Towards Practical Mean Bounds for Small Samples

Anonymous Authors[1]

Abstract

Bounds on the mean of an unknown distribution and their associated confidence intervals are widely used in the scientific literature to characterize uncertainty of mean estimates. 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 Hoeffding's whose only assumption is that the support of the distribution falls in an interval $(-\infty, b]$. For the first time since then, we present a new family of bounds, based on the same assumptions, at least one of which *dominates* Anderson's, meaning that it is always as tight or tighter than Anderson's. We prove that our bounds have *guaranteed coverage*, i.e., they hold with probability at least $1-\alpha$ for all distributions on an interval $(-\infty, b]$. 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 $\mathbf{X} = X_1, X_2, \ldots, X_n$, and the closely related problems of producing upper or lower confidence bounds on the mean. To produce a non-trivial upper bound, one must make assumptions about F, such as finite variance, sub-Gaussianity, or that its support is contained on a finite interval $[a, b]$ or a semi-infinite interval $(-\infty, b]$. We adopt this last assumption. For simplicity, we focus on upper bounds, but the development for lower bounds and confidence intervals is similar.

An upper confidence bound 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] This question is what to do when the sample size is smaller than such a threshold. Our bounds provide a new and better option for guaranteed coverage with small sample sizes–our bounds our tighter 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. Below we review bounds with coverage guarantees, those that do *not* exhibit guaranteed coverage, and those for which the result is unknown.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Adequate sample size for 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.

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, $t \geq 0$:

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

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 $1-\alpha$ upper confidence bound on the mean of

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

Maurer and Pontil. One limitation of Hoeffding's inequality is that the amount added to the sample mean to obtain the upper confidence bound 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 an upper confidence bound for the variance of a random variable, and suggest combining this with Bennet's inequality (via the union bound) to obtain the following $1-\alpha$ upper confidence bound on the mean

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

Notice that Maurer and Pontil's upper confidence bound 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 upper confidence bound 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's bound (Anderson, 1969a)[3] is based upon the well-known Kolmogorov-Smirnov (KS) statistic, which gives the maximum distance between a continuous distribution's CDF and the empirical CDF of a sample from that distribution. Using the distribution of the KS-statistic (which is independent of the underlying distribution F, as long as it is continuous), one may define an 'envelope' around the empirical CDF 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.

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

A variety of variations on the theme of Anderson's bound have been proposed, including those by Diouf & Dufour (2005), Learned-Miller & DeStefano (2008), and 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

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

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. Upper-case letters denote random variables and lower-case 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 $F(\mathbf{x}) \stackrel{\text{def}}{=} (F(x_1), \cdots, F(x_n)) \in [0,1]^n$. 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 an upper confidence bound 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 $T : D^n \rightarrow \mathcal{R}$ 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 $s_D \stackrel{\text{def}}{=} \sup\{x : x \in D\}$. 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 $x_{(n+1)} \stackrel{\text{def}}{=} s_D$. Consider the step function $G_{\mathbf{x},\boldsymbol{\ell}} : \mathcal{R} \rightarrow [0,1]$ 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 $\boldsymbol{\ell} = (1/n, \ldots, n/n)$, $G_{\mathbf{x},\boldsymbol{\ell}}$ 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.

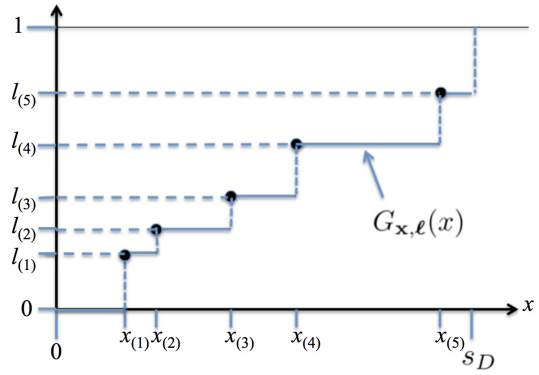


Figure 1. The stairstep function $G_{\mathbf{x},\boldsymbol{\ell}}$, which is a function of the sample $\mathbf{x}$ and a vector $\boldsymbol{\ell}$ of values between 0 and 1. When $\boldsymbol{\ell} = (1/n, \ldots, n/n)$, $G_{\mathbf{x},\boldsymbol{\ell}}$ becomes the empirical CDF.

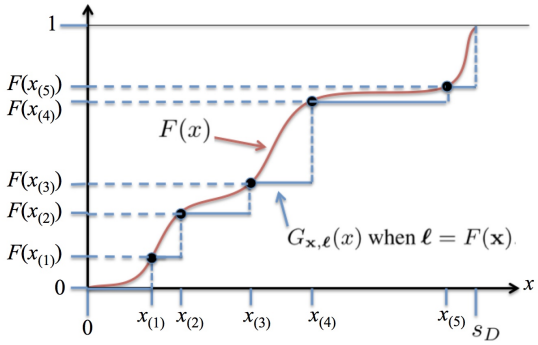


Figure 2. The CDF of a distribution F in red, with a random sample of five order statistics on the x-axis. The blue stairstep function shows the function $G_{\mathbf{x},\boldsymbol{\ell}}(x)$ when $\boldsymbol{\ell} = F(\mathbf{x})$. Notice that for all x, $G_{\mathbf{x},\boldsymbol{\ell}}(x) \leq F(x)$.

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

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

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

where $\ell_{(0)} \stackrel{\text{def}}{=} 0$, $\ell_{(n+1)} \stackrel{\text{def}}{=} 1$ and $x_{(n+1)} \stackrel{\text{def}}{=} s_D$. 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* for the sample $\mathbf{x}$ by the vector $\boldsymbol{\ell}$.

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 $T : \mathcal{R}^n \rightarrow \mathcal{R}$, 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 $\mathbf{u} \stackrel{\text{def}}{=} (u_1, \cdots, u_n)$ being a particular sample of $\mathbf{U}$.

Now consider the random quantity

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

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.,

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

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* $1 - \alpha$*:*

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

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 upper confidence bound 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 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). If there exists a sample $\mathbf{x}_{\max}$ with the largest $b^{\alpha}(\mathbf{x})$ such that $b^{\alpha}(\mathbf{x}) < \mu$, then the probability a sample $\mathbf{Z}$ outputs $b^{\alpha}(\mathbf{Z}) < \mu$ is equal to the probability $\mathbf{Z}$ outputs $b^{\alpha}(\mathbf{Z}) \leq b^{\alpha}(\mathbf{x}_{\max})$ (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 $\mathbf{U} \stackrel{\text{def}}{=} F(\mathbf{Z})$ and $\mathbf{u} \stackrel{\text{def}}{=} F(\mathbf{z})$. Then U_i is uniformly distributed on $[0, 1]$. If F is invertible, we can transform the region $\{\mathbf{z} : b^{\alpha}(\mathbf{z}) \leq b^{\alpha}(\mathbf{x}_{\max})\}$ to $\{\mathbf{u} : b^{\alpha}(F^{-1}(\mathbf{u})) \leq b^{\alpha}(\mathbf{x}_{\max})\}$ where $F^{-1}(\mathbf{u}) \stackrel{\text{def}}{=} (F^{-1}(u_1), \ldots, F^{-1}(u_n))$ (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 $\{\mathbf{u} : b^{\alpha}(F^{-1}(\mathbf{u})) \leq b^{\alpha}(\mathbf{x}_{\max})\}$ is a subset of the striped region $\{\mathbf{u} : b(\mathbf{x}_{\max}, \mathbf{u}) \geq \mu\}$.

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

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

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.

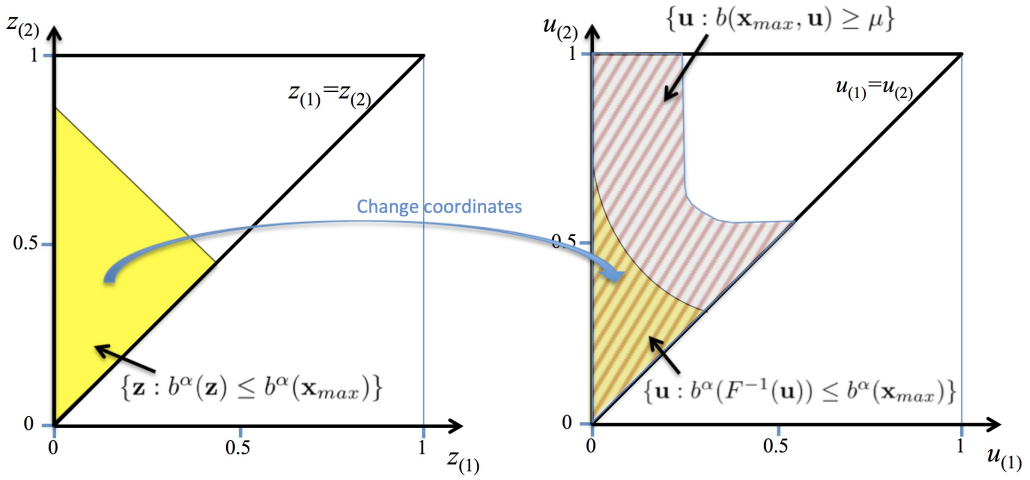


Figure 3. Illustrations of Section 2.1.1. **Left.** The yellow region shows samples of $\mathbf{z} = [z_{(1)}, z_{(2)}]$ such that $b^\alpha(\mathbf{z}) \leq b^\alpha(\mathbf{x}_{max})$. **Right.** The same yellow region, but in the coordinates $\mathbf{u} = F^{-1}(\mathbf{z})$. We will show that the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

Lemma 2.2. *Let X be a random variable with CDF F and $Y \stackrel{\text{def}}{=} F(X)$, 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 $\mathbf{x} \in D^n$,*

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

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 $\alpha \in (0, 1)$,*

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

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

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

Therefore,

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

□

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 $\alpha \in (0, 1)$,*

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

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

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

□

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 $\alpha \in (0, 1)$:*

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

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

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

□

3. Computation

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:

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) $T(\mathbf{y}) \leq T(\mathbf{x})$.
    2) $\forall i : 1 \leq i \leq n, 0 \leq y_{(i)} \leq 1$.
    3) $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(n)}$.
  $\mathbf{ms}[i] = M$.
end for
Sort(**ms**, ascending).
Return $\mathbf{ms}[\lceil (1 - \alpha)l \rceil]$.

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

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

We can compute the $1 - \alpha$ quantile of a random variable M using Monte Carlo simulation, sampling M l times. Letting $m_{(1)} \leq ... \leq m_{(l)}$ 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 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 at most 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, we can set $D^+ = (-\infty, b]$ and our method still dominates Anderson's and Hoeffding's. Our bound becomes tighter or remains constant as the lower support bound increases, whereas Anderson's remains constant because it only uses the upper support bound as an input.

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 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,*

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

Then $H_{\mathbf{X}}$ is called a $(1-\alpha)$ 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*

$$\mathbb{P}_{\mathbf{X}}(m(\mathbf{X}, \boldsymbol{\ell}) \geq \mu) \geq 1 - \alpha. \tag{45}$$

Anderson defined a $1-\alpha$ lower CDF confidence bound $G_{\mathbf{X},\mathbf{u}^{And}}$ where $\mathbf{u}^{And} \in [0,1]^n$ is defined as

$$u_i^{\text{And}} \stackrel{\text{def}}{=} \max\left\{0, i/n - \beta(n)\right\}. \tag{46}$$

Anderson identifies $\beta(n)$ as the one-sided Kolmogorov-Smirnov statistic that results in equality for Eq. 44. Using the fact that $G_{\mathbf{X},\mathbf{u}^{And}}$ is a $1-\alpha$ lower confidence bound for the CDF, and letting $D^+ = (-\infty, b]$, Anderson's $1-\alpha$ upper bound for the mean is

$$b^{\alpha,\text{Anderson}}(\mathbf{x}) \stackrel{\text{def}}{=} m_{D^+}(\mathbf{x}, \mathbf{u}^{\text{And}}). \tag{47}$$

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

Lemma 4.2 (Theorem 2 from (Learned-Miller & Thomas, 2019))**.** *For any sample size n, for any sample value $\mathbf{x} \in D^n$, for all $\alpha \in (0,1)$:*

$$b^{\alpha,Anderson}(\mathbf{x}) \leq b^{\alpha,Hoeffding}(\mathbf{x}). \tag{48}$$

We now compare our method with Anderson's bound:

Theorem 4.3. *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 $T(\mathbf{x}) = b^{\alpha,Anderson}(\mathbf{x})$ yields:*

$$b^{\alpha}_{D^+,T}(\mathbf{x}) \leq b^{\alpha,Anderson}(\mathbf{x}). \tag{49}$$

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

$$\mathbb{P}_{\mathbf{U}}(\forall i, U_{(i)} \geq \mathbf{u}^{And}_{(i)}) \geq 1 - \alpha, \tag{50}$$

where $U_{(i)}$'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 $\forall i, U_{(i)} \geq \mathbf{u}^{And}_{(i)}$ then $b(\mathbf{x}, \mathbf{U}) \leq b(\mathbf{x}, \mathbf{u}^{And})$. Therefore with probability at least $1-\alpha$, $b(\mathbf{x}, \mathbf{U}) \leq b(\mathbf{x}, \mathbf{u}^{And})$. So $b(\mathbf{x}, \mathbf{u}^{And})$ is at least the $1-\alpha$ quantile of $b(\mathbf{x}, \mathbf{U})$, which is the value of our bound. Therefore $b(\mathbf{x}, \mathbf{u}^{And})$ is at least the value of our bound.

Finally, if T is Anderson's bound, through some calculations we can show that $b_{D^+,T}(\mathbf{x}, \mathbf{u}^{And}) = m_{D^+}(\mathbf{x}, \mathbf{u}^{And})$, 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 $\alpha \in (0,1)$, using $T(\mathbf{x}) = b^{\alpha,Anderson}(\mathbf{x})$ yields:*

$$b^{\alpha}_{D^+,T}(\mathbf{x}) \leq b^{\alpha,Hoeffding}(\mathbf{x}). \tag{51}$$

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. Diouf & Dufour (2005) present several such bounds with different $\boldsymbol{\ell}$ computed from the Anderson-Darling or the Eicker statistics (Theorem 4, 5 and Theorem 6 with constant ϵ). Theorem 4.3 is also true for other value of $\beta(n)$ used to compute $\mathbf{u}^{\text{And}}$ as long as $G_{\mathbf{X},\mathbf{u}^{And}}$ is a $1-\alpha$ lower CDF confidence bound. 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,

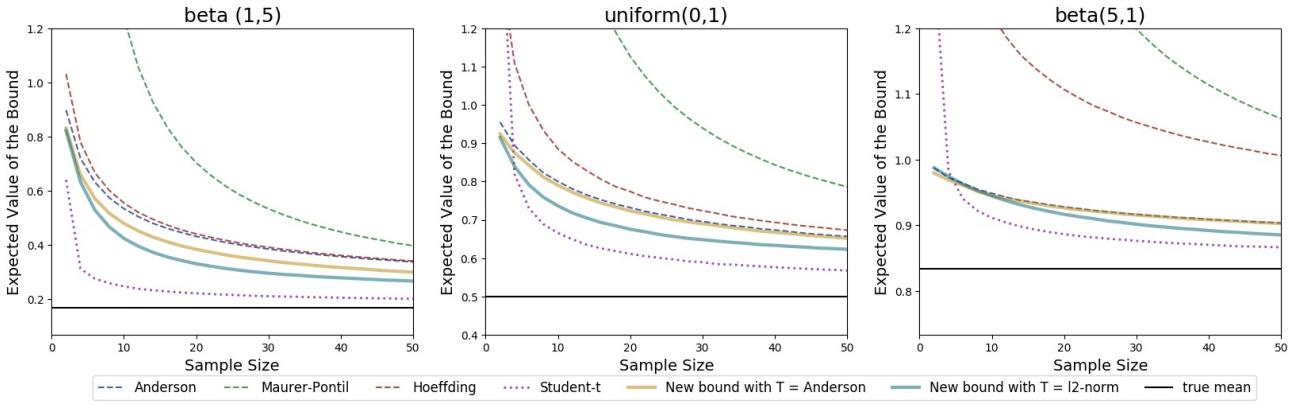


Figure 4. The expected value of the bounds. For each sample size, we sample $\mathbf{X}$ 10,000 times, compute the bound for each sample, and take the average. Our new bound with T being Anderson's bound consistently has lower expected value than Anderson's (Theorem 4.3), Hoeffding's (Theorem 4.4) and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

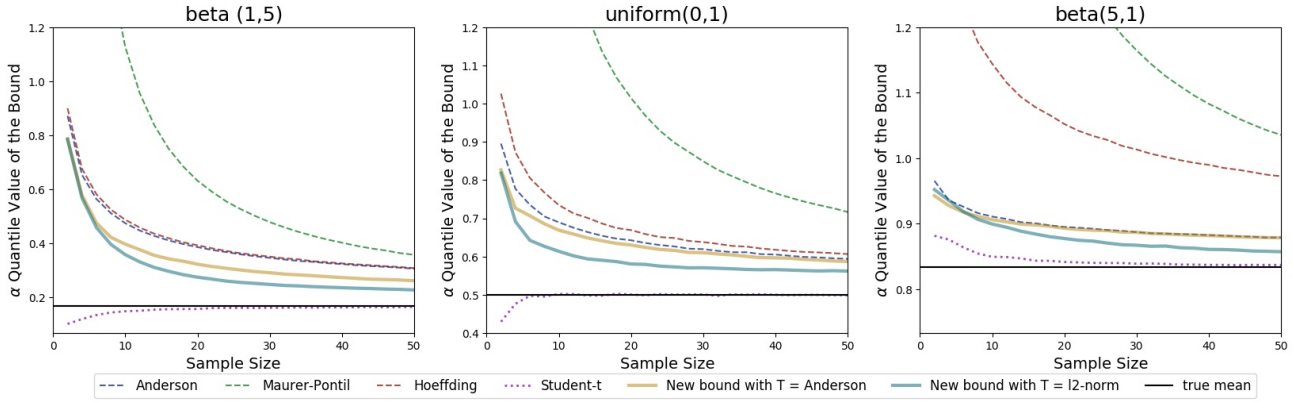


Figure 5. The α-quantile of the bound distribution. For each sample size, we sample $\mathbf{X}$ 10,000 times, compute the bound for each sample, and take the α quantile. If the α-quantile is below the true mean, the bound does not have guaranteed coverage. For the $\texttt{uniform}(0, 1)$ and $\texttt{beta}(1,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

and Student-t's bound (Student, 1908), the latter being

$$b^{\alpha,\text{Student}}(\mathbf{X}) \stackrel{\text{def}}{=} \bar{X}_n + \sqrt{\frac{\hat{\sigma}^2}{n}} t_{1-\alpha,n-1}. \tag{52}$$

For Anderson's bound, we use the Dvoretsky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956) to define the $1-\alpha$ CDF lower bound via $\beta(n) = \sqrt{\ln(1/\alpha)/(2n)}$. We use $D = [0, 1]$ and $l = 10{,}000$ Monte Carlo samples. We consider two functions T:

1. Anderson: $T(\mathbf{x}) = b^{\alpha,\text{Anderson}}(\mathbf{x})$, again with $\beta(n) = \sqrt{\ln(1/\alpha)/(2n)}$. 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 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 left, and becomes similar to Anderson's in right-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).

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.

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

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.

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.

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.

Supplementary Material: Towards Practical Mean Bounds for Small Samples

- In Section A we present more experiments.
- In Section B, as noted in Section 1, we show a log-normal distribution where the sample mean distribution is visibly skewed when $n = 80$.
- In Section C we present the proofs of Section 2.1.2.
- In Section D we discuss the Monte Carlo convergence result of our approximation in Section 3.
- In Section E.1 we show that our bound reduces to the Clopper-Pearson bound for binomial distributions as mentioned in Section 4.1. In Section E.2 we present the proofs of Section 4.2.

A. Other Experiments

In this section we perform experiments to find an upper bound of the mean of distributions given a finite upper bound of the support, or to a lower bound of the mean of distributions given a finite lower bound of the support. We find the lower bound of the mean of a random variable X by finding the upper mean bound of $-X$ and negating it to obtain the lower mean bound of X.

For each experiment we plot the following:

- The expected value of the bounds versus the sample size. For each sample size, we draw 10,000 samples of $\mathbf{x}$, compute the bound for each $\mathbf{x}$ and compute the average.
- For the upper bound of the mean, we plot the α-quantile of the bound distribution versus the sample size. For each sample size, we draw 10,000 samples of $\mathbf{x}$, compute the bound for each $\mathbf{x}$ and take the α quantile. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

  For the lower bound of the mean, we plot the $1-\alpha$-quantile of the bound distribution versus the sample size. For each sample size, we draw 10,000 samples of $\mathbf{x}$, compute the bound for each $\mathbf{x}$ and take the $1-\alpha$ quantile. If the $1-\alpha$-quantile is above the true mean, the bound does not have guaranteed coverage.
- Coverage of the bounds. For each value of α from 0.02 to 1 with a step size of 0.02, we draw 10,000 samples of $\mathbf{x}$, compute the bound for each $\mathbf{x}$ and plot the percentage of the bounds that are greater than or equal to the true mean (denoted *coverage*). If this percentage is larger than $1-\alpha$, the bound has guaranteed coverage.

We perform the following experiments:

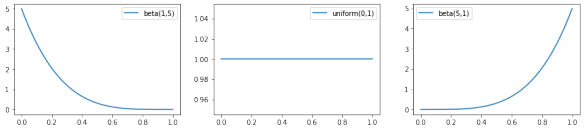
1. For the case in which we know a superset D^+ of the distribution's support with a finite lower bound and a finite upper bound, we compare the following bounds:
   - Anderson's bound.
   - New bound with T being Anderson's bound.
   - Student's t.
   - Hoeffding's bound.
   - Maurer and Pontil's bound.

   We find an upper bound of the mean for the following distributions:
   - $\beta(1,5)$, $\texttt{uniform}(0,1)$ and $\beta(5,1)$. The known superset of the support is $[0,1]$. The result is in Figure 6.
   - $\beta(0.5,0.5)$, $\beta(1,1)$ and $\beta(2,2)$. The known superset of the support is $[0,1]$. The result is in Figure 7.
   - $\texttt{binomial}(10,0.1)$, $\texttt{binomial}(10,0.5)$ and $\texttt{binomial}(10,0.9)$. The known superset of the support is the interval $[0,10]$. The result is in Figure 8.
2. We also consider the case in which we want an upper bound of the mean without knowing the lower bound of the support (or to find a lower bound without knowing an upper bound of the support). Since Hoeffding's and Maurer and Pontil's bounds require knowing both a finite lower bound and upper bound, they are not applicable in this setting. We compare the following bounds:
   - Anderson's bound.
   - New bound with T being Anderson's bound.
   - Student's t

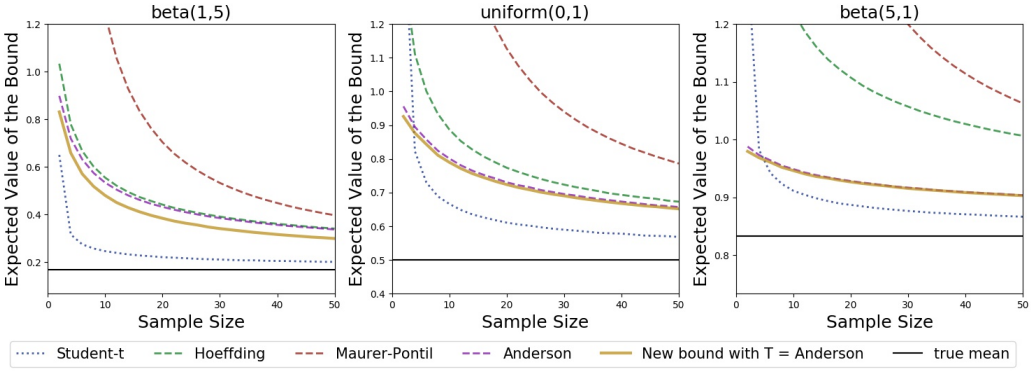
   We address the following distributions:
   - $\beta(1,5)$, $\texttt{uniform}(0,1)$ and $\beta(5,1)$. The known superset of the support is $(-\infty,1]$. We find the upper bound of the mean. The result is in Figure 9.
   - $\texttt{binomial}(10,0.1)$, $\texttt{binomial}(10,0.5)$ and $\texttt{binomial}(10,0.9)$. The known superset of the support is $(-\infty,10]$. We find the upper bound of the mean. The result is in Figure 10.
   - $\texttt{poisson}(2)$, $\texttt{poisson}(10)$ and $\texttt{poisson}(50)$. The known superset of the support is $[0,\infty)$. We find the lower bound of the mean. The result is in Figure 11.

All the experiments confirm that our bound has guaranteed coverage and is tighter than Anderson's and Hoeffding's.

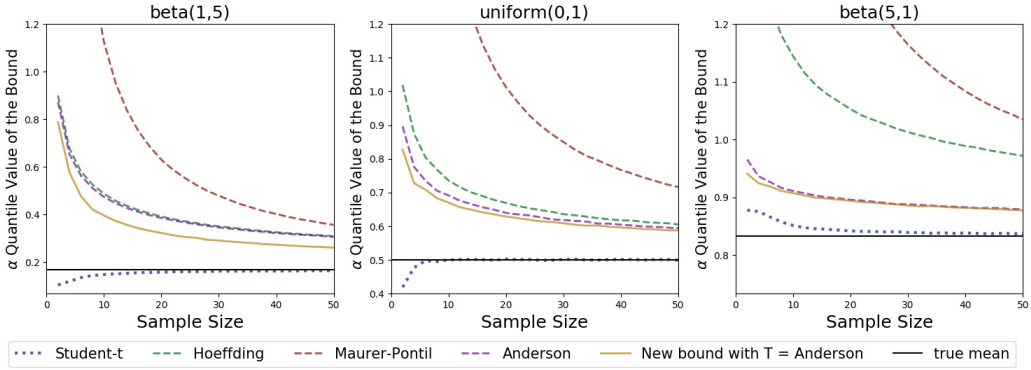
From the experiments, our upper bound performs the best in distributions that are skewed right (respectively, our lower bound will perform the best in distributions that are skewed left), when we know a tight lower bound and upper bound of the support.



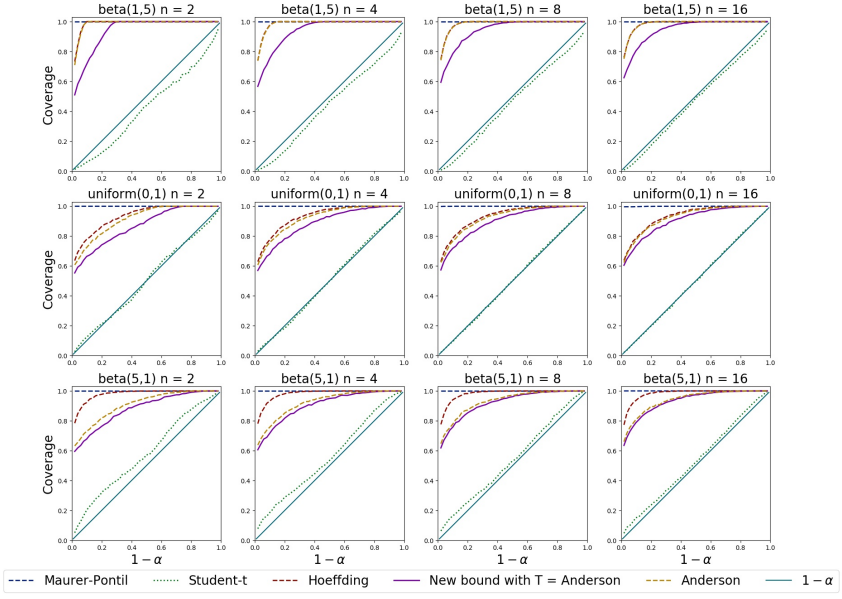
(a) The PDFs of the test distributions.



(b) Expected values of the bounds versus sample size.

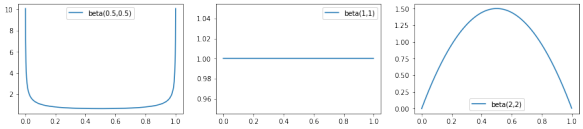


(c) The α-quantiles of bound distributions. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

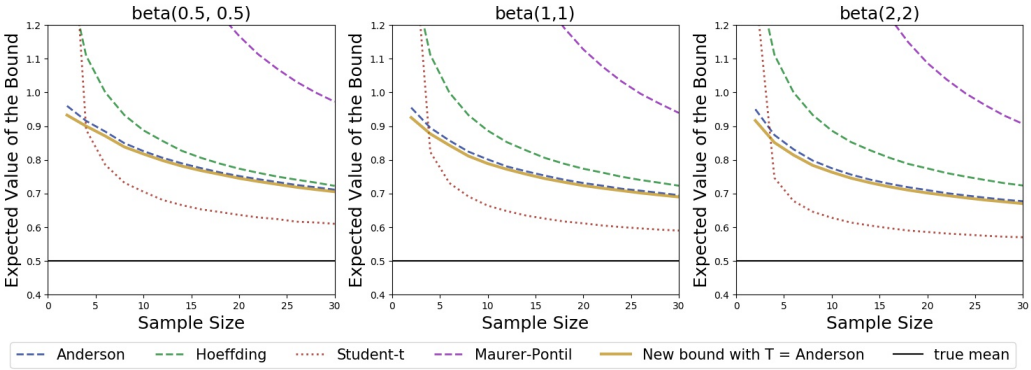


(d) The coverage of the bound. If the coverage is below the line $1 - \alpha$, the bound does not have guaranteed coverage.

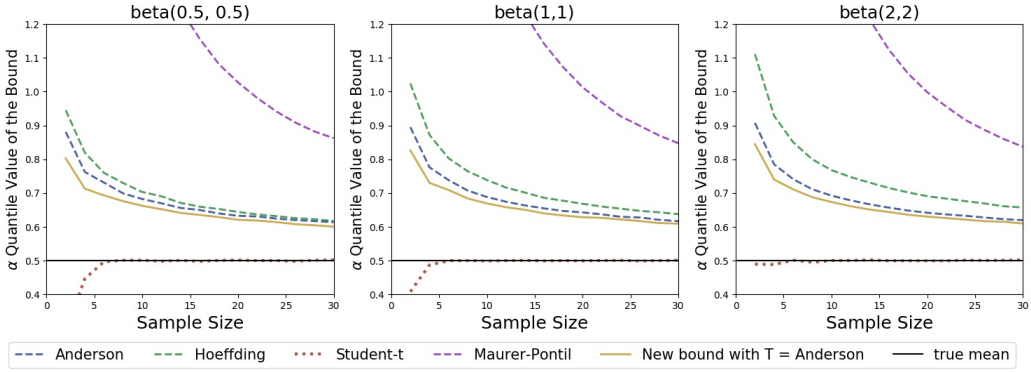
Figure 6. Finding the upper bound of the mean with $D^+ = [0, 1]$



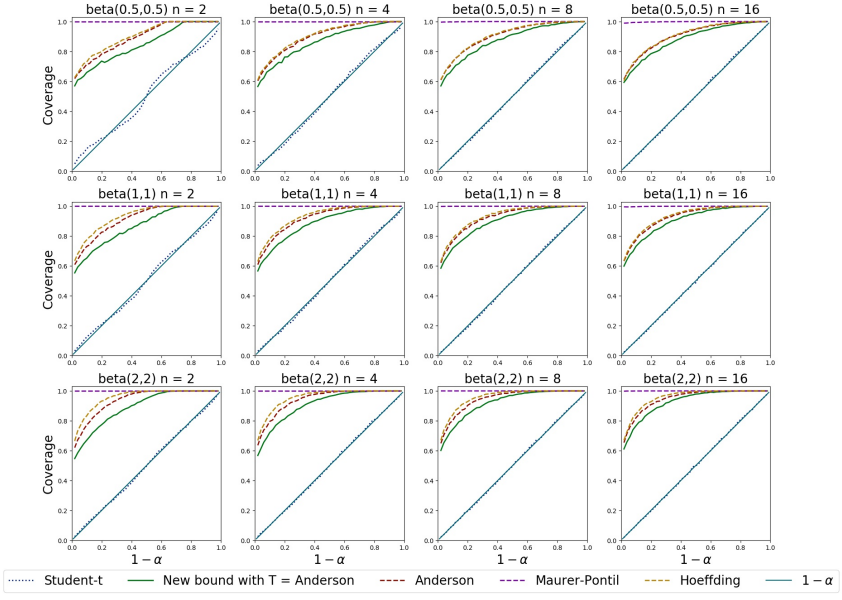
(a) The PDFs of the test distributions.



(b) Expected values of bounds versus sample size.

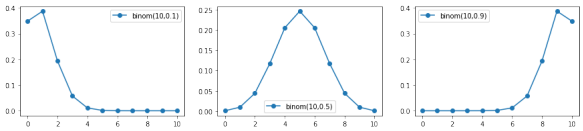


(c) The α-quantiles of bound distributions. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

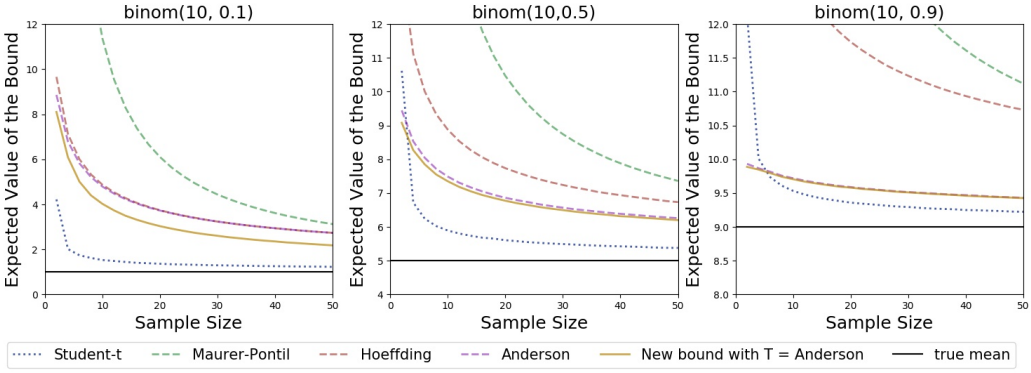


(d) The coverage of the bound. If the coverage is below the line $1 - \alpha$, the bound does not have guaranteed coverage.

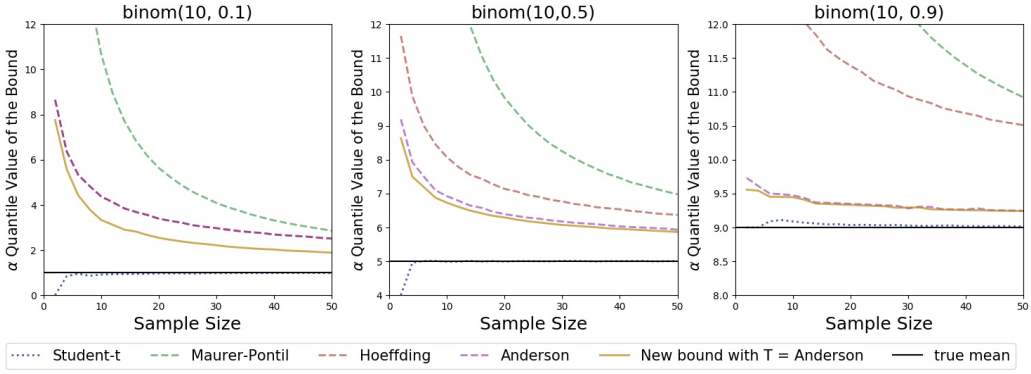
Figure 7. Finding the upper bound of the mean with $D^+ = [0, 1]$



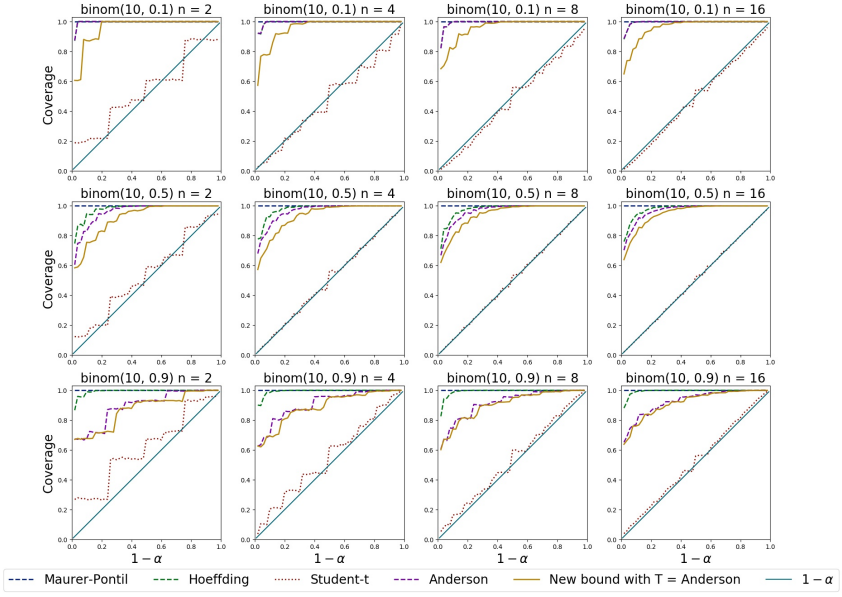
(a) The PMFs of the test distributions.



(b) Expected values of bounds versus sample size.

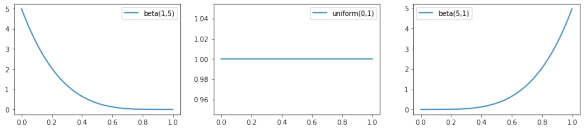


(c) The α-quantiles of bound distributions. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

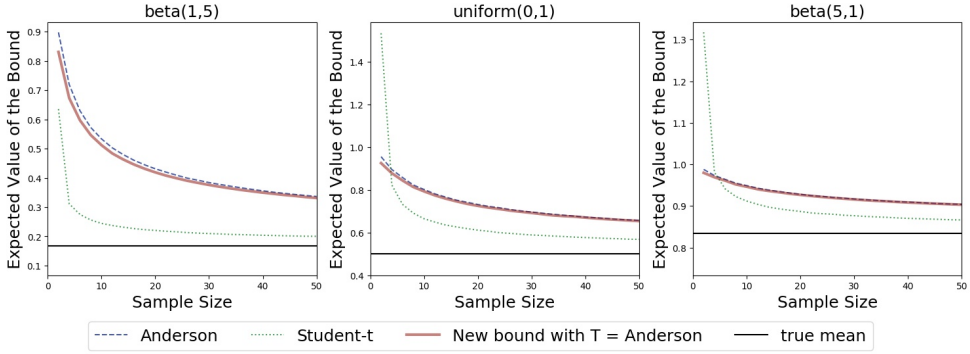


(d) The coverage of the bound. If the coverage is below the line $1 - \alpha$, the bound does not have guaranteed coverage.

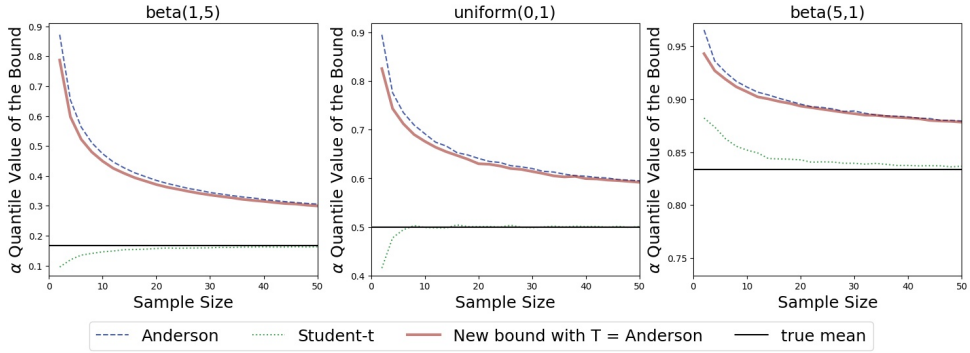
Figure 8. Finding the upper bound of the mean with $D^+ = [0, 10]$



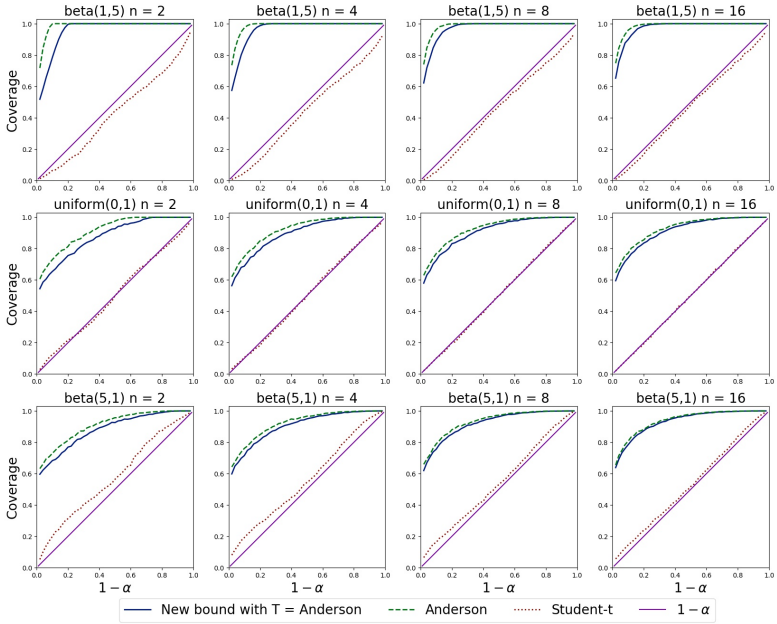
(a) The PDFs of the test distributions.



(b) The expected value of the bounds.

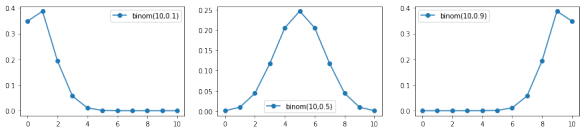


(c) The α-quantile of the bound distribution. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

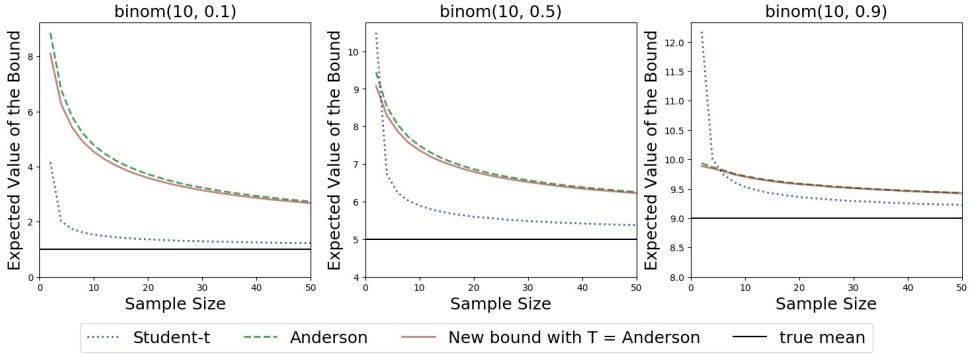


(d) The coverage of the bound. If the coverage is below the line $1 - \alpha$, the bound does not have guaranteed coverage.

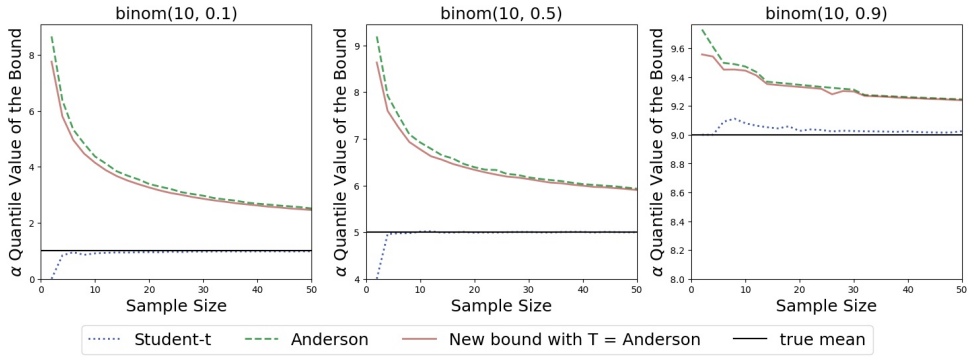
Figure 9. Finding the upper bound of the mean with $D^+ = (-\infty, 1]$



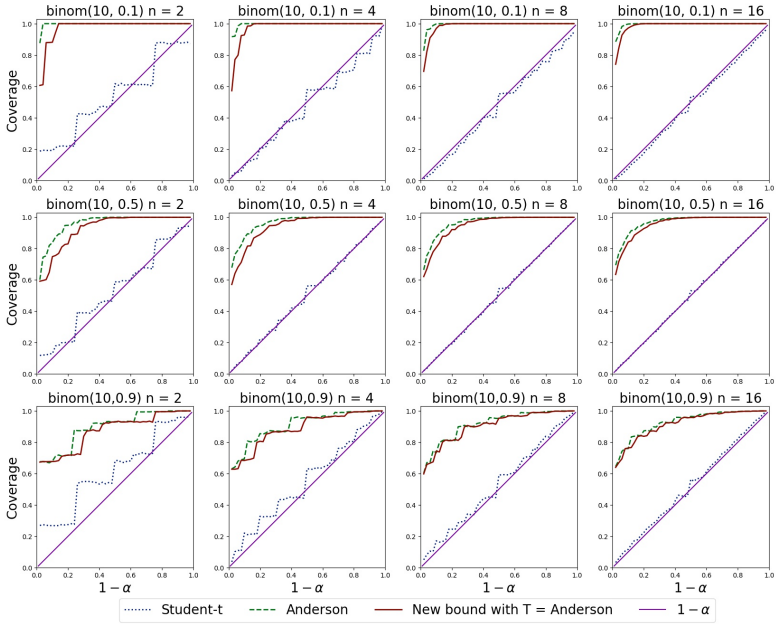
(a) The PMF of the test distributions.



(b) The expected value of the bounds.

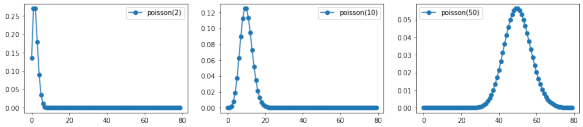


(c) The α-quantile of the bound distribution. If the α-quantile is below the true mean, the bound does not have guaranteed coverage.

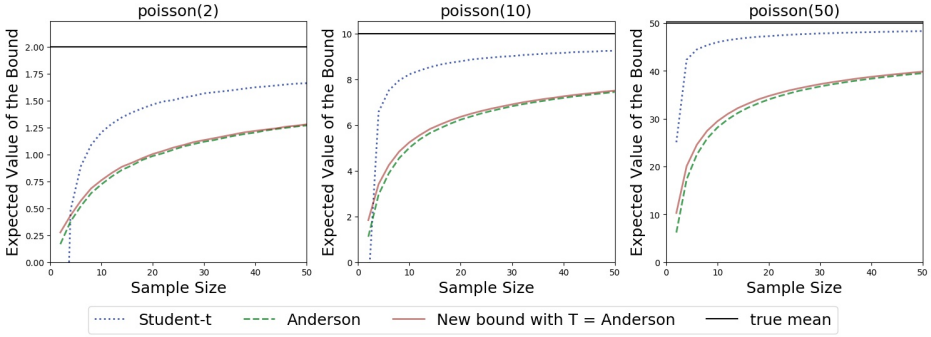


(d) The coverage of the bound. If the coverage is below the line $1 - \alpha$, the bound does not have guaranteed coverage.

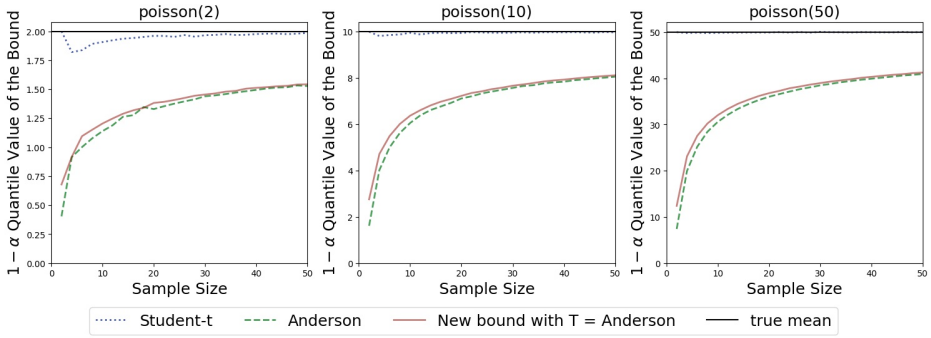
Figure 10. Finding the upper bound of the mean with $D^+ = (-\infty, 10]$



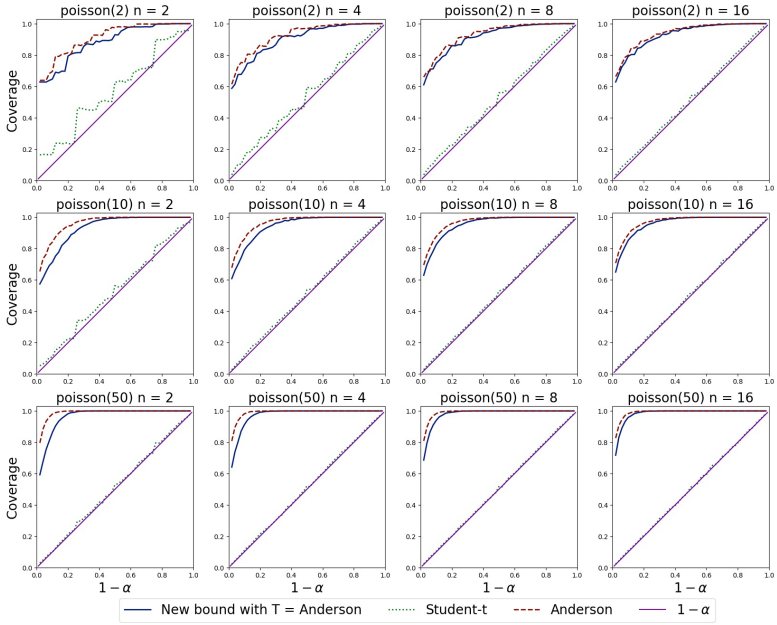
(a) The PMF of the test distributions.



(b) The expected value of the bounds.



(c) The $1-\alpha$-quantile of the bound distribution. If the $1-\alpha$-quantile is above the true mean, the bound does not have guaranteed coverage.



(d) The coverage of the bound. If the coverage is below the line $1-\alpha$, the bound does not have guaranteed coverage.

Figure 11. Finding the lower bound of the mean with $D^+ = [0, \infty)$

B. Discussion on Section 1: Skewed Sample Mean Distribution with $n = 80$

In this section, as noted in Section 1, we show a log-normal distribution where the sample mean distribution is visibly skewed when $n = 80$ (Figure 12). Student's t is not a good candidate in this case because the sample mean distribution is not approximately normal. This example is a variation on the one provided by Frost (2021).

While the log-normal distribution is an extreme example of skew, this example illustrates the danger of assuming the validity of arbitrary thresholds on the sample size, such as the traditional threshold of $n = 30$, for using the Student's t method. Clearly there are cases where such a threshold, and even much larger thresholds, are not adequate.

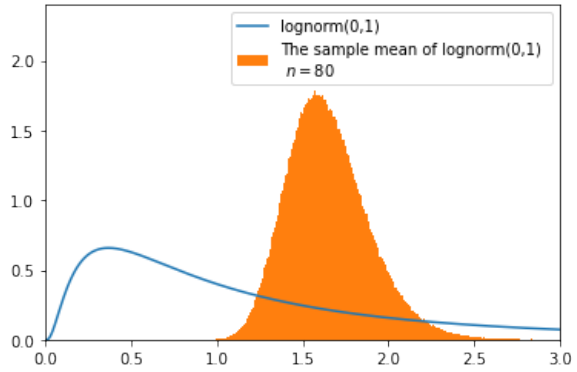


Figure 12. The PDFs of $lognorm(0, 1)$ and the sample mean distribution of $lognorm(0, 1)$. The sample mean distribution of $lognorm(0, 1)$ is visibly skewed when the sample size $n = 80$.

C. Proof of Section 2.1.2

We restate the lemma and theorem statements for convenience.

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

$$\mathbb{P}(Y \leq y) \leq \mathbb{P}(U \leq y). \tag{53}$$

If F is continuous, then Y is uniformly distributed on $(0, 1)$.

Proof. Since $\mathbb{P}(U \leq y) = y$, we will show that $\mathbb{P}(Y \leq y) \leq y$.

We will first show that if $F(x) \leq y$, then $x \leq \sup\{z : F(z) \leq y\}$. Suppose that $x > \sup\{z : F(z) \leq y\}$. Then, $F(x) > y$. Therefore,

$$F(x) \leq y \text{ implies } x \leq \sup\{z : F(z) \leq y\}. \tag{54}$$

Now we have

$$\begin{aligned}
\mathbb{P}(Y \leq y) &= \mathbb{P}(F(X) \leq y) && (55)\\
&\leq \mathbb{P}(X \leq \sup\{z : F(z) \leq y\}) && (56)\\
&= F(z^*) \text{ where } z^* = \sup\{z : F(z) \leq y\} && (57)\\
&\leq y. && (58)
\end{aligned}$$

If F is continuous, (Angus, 1994) shows that Y is uniformly distributed on $(0, 1)$. □

Lemma C.2 (Lemma 2.3)**.** *For any $\mathbf{x} \in D^n$,*

$$m_D(\mathbf{x}, F(\mathbf{x})) \geq \mu. \tag{59}$$

Proof. We have

$$m_D(\mathbf{x}, F(\mathbf{x})) = \sum_{i=1}^{n+1} x_{(i)}(F(x_{(i)}) - F(x_{(i-1)})) \tag{60}$$

where $x_{(0)} \stackrel{\text{def}}{=} -\infty$ and $x_{(n+1)} \stackrel{\text{def}}{=} s_D$. Then:

$$\begin{aligned}
\mu &= \int_x x \, dF(x) && (61)\\
&= \sum_{i=1}^{n+1} \int_{x=x_{(i-1)}}^{x_{(i)}} x \, dF(x) && (62)\\
&\leq \sum_{i=1}^{n+1} \int_{x=x_{(i-1)}}^{x_{(i)}} x_{(i)} \, dF(x) && (63)\\
&= \sum_{i=1}^{n+1} x_{(i)}(F(x_{(i)}) - F(x_{(i-1)})) && (64)\\
&= m_D(\mathbf{x}, F(\mathbf{x})). && (65)
\end{aligned}$$

□

Lemma C.3 (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 $\mathbf{x} \in D^n$:*

$$\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{x})) \leq \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \geq \mu). \tag{66}$$

Proof. Let $\cup$ denote the union of events and $\{\}$ denote an event. Let $\mathbf{Z}$ be a sample from F. Then for any sample $\mathbf{x}$:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{x})) && (67)\\
&= \mathbb{P}_{\mathbf{Z}}(\mathbf{Z} \in \mathbb{S}(\mathbf{x})) && (68)\\
&= \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})}\{\mathbf{Z} = \mathbf{y}\}) && (69)\\
&\leq \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})}\mathbf{Z} \preceq \mathbf{y}) \text{ because } \mathbf{Z} = \mathbf{y} \text{ implies } \mathbf{Z} \preceq \mathbf{y} && (70)\\
&\leq \mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y} \in \mathbb{S}(\mathbf{x})}\{F(\mathbf{Z}) \preceq F(\mathbf{y})\}) && (71)
\end{aligned}$$

because F is non-decreasing, so $Z_{(i)} \leq y_{(i)}$ implies $F(Z_{(i)}) \leq F(y_{(i)})$. Let $U_1, \cdots, U_n$ be n samples from the uniform distribution on $(0,1)$. From Lemma 2.2, for any $u \in (0,1)$, $\mathbb{P}(F(Z_i) \leq u) \leq \mathbb{P}(U_i \leq u)$. Therefore:

$$\mathbb{P}_{\mathbf{Z}}(\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{F(\mathbf{Z}) \preceq F(\mathbf{y})\}) \tag{72}$$
$$\leq \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{\mathbf{U} \preceq F(\mathbf{y})\}). \tag{73}$$

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 $\forall i, U_{(i)} \leq F(y_{(i)})$ then $m_D(\mathbf{y}, \mathbf{U}) \geq m_D(\mathbf{y}, F(\mathbf{y}))$:

$$\mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{\mathbf{U} \preceq F(\mathbf{y})\}) \tag{74}$$
$$\leq \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{m_D(\mathbf{y}, \mathbf{U}) \geq m_D(\mathbf{y}, F(\mathbf{y}))\}) \tag{75}$$
$$\leq \mathbb{P}_{\mathbf{U}}(\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{m_D(\mathbf{y}, \mathbf{U}) \geq \mu\}), \text{by Lemma 2.3} \tag{76}$$
$$\leq \mathbb{P}_{\mathbf{U}}(\sup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})} m_D(\mathbf{y}, \mathbf{U}) \geq \mu) \tag{77}$$
$$= \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \geq \mu) \tag{78}$$

The inequality in Eq. 77 is because if there exists $\mathbf{y} \in \mathbb{S}(\mathbf{x})$ such that $m_D(\mathbf{y}, \mathbf{U}) \geq \mu$, then $\sup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})} m_D(\mathbf{y}, \mathbf{U}) \geq \mu$. Therefore the event $\cup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})}\{m_D(\mathbf{y}, \mathbf{U}) \geq \mu\}$ is a subset of the event $\sup_{\mathbf{y}\in\mathbb{S}(\mathbf{x})} m_D(\mathbf{y}, \mathbf{U}) \geq \mu$, and Eq. 77 follows.

From Eqs. 71, 73 and Eq. 78:

$$\mathbb{P}_{\mathbf{Z}}(T(\mathbf{Z}) \leq T(\mathbf{x})) \tag{79}$$
$$\leq \mathbb{P}_{\mathbf{U}}(b(\mathbf{x}, \mathbf{U}) \geq \mu). \tag{80}$$

□

D. Discussion on Section 3: Monte Carlo Convergence

In Section 3, we discuss the use of Monte Carlo sampling of the induced mean function $m(\mathbf{z}, U)$, via sampling of the uniform random variable U, to approximate the $1-\alpha$ quantile of $m(\mathbf{z}, U)$.

The Monte Carlo approximation error is quantified in the following lemma due to Serfling (1980). Let $F(m-) \stackrel{\text{def}}{=} \lim_{x\to m^-} F(x)$.

Lemma D.1 (Theorem 2.3.2 in Serfling (1980))**.** *Let $0 < p < 1$. If $Q(p, M)$ is the unique solution m of $F(m-) \leq p \leq F(m)$, then for every $\epsilon > 0$,*

$$\mathbb{P}(|M_{\lceil pl \rceil} - Q(p, M)| > \epsilon) \leq 2e^{-2l\delta}, \tag{81}$$

where

$$\delta = \min\left(p - F(Q(p, M) - \epsilon), F(Q(p, M) + \epsilon) - p\right).$$

Note that when the condition that $Q(p, M)$ is the unique solution m of $F(m-) \leq p \leq F(m)$ is satisfied, $\delta > 0$. Let $M \stackrel{\text{def}}{=} b_{D,T}(\mathbf{x}, \mathbf{U}) \in [0, 1]$. In Lemma D.2 we will show that the CDF of M satisfies the condition in Lemma D.1. Therefore the error incurred by computing the bound via Monte Carlo sampling can be decreased to an arbitrarily small value by choosing a large enough number of Monte Carlo samples l. The Monte Carlo estimation of $b^{\alpha}_{D^+,T}(\mathbf{x})$ where $D^+ = [0, 1]$ is presented in Algorithm 1.

Lemma D.2. *Let $M \stackrel{\text{def}}{=} b(\mathbf{x}, \mathbf{U})$. Let F_M be the CDF of M.*

For any $\mathbf{x}$, for any scalar function T, either:

1. *M is a constant, or*
2. *For any $p \in (0, 1)$, $Q(p, M)$ is the unique solution m of $F_M(m-) \leq p \leq F_M(m)$.*

Proof. We will show that for any $\mathbf{x}$, for any T, for any $p \in (0, 1)$, $F_M(m-) \leq p \leq F_M(m)$ has a unique solution by showing that for any $\mathbf{x}$ and T, F_M is strictly increasing on its support. To do so, for any c_1, c_2 in the support such that $c_1 < c_2$ we will show that

$$F_M(c_2) - F_M(c_1) > 0. \tag{82}$$

Recall the definition of the induced mean as

$$b(\mathbf{x}, \boldsymbol{\ell}) = \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} \sum_{i=1}^{n+1} z_{(i)}(\ell_{(i)} - \ell_{(i-1)}), \tag{83}$$
$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - \sum_{i=1}^{n} \ell_{(i)}(z_{(i+1)} - z_{(i)}), \tag{84}$$

where $\ell_{(0)} \stackrel{\text{def}}{=} 0$, $\ell_{(n+1)} \stackrel{\text{def}}{=} 1$ and $z_{(n+1)} \stackrel{\text{def}}{=} s_D$.

We now find the support of M. Let $\phi \stackrel{\text{def}}{=} \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(1)}$. We will show that for any $\mathbf{u}$ where $0 \leq u_i \leq 1$, we have $\phi \leq b(\mathbf{x}, \mathbf{u}) \leq s_D$, and therefore the support of M is a subset of $[\phi, s_D]$. We have

$$b(\mathbf{x}, \mathbf{u}) = \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - \sum_{i=1}^{n} u_{(i)}(z_{(i+1)} - z_{(i)}) \tag{85}$$
$$\leq \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - \sum_{i=1}^{n} 0(z_{(i+1)} - z_{(i)}) \tag{86}$$
$$= s_D. \tag{87}$$

Similarly we have

$$b(\mathbf{x},\mathbf{u}) = \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - \sum_{i=1}^{n} u_{(i)}(z_{(i+1)} - z_{(i)}) \tag{88}$$

$$\geq \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - \sum_{i=1}^{n} 1(z_{(i+1)} - z_{(i)}) \tag{89}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - (z_{(n+1)} - z_{(1)}) \tag{90}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(1)} \tag{91}$$

$$= \phi. \tag{92}$$

Therefore $M = b(\mathbf{x},\mathbf{U}) \in [\phi, s_D]$. We consider two cases: where $\phi = s_D$ and where $\phi < s_D$.

Case 1: $\phi = s_D$.

Then for all i, $\sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(i)} \geq \phi = s_D$. Since $z_{(i)} \leq s_D$, we have for all i, $\sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(i)} = s_D$. Therefore

$$b(\mathbf{x},\boldsymbol{\ell}) = \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} \sum_{i=1}^{n+1} z_{(i)}(\ell_{(i)} - \ell_{(i-1)}) \tag{93}$$

$$= \sum_{i=1}^{n+1} s_D(\ell_{(i)} - \ell_{(i-1)}) \tag{94}$$

$$= s_D. \tag{95}$$

Therefore $M = b(\mathbf{x},\mathbf{U})$ is a constant s_D, and the $1-\alpha$ quantile of M is s_D.

Case 2: $\phi < s_D$.

Let $c_1, c_2 \in \mathcal{R}$ be such that $\phi \leq c_1 < c_2 \leq s_D$. We will now show that

$$F_M(c_2) - F_M(c_1) > 0. \tag{96}$$

Let $v \overset{\text{def}}{=} \frac{s_D - c_2}{s_D - \phi}$ and $w \overset{\text{def}}{=} \frac{s_D - c_1}{s_D - \phi}$. If $\phi \leq c_1 < c_2 \leq s_D$ then $v < w$ and $v, w \in [0,1]$.

Let $\mathbf{v} \overset{\text{def}}{=} (v_1, \cdots, v_n)$ and $\mathbf{w} \overset{\text{def}}{=} (w_1, \cdots, w_n)$ where $\forall i, v_i = v$ and $w_i = w$. Then

$$b(\mathbf{x},\mathbf{v}) \tag{97}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} \sum_{i=1}^{n+1} z_{(i)}(v_{(i)} - v_{(i-1)}) \tag{98}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(n+1)}(v_{(n+1)} - v_{(n)}) + z_{(1)}(v_{(1)} - v_{(0)}) \tag{99}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D(1-v) + z_{(1)}(v-0) \tag{100}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - (s_D - z_{(1)})\frac{s_D - c_2}{s_D - \phi} \tag{101}$$

$$= s_D - (s_D - \phi)\frac{s_D - c_2}{s_D - \phi} \text{ because } \frac{s_D - c_2}{s_D - \phi} \geq 0 \tag{102}$$

$$= c_2. \tag{103}$$

Similarly,

$$b(\mathbf{x},\mathbf{w}) \tag{104}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} \sum_{i=1}^{n+1} z_{(i)}(w_{(i)} - w_{(i-1)}) \tag{105}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} z_{(n+1)}(w_{(n+1)} - w_{(n)}) + z_{(1)}(w_{(1)} - w_{(0)}) \tag{106}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D(1-w) + z_{(1)}(w-0) \tag{107}$$

$$= \sup_{\mathbf{z}\in\mathbb{S}(\mathbf{x})} s_D - (s_D - z_{(1)})\frac{s_D - c_1}{s_D - \phi} \tag{108}$$

$$= s_D - (s_D - \phi)\frac{s_D - c_1}{s_D - \phi} \text{ because } \frac{s_D - c_1}{s_D - \phi} \geq 0 \tag{109}$$

$$= c_1. \tag{110}$$

Since $b(\mathbf{x},\mathbf{u})$ is constructed from linear function of $\mathbf{u}$ with non-positive coefficients, for any $\mathbf{u}$ such that $v \leq u_{(1)} \leq \cdots \leq u_{(n)} < w$ we have:

$$b(\mathbf{x},\mathbf{w}) < b(\mathbf{x},\mathbf{u}) \leq b(\mathbf{x},\mathbf{v}), \tag{111}$$

which is equivalent to:

$$c_1 < b(\mathbf{x},\mathbf{u}) \leq c_2. \tag{112}$$

So we have $v \leq u_{(1)} \leq \cdots \leq u_{(n)} < w$ implies $c_1 < b(\mathbf{x},\mathbf{u}) \leq c_2$. Therefore for any c_1, c_2 such that $\phi \leq c_1 < c_2 \leq s_D$:

$$F_M(c_2) - F_M(c_1) \tag{113}$$

$$= \mathbb{P}(c_1 < M \leq c_2) \tag{114}$$

$$= \mathbb{P}_{\mathbf{U}}(c_1 < b(\mathbf{x},\mathbf{U}) \leq c_2) \tag{115}$$

$$\geq \mathbb{P}_{\mathbf{U}}(v \leq U_{(1)} \leq \cdots \leq U_{(n)} < w) \tag{116}$$

$$> 0 \text{ because } 0 \leq v < w \leq 1. \tag{117}$$

Since the support of M is in $[\phi, s_D]$ we have that F_M is strictly increasing on the support. □

In summary, the Monte Carlo estimate of our bound will converge to the correct value as the number of samples grows, allowing the user to make the error arbitrarily small.

E. Discussion on Section 4

E.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, we will show that our bound becomes the same as the Clopper-Pearson confidence bound for binomial distributions (Clopper & Pearson, 1934).

If $\mathbf{x}, \mathbf{z} \in \{0,1\}^n$ and $T(\mathbf{z}) \leq T(\mathbf{x})$ then $m(\mathbf{z}, \mathbf{u}) \leq m(\mathbf{x}, \mathbf{u})$. Therefore for any $\mathbf{u} \in [0,1]^n$,

$$b_{D,T}(\mathbf{x}, \mathbf{u}) = \sup_{\mathbf{z} \in \{0,1\}^n : T(\mathbf{z}) \leq T(\mathbf{x})} m_D(\mathbf{z}, \mathbf{u}) = m_D(\mathbf{x}, \mathbf{u}). \tag{118}$$

Let $p_{\mathbf{x}}$ be the number of 0's in $\mathbf{x}$. Therefore the bound becomes the $1-\alpha$ quantile of $m_D(\mathbf{x}, \mathbf{U})$ where

$$m_D(\mathbf{x}, \mathbf{U}) = 1 - \sum_{i=1}^{n} U_{(i)}(x_{(i+1)} - x_{(i)}) = 1 - U_{(p_{\mathbf{x}})}. \tag{119}$$

Therefore the bound is the $1-\alpha$ quantile of $1 - U_{(p_{\mathbf{x}})}$. Then

$$\mathbb{P}(U_{(p_{\mathbf{x}})} \leq 1 - b^{\alpha}(\mathbf{x})) = \mathbb{P}(1 - U_{(p_{\mathbf{x}})} \geq b^{\alpha}(\mathbf{x})) = \alpha. \tag{120}$$

Let $\beta(i, j)$ denote a beta distribution with parameters i and j. We use the fact that the order statistics of a uniform distribution are beta-distributed. Since $U_{(p_{\mathbf{x}})} \sim \beta(p_{\mathbf{x}}, n + 1 - p_{\mathbf{x}})$, we have $1 - U_{(p_{\mathbf{x}})} \sim \beta(n - p_{\mathbf{x}} + 1, p_{\mathbf{x}})$

$$b^{\alpha}(\mathbf{x}) = Q(1-\alpha, \beta(n - p_{\mathbf{x}} + 1, p_{\mathbf{x}})). \tag{121}$$

This is the same as the Clopper-Pearson upper confidence interval for binomial distributions.

E.2. Proof of Section 4.2

Lemma E.1 (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*

$$\mathbb{P}_{\mathbf{X}}(m(\mathbf{X}, \boldsymbol{\ell}) \geq \mu) \geq 1 - \alpha. \tag{122}$$

Proof. If $\forall y \in \mathcal{R}, F(y) \geq G_{\mathbf{X},\boldsymbol{\ell}}(y)$ then

$$\forall i : 1 \leq i \leq n, F(X_{(i)}) \geq \ell_{(i)}. \tag{123}$$

Recall that $m_D(\mathbf{X}, \boldsymbol{\ell}) = s_D - \sum_{i=1}^{n} \ell_{(i)}(z_{(i+1)} - z_{(i)})$. Therefore if $\forall i : 1 \leq i \leq n$, $F(X_{(i)}) \geq \ell_{(i)}$ then $m(\mathbf{X}, \boldsymbol{\ell}) \geq m(\mathbf{X}, F(\mathbf{X}))$.

From Lemma 2.3, $m(\mathbf{X}, F(\mathbf{X})) \geq \mu$. Therefore $m(\mathbf{X}, \boldsymbol{\ell}) \geq \mu$. And hence, finally,

$$\mathbb{P}(m(\mathbf{X}, \boldsymbol{\ell}) \geq \mu) \geq \mathbb{P}_{\mathbf{X}}\left(\forall y \in \mathcal{R}, F(y) \geq G_{\mathbf{X},\boldsymbol{\ell}}(y)\right) \tag{124}$$

$$= 1 - \alpha. \tag{125}$$

□

We now show that if $G_{\mathbf{X},\boldsymbol{\ell}}$ (Figure 1) is a lower confidence bound, then the order statistics of $\boldsymbol{\ell}$ are element-wise smaller than the order statistics of a sample of size n from the uniform distribution with high probability:

Lemma E.2. *Let $\mathbf{U} = U_1, ..., U_n$ be a sample of size n from the continuous uniform distribution on $[0,1]$. Let $\boldsymbol{\ell} \in [0,1]^n$ and $\alpha \in (0,1)$. If $G_{\mathbf{X},\boldsymbol{\ell}}$ is a $(1-\alpha)$ lower confidence bound for the CDF then:*

$$\mathbb{P}_{\mathbf{U}}(\ \forall i : 1 \leq i \leq n, U_{(i)} \geq \ell_{(i)}) \geq 1 - \alpha. \tag{126}$$

Proof. Let K be the CDF of a distribution such that K is continuous and strictly increasing (such as the standard Gaussian distribution). Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a sample of size n from the distribution with CDF K. By Lemma 2.2, $K(X)$ is uniformly distributed on $[0,1]$.

By the definition of $G_{\mathbf{X},\boldsymbol{\ell}}$, if $\forall x \in C, K(y) \geq G_{\mathbf{X},\boldsymbol{\ell}}(y)$ then:

$$K(y) \geq 0, \qquad \text{if } y < X_{(1)} \tag{127}$$

$$K(y) \geq \ell_{(i)}, \qquad \text{if } X_{(i)} \leq y < X_{(i+1)} \tag{128}$$

$$K(y) \geq 1, \qquad \text{if } y \geq s_C. \tag{129}$$

which is equivalent to:

$$\forall i : 1 \leq i \leq n, K(y) \geq \ell_{(i)}, \text{if } X_{(i)} \leq y < X_{(i+1)} \tag{130}$$

Since $K(y)$ is non-decreasing, this is equivalent to:

$$\forall i : 1 \leq i \leq n, K(X_{(i)}) \geq \ell_{(i)} \tag{131}$$

Since $G_{\mathbf{X},\boldsymbol{\ell}}$ is a lower confidence bound,

$$1 - \alpha \leq \mathbb{P}_{\mathbf{X}}(\ \forall y \in \mathcal{R}, K(y) \geq G_{\mathbf{X},\boldsymbol{\ell}}(y)) \tag{132}$$

$$= \mathbb{P}_{\mathbf{X}}(\forall i : 1 \leq i \leq n, K(X_{(i)}) \geq \ell_{(i)}) \tag{133}$$

$$= \mathbb{P}_{\mathbf{U}}(\forall i : 1 \leq i \leq n, U_{(i)} \geq \ell_{(i)}) \text{ by Lemma 2.2.} \tag{134}$$

□

To prove Theorem 4.3, we prove the more general version. In the following theorem, we compare $b^{\alpha}_{D^+,T}$ to $m_{D^+}(\mathbf{x}, \boldsymbol{\ell})$ if $G_{\mathbf{X},\boldsymbol{\ell}}$ is a lower confidence bound for the CDF.

Theorem E.3. *Let $\boldsymbol{\ell} \in [0,1]^n$. Let $D^+ = [-\infty, b]$. If $G_{\mathbf{X},\boldsymbol{\ell}}$ is a $1-\alpha$ lower confidence bound for the CDF, then for any sample size n, for all sample values $\mathbf{x} \in D^n$ and all $\alpha \in (0,1)$, using $T(\mathbf{x}) = m_{D^+}(\mathbf{x}, \boldsymbol{\ell})$ to compute $b^{\alpha}_{D^+,T}(\mathbf{x})$ yields:*

$$b^{\alpha}_{D^+,T}(\mathbf{x}) \leq m_{D^+}(\mathbf{x}, \boldsymbol{\ell}). \tag{135}$$

Proof. Since $G_{\mathbf{X},\boldsymbol{\ell}}$ is a lower confidence bound for the cdf F, from Lemma E.2,

$$\mathbb{P}(\forall i, U_{(i)} \geq \ell_{(i)}) \geq 1 - \alpha. \tag{136}$$

First we note that

$$b_{D^+,T}(\mathbf{x},\boldsymbol{\ell}) = \sup_{\mathbf{y}:\mathbf{y}\in\mathbb{S}_{D^+,T}(\mathbf{x})} m_{D^+}(\mathbf{y},\boldsymbol{\ell}) \tag{137}$$

$$= \sup_{m_{D^+}(\mathbf{y},\boldsymbol{\ell})\leq m_{D^+}(\mathbf{x},\boldsymbol{\ell})} m_{D^+}(\mathbf{y},\boldsymbol{\ell}) \tag{138}$$

$$= m_{D^+}(\mathbf{x},\boldsymbol{\ell}). \tag{139}$$

Recall that $b^{\alpha}_{D^+,T}(\mathbf{x})$ is the $1-\alpha$ quantile of $b_{D^+,T}(\mathbf{x},\mathbf{U})$. In order to show that $b^{\alpha}_{D^+,T}(\mathbf{x}) \leq b_{D^+,T}(\mathbf{x},\boldsymbol{\ell})$, we will show that

$$\mathbb{P}(b_{D^+,T}(\mathbf{x},\mathbf{U}) \leq b_{D^+,T}(\mathbf{x},\boldsymbol{\ell})) \geq 1-\alpha. \tag{140}$$

Recall that $b_{D^+,T}(\mathbf{x},\mathbf{U}) = \sup_{\mathbf{y}\in\mathbb{S}_T(\mathbf{x})} t_{D^+} - \sum_{i=1}^{n} U_{(i)}(x_{(i+1)} - x_{(i)})$. Then if $\forall i, U_{(i)} \geq \ell_{(i)}$ then $b_{D^+,T}(\mathbf{x},\mathbf{U}) \leq b_{D^+,T}(\mathbf{x},\boldsymbol{\ell})$. Therefore,

$$\mathbb{P}(b_{D^+,T}(\mathbf{x},\mathbf{U}) \leq b_{D^+,T}(\mathbf{x},\boldsymbol{\ell})) \tag{141}$$

$$\geq \mathbb{P}(\forall i, U_{(i)} \geq \ell_{(i)}) \tag{142}$$

$$\geq 1-\alpha, \text{ by Lemma E.2.} \tag{143}$$

□

We can now show the comparison with Anderson's bound and Hoeffding's bound.

Theorem E.4 (Theorem 4.3)**.** *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 $T(\mathbf{x}) = b^{\alpha,Anderson}(\mathbf{x})$ yields:*

$$b^{\alpha}_{D^+,T}(\mathbf{x}) \leq b^{\alpha,Anderson}(\mathbf{x}). \tag{144}$$

Proof. We have $b^{\alpha,\text{Anderson}}(\mathbf{x}) = m_{D^+}(\mathbf{x},\mathbf{u}^{And})$ where $\mathbf{u}^{And}$ satisfies $G_{\mathbf{X},\mathbf{u}^{And}}$ is a $1-\alpha$ lower confidence bound for the CDF. Therefore applying Theorem E.3 with $\boldsymbol{\ell} = \mathbf{u}^{And}$ yields the result. □

Theorem E.5 (Theorem 4.4)**.** *Let $D^+ = (-\infty, b]$. For any sample size n, for any sample value $\mathbf{x} \in D^n$,for all $\alpha \in (0,1)$, using $T(\mathbf{x}) = b^{\alpha,Anderson}(\mathbf{x})$ yields:*

$$b^{\alpha}_{D^+,T}(\mathbf{x}) \leq b^{\alpha,Hoeffding}(\mathbf{x}). \tag{145}$$

Proof. The proof follows directly from Lemma 4.2 and Theorem 4.3 □