Statistical Methodology 20 (2014) 40–62

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Sharp fixed *n* bounds and asymptotic expansions for the mean and the median of a Gaussian sample maximum, and applications to the Donoho–Jin model

Anirban DasGupta[a], S.N. Lahiri[b,*], Jordan Stoyanov[c]

[a] *Purdue University, United States*
[b] *North Carolina State University, United States*
[c] *Newcastle University, UK*

ARTICLE INFO

Article history:
Received 24 January 2014
Accepted 25 January 2014

Dedicated to the Fond Memories of Kesar Singh

MSC:
60E05
62E20

Keywords:
Standard normal
Maximum
Mean
Median
Bounds
False discovery
Sparse
Mixture

ABSTRACT

We are interested in the sample maximum $X_{(n)}$ of an i.i.d standard normal sample of size n. First, we derive two-sided bounds on the mean and the median of $X_{(n)}$ that are valid for any fixed $n \geq n_0$, where n_0 is 'small', e.g. $n_0 = 7$. These fixed n bounds are established by using new very sharp bounds on the standard normal quantile function $\Phi^{-1}(1-p)$. The bounds found in this paper are currently the best available explicit nonasymptotic bounds, and are of the correct asymptotic order up to the number of terms involved.

Then we establish exact three term asymptotic expansions for the mean and the median of $X_{(n)}$. This is achieved by reducing the extreme value problem to a problem about sample means. This technique is general and should apply to suitable other distributions. One of our main conclusions is that the popular approximation $E[X_{(n)}] \approx \sqrt{2\log n}$ should be discontinued, unless n is fantastically large. Better approximations are suggested in this article. An application of some of our results to the Donoho–Jin sparse signal recovery model is made.

The standard Cauchy case is touched on at the very end.

* Corresponding author. Tel.: +1 9794502228.
E-mail addresses: snlahiri@ncsu.edu, snlahiri@stat.tamu.edu (S.N. Lahiri).

http://dx.doi.org/10.1016/j.stamet.2014.01.002
1572-3127/© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Let $X_{(n)}$ denote the maximum of an i.i.d. sample $X_1, \dots, X_n$ from a univariate standard normal distribution. A knowledge of distributional properties of $X_{(n)}$, and especially of its mean value $E[X_{(n)}]$, became important in several frontier areas in theory as well as applications. A few instances are the widespread use of properties of $X_{(n)}$ for variable selection in sparse high dimensional regression, in studying the hard and soft thresholding estimators of Donoho and Johnstone [7] for sparse signal detection and false discovery, and in analyzing or planning for extreme events of a diverse nature in practical enterprises, such as climate studies, finance, and hydrology.

We do know quite a bit about distributional properties of $X_{(n)}$ already. For example, we know the asymptotic distribution on suitable centering and norming; see [8]. We know that up to the first order, the mean, the median, and the mode of $X_{(n)}$ are all asymptotically of order $\sqrt{2\log n}$, and that convergence to the asymptotic distribution is very slow; see [10]. There is also very original and useful nonasymptotic work by Lai and Robbins [13] on the mean of $X_{(n)}$. The Lai–Robbins bounds were generalized to other order statistics, including the maximum, in [9] and in [15], who considers general L-estimates.

In this article, we first provide the currently best available nonasymptotic bounds on the mean and the median of $X_{(n)}$. The bounds resemble the Edgeworth expansions of sample means. However, here the bounds are valid for fixed n. For example, Theorem 4.1 shows that for $n \geq 7$, we have

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 2}{2\sqrt{2\log n}} + \frac{K(\log 4\pi + \log\log n)}{(2\log n)^{3/2}},$$

where the constant K is explicit and can be taken to be 1.5 (or anything bigger). If simpler nonasymptotic bounds, although numerically less accurate, were desired, they could be easily extracted out from the above bound. In particular, it follows that for all $n \geq 10$,

$$E[X_{(n)}] \leq \sqrt{2\log n} - \frac{\log 4\pi + \log\log n - 6}{6\sqrt{2\log n}}.$$

While valid for fixed n, the successive terms in the above bound go down at increasing powers of $1/\sqrt{2\log n}$, just as the successive terms in an Edgeworth expansion for the CDF of a sample mean go down at increasing powers of $1/\sqrt{n}$. However, the greatest utility of our bounds is that they are nonasymptotic, the sharpest ones available to date, and moreover, they cover both the mean and the median of $X_{(n)}$. Bounds on the median are stated in Section 3, and the bounds on the mean in Section 4.

On the technical side, there are a few principal ingredients for the bounds on the mean and the median. One is the following bound for the standard normal distribution function Φ and its density φ: there are nonnegative constants a, b, c, d such that

$$1 - \Phi(x) \leq \frac{a\varphi(x)}{bx + \sqrt{c^2x^2 + d}} \quad \text{for all } x > 0. \text{ (See [16].)}$$

The next important result we need is the following general inequality: if $X_{(n)}$ is the maximum of n i.i.d. random variables drawn from a distribution F, whose inverse function is denoted by F^{-1}, then

$$E[X_{(n)}] \leq F^{-1}\left(1 - \frac{1}{n}\right) + n\int_{F^{-1}\left(1-\frac{1}{n}\right)}^{\infty} (1 - F(x))dx. \quad \text{(See [13].)}$$

We are going to use also several analytic inequalities, e.g. Jensen, etc.

Our principal strategy is to first analytically bound the standard normal quantile function $z_p = \Phi^{-1}(1-p)$ and then transform those to bounds on the mean and the median of $X_{(n)}$. For example, it is proved in this article (Corollary 2.1) that for $p \leq 0.1$,

$$z_p \leq \sqrt{2\log t} - \frac{\log 4\pi + \log\log t}{2\sqrt{2\log t}}\left(1 - \frac{K}{\log t}\right),$$