Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 126 (2016) 3701–3715

stochastic processes and their applications

www.elsevier.com/locate/spa

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

Evan Greene, Jon A. Wellner*

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

Received 17 December 2014; accepted 26 November 2015
Available online 6 May 2016

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) [3] and by Massart (1990) for one-sample versions of these statistics.

MSC: 60F15; 60F17; 62E20; 62F12; 62G20

Keywords: Bennett inequality; Finite sampling; Hoeffding inequality; Hypergeometric distribution; Two-samples; Kolmogorov–Smirnov statistics; Exponential bounds

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 labelled with the numbers $c_1, \ldots, c_N$. Some notation:

* Corresponding author.
E-mail address: jaw@stat.washington.edu (J.A. Wellner).

http://dx.doi.org/10.1016/j.spa.2016.04.020

we let

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

It is well-known (see e.g. [19, Theorem B, page 208]) that $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ satisfies $E(\overline{Y}_n) = \mu_N$ and

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

Serfling [20, 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 [9] 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 [11–13], 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

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

and the conjectured improvement is

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

Despite related results due to Chvátal [2] and Hush and Scovel [10] it seems that a bound of the form in the last display remains unknown.