Bounding the Mean and the Variance: Fewer Assumptions and Tighter Bounds

Technical Report UM-CS2012-008

Erik Learned-Miller ELM@CS.UMASS.EDU
Benjamin Mears BMEARS@CS.UMASS.EDU
University of Massachusetts Amherst
140 Governors Drive, Amherst, MA 01003

Abstract

A novel probabilistic bound on statistics of a one-dimensional distribution, given bounds on the support of the distribution and a sample from that distribution, is presented. No knowledge beyond bounds on the support of the unknown distribution is required. Among the sample statistics that may be bounded include the mean, higher order moments, the variance, and arbitrary percentiles. The bounds are simple functions of the sample points and are derived based on distribution-free bounds on the cumulative distribution function. We show that we can produce bounds on the mean that are guaranteed to be at least as tight as the Chernoff-Hoeffding bound. And for many common distributions, we show that our bounds are significantly tighter than the Chernoff-Hoeffding bound.

1. Introduction

In almost every scientific field, when the mean of a random variable is estimated from samples that are assumed to be independent and identically distributed (i.i.d.), it is common to accompany the estimate of the mean with confidence intervals. Given that the assumptions associated with a given method of computing a confidence interval hold, the probability that the confidence interval will contain the population mean is α, which is known as the confidence.

There are three common ways of generating confidence intervals. They each have significant drawbacks. The most common is to assume that the distribution of the sample mean is Gaussian. While this is true in the limit of infinite sample sizes, it is not true for finite sample sizes unless the population distribution is Gaussian. Thus the guarantees provided by these confidence intervals are frequently violated. The second method is to invoke the Chernoff-Hoeffding bound, which does not make assumptions about the population distribution other than to bracket its support. While the assumptions associated with the Chernoff-Hoeffding bound are a much better match to realistic applications, these bounds are disappointingly loose. Finally, there are bootstrap methods, which are appealing for their simplicity and applicability to a wide range of statistics, but which rely on asymptotic properties and so are not gauranteed to produce valid intervals for finite sample sizes.

In this report we present a significantly improved method for producing confidence intervals in a general setting. In particular, we present a novel method for producing confidence intervals for the mean that has the following properties.

1. It does not make any assumptions about the distribution of the sample mean.
2. It relies only on the assumption that the support of the population distribution can be bracketed with upper and lower values and that the data samples are i.i.d. In this sense it is like the Chernoff-Hoeffding bound.
3. It is *guaranteed* to be at least as tight, and is often significantly tighter, than the Chernoff-Hoeffding bound.

This bound is a drop-in replacement for any of the bounds currently used, and overcomes all of the problems discussed with the previous bounds. As such, we believe it has very wide applicability in practical applications.

Finally, we show how the same basic ideas can be used to bound the variance of an unknown distribution, the median, arbitrary percentiles, and a variety of other moments. It is hence a very general methodology with a large range of applications.

2. Background and Prior Work

In this section, we first review the three most common approaches for producing confidence intervals for the mean. Since our approach depends on bounds on the cumulative distribution function (CDF) given an empirical CDF, we then describe two existing bounds on the CDF. Finally, we review related work on bounding statistics of a random variable given information about the underlying distribution.

2.1. Previous Methods for constructing Confidence Intervals for the Mean

Here we review previous methods for constructing confidence intervals for the mean, starting with the Central Limit Theorem-based approach and then reviewing the Chernoff-Hoeffding bound and bootstrap methods.

2.1.1. Central Limit Theorem-Based Approach

The Central Limit Theorem states that under mild assumptions on the underlying distribution, the sample mean of i.i.d. samples has a limiting normal distribution (Casella and Berger (2001)). More precisely, if $X_1, X_2, \ldots$ are a sequence of i.i.d. random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to N(0,1), \tag{1}$$

in distribution.

Based on the Central Limit Theorem, an approximate α-confidence interval for the mean is

$$[\bar{X}_n - z_{(1-\alpha)/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{(1-\alpha)/2}\frac{\sigma}{\sqrt{n}}], \tag{2}$$

where $P(Z > z_{(1-\alpha)/2}) = (1-\alpha)/2$ when Z is a standard normal random variable.

For finite samples, the confidence interval based on the Central Limit Theorem is only an approximation and the quality of the approximation depends on both the sample size and the underlying distribution.

2.1.2. Chernoff-Hoeffding Bound

Unlike approaches based on the Central Limit Theorem, the Chernoff-Hoeffding inequality (Hoeffding (1963)) makes no asymptotic assumptions and so is valid for finite sample sizes. Let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with support on $[0, 1]$ and mean μ. Let $\bar{X}_n$ be the sample mean. The Chernoff-Hoeffding inequality states that for $0 < \epsilon < 1 - \mu$

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2} \equiv 1 - \alpha. \tag{3}$$

The Chernoff-Hoeffding bound is a function of the sample size and does not use any additional properties of the set of samples. Bennett's inequality (Bennett (1962)) is another well-known bound and is a function of both the sample size and the true variance of the distribution. But in many practical applications, the true variance is not known and so either the Bennett inequality cannot be used or the sample variance must be substituted for the true variance. In the latter case, the guarantees made by the Bennett inequality are lost.

2.1.3. Bootstrapping

Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued random variables with cumulative distribution function $F(x) = P(X \leq x)$. The empirical cumulative distribution function is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}}, \tag{4}$$

where $I_{\{E\}}$ is the indicator function which takes on a value of 1 when E is true, and 0 otherwise.

Given n i.i.d. samples from a distribution F, the bootstrap method (Efron and Tibshirani (1993)) can be used to find confidence intervals for statistical functionals of the form $T(F)$. To do so, the bootstrap approximates the distribution of $T(F)$ with the distribution of $T(F_n)$ by drawing B bootstrap samples from the original set of samples. Various approaches can then be used to derive confidence intervals for $T(F)$ using the approximate distribution of $T(F_n)$. Since the bootstrap uses a two-stage approximation, first approximating $T(F)$ with $T(F_n)$ and then approximating $T(F_n)$ with B bootstrap samples, the results are not guaranteed to be valid for finite B and n. While the quality of the second approximation can be controlled by increasing the number of bootstrap samples, the quality of the first approximation depends on the size of the available sample and often this size cannot be easily increased.

2.2. Bounds on the CDF

Like the bootstrap method, our approach also relies on the empirical CDF. Yet, our approach uses probabilistic bounds on the CDF that are valid for finite sample sizes and thus requires no asymptotic assumptions. In the following two sections, we review two probabilistic bounds on the CDF that can be used with our approach.

2.2.1. Massart Bound

Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued random variables with cumulative distribution function $F(x) = P(X \leq x)$ and $F_n(x)$ be the empirical distribution function defined by Eq. (4). The following bound on $F(x)$ is due to Dvoretzky et al. (1956) and its constant was determined by Massart (1990):

$$P(\sup_x |F(x) - F_n(x)| > \epsilon) \leq 2e^{-2n\epsilon^2} \equiv 1 - \alpha. \tag{5}$$

We refer to this as the Massart bound and it says that with probability at least α, the maximum absolute difference between the cumulative distributive function and the empirical distribution (the Kolmogorov-Smirnov statistic) is at most ϵ.

2.2.2. Order Statistics Bounds

Learned-Miller and DeStefano (2008) note that the bound given by Eq. (5) gives the same margin of uncertainty at all points of the empirical CDF but intuitively, the distribution should be able to be bounded more tightly near the ends of the support than in the middle. Based on this observation and subsequent empirical validation, Learned-Miller and DeStefano proposed an alternative bound based on the order statistics of the sample data.

Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$. Note that for samples X_i from a continuous distribution $F(x)$, the values $F(X_i)$ are uniformly distributed on $[0, 1]$ (Casella and Berger (2001)). The random variable $F(X_{(i)})$ is therefore beta distributed with parameters i and $n - i + 1$ (See, for example, Arnold et al. (2008)). This means that for each i and $\frac{1}{2} \leq \delta \leq 1$,

$$P\left(F\left(X_{(i)}\right) \in \left(\beta^{-1}_{i,n-i+1}\left(\frac{1-\delta}{2}\right), \beta^{-1}_{i,n-i+1}\left(\frac{1+\delta}{2}\right)\right)\right) = \delta, \tag{6}$$

where $\beta^{-1}_{i,n-i+1}$ is the inverse CDF of the beta distribution with parameters i and $n - i + 1$. For fixed n and δ, define $a_i = \beta^{-1}_{i,n-i+1}\left(\frac{1-\delta}{2}\right)$ and $b_i = \beta^{-1}_{i,n-i+1}\left(\frac{1+\delta}{2}\right)$. Note that while Eq. (6) is valid, the interval (a_i, b_i) is not the smallest interval that contains δ probability mass since the mode of the beta distribution does not necessarily correspond to the median and the beta distribution is also not necessarily symmetric about the median.

For fixed n and δ, Learned-Miller and DeStefano then use Monte Carlo simulations to determine the α such that

$$\alpha = P\left(F(X_{(i)}) \in (a_i, b_i), \forall i \in \{1, \ldots, n\}\right).$$

In practice, we need to compute the δ that produces a particular set of values for α and n. This is computationally expensive as we need to search for the corresponding δ and at each step in the search run a Monte Carlo simulation. Fortunately, the triplets (δ, α, n) do not depend on the particular set of sample points and so can be precomputed and stored.

It is important to note a significant difference between the bounds given by Eq. (5) and the order statistics bounds. While the former holds at all points in the support of the

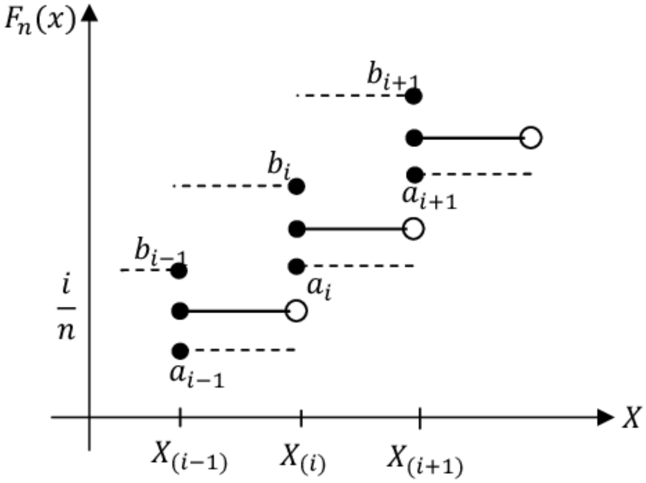

Figure 1: Extending the order statistics bounds does not add any restrictions on the CDFs beyond the requirement that they pass within the bounds at each sample point. In the diagram, the lines with the closed and open circles represent the empirical CDF. The a_i and b_i represent the upper and lower bounds obtained for the CDF at each of the sample points using the order statistics method. The dotted lines show how the upper and lower order statistics bounds may be extended past sample points.

distribution, the latter holds only at the sample points. Learned-Miller and DeStefano note that because of the monotonicity of CDFs, when $F(X_{(i)}) \in (a_i, b_i)$, for $w \in (X_{(i-1)}, X_{(i)}]$, $F(w) < b_i$. Similarly, when $F(X_{(i)}) \in (a_i, b_i)$, for $w \in [X_{(i)}, X_{(i+1)})$, $F(w) > a_i$. While these extended bounds are indeed valid, because of the monotonicity of the CDF, they do not add any restrictions on the CDFs beyond the requirement that they pass within the bounds at each sample point. In other words, in bounding the CDF we can restrict our attention to the bounds at the sample points without loosening the bounds. See Figure 1.

2.3. Related Approaches

Our approach to bounding the mean and other statistics relies on probabilistic bounds on the CDF, such as those given by the Massart and order statistics bounds. Previous work has also used bounds on either the probability distribution or cumulative distribution function to derive conclusions about statistics of the random variable. For example, given uncertainty about a prior distribution and the family of sampling distributions, Lavine (1991) developed methods for computing bounds on the posterior expectations over classes of sampling distributions that are "similar" to a given parametric family.

Smith (1995) described how bounds on the expected value of arbitrary objective functions can be computed given only limited information about the underlying distribution.

This information included properties of the distribution such as moments, entropy constraints, and bounds on the density of of the distribution.

Langewisch and Choobineh (2004) show how bounds on the mean and variance can be computed when there is known uncertainty about the underlying random variable. They define various classes of ill-specified random variables in which there is uncertainty about the sets of points or bounded intervals for which probability mass is assigned and/or the probability mass assigned to each of those sets. They then show how bounds on the variance and mean can be computed for distributions from each of these classes.

Williamson and Downs (1990) use bounds on the CDF to derive corresponding bounds for statistical functionals. While our approach also uses bounds on the CDF to derive bounds on statistics of a random variable, rather than assume the bounds are given, we derive our CDF bounds in a probabilistic manner. As a result, probabilistic statements can be made about the corresponding derived bounds. In this regard, our approach is most similar to that of Learned-Miller and DeStefano (2008). Given bounds on the CDF computed using Eq. (5), they use an efficient graph algorithm, which they term the "string-tightening algorithm," to find the CDF with maximum entropy among all valid CDFs that obey the bounds.

In summary, rather than being a function of a limited set of properties of the sample, such as the sample mean and sample size, or making assumptions about the underlying distribution, our bounds utilize the entire sample and make no assumptions other than that the distribution has known bounded support. Further, even though our bounds are a function of the entire sample, they are still efficient to compute and in the case of the mean, are guaranteed to be as tight as the Chernoff-Hoeffding bound.

3. Deriving Concentration Inequalities from CDF bounds

Given bounds on the CDF and support of a probability distribution, bounds on statistics such as the mean, variance, higher order moments, and percentiles can be computed. Note that when we require known bounds on the support of a probability distribution, all we require is an interval $[a, b]$ that contains all the non-zero probability mass of the distribution. In particular, this interval may be larger than the true support of the distribution.

In this section, we first begin by showing how the Chernoff-Hoeffding inequality can be derived from the Massart bound and then describe our method for computing probabilistic bounds on statistics of a distribution.

3.1. Relationship between Massart Bound and Chernoff-Hoeffding Inequality

Examining the inequalities (3) and (5), it appears that the two are closely related as they share the term $2e^{-2n\epsilon^2}$. To illustrate the relationship between the two, we use the following lemma.

Lemma 1 *Let X be a random variable with CDF $F(x)$, and support contained in $[a, b], 0 \leq a \leq b$. Then*

$$E[X] = \int_a^b (1 - F(x))dx. \tag{7}$$

Algorithm 1 Compute Massart Bounds

```
Input: $n$ = number of sample points
    $\{x_n\}$ = set of sample points
    $a, b$ = end points of support of R.V.
    $\alpha$ = confidence
Output: $m_U$ {Upper bound on Mean}
    $m_L$ {Lower Bound on Mean}
$\{x_{(n)}\}$ = order statistics of $\{x_n\}$
$\epsilon = \sqrt{\frac{\ln((1-\alpha)/2)}{-2 \cdot n}}$
$m_L = a \cdot \epsilon + x_{(1)} \cdot (1/n)$
$m_U = x_{(1)} \cdot \max(0, 1/n - \epsilon)$
for $i = 2 \to n$ do
    $m_U += (\max(0, i/n - \epsilon) - \max(0, (i-1)/n - \epsilon)) \cdot x_{(i)}$
    $m_L += (\min(1, i/n + \epsilon) - \min(1, (i-1)/n + \epsilon)) \cdot x_{(i)}$
end for
$m_U += \epsilon \cdot B$
```

The proof of Lemma 1 is included in Appendix A for reference.

Recall that the Massart bound places confidence envelopes that lie above and below and run parallel to the empirical CDF. Let $F(x)$ be a distribution with support on $[0, 1]$ and let $X_1, .., X_n$ be i.i.d. samples. Using Lemma 1, the absolute value of the difference between the empirical mean and the true mean, μ, is given by

$$\left|E[X] - \bar{X}_n\right| = \left|\int_0^1 (1 - F(t))dt - \int_0^1 (1 - F_n(t))dt\right| = \left|\int_0^1 F_n(t) - F(t)dt\right|.$$

The Massart bound asserts that the probability that $F_n(t)$ and $F(t)$ differ by more than ϵ is less than $2e^{-2n\epsilon^2}$. Therefore

$$P\left(\left|E[X] - \bar{X}_n\right| = \left|\int_0^1 F_n(t) - F(t)dt\right| \geq \left|\int_0^1 (F(t) + \epsilon) - F(t)dt\right| = \epsilon\right) \leq 2e^{-2n\epsilon^2},$$

thus recovering the Chernoff-Hoeffding inequality. The intuition behind the relationship between the Massart bound and the Chernoff-Hoeffding inequality is that a bound on a CDF can be translated into a bound on the mean. Using this intuition, we proceed in the following sections to derive additional bounds on the mean along with bounds on other statistics.

3.2. Bounding the Mean

Algorithm 1 presents pseudocode for computing an upper bound, m_U, and lower bound, m_L, on the mean for a given confidence level α. The algorithm uses the Massart bound together with the intuition, which is formalized in the following lemma, that given bounds on the CDF of a probability distribution, the CDF that maximizes the mean is the one that runs along the lower envelope of the bound. Similarly, the CDF that minimizes the mean is the one that runs along the upper envelope.

Lemma 2 *Let X be a random variable with CDF $F(x)$, and support contained in $[a, b], 0 \leq a \leq b$. Let $L(x)$ and $U(x)$ be lower and upper bounds on $F(x)$. Then among the CDFs that obeys the bounds, the CDF with the maximum mean is given by $L(x)$ and the CDF with the minimum mean is given by $U(x)$.*

Proof To maximize the mean, we wish to maximize Eq. (7). This is equivalent to minimizing $\int_a^b F(t)dt$. The CDF that obeys the bounds given by $L(x)$ and $U(x)$ and minimizes this term is given by $L(x)$.

Similarly, to minimize the mean, we wish to minimize Eq. (7). This is equivalent to maximizing $\int_a^b F(t)dt$. The CDF that obeys the bounds given by $L(x)$ and $U(x)$ and maximizes this term is given by $U(x)$. ■

Using Lemma 2, we can prove the correctness of Algorithm 1.

Theorem 3 *Let X be a random variable with CDF $F(x)$, support contained in $[a, b], 0 \leq a \leq b$, and mean μ. Let $\{x_n\}$ be a set of n i.i.d. samples from $F(x)$ and m_U and m_L be the upper and lower bounds, respectively, computed using Algorithm 1 with confidence level α. Then*

$$P(\mu \in (m_L, m_U)) \geq \alpha.$$

Proof Denote $F_n(x)$ to be the empirical distribution of the $\{x_n\}$ and $L(x)$ and $U(x)$ to be the upper and lower bounds on $F(x)$ computed using the Massart bound with confidence level α. By construction, m_L is computed from $U(x)$ and m_U is computed from $L(x)$. Thus, by Lemma 2, m_L and m_U correspond to the means of the CDFs that have the minimum and maximum means, respectively, among all CDFs that obey the bounds given by $L(x)$ and $U(x)$. And since by construction $F(x)$ lies within the bounds given by $L(x)$ and $U(x)$, we have

$$P(\mu \in (m_L, m_U)) \geq \alpha.$$

■

Note that while Lemmas 1 and 2 and Theorem 3 require that the random variable takes on only nonnegative values, they can be extended to random variables that take on negative values by defining an auxiliary random variable $Y = X - a$, computing the bounds on $E[Y]$ and then using the linearity of expectation to transform these bounds on Y to bounds on X.

The next theorem shows that for a given confidence level, the bounds computed using Algorithm 1 are at least as tight as those obtained from the Chernoff-Hoeffding inequality. We start with the following lemma.

Lemma 4 *Let X be a random variable with CDF $F(x)$, and support contained in $[0, 1]$. Let $\{x_n\}$ be a set of $n \geq 3$ i.i.d. samples, $\bar{x}_n$ be the sample mean of $\{x_n\}$, $\alpha \in (0, 1)$, and let*

$$\epsilon = \sqrt{\ln((1 - \alpha)/2)/(-2n)}. \tag{8}$$

Let $[m_L, m_U]$ be the lower and upper bounds obtained using Algorithm 1. Then

$$\max(\bar{x}_n - m_L, m_U - \bar{x}_n) \leq \epsilon. \tag{9}$$

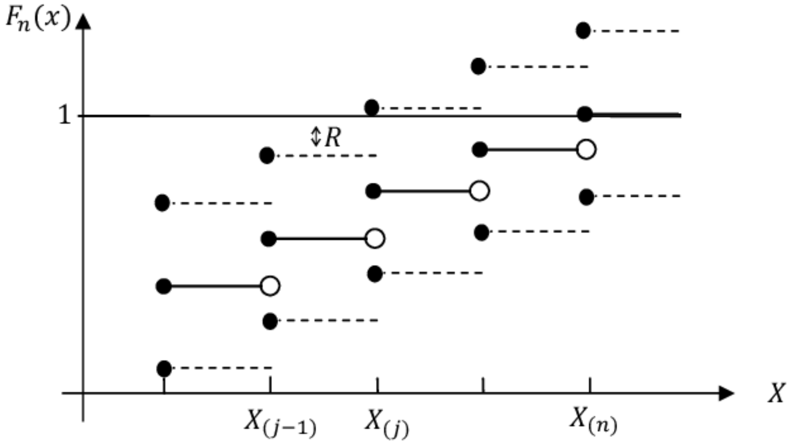

Figure 2: An illustration of the quantity R used in the proof of Theorem 4. The central solid lines with the open and closed circles represent the empirical CDF while the upper and lower dotted lines represent the upper and lower Massart bounds on the CDF.

Proof We show $\bar{x}_n - m_L \leq \epsilon$. The reasoning for $m_U - \bar{x}_n \leq \epsilon$ is similar.

First note that the ϵ computed in line 2 of Algorithm 1 is given by Eq. (8). Also note that

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}.$$

After line 3, $m_L = (1/n)x_{(1)}$. Let

$$j = \min\left(\{j \in [2, n-1]; (\min(1, j/n + \epsilon) - \min(1, (j-1)/n + \epsilon)) < 1/n\}\right).$$

Such a j exists because, as shown by Learned-Miller and DeStefano (2008), if $\epsilon = \sqrt{\ln((1-\alpha)/2)/(-2n)}$, then $\epsilon > \frac{1}{n}$, $\forall n \geq 3$. Then define

$$\sum_{i=j}^{n} (\min(1, i/n + \epsilon) - \min(1, (i-1)/n + \epsilon))$$
$$= (\min(1, j/n + \epsilon) - \min(1, (j-1)/n + \epsilon)) = R$$

and so

$$\sum_{i=j}^{n} \frac{1}{n} = \epsilon + R.$$

See Figure 2 for a pictorial representation of the quantity R. Now we have

$$\begin{aligned}
\bar{x}_n - m_L &= \frac{1}{n}\sum_{i=1}^{n} x_{(i)} - m_L && (10)\\
&= \frac{1}{n}\sum_{i=1}^{j-1} x_{(i)} + \frac{1}{n}\sum_{i=j}^{n} x_{(i)} - \\
&\quad \left(\frac{1}{n}\sum_{i=1}^{j-1} x_{(i)} + \sum_{i=j}^{n} \left(\min\left(1, i/n+\epsilon\right) - \min\left(1, (i-1)/n+\epsilon\right)\right) x_{(i)}\right) && (11)\\
&= \sum_{i=j}^{n} \left(\frac{1}{n} - \left(\min\left(1, i/n+\epsilon\right) - \min\left(1, (i-1)/n+\epsilon\right)\right) x_{(i)}\right. && (12)\\
&\leq \sum_{i=j}^{n} \left(\frac{1}{n} - \left(\min\left(1, i/n+\epsilon\right) - \min\left(1, (i-1)/n+\epsilon\right)\right)\right. && (13)\\
&= \left(\sum_{i=j}^{n} \frac{1}{n}\right) - \left(\sum_{i=j}^{n} \left(\min\left(1, i/n+\epsilon\right) - \min\left(1, (i-1)/n+\epsilon\right)\right)\right) && (14)\\
&= \epsilon + R - R = \epsilon. && (15)
\end{aligned}$$

In going from Eq. (12) to Eq. (13) we use the fact that by construction the terms in the summation in Eq. (13) are positive and $x_i \leq 1, \forall i \in [1, \ldots, n]$. ■

With this lemma, it can be shown that our Massart-based bounds are at least as tight as those provided by the Chernoff-Hoeffding inequality.

Theorem 5 *For a given confidence level α, the bounds computed using Algorithm 1 are at least as tight as the bounds provided by the Chernoff-Hoeffding inequality.*

Proof Let $\{x_n\}$ be a set of n i.i.d. samples from a distribution with CDF $F(x)$, support contained in $[0, 1]$, and mean μ. Let $\alpha \in (0, 1)$ with corresponding ϵ given by Eq. (8). Note that this ϵ is equal to the ϵ obtained by solving Eq. (3) for ϵ given fixed α. From the Chernoff-Hoeffding inequality, we have

$$P(\mu \in (\bar{x}_n - \epsilon, \bar{x}_n + \epsilon)) \geq \alpha.$$

which is equivalent to the probabilistic interpretation of the output of Algorithm 1 using Theorem 3 and Lemma 4 when we have equality in Eq. (9) with

$$\bar{x}_n - m_L = m_U - \bar{x}_n = \epsilon.$$

■

3.3. Bounding the Variance

Compared to bounds on the mean of a distribution, relatively few bounds on the variance of unrestricted distributions exist. In this section, we derive a bound that requires only knowledge of a finite interval, not necessarily tight, that contains the support of the distribution. Knowing that the support of a distribution is contained in $[a, b]$ results in an upper bound of $(b-a)^2/4$ corresponding to probability mass equally split between the two end points of the interval. Jacobson (1969) showed that if it is also known the distribution is unimodal, then this bound can be tightened to $(b-a)^2/9$.

Our variance bound is derived from bounds on the CDF and the following lemma.

Lemma 6 *Let X be a random variable with CDF $F(x)$, and support contained in $[0, 1]$. Then the variance of X is given by*

$$2\int_0^1 x(1-F(x))dx - \left(\int_0^1 (1-F(x))dx\right)^2. \tag{16}$$

The proof of Lemma 6 is included in Appendix B for reference. Using this lemma, we now prove the following theorem.

Theorem 7 *Let X be a random variable with CDF $F(x)$, and support contained in $[0, 1]$. Let $L(x)$ and $U(x)$ be lower and upper bounds on the CDF. Then the variance-maximizing or -minimizing CDF that obeys these bounds is either identically $L(x)$, identically $U(x)$, begins on $L(x)$ and then has a jump discontinuity to $U(x)$, or begins on $U(x)$ and then moves horizontally to $L(x)$.*

Proof We want to find the extrema of the functional specified by Eq. (16) subject to the constraints

$$F'(x) \geq 0, \tag{17}$$

$$F(x) \geq L(x), \tag{18}$$

$$F(x) \leq U(x). \tag{19}$$

Forming the Lagrangian gives

$$\begin{aligned} L = {} & 2\int_0^1 x(1-F(x))dx - \left(\int_0^1 (1-F(x))dx\right)^2 \\ & + \int_0^1 \lambda_{1,x}(L(x)-F(x))dx + \int_0^1 \lambda_{2,x}(F(x)-U(x))dx - \int_0^1 \lambda_{3,x}F'(x)dx. \end{aligned}$$

At an extremum, the derivative of L with respect to each $F(x)$ is zero:

$$\begin{aligned} \frac{\partial L}{\partial F(x)} &= -2x + 2\int_0^1 (1-F(u))du - \lambda_{1,x} + \lambda_{2,x} - \lambda_{3,x}F^{(2)}(x) \\ &= 2(1-x) - 2\int_0^1 F(u)du - \lambda_{1,x} + \lambda_{2,x} - \lambda_{3,x}F^{(2)}(x) = 0 \end{aligned}$$

Further, the Karush-Kuhn-Tucker conditions must be satisfied:

$$\begin{aligned}\lambda_{1,x}(L(x) - F(x)) &= 0,\\ \lambda_{2,x}(F(x) - U(x)) &= 0,\\ \lambda_{3,x}F'(x) &= 0,\\ \lambda_{i,x} &\geq 0.\end{aligned}$$

First note that if $\lambda_{1,x} = 0$ and $\lambda_{2,x} = 0$, then

$$2(1 - x) = 2\int_0^1 F(u)du + \lambda_{3,x}F^{(2)}(x).$$

Note that in particular, there is only one x for which this is true and $\lambda_{3,x} = 0$ which in turn implies that there can be at most one point on the variance-maximizing or -minimizing CDF that lies inside the bounds and for which $F'(x)$ is positive. So other than this point, if the variance-maximizing or -minimizing CDF is within (but not on) the bounds, the derivative of the CDF must be zero.

There are thus only four possible cases we must consider for the variance-maximizing or minimizing CDF.

1. It starts and stays on the lower bound.
2. It starts and stays on the upper bound.
3. It starts on the lower bound and then jumps to the upper bound.
4. It starts on the upper bound and then moves horizontly to the lower bound, proceeding to follow the lower bound.

■

For Case 3, we must solve a single variable optimization problem at each of the horizontal edges of the lower bound on the CDF. Let $x_{(i)}$ and $x_{(i+1)}$ be two consecutive order statistics. Then for the horizontal edge between $x_{(i)}$ and $x_{(i+1)}$ we must optimize the following with respect to t, where $x_{(i)} \leq t < x_{(i+1)}$.

$$2\int_0^t x(1 - L(x))dx + 2\int_t^1 x(1 - U(x))dx - \left(\int_0^t (1 - L(x))dx + \int_t^1 (1 - U(x))dx\right)^2 \tag{20}$$

Eq. 20 can be rewritten as

$$\begin{aligned}&2\int_0^{x_{(i)}} x(1 - L(x))dx + 2\int_{x_{(i)}}^t x(1 - L(x_{(i)}))dx + 2\int_t^{x_{(i+1)}} x(1 - U(x_{(i)}))dx + 2\int_{x_{(i+1)}}^1 x(1 - U(x))dx - \\ &\qquad \left(\int_0^{x_{(i)}} (1 - L(x))dx + \int_{x_{(i)}}^t (1 - L(x_{(i)}))dx + \int_t^{x_{(i+1)}} (1 - U(x_{(i)}))dx + \int_{x_{(i+1)}}^1 (1 - U(x))dx\right)^2.\end{aligned} \tag{21}$$

Now explicitly computing the integrals that depend on t.

$$2\int_0^{x_{(i)}} x(1-L(x))dx + (t^2 - x_{(i)}^2)(1-L(x_{(i)})) + (x_{(i+1)}^2 - t^2)(1-U(x_{(i)})) + 2\int_{x_{(i+1)}}^1 x(1-U(x))dx - \left(\int_0^{x_{(i)}} (1-L(x))dx + (t-x_{(i)})(1-L(x_{(i)})) + (x_{(i+1)}-t)(1-U(x_{(i)})) + \int_{x_{(i+1)}}^1 (1-U(x))dx\right)^2. \tag{22}$$

Now taking the derivative of Eq. 22 w.r.t. t, setting equal to zero, and solving for t gives:

$$t = \frac{\int_0^{x_{(i)}}(1-L(x))dx + \int_{x_{(i+1)}}^1 (1-U(x))dx + (x_{(i+1)} - x_{(i)}) + (x_{(i)}L(x_{(i)}) - x_{(i+1)}U(x_{(i)}))}{1-(U(x_{(i)}) - L(x_{(i+1)}))} \tag{23}$$

In addition to checking the point t if $x_{(i)} < t < x_{(i+1)}$, the endpoints, $x_{(i)}$ and $x_{(i+1)}$ must also be checked.

Similarly for Case 4, we must simply solve a single variable optimization problem at each of the sample points to find the horizontal transition.

3.4. Bounding Higher-Order Moments

In order to derive bounds on higher-order moments from bounds on the CDF, the moments can be categorized into four cases:

1. *Odd powers*-The CDF that maximizes the moment follows the lower envelope and the CDF that minimizes the moment follows the upper envelope.

2. *Even power and lower bound of the distribution's support is nonnegative*-Once again, the CDF that maximizes the moment follows the lower envelope and the CDF that minimizes the moment follows the upper envelope.

3. *Even power and upper bound of the distribution's support is nonpositive*-The CDF that maximizes the moment follows the upper envelope and the CDF that minimizes the moment follows the lower envelope.

4. *Even power and lower and upper bounds of the distribution's support are negative and postive, respectively*-The CDF that maximizes the variance begins on the lower envelope and jumps to the upper envelope at $x = 0$. The CDF that minimizes the variance begins on the upper envelope and transitions horizontally to the lower envelope. The proof is very similar to the proof of the variance bound and hence is omitted.

3.5. Bounding Arbitrary Percentiles

From bounds on the CDF, corresponding bounds may also be placed on any percentile. For a distribution with CDF $F(x)$, the percentile $\mathcal{P}_i$ for $i \in [0,1]$ can be defined as $F^{-1}(i)$. Given an upper bound, $U(x)$, and lower bound, $L(x)$, on the CDF of a distribution, it is trivial to bound the value of a percentile. Define

$$\mathcal{L}_i = \inf\{x|U(x) = i\} \qquad \mathcal{U}_i = \sup\{x|L(x) = i\}.$$

If the upper and lower bounds on the CDF hold with probability at least α, then

$$P(\mathcal{P}_i \in (\mathcal{L}_i, \mathcal{U}_i)) \geq \alpha.$$

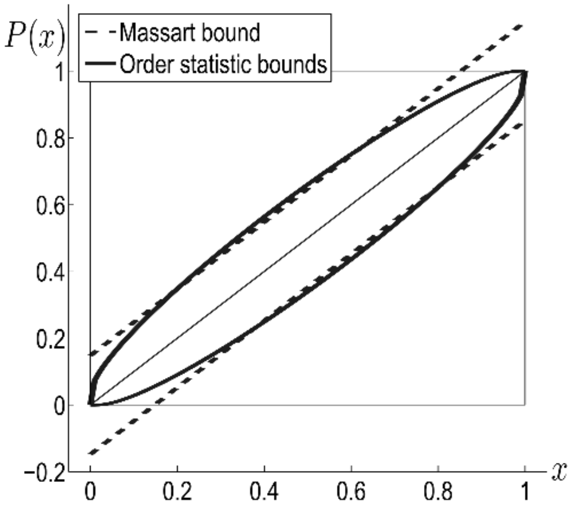

Figure 3: A comparison of the shape of the Massart bound and order statistics bound on a CDF for a uniform distribution. The Massart bound is equally tight across all values of the domain while the order statistics bound is tighter near the ends of the support of the distribution and looser in the middle. Figure reproduced with permision from Learned-Miller and DeStefano (2008).

3.6. One-Sided Bounds

In many cases, bounds may be tightened if we only wish to upper or lower limit the statistic of interest. For example, if we only want to upper bound the mean, μ, of a random variable with support contained in $[0,1]$ given n i.i.d. samples, the Chernoff-Hoeffding inequality states that for $0 < \epsilon < 1 - \mu$

$$P(\bar{X}_n - \mu \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

For some of the bounds on the statistics discussed above, the computation of the upper or lower limits depends only on either the upper or lower bound on the CDF. Thus, one-sided versions of the Massart and order statistics bounds on the CDFs can be used to produce tighter upper or lower bounds for these statistics. For the order statistics bound, we can also adjust the confidence intervals at each sample point in order to get tighter bounds on the statistics. The bound on the CDF discussed in Section 2.2.2 has an elongated-oval shape since when the order statistics are bounded with equal confidence, the bounds are smaller at the ends of the support and larger in the middle of the support (Figure 3). For certain statistics though, we can alter this shape to result in a tighter bound. For example, to upper bound the mean it is better to have a tighter bound for larger values in the domain at the expense of looser bounds for smaller values.

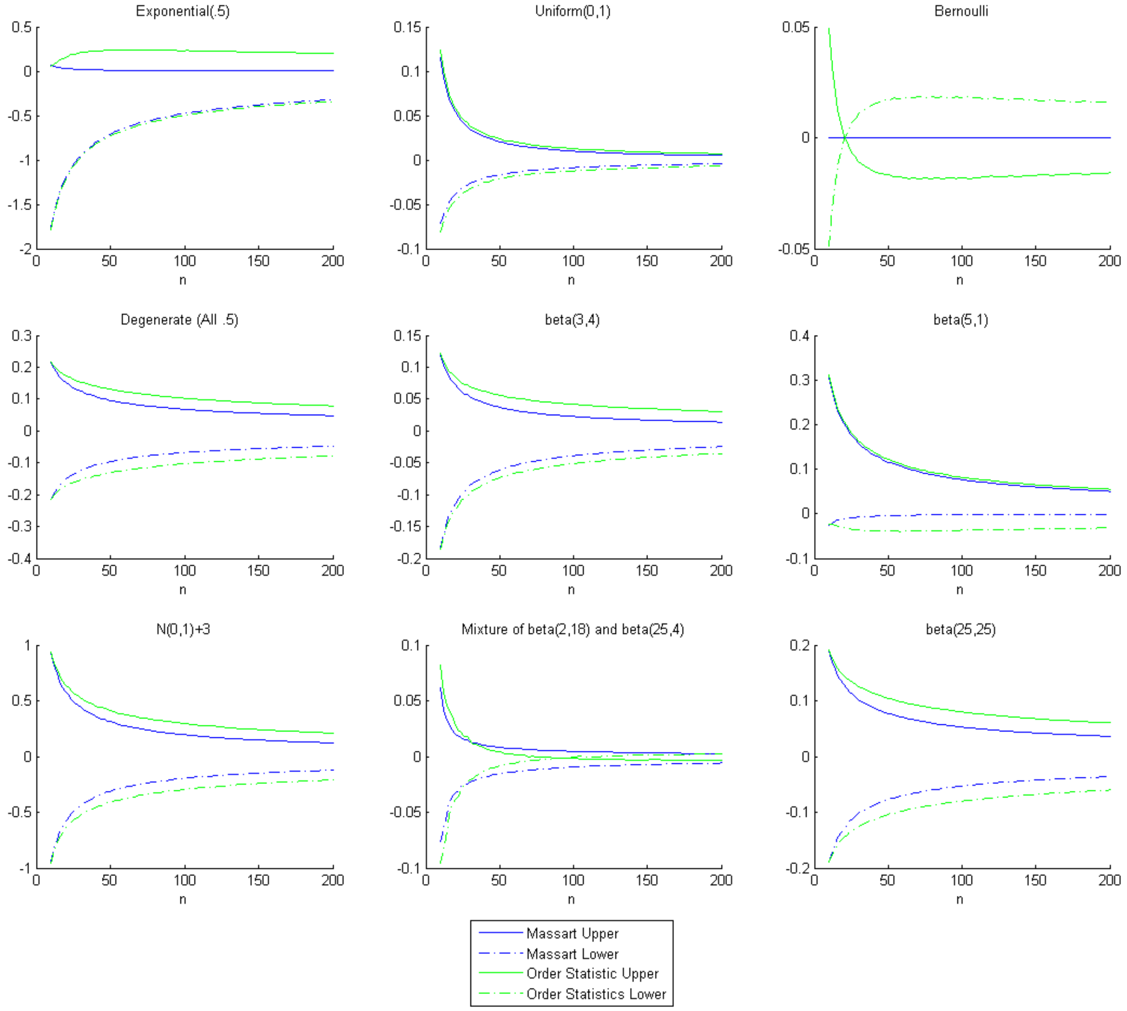

Figure 4: Plots of the average difference between the Chernoff-Hoeffding bound and the corresponding CDF-based bound as a function of the sample size. We set $\alpha = .95$ in these experiments. For upper bounds, larger values indicate tighter bounds relative to the Chernoff-Hoeffding bound and for lower bounds, smaller values indicate tighter bounds relative to the Chernoff-Hoeffding bound. Note that the curve for the Massart-based upper bound cannot be seen in the Bernoulli plot because it is covered by the curve for the Massart-based lower bound.

4. Experimental Results

To evaluate empirically the tightness of each of the bounds, we ran experiments on a variety of distributions. For each distribution and each sample size n, we sampled n points from the distribution and computed the corresponding bounds over multiple trials and then plotted the difference between the Chernoff-Hoeffding bound and the corresponding CDF-based bound. Thus for upper bounds, larger values indicate tighter bounds relative to the Chernoff-Hoeffding bound and for lower bounds, smaller values indicate tighter bounds relative to the Chernoff-Hoeffding bound. The results are shown in Figure 4. Note the different scales on the y-axis. For distributions without finite support, we used rejection sampling to reject points that were outside user-specified bounds.

Note that in many cases the bound based on order statistics is as tight or tighter than the other bounds, although it is possible for the bound to be looser than the Massart-based and Chernoff-Hoeffding bounds for certain distributions (most notably, the Bernoulli distribution). Also notice that the Massart-based bound is always at least as tight as the Chernoff-Hoeffding bound and usually tighter. It is also important to note that both of the CDF-based bounds tend to be significantly tighter than the Chernoff-Hoeffding bound for small sample sizes.

In Figure 5 we show the results of a similar set of experiments, this time varying α with fixed n. The results show that the relationship between the relative tightness of the different bounds tends to hold across different confidence levels.

To evaluate empirically the tightness of the variance bounds, we ran experiments on a variety of distributions. For each distribution and each sample size n, we sampled n points from the distribution and computed the corresponding bounds over multiple trials and then plotted the average upper and lower bounds. For comparison, we also plotted the true variance and the trivial upper bound $(b-a)^2/4$. The results are shown in Figure 6.

5. Discussion and Conclusion

We have shown that by using probabilistic bounds on the CDF of a distribution, corresponding bounds can be computed for statistics such as the mean, the variance, higher-order moments, and percentiles. While classical bounds are a function of a relatively small number of inputs, such as the sample size, the sample mean, and if known, the variance of the distribution, our bounds use properties of the entire sample.

For the Massart-based bound on the mean, our bounds are guaranteed to be at least as tight as those produced by the Chernoff-Hoeffding inequality. And since the algorithm to compute these bounds is simple and efficient, with the most expensive step being the sorting of the input samples, it can easily be substituted for the Chernoff-Hoeffding inequality in applications that presently utilize the latter, such as randomized algorithms.

References

B.C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics*, volume 54. Society for Industrial Mathematics, 2008.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, pages 33–45, 1962.

G. Casella and R.L. Berger. *Statistical Inference.* Duxbury Press, 2001.

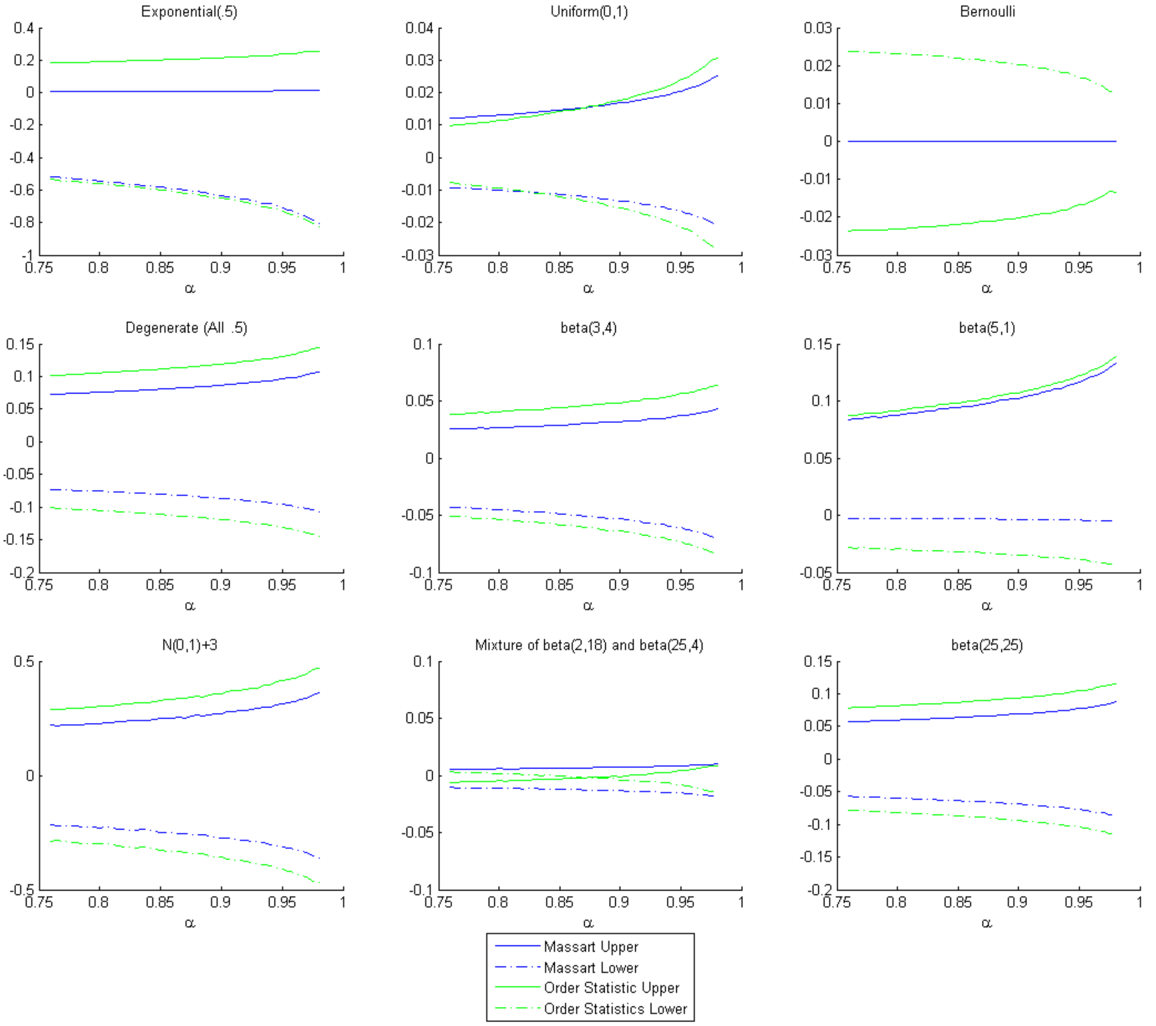

Figure 5: Plots of the average difference between the Chernoff-Hoeffding bound and the corresponding CDF-based bound as a function of α. We set $n = 50$ in these experiments. For upper bounds, larger values indicate tighter bounds relative to the Chernoff-Hoeffding bound and for lower bounds, smaller values indicate tighter bound relatives to the Chernoff-Hoeffding bound. Note that the curve for the Massart-based upper bound cannot be seen in the Bernoulli plot because it is covered by the curve for the Massart-based lower bound.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

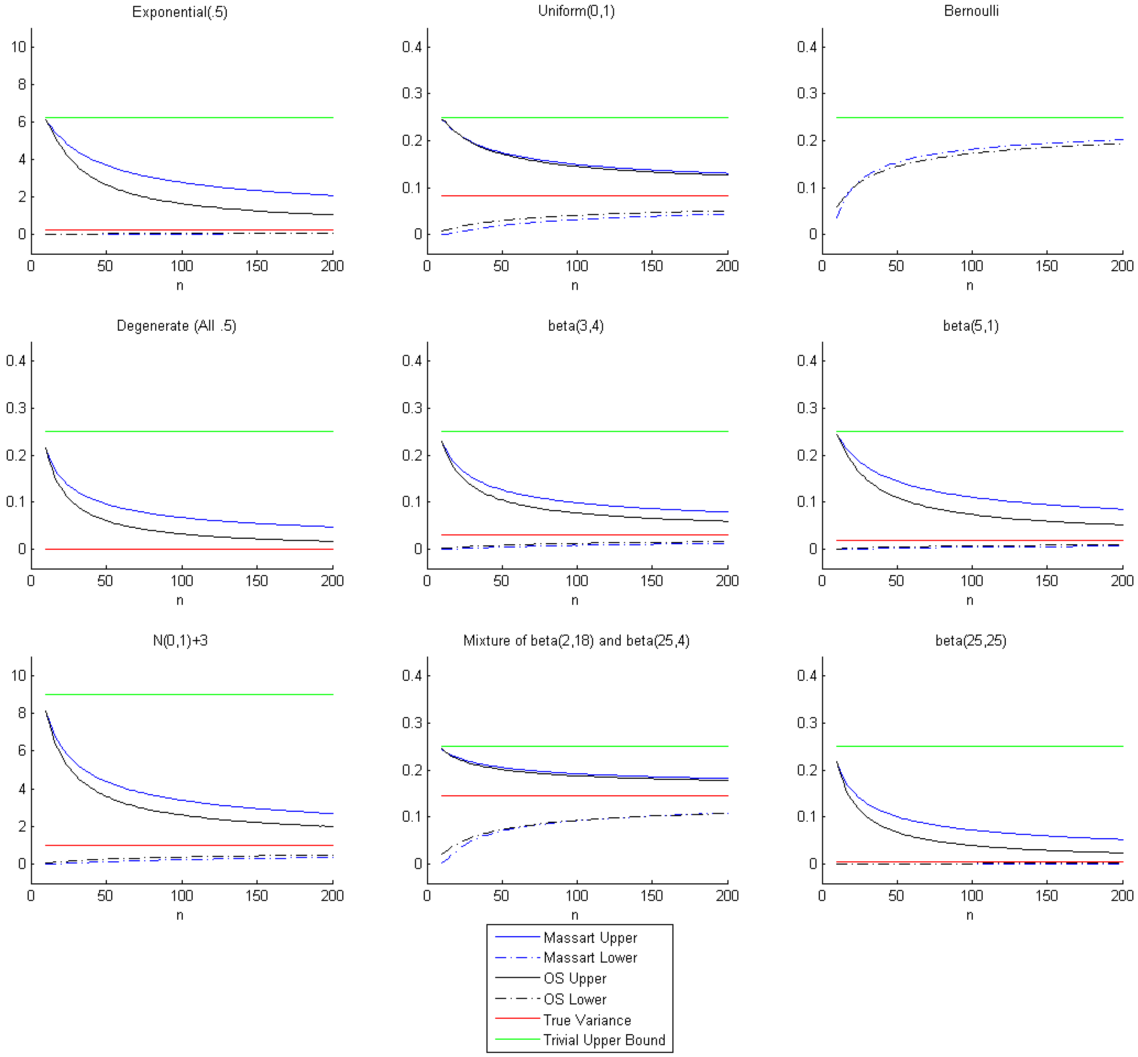

Figure 6: Plots of the average upper and lower bounds on the variance as a function of the sample size. We set $\alpha = .95$ in these experiments. For comparison, we also plotted the true variance and the trivial upper bound $(b-a)^2/4$. Note that the lines for the lower bounds for the degenerate distribution and the lines for the upper bounds for the Bernoulli distribution cannot be seen in the plot because they are hidden by the lines for the true variances.

B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1993.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30, 1963.

H.I. Jacobson. The maximum variance of restricted unimodal distributions. *The Annals of Mathematical Statistics*, 40(5):1746–1752, 1969.

A.T. Langewisch and F.F. Choobineh. Mean and variance bounds and propagation for ill-specified random variables. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 34(4):494–506, 2004.

M. Lavine. Sensitivity in Bayesian statistics: The prior and the likelihood. *Journal of the American Statistical Association*, pages 396–399, 1991.

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

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

J.E. Smith. Generalized Chebychev inequalities: Theory and applications in decision analysis. *Operations Research*, pages 807–825, 1995.

R.C. Williamson and T. Downs. Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158, 1990.

Appendix A. Proof of Lemma 1

We show the proof for continuous distributions. The proof for discrete distributions is similar with summations replacing integrals.

$$\begin{aligned}
\int_a^b (1-F(x))dx &= \int_0^b P(X \geq x)dx \\
&= \int_0^b \int_x^b f(t)dt\ dx \\
&= \int_0^b \int_0^t f(t)dx\ dt \\
&= \int_0^b tf(t)dt = E[X].
\end{aligned}$$

Appendix B. Proof of Lemma 6

We show the proof for continuous distributions. The proof for discrete distributions is similar with summations replacing integrals.

Since $V(X) = E[X^2] - E[X]^2$ and from Lemma 1,

$$E[X] = \int_0^1 (1-F(x))dx,$$

we must only show that

$$E[X^2] = 2\int_0^1 x(1-F(x))dx.$$

$$\begin{aligned}
2\int_0^1 x(1-F(x))dx &= 2\int_0^1 xP(X \geq x)dx \\
&= 2\int_0^1 \int_x^1 xf(t)dt\ dx \\
&= 2\int_0^1 \int_0^t xf(t)dx\ dt \\
&= \int_0^1 t^2 f(t)dt = E[X^2].
\end{aligned}$$