefit from a lower support, we are tighter than Anderson's.

Anderson's bound constructs an upper bound for the mean by constructing a lower bound for the CDF. We defined a lower bound for the CDF as follows.

Definition 4.1 (Lower confidence bound for the CDF)**.** *Let* $\mathbf{X} = (X_1, \cdots, X_n)$ *be a sample of size* n *from the distribution on* $\mathcal{D}^+$ *with unknown CDF* F*. Let* $\alpha \in (0,1)$*. Let* $H_{\mathbf{X}} : \mathcal{R} \to [0,1]$ *be a function computed from the sample* $\mathbf{X}$ *such that for any CDF* F*,*

$$\mathbb{P}_{\mathbf{X}}(\,\forall x \in R, F(x) \ge H_{\mathbf{X}}(x)) \ge 1-\alpha. \tag{44}$$

Then $H_{\mathbf{X}}$ *is called a* $(1-\alpha)$ *lower confidence bound for the CDF.*

If there exists a CDF F *such that*

$$\mathbb{P}_{\mathbf{X}}(\,\forall x \in R, F(x) \ge H_{\mathbf{X}}(x)) = 1-\alpha, \tag{45}$$

then $H_{\mathbf{X}}$ *is called an* exact $(1-\alpha)$ *lower confidence bound for the CDF.*

In Figs. 1 and 2, it is easy to see that if the stairstep function $G_{\mathbf{X},\boldsymbol{\ell}}$ is a lower confidence bound for the CDF then its induced mean $m(\mathbf{X}, \boldsymbol{\ell})$ is an upper confidence bound for μ.

Lemma 4.1. *Let* $\mathbf{X} = (X_1, \cdots, X_n)$ *be a sample of size* n *from a distribution with mean* μ*. Let* $\boldsymbol{\ell} \in [0,1]^n$*. If* $G_{\mathbf{X},\boldsymbol{\ell}}$ *is a* $(1-\alpha)$ *lower confidence bound for the CDF then*

$$\mathbb{P}_{\mathbf{X}}(m(\mathbf{X}, \boldsymbol{\ell}) \ge \mu) \ge 1-\alpha. \tag{46}$$

Let $U_{(i)}, 1 \le i \le n$ be the order statistics of the uniform distribution. Note that for any CDF F:

$$\begin{aligned} &\mathbb{P}_{\mathbf{X}}(\,\forall x \in \mathcal{R}, F(x) \ge G_{\mathbf{X},\boldsymbol{\ell}}(x)) && (47)\\ &= \mathbb{P}_{\mathbf{X}}(\forall i : 1 \le i \le n, F(X_{(i)}) \ge \ell_{(i)}) && (48)\\ &\ge \mathbb{P}_{\mathbf{U}}(\forall i : 1 \le i \le n, U_{(i)} \ge \ell_{(i)}) \text{ by Lemma 2.2}, && (49)\end{aligned}$$

where Eq. 49 is an equality if F is the CDF of a continuous random variable. Therefore $G_{\mathbf{X},\boldsymbol{\ell}}$ is an exact $(1-\alpha)$ lower confidence bound for the CDF is equivalent to $\boldsymbol{\ell}$ satisfying:

$$\mathbb{P}_{\mathbf{U}}(\forall i : 1 \le i \le n, U_{(i)} \ge \ell_{(i)}) = 1-\alpha. \tag{50}$$

Anderson (1969a) presents $b_{\boldsymbol{\ell}}^{\alpha,\text{Anderson}}(\mathbf{x}) = m_{D^+}(\mathbf{x}, \boldsymbol{\ell})$ as a UCB for μ where $\boldsymbol{\ell} \in [0,1]^n$ is a vector such that $G_{\mathbf{X},\boldsymbol{\ell}}$ is an exact $(1-\alpha)$ lower confidence bound for the CDF.

In one instance of Anderson's bound, $\boldsymbol{\ell} = \mathbf{u}^{And} \in [0,1]^n$ is defined as

$$u_i^{\text{And}} \overset{\text{def}}{=} \max\left\{0, i/n - \beta(n)\right\}. \tag{51}$$

Anderson identifies $\beta(n)$ as the one-sided Kolmogorov-Smirnov statistic such that $G_{\mathbf{X},\boldsymbol{\ell}}$ is an exact $(1-\alpha)$ lower confidence bound for the CDF when $\boldsymbol{\ell} = \mathbf{u}^{\text{And}}$. $\beta(n)$ can be computed by Monte Carlo simulation (Appendix A).

Learned-Miller & Thomas (2019) show that for any sample $\mathbf{x}$, a looser version of Anderson's bound is better than Hoeffding's:

Lemma 4.2 (from Theorem 2 from (Learned-Miller & Thomas, 2019))**.** *For any sample size n, for any sample value $\mathbf{x} \in D^n$, for all $\alpha \in (0, 0.5]$:*

$$b_{\boldsymbol{\ell}}^{\alpha,\textit{Anderson}}(\mathbf{x}) \leq b^{\alpha,\textit{Hoeffding}}(\mathbf{x}), \tag{52}$$

where $\boldsymbol{\ell}$ is defined[8] *as*

$$\ell_i \overset{\text{def}}{=} \max\left\{0, i/n - \sqrt{\ln(1/\alpha)/(2n)}\right\}. \tag{53}$$

When $\alpha \leq 0.5$, this definition of $\boldsymbol{\ell}$ satisfies $G_{\mathbf{X},\boldsymbol{\ell}}$ is a $(1-\alpha)$ lower confidence bound for the CDF.

The inequality in Eq. 52 is strict for $n \geq 3$.

We show below that our bound is always equal to or tighter than Anderson's bound. Phan et al. (2021) provide a more detailed analysis showing that our bound is equal to Anderson's when the lower bound of the support is too small and can be tighter than Anderson's when the lower bound of the support is large enough.

Theorem 4.3. *Let $\boldsymbol{\ell} \in [0,1]^n$ be a vector satisfying $G_{\mathbf{X},\boldsymbol{\ell}}$ is an exact $(1-\alpha)$ lower confidence bound for the CDF.*

Let $D^+ = (-\infty, b]$. For any sample size n, for any sample value $\mathbf{x} \in D^n$, for all $\alpha \in (0,1)$, using $T(\mathbf{x}) = b_{\boldsymbol{\ell}}^{\alpha,\textit{Anderson}}(\mathbf{x})$ yields

$$b_{D^+,T}^{\alpha}(\mathbf{x}) \leq b_{\boldsymbol{\ell}}^{\alpha,\textit{Anderson}}(\mathbf{x}). \tag{54}$$

We explain briefly why this is true. First, from Figure 2, we can see that if $G_{\mathbf{X},\boldsymbol{\ell}}$ is a lower confidence bound then $\forall i, F(X_{(i)}) \geq \ell_{(i)}$. Note that $G_{\mathbf{X},\boldsymbol{\ell}}$ must be a lower bound for all unknown CDFs F, so we can pick a continuous F where, according to Lemma 2.2, $U \overset{\text{def}}{=} F(X)$ is uniformly distributed on $[0,1]$. Therefore $\boldsymbol{\ell}$ satisfies

$$\mathbb{P}_{\mathbf{U}}(\forall i, U_{(i)} \geq \ell_{(i)}) \geq 1-\alpha, \tag{55}$$

where the $U_{(i)}$'s are the order statistics of the uniform distribution. Since $b(\mathbf{x}, \mathbf{U})$ is defined from linear functions of $\mathbf{U}$ with negative coefficients (Eq. 6), if $\forall i, U_{(i)} \geq \ell_{(i)}$ then $b(\mathbf{x}, \mathbf{U}) \leq b(\mathbf{x}, \boldsymbol{\ell})$. Therefore with probability at least $1-\alpha$, $b(\mathbf{x}, \mathbf{U}) \leq b(\mathbf{x}, \boldsymbol{\ell})$. So $b(\mathbf{x}, \boldsymbol{\ell})$ is at least the $1-\alpha$ quantile of $b(\mathbf{x}, \mathbf{U})$, which is the value of our bound. Therefore $b(\mathbf{x}, \boldsymbol{\ell})$ is at least the value of our bound.

Finally, if T is Anderson's bound, through some calculations we can show that $b_{D^+,T}(\mathbf{x}, \boldsymbol{\ell}) = m_{D^+}(\mathbf{x}, \boldsymbol{\ell})$, which is Anderson's bound. The result follows.

The comparison with Hoeffding's bound follows directly from Lemma 4.2 and Theorem 4.3:

Theorem 4.4. *Let $D^+ = (-\infty, b]$. For any sample size n, for any sample value $\mathbf{x} \in D^n$, for all $\alpha \in (0, 0.5]$, using $T(\mathbf{x}) = b_{\boldsymbol{\ell}}^{\alpha,\textit{Anderson}}(\mathbf{x})$ where $\boldsymbol{\ell} = \mathbf{u}^{\textit{And}}$ yields:*

$$b_{D^+,T}^{\alpha}(\mathbf{x}) \leq b^{\alpha,\textit{Hoeffding}}(\mathbf{x}), \tag{56}$$

where the inequality is strict when $n \geq 3$.

Diouf & Dufour (2005) present several instances of Anderson's bound with different $\boldsymbol{\ell}$ computed from the Anderson-Darling or the Eicker statistics (Theorem 4, 5 and Theorem 6 with constant ϵ).

Note that the result from Theorem 4.3 can be generalized for bounds $m(\mathbf{X}, \boldsymbol{\ell})$ constructed from a $(1-\alpha)$ confidence lower bound $G_{\mathbf{X},\boldsymbol{\ell}}$ using Lemma 4.1. We show the general case in the supplementary material.

5. Simulations

We perform simulations to compare our bounds to Hoeffding's inequality, Anderson's bound, Maurer and Pontil's, and Student-t's bound (Student, 1908), the latter being

$$b^{\alpha,\text{Student}}(\mathbf{X}) \overset{\text{def}}{=} \bar{X}_n + \sqrt{\frac{\hat{\sigma}^2}{n}}t_{1-\alpha,n-1}. \tag{57}$$

We compute Anderson's bound with $\boldsymbol{\ell} = \mathbf{u}^{And}$ defined in Eq. 51 through Monte Carlo simulation (described in Appendix A). We use $\alpha = 0.05$, $D^+ = [0,1]$ and $l =$ 10,000 Monte Carlo samples. We consider two functions T:

1. Anderson: $T(\mathbf{x}) = b_{\boldsymbol{\ell}}^{\alpha,\text{Anderson}}(\mathbf{x})$, again with $\boldsymbol{\ell} = \mathbf{u}^{And}$. Because this T is linear in $\mathbf{x}$, it can be computed with the linear program in Eq. 42.
2. $l2$ norm: $T(\mathbf{x}) = (\sum_{i=1}^n x_i^2)/n$. In this case, T requires the optimization of a linear functional over a

[8] Although Anderson's bound $b_{\boldsymbol{\ell}}^{\alpha,\text{Anderson}}(\mathbf{x})$ is only defined when $G_{\mathbf{X},\boldsymbol{\ell}}$ is an exact $(1-\alpha)$ lower confidence bound for the CDF, here we re-use the same notation for the case when $G_{\mathbf{X},\boldsymbol{\ell}}$ is a $(1-\alpha)$ lower confidence bound for the CDF.

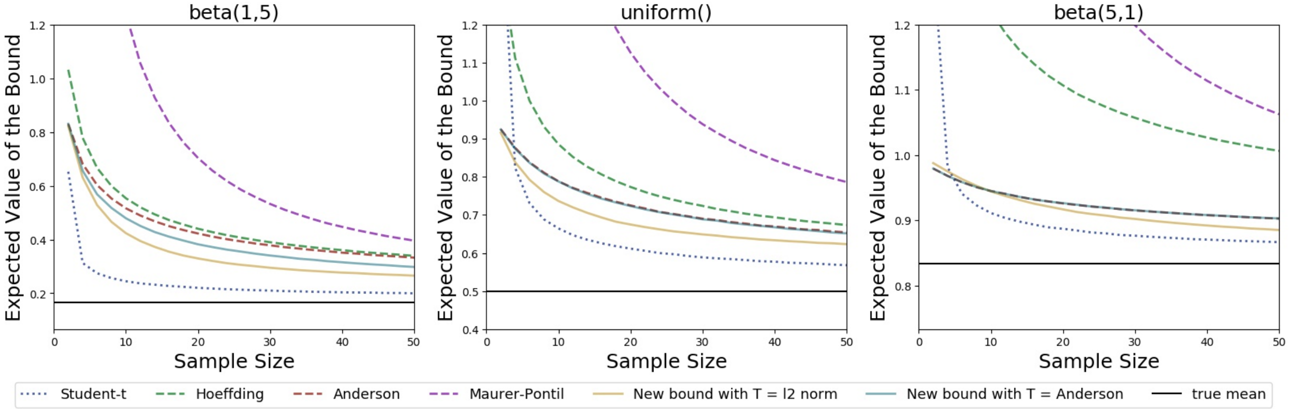

Figure 4. The expected value of the bounds for $\alpha = 0.05$ and $D^+ = [0, 1]$. For each sample size, we sample $\mathbf{X}$ 10,000 times, compute the bound for each sample, and take the average. Our new bound with T being Anderson's bound consistently has lower expected value than Anderson's (Theorem 4.3), Hoeffding's (Theorem 4.4) and Maurer and Pontil's. With T being the $l2$-norm, the bound is substantially tighter in these examples, and also has guaranteed coverage.

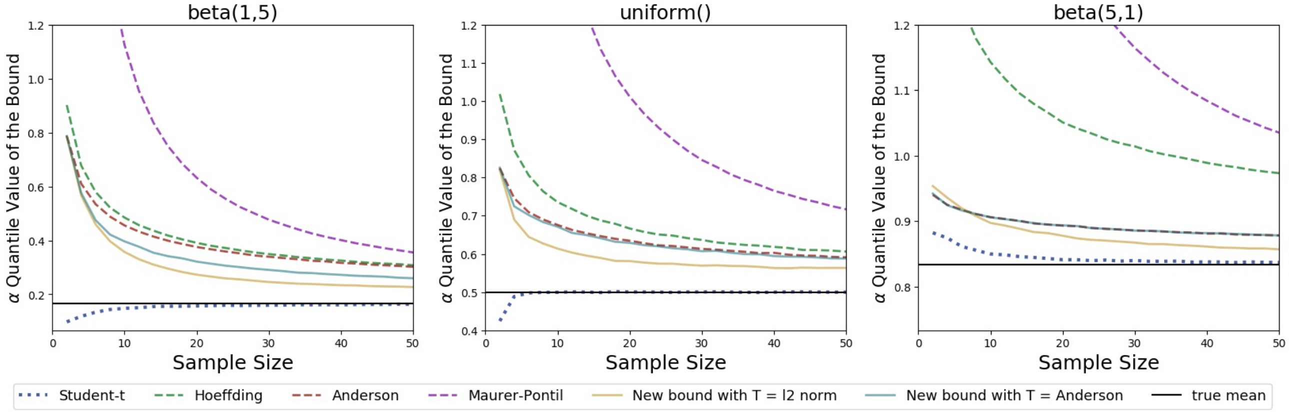

Figure 5. The α-quantile of the bound distribution for $\alpha = 0.05$ and $D^+ = [0, 1]$. For each sample size, we sample $\mathbf{X}$ 10,000 times, compute the bound for each sample, and take the α quantile. If the α-quantile is below the true mean, the bound does not have guaranteed coverage. For the $\texttt{uniform}(0, 1)$ and $\texttt{beta}(1, 5)$ distribution, when the sample size is small, Student-t does not have guarantee.

convex region, which results in a simple convex optimization problem.

We perform experiments on three distributions: $\texttt{beta}(1, 5)$ (skewed right), $\texttt{uniform}(0, 1)$ and $\texttt{beta}(5, 1)$ (skewed left). Their PDFs are included in the supplementary material for reference. Additional experiments are in the supplementary material.

In Figure 4 and Figure 5 we plot the expected value and the α-quantile value of the bounds as the sample size increases. Consistent with Theorem 4.3, our bound with T being Anderson's bound outperforms Anderson's bound. Our new bound performs better than Anderson's in distributions that are skewed right, and becomes similar to Anderson's in left-skewed distributions. Our bound outperforms Hoeffding and Maurer and Pontil's for all three distributions. Student-t fails (the error rate exceeds α) for $\texttt{beta}(1, 5)$ and $\texttt{uniform}(0, 1)$ when the sample size is small (Figure 5).

Acknowledgements

This work was partially supported by DARPA grant FA8750-18-2-0117. Research reported in this paper was sponsored in part by NSF award #2018372 and the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Anderson, T. W. 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, 1969a.

Anderson, T. W. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Technical Report Number 1, Department of Statistics, Stanford University*, 1969b.

Angus, J. E. The probability integral transform and related results. *SIAM Rev.*, 36(4):652–654, December 1994. ISSN 0036-1445. doi: 10.1137/1036146. URL https://doi.org/10.1137/1036146.

Bennett, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Clopper, C. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Diouf, M. A. and Dufour, J. M. 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.

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

Fienberg, S. E., Neter, J., and Leitch, R. A. Estimating the total overstatement error in accounting populations. *Journal of the American Statistical Association*, 72(358): 295–302, 1977.

Frost, J. Statistics by Jim: Central limit theorem explained, January 2021. URL https://statisticsbyjim.com/basics/central-limit-theorem/.

Gaffke, N. Three test statistics for a nonparametric one-sided hypothesis on the mean of a nonnegative variable. *Mathematical Methods of Statistics*, 14(4):451–467, 2005.

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

Institute of Medicine. *Small clinical trials: Issues and challenges*. The National Academies Press, 2001.

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

Learned-Miller, E. and Thomas, P. S. A new confidence interval for the mean of a bounded random variable. *arXiv preprint arXiv:1905.06208*, 2019.

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

Pap, G. and van Zuijlen, M. C. A. The Stringer bound in case of uniform taintings. *Computers and Mathematics with Applications*, 29(10):51–59, 1995.

Phan, M., Thomas, P. S., and Learned-Miller, E. Towards practical mean bounds for small samples, 2021.

Romano, J. P. and Wolf, M. Finite sample nonparametric inference and large sample efficiency. *Annals of Mathematical Statistics*, 28(3):756–778, 2000.

Serfling, R. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980. ISBN 0471024031. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+024353353&sourceid=fbw_bibsonomy.

Stringer, K. W. Practical aspects of statistical sampling. *Proceedings of Business and Economic Statistics Section, American Statistical Association*, 1963.

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

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *AAAI*, 2015.