UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2014, Vol. 7, No. 1, pp. 11–17

Should we think of a different median estimator?

¿Debemos pensar en un estimator diferente para la mediana?

Jorge Iván Vélez[a]
jorgeivanvelez@gmail.com

Juan Carlos Correa[b]
jccorrea@unal.edu.co

Resumen

La mediana, una de las medidas de tendencia central más populares y utilizadas en la práctica, es el valor numérico que separa los datos en dos partes iguales. A pesar de su popularidad y aplicaciones, muchos desconocen la existencia de diferentes expresiones para calcular este parámetro. A continuación se presentan los resultados de un estudio de simulación en el que se comparan el estimador clásico y el propuesto por Harrell & Davis (1982). Mostramos que, comparado con el estimador de Harrell–Davis, el estimador clásico no tiene un buen desempeño para tamaños de muestra pequeños. Basados en los resultados obtenidos, se sugiere promover la utilización de un mejor estimador para la mediana.

***Palabras clave*:** mediana, cuantiles, estimador Harrell-Davis, simulación estadística.

Abstract

The median, one of the most popular measures of central tendency widely-used in the statistical practice, is often described as the numerical value separating the higher half of the sample from the lower half. Despite its popularity and applications, many people are not aware of the existence of several formulas to estimate this parameter. We present the results of a simulation study comparing the classic and the Harrell-Davis (Harrell & Davis 1982) estimators of the median for eight continuous statistical distributions. It is shown that, relatively to the latter, the classic estimator performs poorly when the sample size is small. Based on these results, we strongly believe that the use of a better estimator of the median must be promoted.

***Keywords*:** median, quantiles, Harrell–Davis estimator, statistical simulation.

[a] Translational Genomics Group, Genome Biology Department, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia. Grupo de Neurociencias de Antioquia, Universidad de Antioquia, Colombia. Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín.

[b] Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín. Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with absolutely continuous distribution function F, and let $X_{(i)}$ be the ith order statistic ($i = 1, 2, \ldots, n$), e.g., $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Denote θ as the *true* median (a parameter) and any estimator of θ as $\hat{\theta}$. The most common estimator of the median is

$$\hat{\theta}_1 = \begin{cases} X_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2)+1}\right) & \text{if } n \text{ is even.} \end{cases} \tag{1}$$

Harrell & Davis (1982) proposed a new distribution-free estimator of the pth percentile, denoted as Q_p. For the median, the estimator is given by:

$$\hat{\theta}_2 = Q_{1/2} = \sum_{i=1}^{n} W_{n,i}\, X_{(i)} \tag{2}$$

with

$$W_{n,i} = \frac{\Gamma\left(n+1\right)}{\Gamma\left(\frac{n+1}{2}\right)^2} \int_{(i-1)/n}^{i/n} \left[z\left(1-z\right)\right]^{(n-1)/2}\, dz.$$

Other estimators for the median have also been proposed in the literature, but their complexity and dependence on arbitrary constants make them less appealing and difficult to implement (see Ekblom, 1973). Comparative studies have been performed to evaluate the equivalency and asymptotic properties of $\hat{\theta}_1$ and $\hat{\theta}_2$, with the work by Yoshizawa (1984) being the first of them. The author showed that both estimators are asymptotically equivalent, and gave regularity conditions to guarantee the asymptotic normality of each of them. On the other hand, Bassett (1991) showed that the traditional estimator of the median is the only equivariant and monotonic with 50 % breakdown, and Zielinski (1995) concluded the $\hat{\theta}_1$ is not a good estimator under asymmetric distributions.

In this paper we compare the performance of $\hat{\theta}_1$ and $\hat{\theta}_2$ for several continuous distributions when the sample size n is small, and by considering the skewness as the main factor (measure) to control. As explained further below, this measure represents the relative efficiency of one of the estimators when B samples of size n are draw from a specific distribution F.

2. Simulation Study and Results

2.1. Simulation set up

In order to compare the performance of $\hat{\theta}_1$ and $\hat{\theta}_2$, we carried out a simulation study in which eight continuous distributions were considered (see Table 1). These distributions represent those most frequently encountered in the statistical practice. For each of these distributions, a total of $B = 5000$ samples of size $n = \{5, 10, 15, \ldots, 200\}$ were generated. The choice of theses sample sizes was driven because of what is often seen in real-world applications.

Tabla 1: *Probability distributions considered in this study. Source: compiled by authors.*

Distribution	$F(\cdot)$	Parameters	Median ($\dot{\theta}$)
Uniform	$\frac{1}{b-a}$	a, b	$\frac{a+b}{2}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ, σ	μ
Laplace	$\frac{1}{2\tau}e^{-\frac{\lvert x-\mu\rvert}{\tau}}$	μ, τ	μ
Cauchy	$\frac{1}{\pi(1+x^2)}$	–	0
$t-$Student	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)\sqrt{\nu\pi}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	ν	0
Exponential	$\lambda e^{-\lambda x}$	λ	$\lambda\log(2)$
Gamma	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}$	α, β	No closed form
Weibull	$\frac{\beta}{\alpha^\beta}x^{\beta-1}e^{-\left(\frac{x}{\alpha}\right)^\beta}$	α, β	$\alpha(\log(2))^{\frac{1}{\beta}}$

We compare the performance of $\hat{\theta}_1$ and $\hat{\theta}_2$ using the following measure of relative efficiency

$$\gamma = \frac{\text{MSE}_1}{\text{MSE}_2} \tag{3}$$

with

$$\text{MSE}_j = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_{ij} - \dot{\theta})^2$$

the mean squared error (MSE) for the jth estimator ($j = 1, 2$), $\dot{\theta}$ the true median, and B the number of samples of size n that are draw from a specific distribution function F (see Table 1). Note that the lower the MSE, the better the estimator. Here, $\gamma = 1$ indicates that both estimators perform equally well; $\gamma < 1$ indicates that $\hat{\theta}_1$ outperforms $\hat{\theta}_2$; and $\gamma > 1$ indicates that $\hat{\theta}_2$ outperforms $\hat{\theta}_1$. In general, it is possible to derive closed-form expressions for calculating $\dot{\theta}$ provided F. However, when this is not the case, the use of computational routines is required. In our case, the `qgamma()` function in R (R Core Team 2013) was utilised for estimating $\dot{\theta}$ for the Gamma(α, β) distribution.

For our simulation study, we implemented the following algorithm in R:

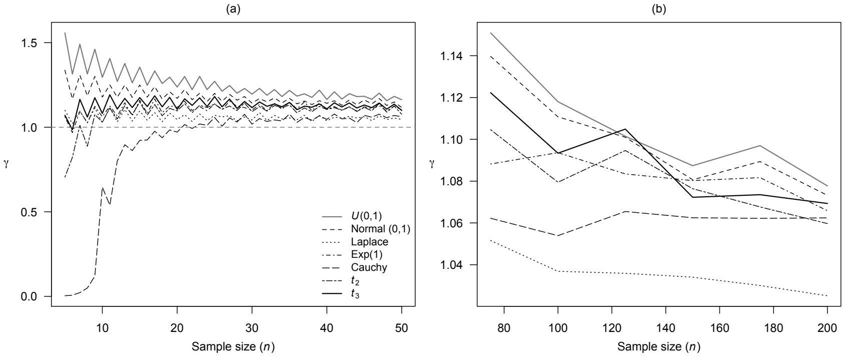

Figura 1: γ *as a function of the sample size when (a)* $n \leq 50$ *and (b)* $n > 50$ *for the first six distributions in Table 1. Here, the dotted horizontal line represents a comparable performance between the classic and the Harrell–Davis estimators. Note that all probability distributions but the Exponential are symmetric. Source: elaborated by authors.*

1. Generate a sample of size n from F (see Table 1 for details).
2. Calculate $\hat{\theta}_1$ as in (1), and $\hat{\theta}_2$ as in (2).
3. Repeat 1–2, B times, calculate the MSE for each estimator and then the ratio of the resulting quantities.

2.2. Results

The results of our simulation study are presented in figures 1 and 2. Figure 1 depicts the value of γ as a function of the sample size n for the first six continuous distributions in Table 1. Figure 2 shows, for fixed n, a 3D representation of γ as a function of α and β, for the Gamma(α, β) and Weibull(α, β) distributions.

As shown in figure 1, γ is always greater than one except for the t_2 distribution when $n < 10$, and the t_3 distribution when $n < 25$. Another interesting finding is that, regardless of n, the highest values of γ were obtained for the $U(0, 1)$ followed by the $N(0, 1)$ and the Laplace distributions. It is intriguing that, despite not being a symmetric distribution, the values of γ for the exponential distribution with parameter $\lambda = 1$ were the forth highest. In addition, note that $\gamma \to 1$ as $n \to \infty$, which is consistent with the assymptotic equivalency of both estimators described by Yoshizawa (1984).

In figure 2 we present the results for the Gamma(α, β) and Weibull(α, β) distributions for different values of α and β for n is fixed. These results suggest that, regardless of n, the Harrell–Davis estimator outperforms the classic estimator, e.g., $\gamma > 1$. On the other hand, the higher γ values were obtained when $n = 5$, and the lowest when $n = 200$, supporting the assymptotic equivalency of both estimators

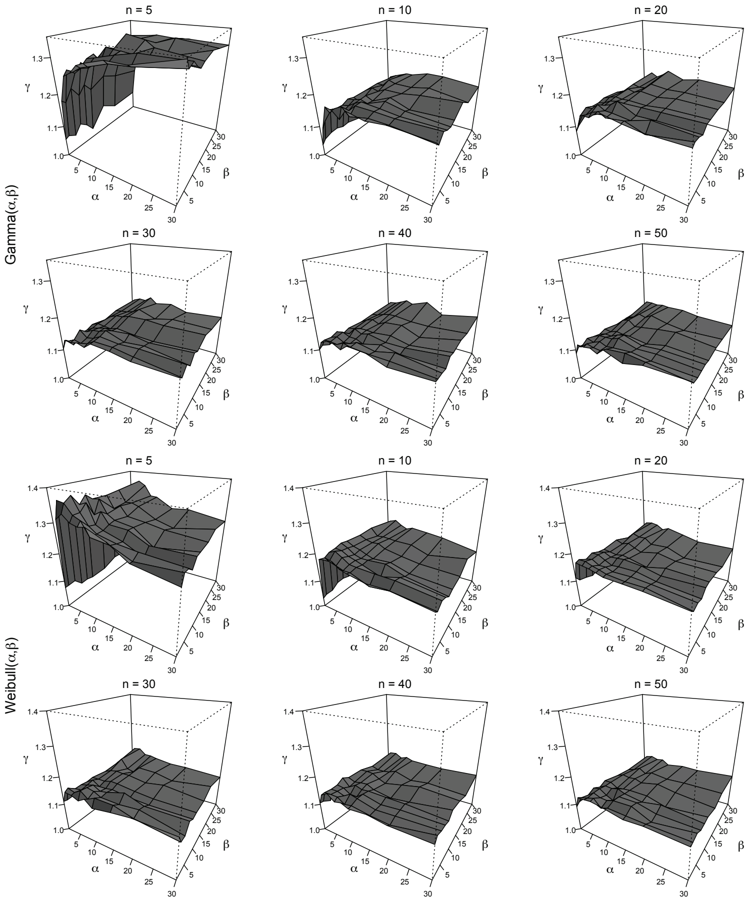

Figura 2: γ *as a function of* n *and the parameters* (α, β) *for the Gamma*(α, β) *and Weibull*(α, β) *distributions. Note that* $\gamma > 1$ *regardless of* n*,* α *and* β*, showing that the Harrell–Davis estimator of the median outperforms the traditional estimator. Source: elaborated by authors.*

(Yoshizawa 1984).

3. Conclusions

We have shown under a large number of scenarios that the Harrell–Davis estimator of the median behaves better than the traditional estimator in terms of the MSE. In particular, it is found that, for small sample sizes, the MSE of the Harrell–Davis estimator of the median is lower than that of the traditional estimator for most of the continuous statistical distributions considered in this study, and often seen by data analysts. Despite the use and popularity of the traditional estimator of the median, and the fact that it is taught in most of statistics textbooks, we strongly believe that, with the current computational capability, the use of a better estimator must be promoted. In Appendix A we provide R code to facilitate this process.

4. Acknowledgments

We thank Dr. Freddy Hernández Barajas for critical reading of this manuscript, and one anonymous reviewer for his useful comments and suggestions. JIV was supported by the Eccles Scholarship in Medical Sciences, the Fenner Merit Scholarship and the Australian National University High Degree Research Scholarship. JIV thanks Dr. Mauricio Arcos-Burgos for his support.

Recibido: 5 de julio de 2013
Aceptado: 9 de septiembre de 2013

Referencias

Bassett, J. G. W. (1991), ‘Equivariant, monotonic, 50 % breakdown estimators’, *The American Statistician* **45**(2), 135–137.

Harrell, F. E. & Davis, C. E. (1982), ‘A new distribution-free quantile estimator’, *Biometrika* **69**(3), 635–640.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org/

Yoshizawa, C. N. (1984), *Some Symmetry Tests*, Institute of Statistics, Mimeo Series No. 1460. University of North Carolina, Chapel Hill, USA.

Zielinski, R. (1995), ‘Estimating median and other quantiles in nonparametric models’, *Applicationes Mathematicae* **23**(3), 363–370.

A. Harrell–Davis estimator in R

A generalisation of the Harrell–Davis estimator for any quantile $p \in (0, 1)$ can be found in the `Hmisc` package (Harrell, 2012). Our implementation, as follows, deals only with the case $p = 1/2$.

```r
### Harrell-Davis estimator of the median
HD <- function(x, n = length(x)) {
    ## auxiliary function
    prob.beta <- function(limits, n) diff(pbeta(limits, (n + 1)/2, (n + 1)/2))
    i <- 1:n
    limits <- cbind((i - 1)/n, i/n)
    Wi <- apply(limits, 1, prob.beta, n)
    sum(Wi * sort(x))
}
## Example: theoretical median is log(2) = 0.6931472
set.seed(123)  # to replicate the results
x <- rexp(150, 1)  # X~Exp(1)
HD(x)  # Harrell-Davis estimator

## [1] 0.7959

median(x)  # traditional estimator

## [1] 0.8431

## calculating gamma (see section 2.1) using B = 1000, n = 20 and X~Exp(1)
out <- replicate(1000, {
    x <- rexp(20, 1)
    theta1 <- median(x)
    theta2 <- HD(x)
    mse1 <- (theta1 - log(2))^2
    mse2 <- (theta2 - log(2))^2
    c(mse1, mse2)
})
out <- rowMeans(out)
gamma <- out[1]/out[2]
names(gamma) <- "gamma"
gamma

## gamma
## 1.086
```