Accepted Manuscript

Finite sampling inequalities: An application to two-sample Kolmogorov-Smirnov statistics

Evan Greene, Jon A. Wellner

PII: S0304-4149(16)30042-4
DOI: http://dx.doi.org/10.1016/j.spa.2016.04.020
Reference: SPA 2947

To appear in: *Stochastic Processes and their Applications*

Received date: 17 December 2014
Accepted date: 26 November 2015

Please cite this article as: E. Greene, J.A. Wellner, Finite sampling inequalities: An application to two-sample Kolmogorov-Smirnov statistics, *Stochastic Processes and their Applications* (2016), http://dx.doi.org/10.1016/j.spa.2016.04.020

Finite sampling inequalities: an application to two-sample Kolmogorov-Smirnov statistics

Evan Greene

Department of Statistics
University of Washington
Seattle, WA 98195-4322

Jon A. Wellner

Department of Statistics
University of Washington
Seattle, WA 98195-4322

Abstract

We review a finite-sampling exponential bound due to Serfling and discuss related exponential bounds for the hypergeometric distribution. We then discuss how such bounds motivate some new results for two-sample empirical processes. Our development complements recent results by Wei and Dudley (2012) concerning exponential bounds for two-sided Kolmogorov - Smirnov statistics by giving corresponding results for one-sided statistics with emphasis on "adjusted" inequalities of the type proved originally by Dvoretzky et al. (1956) and by Massart (1990) for one-sample versions of these statistics.

Keywords: Bennett inequality, finite sampling, Hoeffding inequality, hypergeometric distribution, two-samples, Kolmogorov-Smirnov statistics, exponential bounds.
2000 MSC: 60F15, 60F17, 62E20, 62F12, 62G20

1. Introduction: Serfling's finite sampling exponential bound

Suppose that $\{c_1, \ldots, c_N\}$ is a finite population with each $c_i \in \mathbb{R}$. For $n \leq N$, let $Y_1, \ldots, Y_n$ be a sample drawn from $\{c_1, \ldots, c_N\}$ without replacement; we can regard the finite population $\{c_1, \ldots, c_N\}$ as an urn containing N balls

labeled with the numbers $c_1, \ldots, c_N$. Some notation: we let

$$\mu_N = N^{-1}\sum_{i=1}^{N} c_i \equiv \overline{c}_N, \qquad \sigma_N^2 = N^{-1}\sum_{i=1}^{N}(c_i - \overline{c}_N)^2,$$
$$a_N \equiv \min_{1\le i\le N} c_i, \qquad b_N \equiv \max_{1\le i\le N} c_i,$$
$$f_n \equiv \frac{n-1}{N-1}, \qquad \text{and} \qquad f_n^* \equiv \frac{n-1}{N}.$$

It is well-known (see e.g. Rice (2007), Theorem B, page 208) that $\overline{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$ satisfies $E(\overline{Y}_n) = \mu_N$ and

$$Var(\overline{Y}_n) = \frac{\sigma_N^2}{n}\left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma_N^2}{n}(1 - f_n). \tag{1}$$

Serfling (1974), Corollary 1.1, shows that for all $\lambda > 0$

$$P(\sqrt{n}(\overline{Y}_n - \mu_N) \ge \lambda) \le \exp\left(-\frac{2\lambda^2}{(1-f_n^*)(b_N - a_N)^2}\right). \tag{2}$$

This inequality is an inequality of the type proved by Hoeffding (1963) for sampling with replacement and more generally for sums of independent bounded random variables. Comparing (1) and (2), it seems reasonable to ask whether the factor f_n^* in (2) can be improved to $f_n \equiv (n-1)/(N-1)$? Indeed Serfling ends his paper (on page 47) with the remark: "(it is) also of interest to obtain (2) with the usual sampling fraction instead of f_n^*". Note that when $n = N$, $\overline{Y}_n = \mu_N$, and hence the probability in (2) is 0 for all $\lambda > 0$, and the conjectured improvement of Serfling's bound agrees with this while Serfling's bound itself is positive when $n = N$.

Despite related results due to Kemperman (1973a,b,c), it seems that a definitive answer to this question is not yet known.

A special case of considerable importance is the case when the numbers on the balls in the urn are all 1's and 0's: suppose that $c_1 = \cdots = c_D = 1$, while $c_{D+1}, \ldots, c_N = 0$. Then $X \equiv n\overline{Y}_n = \sum_{i=1}^{n} Y_i$ is well-known to have a Hypergeometric(n, D, N) distribution given by

$$P\left(\sum_{i=1}^{n} Y_i = k\right) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}, \quad \max\{0, D+n-N\} \le k \le \min\{n, D\}.$$

In this special case $\mu_N = D/N$, $\sigma_N^2 = \mu_N(1-\mu_N)$, while $b_N = 1$ and $a_N = 0$. Thus Serfling's inequality (2) becomes

$$P(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{1-f_n^*}\right) \quad \text{for all } \lambda > 0,$$

and the conjectured improvement is

$$P(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{1-f_n}\right) \quad \text{for all } \lambda > 0.$$

Despite related results due to Chvátal (1979) and Hush and Scovel (2005) it seems that a bound of the form in the last display remains unknown.

We should note that an exponential bound of the Bennett type for the tails of the hypergeometric distribution does follow from results of Vatutin and Mikhaĭlov (1982) and Ehm (1991); see also Pitman (1997).

Theorem 1. *(Ehm, 1991) If* $1 \leq n \leq D \wedge (N-D)$*, then* $\sum_{i=1}^n Y_i \stackrel{d}{=} \sum_{i=1}^n X_i$ *where* $X_i \sim$ *Bernoulli*(π_i)*, with* $\pi_i \in (0,1)$*, are independent.*

It follows from Theorem 1 that

$$n(D/N) = E\left(\sum_1^n Y_i\right) = E\left(\sum_1^n X_i\right) = \sum_{i=1}^n \pi_i,$$
$$n\frac{D}{N}\left(1-\frac{D}{N}\right)(1-f_n) = Var\left(\sum_1^n Y_i\right) = Var\left(\sum_1^n X_i\right) = \sum_{i=1}^n \pi_i(1-\pi_i).$$

Furthermore, by applying Theorem 1 together with Bennett's inequality (Bennett (1962); see also Shorack and Wellner (1986), page 851), we obtain the following exponential bound for the tail of the hypergeometric distribution:

Corollary 1. *If* $1 \leq n \leq D \wedge (N-D)$*, then for all* $\lambda > 0$

$$P(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2\sigma_N^2(1-f_n)}\psi\left(\frac{\lambda}{\sqrt{n}\sigma_N^2(1-f_n)}\right)\right)$$

where $\mu_N \equiv D/N$*,* $\sigma_N^2 \equiv \mu_N(1-\mu_N)$*,* $1-f_n \equiv 1-(n-1)/(N-1)$ *is the finite-sampling correction factor, and* $\psi(y) \equiv 2y^{-2}h(1+y)$ *where* $h(y) \equiv y(\log y - 1) + 1$*.*

Since $\sigma_N^2 = \mu_N(1-\mu_N) \leq 1/4$, the inequality of the corollary yields a further bound which is quite close to the conjectured Hoeffding type improvement of Serfling's bound, and which now has the desired finite-sampling correction factor $1 - f_n$:

Corollary 2.

$$\begin{aligned} P(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda) &\leq \exp\left(-\frac{2\lambda^2}{(1-f_n)}\psi\left(\frac{\lambda}{\sqrt{n}\sigma_N^2(1-f_n)}\right)\right) \\ &\leq \exp\left(-\frac{2\lambda^2}{(1-f_n)}\psi\left(\frac{1}{\sigma_N^2(1-f_n)}\right)\right). \end{aligned}$$

By considerations related to the work of Talagrand (1994) and León and Perron (2003), the authors of the present paper have succeeded in proving the following exponential bound.

Theorem 2. *(Greene and Wellner (2015); Greene (2016)) Suppose that* $\sum_{i=1}^n Y_i \sim$ *Hypergeometric*(n, D, N). *Define* $\mu_N = D/N$ *and suppose* $N > 4$ *and* $2 \leq n < D \leq N/2$. *Then for all* $0 < \lambda < \sqrt{n}/2$ *we have*

$$\begin{aligned} &P\left(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda\right) \\ &\leq \sqrt{\frac{1}{2\pi\lambda^2}}\left(\frac{1}{2}\right)\sqrt{\left(\frac{N-n}{N}\right)\left(\frac{\sqrt{n}+2\lambda}{\sqrt{n}-2\lambda}\right)\left(\frac{N-n+2\sqrt{n}\lambda}{N-n-2\sqrt{n}\lambda}\right)} \\ &\quad \cdot \exp\left(-\frac{2}{1-\frac{n}{N}}\lambda^2\right)\exp\left(-\frac{1}{3}\left(1+\frac{n^3}{(N-n)^3}\right)\frac{\lambda^4}{n}\right). \end{aligned}$$

The proof of this bound, along with a complete analogue for the hypergeometric distribution of a bound of Talagrand (1994) for the binomial distribution, appears in Greene and Wellner (2015) and in the forthcoming Ph.D. thesis of the first author, Greene (2016).

The bound given in Theorem 2 involves a still better finite-sampling correction factor, namely $1 - \overline{f}_n = 1 - n/N$, which has also appeared in Lo (1986) in the context of a Bayesian analysis of finite sampling. Note that as $N \to \infty$, the above bound yields

$$\begin{aligned} &\limsup_{N\to\infty} P\left(\sqrt{n}(\overline{Y}_n - \mu_N) \geq \lambda\right) \\ &\leq \sqrt{\frac{1}{2\pi\lambda^2}}\left(\frac{1}{2}\right)\sqrt{\left(\frac{\sqrt{n}+2\lambda}{\sqrt{n}-2\lambda}\right)} \cdot \exp\left(-2\lambda^2 - \frac{\lambda^4}{3n}\right), \end{aligned}$$

a bound which improves slightly on the bound given by León and Perron (2003) in the case of sums of i.i.d. Bernoulli random variables.

Before leaving this section we begin to make a connection to finite-sampling empirical distributions: Now let $\mathbb{F}_n(t) = n^{-1}\sum_{i=1}^{n} 1_{(-\infty,t]}(Y_i)$ and $F_N(t) = N^{-1}\sum_{i=1}^{N} 1_{(-\infty,t]}(c_i)$. Then it is easily seen that Serfling's bound yields

$$P(\sqrt{n}(\mathbb{F}_n(t) - F_N(t)) \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{(1-(n-1)/N)}\right)$$

for each fixed $\lambda > 0$ and $t \in \mathbb{R}$. Note that since $\mathbb{F}_n(t)$ is equal in distribution to the sample mean of n draws without replacement from an urn containing $NF_N(t)$ 1's and $N(1 - F_N(t))$ 0's, the bound in the last display only involves the hypergeometric special case of Serfling's inequality. This leads to the following conjecture concerning bounds for the finite sampling empirical process $\{\sqrt{n}(\mathbb{F}_n(t) - F_N(t)) : \ t \in \mathbb{R}\}$:

Conjecture: There exist constants $C, D > 0$ (possibly $C = 1$ and $D = 2$?) such that

$$P\left(\sqrt{n}\sup_t(\mathbb{F}_n(t) - F_N(t)) \geq \lambda\right) \leq C\exp\left(-\frac{2\lambda^2}{(1-f_n)}\right), \tag{3}$$

$$P\left(\sqrt{n}\sup_t|\mathbb{F}_n(t) - F_N(t)| \geq \lambda\right) \leq D\exp\left(-\frac{2\lambda^2}{(1-f_n)}\right) \tag{4}$$

for all $\lambda > 0$. The possibility that $D = 2$ is suggested by the corresponding inequality established by Massart (1990) in the case of sampling with replacement.

With these strong indications of the plausibility of an improvement of Serfling's bound and corresponding improvements in exponential bounds for the uniform-norm deviations of the finite-sampling empirical process, we can now turn to an application of the basic idea in the context of two-sample Kolmogorov-Smirnov statistics.

2. Two-sample tests and finite-sampling connections

To connect this with the two-sample Kolmogorov-Smirnov statistics, suppose that $X_1, \ldots, X_m$ are i.i.d. F and $Y_1, \ldots, Y_n$ are i.i.d. G. Let $N = m+n$. Then

for testing $H_c : F = G$ with F continuous versus $K^+ : \ F \geq G$ ($F \prec_s G$), $K^- : \ G \geq F$, ($G \prec_s F$), or $K : \ F \neq G$, the classical K-S test statistics are

$$D^+_{m,n} \equiv \sqrt{\frac{mn}{N}} \sup_x (\mathbb{F}_m(x) - \mathbb{G}_n(x)),$$
$$D^-_{m,n} \equiv \sqrt{\frac{mn}{N}} \sup_x (\mathbb{G}_n(x) - \mathbb{F}_m(x)), \quad \text{and}$$
$$D_{m,n} \equiv \sqrt{\frac{mn}{N}} \sup_x |\mathbb{F}_m(x) - \mathbb{G}_n(x)|,$$

respectively. It is well-known that under H_c we have

$$D^{\pm}_{m,n} \to_d \sup_{0 \leq t \leq 1} \mathbb{U}(t), \qquad D_{m,n} \to_d \sup_{0 \leq t \leq 1} |\mathbb{U}(t)|$$

if $m \wedge n \to \infty$ where $\mathbb{U}$ is a standard Brownian bridge process on $[0, 1]$; see e.g. Hájek and Šidák (1967), pages 189-190, Hodges (1958), and van der Vaart and Wellner (1996), pages 360-366.

Note that with $\lambda_N \equiv m/N$ and

$$\mathbb{H}_N \equiv \lambda_N \mathbb{F}_m + (1 - \lambda_N)\mathbb{G}_n = N^{-1} \sum_{i=1}^{N} 1_{(-\infty,\cdot]}(Z_{(i)})$$

where $Z_{(1)} \leq \cdots \leq Z_{(N)}$ are the order statistics of the pooled sample, we have

$$\begin{aligned} \mathbb{F}_m - \mathbb{H}_N &= \mathbb{F}_m - \lambda_N \mathbb{F}_m - (1 - \lambda_N)\mathbb{G}_n = (1 - \lambda_N)(\mathbb{F}_m - \mathbb{G}_n), \qquad \text{and} \\ \mathbb{G}_n - \mathbb{H}_N &= \mathbb{G}_n - \lambda_N \mathbb{F}_m - (1 - \lambda_N)\mathbb{G}_n = \lambda_N(\mathbb{G}_n - \mathbb{F}_m), \end{aligned}$$

and hence, with $\overline{\lambda}_N = 1 - \lambda_N$,

$$\sqrt{\frac{mn}{N}}(\mathbb{F}_m - \mathbb{G}_n) = \sqrt{N}\sqrt{\lambda_N \overline{\lambda}_N}\frac{1}{\overline{\lambda}_N}(\mathbb{F}_m - \mathbb{H}_N) = \frac{1}{\sqrt{\overline{\lambda}_N}}\sqrt{m}(\mathbb{F}_m - \mathbb{H}_N),$$
$$\sqrt{\frac{mn}{N}}(\mathbb{G}_n - \mathbb{F}_m) = \sqrt{N}\sqrt{\lambda_N \overline{\lambda}_N}\frac{1}{\lambda_N}(\mathbb{G}_n - \mathbb{H}_N) = \frac{1}{\sqrt{\lambda_N}}\sqrt{n}(\mathbb{G}_n - \mathbb{H}_N).$$

Thus, using the independence of the ranks $\underline{R}$ and the order statistics $\underline{Z}$ (both based on the pooled sample),

$$P(D^+_{m,n} \geq t) \ = \ E_Z P_R \left(\sqrt{m} \|(\mathbb{F}_m - \mathbb{H}_N)^+\|_\infty > t\sqrt{1 - \lambda_N} \right)$$

and it would follow from (3) that

$$\begin{aligned} P(D^+_{m,n} \geq t) &\leq C \exp\left(-2\overline{\lambda}_N t^2/(1-f_m)\right) \\ &\leq C \exp\left(-2(n/N)t^2/(n/(N-1))\right) \\ &= C \exp\left(-2\frac{N-1}{N}t^2\right) \end{aligned} \tag{5}$$

for all $t > 0$. Similarly it would also follow from (3) that

$$\begin{aligned} P(D^-_{m,n} \geq t) &\leq C \exp\left(-2\lambda_N t^2/(1-f_n)\right) \\ &\leq C \exp\left(-2(m/N)t^2/(m/(N-1))\right) = C \exp\left(-2\frac{N-1}{N}t^2\right) \end{aligned}$$

for all $t > 0$. Combining the two one-sided inequalities yields a (conjectured) two-sided inequality:

$$\begin{aligned} P(D_{m,n} \geq t) &\equiv P(\sqrt{mn/N}\|\mathbb{F}_m - \mathbb{G}_n\|_\infty > t) \\ &\leq P(D^+_{m,n} > t) + P(D^-_{m,n} > t) \\ &\leq 2C \exp\left(-2\frac{N-1}{N}t^2\right). \end{aligned}$$

In the next section we will prove that bounds of this type with $C = 1$ and $D = 2$ hold in the special case $m = n$. For some results for the two-side two-sample Kolmogorov-Smirnov statistic in the case $m = n$ and computational results for $m \neq n$, see Wei and Dudley (2012). These authors were aiming for a bound of the form $C \exp(-2t^2)$ both for $m = n$ and $m \neq n$. The above heuristics seem to suggest that a bound of the form $C \exp(-2((N-1)/N)t^2)$ might be a natural goal.

3. An exponential bound for $D^+_{m,n}$ when $m = n$

Throughout this section we suppose that the null hypothesis H_c holds: $G = F$ is a continuous distribution function.

From Hodges (1958), (2.3) on page 473 (together with $t = \sqrt{mn/N}d$ and

$d = a/n$ from page 473, line 4), when $m = n$ (so $N = 2n$),

$$\begin{aligned} P(D^+_{n,n} \geq t) &= P\left(\sqrt{\frac{mn}{N}}\sup_x(\mathbb{F}_m(x) - \mathbb{G}_n(x)) \geq \sqrt{\frac{mn}{N}}\frac{a}{n}\right) \\ &= P\left(\sqrt{\frac{n^2}{2n}}\sup_x(\mathbb{F}_n(x) - \mathbb{G}_n(x)) \geq \sqrt{\frac{n^2}{2n}}\frac{a}{n}\right) \\ &= \frac{\binom{2n}{n-a}}{\binom{2n}{n}} \quad \text{for} \quad a = 1, 2, \ldots, n. \end{aligned}$$

We first compare the exact probability from the last display with the possible upper bounds

$$PB_2(n) = \exp\left(-2\frac{2n-1}{2n}\frac{a^2}{2n}\right);$$
$$PB_3(n) = \exp\left(-2\frac{a^2}{2n}\right).$$

For $n = 3$ we find that

a	0	1	2	3
$E(xact)$	1	.75	0.3	0.05
$PB2$	1	0.7574	0.3291	0.0821
$PB2 - E$	0	0.0074	0.0291	0.0321
$PB3$	1	0.7165	0.2636	0.0498
$PB3 - E$	0	−0.0335	−0.0364	−0.0002

Further comparisons for $m = n = 10, 12, 13, 14, 15, 25$ support the validity of the bound involving the finite sampling fraction f_n. These comparisons agree with the following theorem:

Theorem 3. *A. When $m = n$ (so that $N = 2n$) the second bound in (5) holds for all $n \geq 1$ with $C = 1$:*

$$P(D^+_{n,n} \geq t) = P\left(\sqrt{\frac{mn}{N}}\sup_x(\mathbb{F}_m(x) - \mathbb{G}_n(x)) \geq t\right) \tag{6}$$
$$\leq \exp\left(-2\frac{N-1}{N}t^2\right) \quad \textit{for all} \quad t > 0. \tag{7}$$

Equivalently, when $m = n$,

$$P\left(\sqrt{\frac{mn}{N}}\sqrt{\frac{N-1}{N}}\sup_x(\mathbb{F}_m(x) - \mathbb{G}_n(x)) \geq t\right) \leq \exp\left(-2t^2\right) \tag{8}$$

for all $t > 0$.

B. On the other hand, when $m = n$ *(so that* $N = 2n$*), for all* $n \geq 1$ *we have*

$$P(D^+_{n,n} \geq t) > \exp(-2t^2) \quad \textit{for all} \;\; 0 < t < 1.$$

Proof. A. Since the inequality holds trivially for $a = 0$, and can be shown easily by numerical computation for $a \in \{1, 2, 3\}$ (see the Table above), it suffices to show that

$$\frac{\binom{2n}{n-a}}{\binom{2n}{n}} \leq \exp\left(-2\frac{2n-1}{2n}\frac{a^2}{2n}\right)$$

for $a \in \{1, \ldots, n\}$ and $n \geq 4$. Furthermore, we will show that it holds for $a = n$ in a separate argument, and thus it suffices to show that it holds for $a \in \{1, \ldots, n-1\}$ and $n \geq 4$. By rewriting the numerator and denominator on the left side of the last display, the desired inequality can be rewritten as

$$\frac{n!n!}{(n-a)!(n+a)!} \leq \exp\left(-\frac{2n-1}{2n} \cdot \frac{a^2}{n}\right).$$

By taking logarithms we can rewrite this as

$$\log\left(\frac{n!n!}{(n-a)!(n+a)!}\right) + \frac{2n-1}{2n}\frac{a^2}{n} \leq 0. \tag{9}$$

Now by Stirling's formula with bounds (see e.g. Nanjundiah (1959)) we have

$$\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \exp\left(\frac{1}{12k} - \frac{1}{360k^3}\right) \leq k! \leq \sqrt{2\pi k}\left(\frac{k}{e}\right)^k \exp\left(\frac{1}{12k}\right). \tag{10}$$

Using these bounds in (9) we find that the left side is bounded above by

$$\begin{aligned}
&-n\left\{(1-\frac{a}{n})\log(1-\frac{a}{n}) + (1+\frac{a}{n})\log(1+\frac{a}{n})\right\} \\
&\quad - \frac{1}{2}\left(\log\left(1-\frac{a}{n}\right) + \log\left(1+\frac{a}{n}\right)\right) \\
&\quad + \left\{\frac{1}{6n} - \frac{1}{12(n-a)} - \frac{1}{12(n+a)} + \frac{1}{360}\left(\frac{1}{(n-a)^3} + \frac{1}{(n+a)^3}\right)\right\} \\
&\quad + \frac{a^2}{n} - \frac{a^2}{2n^2} \\
&\equiv \; I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Note that I_1 and I_2 are as defined in Wei and Dudley (2012) page 640, while I_3 and I_4 differ. From Wei and Dudley (2012) page 640,

$$I_1 \le -\frac{a^2}{n} - \frac{a^4}{6n^3} - \frac{a^6}{15n^5} - \frac{a^8}{28n^7}, \tag{11}$$

(which is proved by Taylor expansion of $(1+x)\log(1+x)+(1-x)\log(1-x)$ about $x=0$), and

$$I_2 \le \frac{a^2}{2n^2} + \frac{a^4}{4n^4} + \frac{a^6}{6n^6(1-a^2/n^2)}. \tag{12}$$

Note that the lead term in the bound (11) for I_1 and lead term of I_4 cancel each other, while the first term of the bound (12) for I_2 cancels the second term of I_4. Adding the bounds yields

$$\begin{aligned}
&I_1 + I_2 + I_3 + I_4 \\
&\le -\frac{a^4}{12n^3} - \frac{a^4}{12n^3} - \frac{a^6}{15n^5} - \frac{a^8}{28n^7} \\
&\quad + \frac{a^4}{4n^4} + \frac{a^6}{6n^6(1-a^2/n^2)} + I_3 \\
&= -\frac{a^4}{n^3}\left(\frac{1}{12} - \frac{1}{4n}\right) - \frac{a^4}{12n^3} - \frac{a^6}{n^5}\left(\frac{1}{15} - \frac{1}{6n(1-a^2/n^2)}\right) - \frac{a^8}{28n^7} \; + \; I_3 \\
&\le -\frac{a^4}{n^3}\left(\frac{1}{12} - \frac{1}{4n}\right) - \frac{a^4}{12n^3} - \frac{a^6}{n^5}\left(\frac{1}{15} - \frac{1}{6(2-1/n)}\right) - \frac{a^8}{28n^7} + I_3 \\
&\le -\frac{a^4}{n^3}\left(\frac{1}{12} - \frac{1}{4n}\right) - \frac{a^4}{12n^3} + \frac{3a^6}{105n^5} - \frac{a^8}{28n^7} + I_3 \\
&= -\frac{a^4}{n^3}\left(\frac{1}{12} - \frac{1}{4n}\right) - \frac{a^4}{12n^3}\left(1 - \frac{36a^2}{105n^2}\right) - \frac{a^8}{28n^7} + I_3 \\
&\le -\frac{a^4}{n^3}\left(\frac{1}{12} - \frac{1}{4n}\right) - \frac{a^4}{21n^3} - \frac{a^8}{28n^7} + I_3 \\
&\equiv R_{12} + I_3.
\end{aligned}$$

Now $R_{12} \le 0$ for $n \ge 4$ and $I_3 \le 0$ for all $n \ge 2$ and $a \in \{1, \ldots, n-1\}$ by

the following argument:

$$
\begin{aligned}
I_3 &= \frac{1}{6n} - \frac{1}{12(n-a)} - \frac{1}{12(n+a)} + \frac{1}{360}\left(\frac{1}{(n+a)^3} + \frac{1}{(n-a)^3}\right) \\
&= -\frac{1}{6}\frac{a^2}{n(n^2-a^2)} + \frac{2}{360}\frac{n(n^2+3a^2)}{(n^2-a^2)^3} \\
&= -\frac{1}{6n(n^2-a^2)}\left\{a^2 - \frac{2}{60}\frac{n^2(n^2+3a^2)}{(n^2-a^2)^2}\right\} \\
&= -\frac{1}{6n(n^2-a^2)}\left\{a^2 - \frac{1}{30}\frac{n^2(n^2-a^2+4a^2)}{(n^2-a^2)^2}\right\} \\
&= -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{2}{15}\frac{n^2}{(n^2-a^2)^2}\right) - \frac{n^2(n^2-a^2)}{30(n^2-a^2)^2}\right\} \\
&\le -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{2}{15}\frac{1}{3}\right) - \frac{n^2}{30(n^2-a^2)}\right\} \\
&\quad \text{by using } \ a \le n-1, \text{ so } \ n^2-a^2 \ge n^2-(n-1)^2 = (2n-1), \\
&\quad \text{and } n^2/(2n-1)^2 \le 1/3 \ \text{ for } \ n \ge 4, \\
&= -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{2}{3\cdot 15}\right) - \frac{n^2-a^2+a^2}{30(n^2-a^2)}\right\} \\
&= -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{2}{3\cdot 15} - \frac{1}{30(n^2-a^2)}\right) - \frac{1}{30}\right\} \\
&\le -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{2}{3\cdot 15} - \frac{1}{30(2n-1)}\right) - \frac{1}{30}\right\} \\
&\le -\frac{1}{6n(n^2-a^2)}\left\{a^2\left(1 - \frac{31}{630}\right) - \frac{1}{30}\right\}
\end{aligned}
$$

for $n \ge 4$. This is a decreasing function of a for fixed n, and hence to show that it is < 0 it suffices to check it for $a = 1$. But when $a = 1$ the right side above equals

$$
\begin{aligned}
&-\frac{1}{6n(n^2-1^2)}\left\{1 - \frac{31}{630} - \frac{1}{30}\right\} \\
&= -\frac{1}{n(n^2-1)}\left\{\frac{289}{6\cdot 315}\right\} < -\frac{1}{n(n^2-1)}\left\{\frac{280}{6\cdot 315}\right\} = -\frac{4}{27n(n^2-1)} < 0,
\end{aligned}
$$

so we conclude that $I_3 < 0$ for $a \in \{1, \ldots, n-1\}$ and $n \ge 4$. It remains only

to show that the desired bound holds for $a = n$; that is we have

$$\frac{1}{\binom{2n}{n}} \leq \exp(-(n - 1/2)).$$

But this can easily be shown via the Stirling formula bounds (10).

Thus

$$\exp(I_1 + I_2 + I_3) \leq \exp(-I_4) = \exp\left(-\frac{2n-1}{2n}\frac{a^2}{n}\right),$$

and the claimed inequality holds for all $n \geq 4$. Since the bounds hold for $n = 1, 2, 3$ by direct numerical computation, the claim follows.

B. We first define

$$\begin{aligned} r_n(a) &\equiv \log\left\{\frac{\binom{2n}{n-a}/\binom{2n}{n}}{\exp(-2a^2/(2n))}\right\} \\ &= \log\binom{2n}{n-a} - \log\binom{2n}{n} + \frac{a^2}{n}. \end{aligned}$$

Since we can take $t = a/\sqrt{2n}$, it suffices to show that $r_n(a) > 0$ for $1 \leq a \leq \lfloor\sqrt{2n}\rfloor$. We will first show this for $n \geq 31$. Then the proof will be completed by checking the inequality numerically for $1 \leq a \leq \lfloor\sqrt{2n}\rfloor$ and $n \in \{1, \ldots, 30\}$.

By using the Stirling formula bounds of (10) as in the proof of A, but now with upper bounds replaced by lower bounds, we find that

$$\begin{aligned} r_n(a) &= 2\log(n!) - \log(n-a)! - \log(n+a)! + \frac{a^2}{n} \\ &\geq -n\left\{\left(1-\frac{a}{n}\right)\log\left(1-\frac{a}{n}\right) + \left(1+\frac{a}{n}\right)\log\left(1+\frac{a}{n}\right)\right\} \\ &\quad - \frac{1}{2}\left\{\log\left(1-\frac{a}{n}\right) + \log\left(1+\frac{a}{n}\right)\right\} \\ &\quad + \frac{1}{6n} - \frac{1}{180n^3} - \frac{1}{12(n-a)} - \frac{1}{12(n+a)} \\ &\quad + \frac{a^2}{n} \\ &\equiv L_1 + L_2 + L_3 + L_4. \end{aligned}$$

As in (11) and (12) and the displays following them, we find that

$$\begin{aligned} L_1 &\geq -n\left\{\frac{a^2}{n^2}+\frac{a^4}{6n^4}+\frac{a^6}{15n^6}+\frac{a^8}{28n^8}\left(\frac{n^2}{n^2-a^2}\right)\right\}, \\ L_2 &\geq \frac{a^2}{2n^2}+\frac{a^4}{4n^4}+\frac{a^6}{6n^6}, \\ L_3 &= -\frac{a^2}{6n(n^2-a^2)}-\frac{1}{180n^3}, \\ L_4 &= \frac{a^2}{n}. \end{aligned}$$

Putting these pieces together and rearranging we find that

$$\begin{aligned} r_n(a) \geq & \left[\frac{31a^2}{64n^2}+\frac{a^4}{4n^4}+\frac{a^6}{6n^6}-\frac{a^4}{6n^3}-\frac{a^6}{15n^5}-\frac{a^8}{28n^5}\left(\frac{1}{n^2-a^2}\right)\right] \\ & +\left[\frac{a^2}{64n^2}+\frac{1}{6n}-\frac{1}{180n^3}-\frac{1}{12(n+a)}-\frac{1}{12(n-a)}\right] \qquad (13) \\ & =: K_1+K_2>0 \qquad (14) \end{aligned}$$

will prove the claim. Note in (13) that the a^2/n term cancelled by virtue of the lower bound estimate based on the Taylor expansion of $(1+x)\log(1+x)+(1-x)\log(1-x)$. First note that

$$\begin{aligned} K_2 &= \frac{a^2}{64n^2}+\frac{1}{6n}-\frac{1}{180n^3}-\frac{1}{12(n+a)}-\frac{1}{12(n-a)} \\ &= \frac{a^2[28n^3-45a^2n]+a^2[16n^3-480n^2]+[a^2n^3+16a^2-16n^2]}{2880n^3(n-a)(n+a)} \end{aligned}$$

The denominator of the right-hand-side is clearly positive for $a \in \{1,2,\ldots,\lfloor\sqrt{2n}\rfloor\}$. By inspection, we can see the term $a^2n^3+16a^2-16n^2$ in the numerator is increasing in a. Picking $a=1$, we then see $n^3+16-16n^2>0$ for $n\geq 31$, and thus $a^2n^3+16a^2-16n^2>0$ for all admissible a. Next, the polynomial $28n^3-45a^2n$ is decreasing in the admissible a. For any fixed n, the minimum value it can attain is then larger than $28n^3-90n^2$. For $n\geq 31$, this quantity is positive. Therefore, $28n^3-45a^2n>0$ for all admissible a when $n\geq 31$. Finally, note that $16n^3-480n^2=16n^2(n-30)>0$ for $n\geq 31$. Hence we have shown $K_2>0$.

We next have

$$\begin{aligned} K_1 &= \left[\frac{31a^2}{64n^2} - \frac{a^4}{6n^3}\right] + \left[\frac{a^4}{4n^4} - \frac{a^6}{15n^5}\right] + \left[\frac{a^6}{6n^6} - \frac{a^8}{28n^5}\left(\frac{1}{n^2-a^2}\right)\right] \\ &= \left[\left(\frac{a^2}{192n^3}\right)(93n - 32a^2)\right] + \left[\left(\frac{a^4}{60n^5}\right)(15n - 4a^2)\right] \\ &\quad + \left[\left(\frac{a^6}{84n^6\,(n^2-a^2)}\right)(14n^2 - 3a^2n - 14a^2)\right] \\ &\equiv \left[(\alpha)\,(93n - 32a^2)\right] + \left[(\beta)\,(15n - 4a^2)\right] \\ &\quad + \left[(\gamma)\,(14n^2 - 3a^2n - 14a^2)\right]\ . \end{aligned} \tag{15}$$

Again since $a \in \left\{1, \ldots, \lfloor\sqrt{2n}\rfloor\right\}$, it is clear that α, β, and γ in (15) are positive for all admissible choices of a. Hence, the sign of each bracketed term will be dictated by the remaining polynomial in a. It is also clear from their form that each polynomial is decreasing in a; hence we need only evaluate at the endpoints to determine positivity. But $93n - 32(\sqrt{2n})^2 = 29n > 0$, $15n - 4(\sqrt{2n})^2 = 15n - 8n = 7n > 0$, and $14n^2 - 3(\sqrt{2n})^2 n - 14(\sqrt{2n})^2 = 14n^2 - 6n^2 - 28n = 4n(2n-7) > 0$ with the final inequality following as $n \geq 31$. Hence all terms in (15) are positive and so $K_1 > 0$. Together with $K_2 > 0$ as proved above, the claim is proved for $n \geq 31$.

Since the bound holds for $a \in \{1, \ldots, \lfloor\sqrt{2n}\rfloor\}$ and $n \in \{1, \ldots, 30\}$ by direct numerical computation, the claim follows. □

4. Some comparisons and connections

4.1. Comparisons: two-sided tail bounds

Here we compare and contrast our results with those of Wei and Dudley (2012). As in Wei and Dudley (2012) (see also Wei and Dudley (2011)), we say that *the DKW inequality holds for given m, n and C* if

$$P(D_{m,n} \geq t) \leq C\exp(-2t^2) \quad \text{for all} \quad t > 0,$$

and we say that *the DKWM inequality holds for given m, n* if the inequality in the last display holds with $C = 2$. Wei and Dudley (2012) prove the following theorem:

Theorem 4. *(Wei and Dudley, 2012) For $m = n$ in the two sample case: (a) The DKW inequality always holds with $C = e \doteq 2.71828$.*

(b) For $m = n \geq 4$, the smallest n such that H_c can be rejected at level 0.05*, the DKW inequality holds with* $C = 2.16863$.
(c) The DKWM inequality holds for all $m = n \geq 458$.
(d) For each $m = n < 458$, the DKWM inequality fails for some t of the form $t = k/\sqrt{2n}$.
(e) For each $m = n < 458$, the DKW inequality holds for $C = 2(1 + \delta_n)$ for some $\delta_n > 0$ where, for $12 \leq n \leq 457$,

$$\delta_n < -\frac{0.07}{n} + \frac{40}{n^2} - \frac{400}{n^3}.$$

For comparison, the following theorem follows from Theorem 3. We say that *the modified DKWM inequality holds for given m, n* if

$$P(D_{m,n} \geq t) \leq 2\exp\left(-2\left(\frac{N-1}{N}\right)t^2\right) \quad \text{for all} \quad t > 0,$$

Theorem 5. *For $m = n$ in the two sample case:*
(a) For all $n \geq 1$ the modified DKWM inequality holds.
(b) Alternatively, for the modified Kolmogorov statistic given by

$$D_{m,n}^{mod} \equiv \sqrt{\frac{N-1}{N}}\sqrt{\frac{mn}{N}}\|\mathbb{F}_m - \mathbb{G}_n\|_\infty,$$

the DKWM inequality holds for all $n \geq 1$.

We are not claiming that our "modified" version of the DKWM inequality improves on the results of Wei and Dudley (2012): it is clearly worse for $m = n > 458$. On the other hand, it may provide a useful clue to the formulation of DKWM type exponential bounds for two-sample Kolmogorov statistics when $m \neq n$. In this direction we have the following conjecture:
Conjecture: For any $m \neq n$,

$$P\left(D_{m,n}^{+} > t\right) \leq \exp\left(-2\left(\frac{N-1}{N}\right)t^2\right) \quad \text{for all} \quad t > 0 \tag{16}$$

$$P\left(D_{m,n} > t\right) \leq 2\exp\left(-2\left(\frac{N-1}{N}\right)t^2\right) \quad \text{for all} \quad t > 0. \tag{17}$$

That is, we conjecture that the modified DKWM inequality holds for all $m, n \geq 1$. This is supported by all the numerical experiments we have conducted so far.

4.2. Comparisons: one-sided tail bounds

Wei and Dudley (2012) do not treat bounds for the one-sided statistics. Here we summarize our results with a theorem which parallels their Theorem 4 above. In analogy with their terminology, we say that *the one-sided DKW inequality holds for given* m, n *and* C if

$$P(D^+_{m,n} \geq t) \leq C \exp(-2t^2) \quad \text{for all} \quad t > 0,$$

and we say that *the one-sided DKWM inequality holds for given* m, n if the inequality in the last display holds with $C = 1$. Moreover, we say that *the modified one-sided DKWM inequality holds for given* m, n if

$$P(D^+_{m,n} \geq t) \leq \exp\left(-2\left(\frac{N-1}{N}\right)t^2\right) \quad \text{for all} \quad t > 0.$$

Theorem 6. *For* $m = n$ *in the two sample case:*
(a) The one-sided DKW inequality holds for all $n \geq 1$ *with* $C = e/2 \doteq 2.71828/2 = 1.35914$. *For this range of* n, $C = e/2$ *is sharp since equality occurs for* $n = 1$ *and* $t = 1/\sqrt{2}$ *(or* $a = t\sqrt{2n} = 1$*).*
(b) For $m = n \geq 5$, *the one-sided DKW inequality holds with* $C = 2.16863/2 = 1.084315$.
(c) The one-sided DKWM inequality fails for all $m = n \geq 1$.
(d) The modified one-sided DKWM inequality holds for all $m = n \geq 1$.

Proof. (c) follows from Theorem 3-B. (d) follows from Theorem 3-A. It remains only to prove (a) and (b).

To prove (a), we first note that Wei and Dudley (2012) showed that for $n \geq 108$ we have

$$\begin{aligned} \frac{\binom{2n}{n+a}}{\binom{2n}{n}} &< \exp(-a^2/n) \quad \text{for} \quad \sqrt{3n} \leq a \leq n \\ &< (e/2)\exp(-a^2/n). \end{aligned}$$

Thus to prove that the claimed inequality holds for $n \geq 108$, it suffices to show that it holds for $t_0\sqrt{n} \leq a \leq \sqrt{3}\sqrt{n}$ where $t_0 \equiv \sqrt{(1/2)\log(e/2)}$ is the smallest value of t for which the bound is less than or equal to 1.

Proceeding as in the proof of Theorem 3-A, we find that we want to show that

$$\log \frac{n!n!}{(n+a)!(n-a)!} + \frac{a^2}{n} - \log(e/2) < 0 \quad \text{for} \quad t_0\sqrt{n} \leq a \leq \sqrt{3}\sqrt{n}.$$

By the same arguments used in the proof of Theorem 3-A, we find that the left side in the last display is bounded above by

$$\begin{aligned} &-\frac{a^4}{6n^3} - \frac{a^6}{15n^5} - \frac{a^8}{28n^7} + \frac{a^4}{4n^4} + \frac{a^6}{6n^6(1-a^2/n^2)} + I_3 \\ &\quad + \frac{a^2}{2n^2} - \log(e/2) \\ &\equiv K_1 + K_2. \end{aligned}$$

Now $K_1 \le 0$ for $n \ge 4$ and $a \in \{1, \ldots, n-1\}$ by the previous proof, and

$$K_2 \equiv \frac{a^2}{2n^2} - \log(e/2) < 0 \quad \text{for all} \quad a \le \sqrt{3}\sqrt{n}$$

if

$$\frac{3}{2n} < \log(e/2), \quad \text{or} \quad n > \frac{3}{2\log(e/2)} \doteq 4.888\ldots.$$

This completes the proof for $n \ge 108$. Numerical computation easily shows that the claim holds for all $n \in \{1, \ldots, 107\}$.

The proof of (b) is similar upon replacing $e/2$ by 1.084315, and again computing numerically for $n \in \{1, \ldots, 107\}$. □

Corollary 3. *For $n \ge 5$ and $C = 1.084315$,*

$$\begin{aligned} P(D^+_{n,n} \ge t) &\le \min\left\{\exp\left(-2\,(1-1/N)\,t^2\right),\ C\exp(-2t^2)\right\} \\ &= \begin{cases} C\exp(-2t^2), & t \ge t_0 \equiv \sqrt{n\log C} \doteq .285\sqrt{n}, \\ \exp(-2(1-1/N)t^2), & t \le t_0 \equiv \sqrt{n\log C}. \end{cases} \end{aligned}$$

Figures 1 and 2 illustrate Theorem 6.

Acknowledgements

The second author owes thanks to Werner Ehm for several helpful conversations and to Martin Wells for pointing out the Pitman reference. We also owe thanks to the referee for a number of helpful comments and suggestions.

References

Bennett, G., 1962. Probability inequalities for the sum of independent random variables. J. Amer. Statist. Assoc. 57, 33–45.

Chvátal, V., 1979. The tail of the hypergeometric distribution. Discrete Math. 25 (3), 285–287.

Dvoretzky, A., Kiefer, J., Wolfowitz, J., 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Statist. 27, 642–669.

Ehm, W., 1991. Binomial approximation to the Poisson binomial distribution. Statist. Probab. Lett. 11 (1), 7–16.

Greene, E., 2016. Finite sampling exponential bounds with applications to empirical processes. Ph.D. thesis, University of Washington.

Greene, E., Wellner, J. A., 2015. Exponential bounds for the hypergeometric distribution. Tech. Rep. arXiv:1507.08298.

Hájek, J., Šidák, Z., 1967. Theory of rank tests. Academic Press, New York.

Hodges, Jr., J. L., 1958. The significance probability of the Smirnov two-sample test. Ark. Mat. 3, 469–486.

Hoeffding, W., 1963. Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13–30.

Hush, D., Scovel, C., 2005. Concentration of the hypergeometric distribution. Statist. Probab. Lett. 75 (2), 127–132.

Kemperman, J. H. B., 1973a. Moment problems for sampling without replacement. I. Nederl. Akad. Wetensch. Proc. Ser. A **76**=Indag. Math. 35, 149–164.

Kemperman, J. H. B., 1973b. Moment problems for sampling without replacement. II. Nederl. Akad. Wetensch. Proc. Ser. A **76**=Indag. Math. 35, 165–180.

Kemperman, J. H. B., 1973c. Moment problems for sampling without replacement. III. Nederl. Akad. Wetensch. Proc. Ser. A **76**=Indag. Math. 35, 181–188.

León, C. A., Perron, F., 2003. Extremal properties of sums of Bernoulli random variables. Statist. Probab. Lett. 62 (4), 345–354.

Lo, A. Y., 1986. Bayesian statistical inference for sampling a finite population. Ann. Statist. 14 (3), 1226–1233.

Massart, P., 1990. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. 18 (3), 1269–1283.

Nanjundiah, T. S., 1959. Note on Stirling's formula. Amer. Math. Monthly 66, 701–703.

Pitman, J., 1997. Probabilistic bounds on the coefficients of polynomials with only real zeros. J. Combin. Theory Ser. A 77 (2), 279–303.

Rice, J. A., 2007. Mathematical Statistics and Data Analysis, 3rd Edition. Duxbury Press, Belmont, CA.

Serfling, R. J., 1974. Probability inequalities for the sum in sampling without replacement. Ann. Statist. 2, 39–48.

Shorack, G. R., Wellner, J. A., 1986. Empirical processes with applications to statistics. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Talagrand, M., 1994. Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22 (1), 28–76.

van der Vaart, A. W., Wellner, J. A., 1996. Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York.

Vatutin, V. A., Mikhaĭlov, V. G., 1982. Limit theorems for the number of empty cells in an equiprobable scheme for the distribution of particles by groups. Teor. Veroyatnost. i Primenen. 27 (4), 684–692.

Wei, F., Dudley, R. M., 2011. Dvoretzky-Kiefer-Wolfowitz inequalities for the two-sample case. Tech. rep., MIT, Department of Mathematics.

Wei, F., Dudley, R. M., 2012. Two-sample Dvoretzky-Kiefer-Wolfowitz inequalities. Statist. Probab. Lett. 82 (3), 636–644.

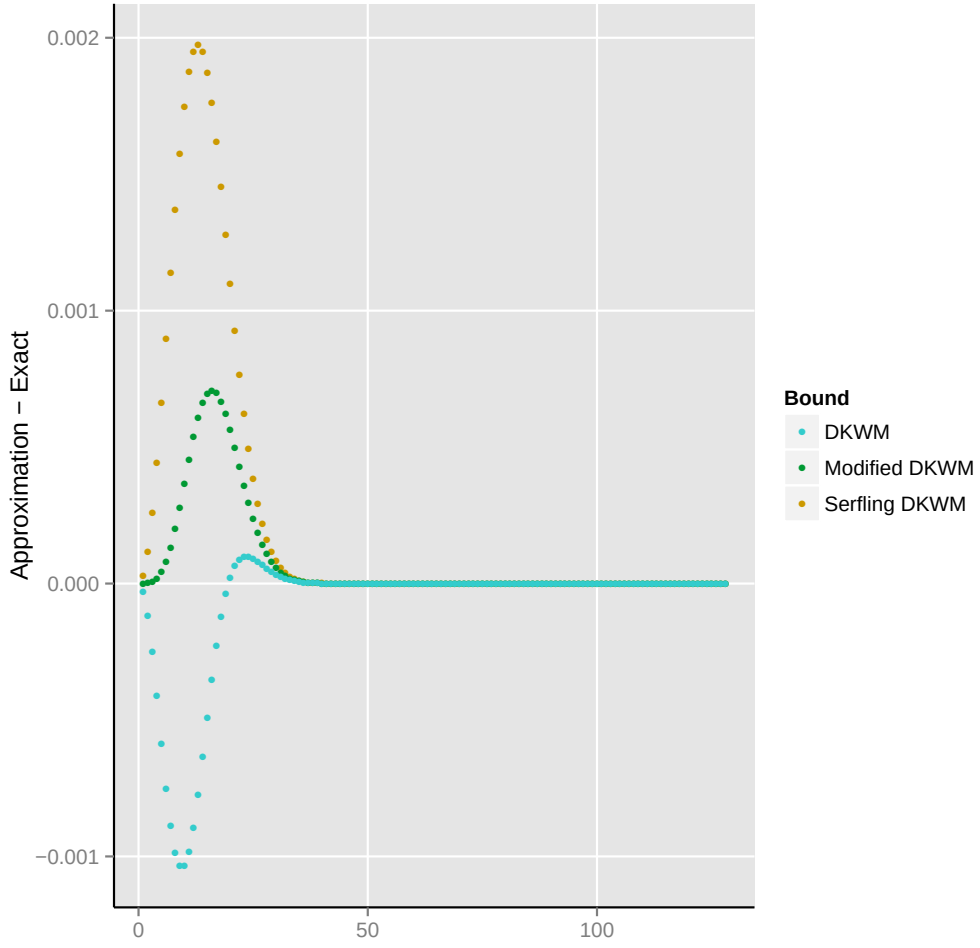

Figure 1: Difference between approximations and exact one-sided probabilities $P\left(D_{n,n}^{+} > t\right)$ for $n = 128$ and $a \in \{1, 2, \ldots, 128\}$. Negative values indicate the exact probability exceeds the approximation. Serfling DKWM is the bound obtained via the heuristic of section 2, using the sampling fraction $1 - f_n^* = (N - n + 1)/N$. Modified DKWM uses the sampling fraction $1 - f_n = (N - n)/(N - 1)$. DKWM uses the fraction from Wei and Dudley.

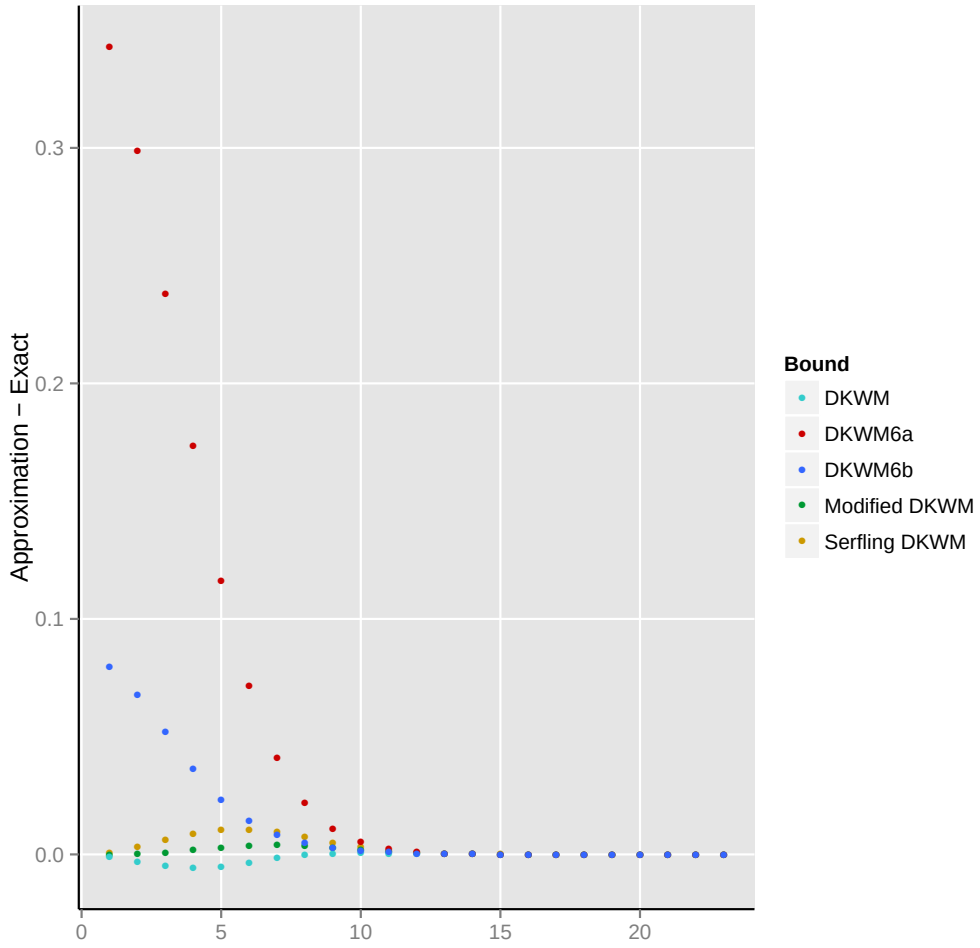

Figure 2: Difference between approximations and exact one-sided probabilities $P\left(D_{n,n}^{+} > t\right)$ for $n = 23$ and $a \in \{1, 2, \ldots, 23\}$. Negative values indicate the exact probability exceeds the approximation. DKWM6a corresponds to the DKWM bound with the constant $e/2$, discussed in Theorem 6(a). DKWM6b corresponds to the DKWM bound with the constant 2.16863/2, discussed in Theorem 6(b).