DigitalCommons @WayneState

Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 28

11-1-2011

# A Pooled Two-Sample Median Test Based on Density Estimation

Vadim Y. Bichutskiy
*George Mason University*, vbichuts@masonlive.gmu.edu

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 692-698

1538 – 9472/11/$95.00

*Emerging Scholars*

# A Pooled Two-Sample Median Test Based on Density Estimation

Vadim Y. Bichutskiy
George Mason University
Fairfax, Virginia

A new method based on density estimation is proposed for medians of two independent samples. The test controls the probability of Type I error and is at least as powerful as methods widely used in statistical practice. The method can be implemented using existing libraries in R.

Key words: Sample median, two-sample hypothesis test, adaptive kernel density estimation.

## Introduction

Let $X_1, X_2, \dots, X_n$ be iid having cdf F and pdf f with $F(\eta) = 1/2$ so that $\eta$ is the population median. Suppose f is continuous at $\eta$ with $f(\eta) > 0$. Denote the sample median by H. It is known that H is asymptotically normal with mean $\eta$ and variance $1/4nf^2(\eta)$. Estimating the asymptotic standard error of the sample median requires an estimate of the population density at the median. Besides being a challenging problem, density estimation was difficult to apply in practice prior to the computer revolution; due to this, several alternative methods for estimating the standard error of the sample median have been developed (Maritz & Jarrett, 1978; McKean & Schrader, 1984; Price & Bonett, 2001; Sheather & Maritz, 1983; Sheather, 1986).

Comparing medians based on two independent samples is a well-studied problem (see Wilcox & Charlin, 1986; Wilcox, 2005; Wilcox, 2006; Wilcox, 2010 also has a good discussion). The methods fall into two main categories. The first uses the bootstrap (Efron, 1979), and the second assumes the sample median or some other estimator of the population median is approximately normal and uses one of several methods for estimating the standard error of the sample median. Virtually all methods are very conservative, particularly for heavy-tailed populations.

A new two-sample test is proposed for comparing medians. When population shapes can be assumed to be the same, a pooled test statistic, analogous to a pooled two-sample Student's t statistic for comparing means, is derived. Computer-intensive Monte Carlo simulations in R (R Development Core Team, 2009) are used to study the properties of the test and compare it to other methods. The method offers several additional benefits to practitioners: (1) a parameter that controls the trade-off between making the test conservative and liberal with a suitable value of the parameter producing a test with a nominal significance level; (2) the test is easy to implement in R using the QUANTREG (Koenker, 2009) library.

Vadim Y. Bichutskiy is a Ph.D. student in the Department of Statistics. This work was completed when he was a M.S. student in the Department of Statistics and Biostatistics at California State University, East Bay (Hayward). Email him at: vbichuts@masonlive.gmu.edu.

## Methodology

### Two-Sample Test Statistic for Difference in Medians

Let $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_m$ be two independent random samples of sizes n and m from populations with densities $f_x$, $f_y$ that are continuous at the medians $\eta_x$, $\eta_y$ with $f_x(\eta_x) > 0$, $f_y(\eta_y) > 0$, respectively. Denote sample medians by $H_x$, $H_y$. The test hypotheses are:

$$H_0 : \eta_x - \eta_y = \Delta$$
$$vs.$$
$$H_1 : \eta_x - \eta_y \neq \Delta,$$

where $\Delta$ is a specified difference in medians, and is often 0.

For sufficiently large n and m:

$$H_x \dot{\sim} N\left(\eta_x \ , 1/4nf_x^2(\eta_x)\right),$$

$$H_y \dot{\sim} N\left(\eta_y \ , 1/4mf_y^2(\eta_y)\right),$$

$$H_x - H_y \dot{\sim} N\left(\eta_x - \eta_y \ , \ \frac{1}{4}\left\{\frac{1}{nf_x^2(\eta_x)} + \frac{1}{mf_y^2(\eta_y)}\right\}\right),$$

$$\frac{H_x - H_y - (\eta_x - \eta_y)}{\frac{1}{2}\sqrt{\frac{1}{nf_x^2(\eta_x)} + \frac{1}{mf_y^2(\eta_y)}}} \dot{\sim} N(0,1).$$

Assuming the normal approximation holds when the standard error of the difference in medians is estimated, then under the null hypothesis, the V statistic is:

$$V = \frac{(H_x - H_y) - \Delta}{\frac{1}{2}\sqrt{\frac{1}{n\hat{f}_x^2(H_x)} + \frac{1}{m\hat{f}_y^2(H_y)}}} \dot{\sim} N(0,1)$$

where $\hat{f}_x(H_x)$ and $\hat{f}_y(H_y)$ are respective population density estimates at the median.

Further, if it is assumed that the two populations have the same shape, possibly with a difference in location, then $f_x(\eta_x) = f_y(\eta_y)$, and the density estimates can be pooled to obtain a pooled test statistic:

$$V_p = \frac{(H_x - H_y) - \Delta}{\frac{1}{2\hat{f}_p(H)}\sqrt{\frac{1}{n} + \frac{1}{m}}} \dot{\sim} N(0,1) \tag{1}$$

where

$$\hat{f}_p(H) = \sqrt{\frac{n\hat{f}_x^2(H_x) + m\hat{f}_y^2(H_y)}{n+m}}$$

is the pooled estimate of the population density at the median.

Simulations

The software R was used to simulate the power of the pooled test statistic (1). Two cases were considered: (i) population shapes are assumed to be known, and (ii) population shapes are unknown. The assumption of known population shapes is analogous to the assumption of known population variances in the z-test for comparing the means of two normal populations since the variance determines the shape of the normal distribution. The goal was to see how the test would perform for samples of moderate size from symmetric heavy-tailed populations. Parent populations investigated were Cauchy, Laplace and Student's t distributions with 2 and 3 degrees of freedom. In all settings, the parent populations were of the same shape, shifted under the alternative, and a two-sided test $H_0$: $\eta_x = \eta_y$ versus $H_1$: $\eta_x \neq \eta_y$ was performed.

Adaptive Kernel Density Estimation

When population shapes are unknown, $f_x(\eta_x)$ and $f_y(\eta_y)$ are estimated with $\hat{f}_x(H_x)$ and $\hat{f}_y(H_y)$, respectively, using adaptive kernel density estimation (AKDE).

Let $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ be a sample from unknown density f. The AKDE is a three step procedure:

1. Find a pilot estimate $\tilde{f}(X)$ that satisfies $\tilde{f}(X_i) > 0$, $i = 1, 2, \ldots, n$.
2. Define local bandwidth factors $\lambda_i = \{\tilde{f}(X_i)/g\}^{-\gamma}$ where $g$ is the geometric mean of the $\tilde{f}(X_i)$ and $0 \le \gamma \le 1$ is the sensitivity parameter.
3. The adaptive kernel estimate is defined by

$$\hat{f}(X) = n^{-1}\sum_{i=1}^{n} h^{-d}\lambda_i^{-d}K\{h^{-1}\lambda_i^{-1}(X - X_i)\}$$

where $K(.)$ is a kernel function and $h$ is the bandwidth.

The AKDE method varies the bandwidth among data points and is better suited for heavy-tailed populations than ordinary KDE (Silverman, 1998, pp. 100-110). Intuitively, the AKDE is based on the idea that for heavy-tailed populations a larger bandwidth is needed for data points in the tails of the distribution (i.e., for outliers). In R, function AKJ in library QUANTREG implements AKDE. Obtaining the pilot estimate requires the use of another density estimation method, such as ordinary KDE. The general view in the literature is that AKDE is fairly robust to the method used for the pilot estimate (Silverman, 1998) and that the choice of the sensitivity parameter $\gamma$ is more critical. When using AKDE with Gaussian kernel, if the parent population has tails close to normal then $\gamma < .5$ should be used, however, if the parent population is heavy-tailed then $\gamma > .5$ should be used. Thus, $\gamma = .5$ is a good choice and has been shown to reduce bias (Abramson, 1982).

## Results

### Case 1: Known Population Shapes

Figure 1 shows the power curves for the pooled test when population shapes are assumed to be known at the 5% level of significance. Each point on the curves is based on 10,000 simulated samples. The Type I error rate is controlled very well.

### Case 2: Unknown Population Shapes

Figure 2 shows the power curves for the pooled test when population shapes are unknown at the 5% level of significance and using AKDE with $\gamma = .5$. Each point on the curves is based on 10,000 simulated samples. The Type I error rate is controlled very well.

### Comparisons with Other Methods

The test was compared to the following methods: (i) Student's t-test; (ii) Mann-Whitney-Wilcoxon (MWW) rank sum test; (iii) bootstrap (Efron & Tibshirani, 1993. p. 221); and (iv) permutation test. Figure 3 shows the receiver operating characteristic (ROC) curves for a balanced design with n = m = 30. The parent populations were of the same shape in each case and the difference in population medians was set to 1. For the bootstrap and the permutation test, the difference in medians was used as the metric. Each point on the curves is based on 10,000 simulated samples.

## Conclusion

Tests for comparing medians tend to be very conservative. The proposed test is able to control the probability of Type I error. It is as powerful as the permutation test and the bootstrap and is more powerful than the MWW test for heavy-tailed populations. The more heavy-tailed the parent population, the greater the power advantage of the proposed test over the MWW test; when the parent population is light-tailed, the MWW test is more powerful than the proposed test.

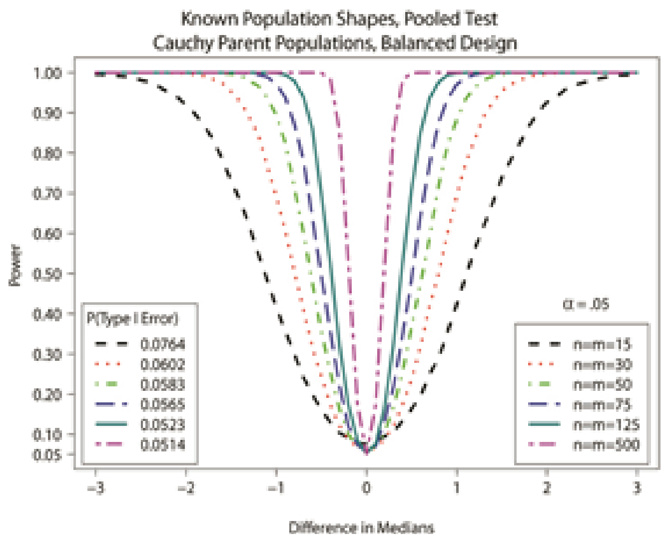
A key precept of the method is that AKDE provides a better estimate of the population density at the median, especially for heavy-tailed populations, than ordinary KDE. As expected, using ordinary KDE makes the test very conservative where the Type I error rate can be as low as 0.02 at the 5% significance level.

These experiments show that the sensitivity parameter $\gamma$ in AKDE controls the trade-off between making the test conservative and liberal, with a suitable value of $\gamma$ producing a test with a nominal significance level. The Type I error rate of the test can be increased (decreased) by increasing (decreasing) $\gamma$.

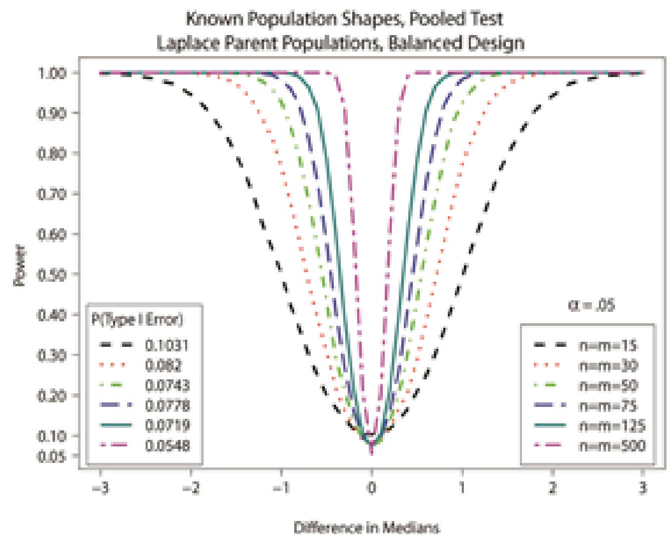
The asymptotic distribution of the sample median has been known for over 50 years (Chu, 1955; Chu & Hotelling, 1955), but it is only now with the improvement in computing power that this theory can be practically employed to derive useful statistical methodology, illustrating the interplay between theory, methodology and computation in the 21st century.

Figure 1: Power Curves for Known Population Shapes (10,000 Simulated Samples)

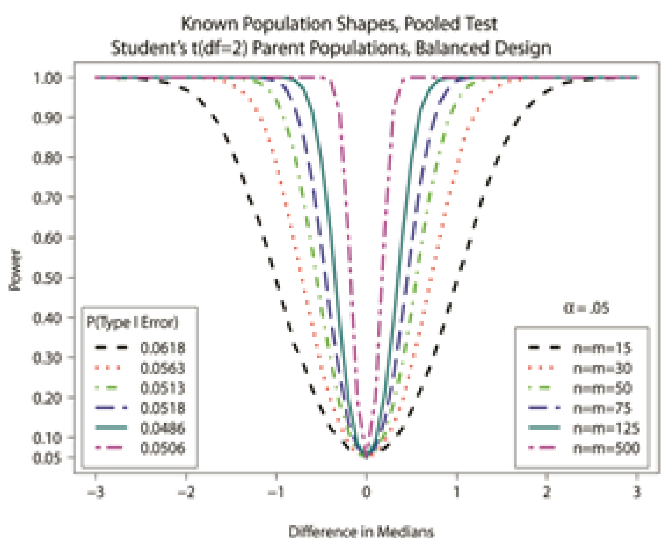
Cauchy




Laplace




Student’s t (df = 2)




Student’s t (df = 3)

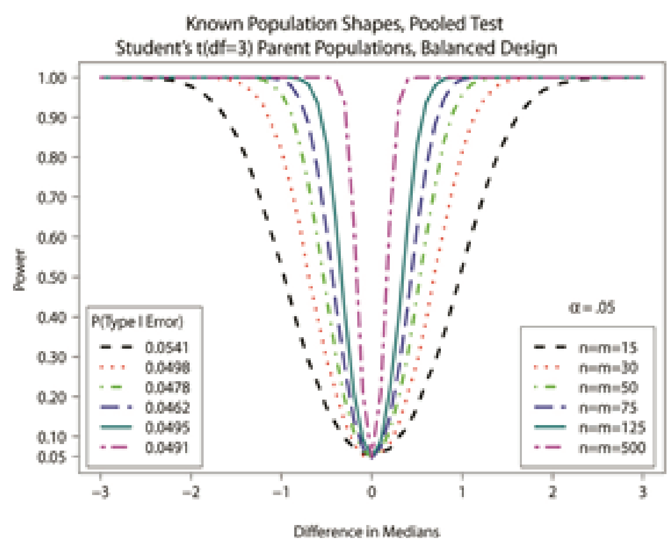
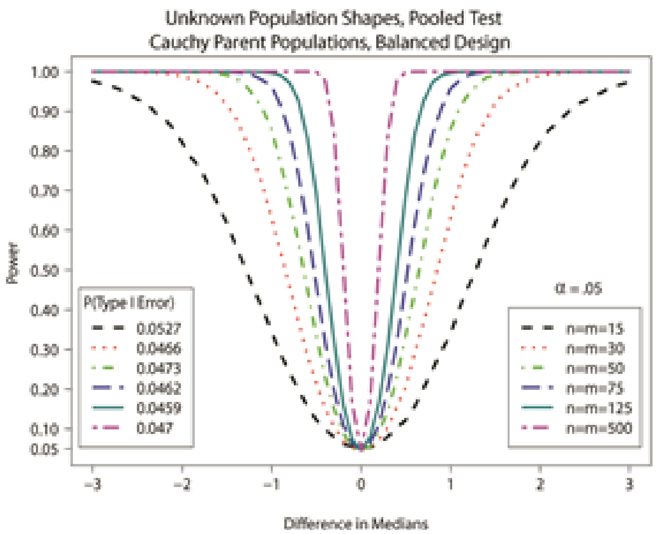
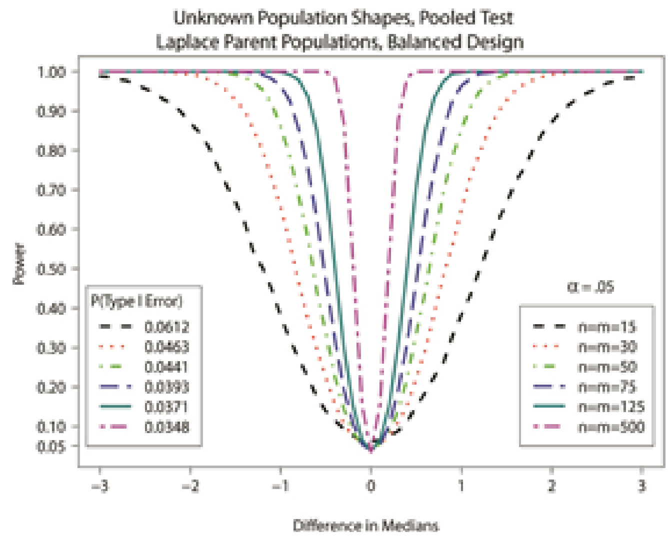


Figure 2: Power Curves for Unknown Population Shapes
(10,000 Simulated Samples, AKDE with $\gamma = .5$ )

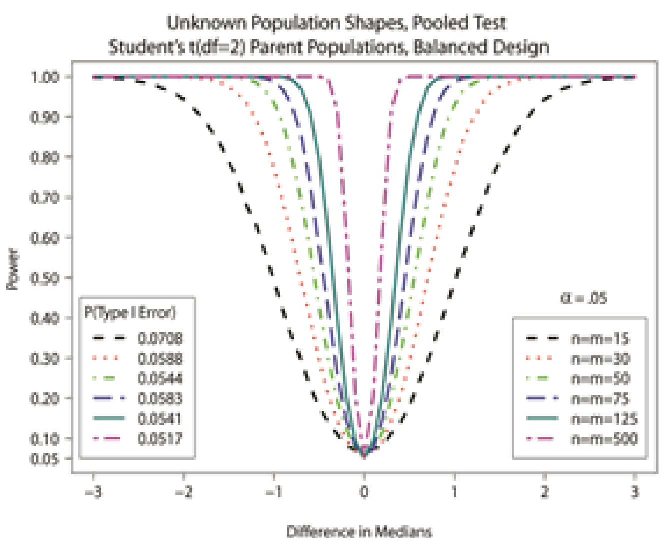
Cauchy




Laplace




Student’s t (df = 2)




Student’s t (df = 3)

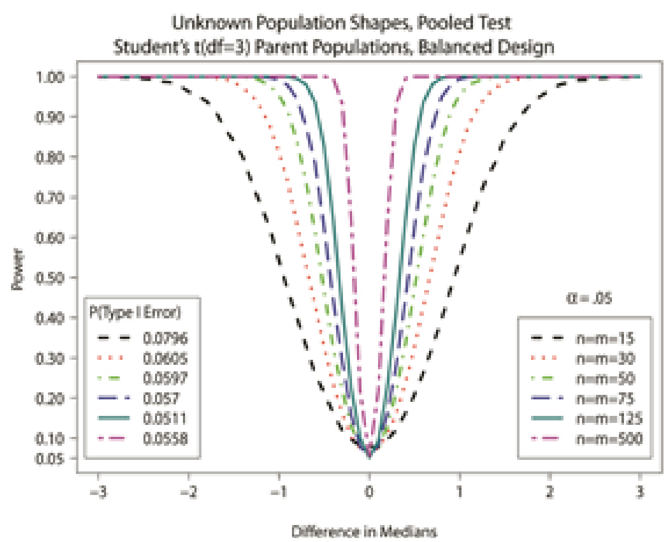
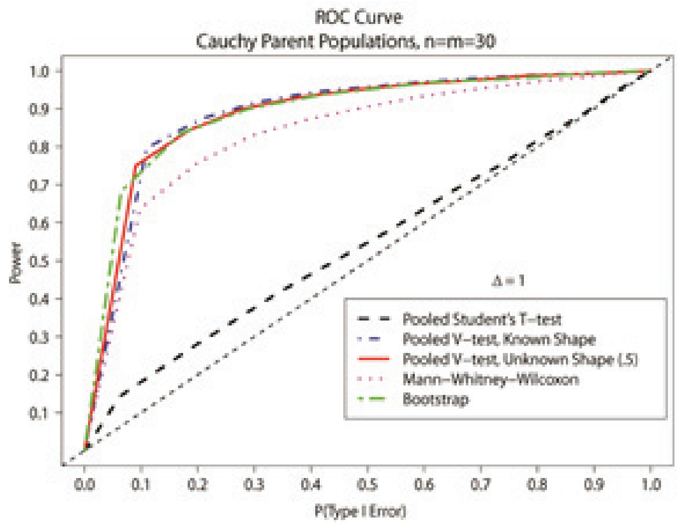
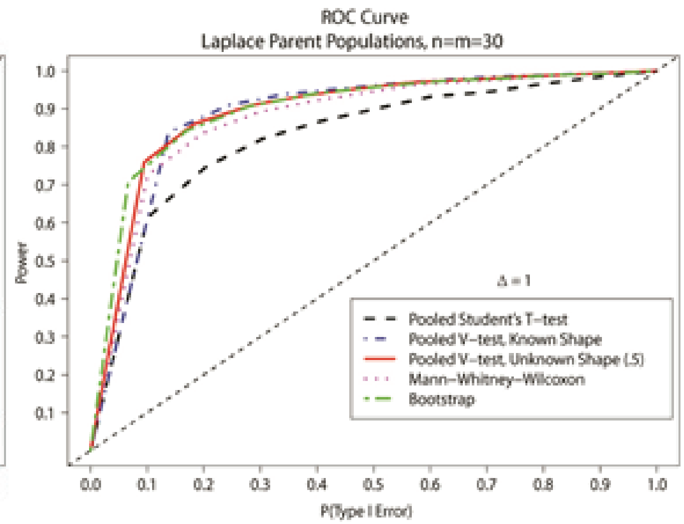


Figure 3: ROC Curves. Balanced Design with n = m = 30 (10,000 Simulated Samples)
(The curves for the permutation test coincide closely with the curves for the proposed test and have been omitted for clarity.)

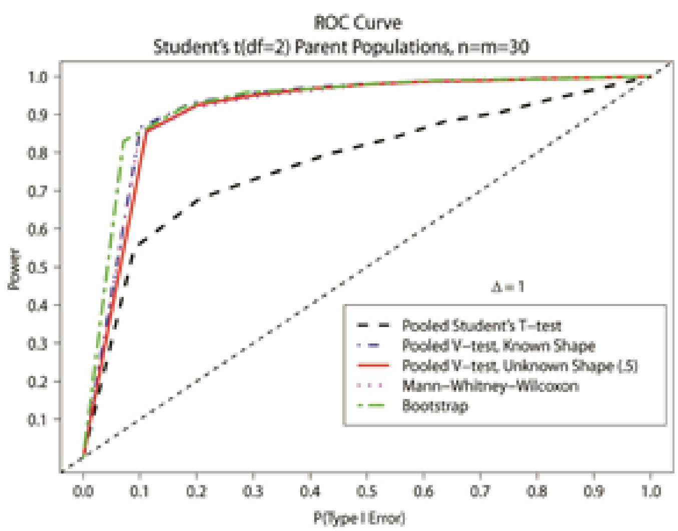
Cauchy



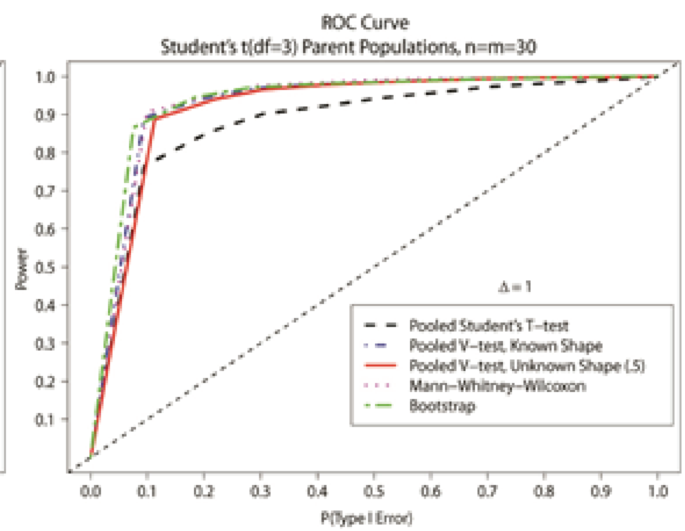

Laplace




Student's t (df = 2)




Student's t (df = 3)



Acknowledgements

The author thanks Professor Emeritus Bruce E. Trumbo, Professor Eric A. Suess and Professor Joshua D. Kerr at California State University, East Bay (Hayward), for helpful discussions and suggestions. The journal staff improved the prose. Earlier versions of this work were contributed at Joint Statistical Meetings (Bichutskiy, Kerr & Trumbo, 2009; Bichutskiy, et al., 2010).

References

Abramson, I. S. (1982). On bandwidth variation in kernel estimates: A square root law. *The Annals of Statistics*, *10*, 1217-1223.

Bichutskiy, V. Y., Kerr, J., & Trumbo, B. E. (2009). Classroom simulation: Investigation of the asymptotic distribution of the sample median. In *JSM Proceedings*, Statistical Education Section, Alexandria, VA: American Statistical Association, 3715-3728.

Bichutskiy, V. Y., Kerr, J. D., Suess, E. A., & Trumbo, B. E. (2010). Classroom derivation and simulation: An asymptotic two-sample test for comparing population medians. In *JSM Proceedings*, Statistical Education Section, Alexandria, VA: American Statistical Association, 4531-4545.

Chu, J. T. (1955). On the distribution of the sample median. *The Annals of Mathematical Statistics*, *26*, 112-116.

Chu, J. T., & Hotelling, H. (1955). The moments of the sample median. *The Annals of Mathematical Statistics*, *26*, 593–606.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(*1*), 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Koenker, R. (2009). *quantreg: Quantile regression*, R package version 4.44. http://CRAN.R-project.org/package=quantreg.

Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, *73*, 194-196.

McKean, J. W., & Schrader, R. M. (1984). A comparison of methods for Studentizing the sample median. *Communications in Statistics – Simulation and Computation*, *13*, 751-773.

Price, R. M., & Bonett, D. G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation*, *68*, 295-305.

R Development Core Team. (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sheather, S. J., & Maritz, J. S. (1983). An estimate of the asymptotic standard error of the sample median. *Australian Journal of Statistics*, *25(1)*, 109-122.

Sheather, S. J. (1986). A finite sample estimate of the variance of the sample median. *Statistics and Probability Letters*, *4*, 337-342.

Silverman, B. W. (1998). *Density estimation for statistics and data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Wilcox, R. R., & Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, *11*(*4*), 263-274.

Wilcox, R. R. (2005). Comparing medians: An overview plus new results on dealing with heavy-tailed distributions. *The Journal of Experimental Education*, *73*(*3*), 249-263.

Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, *51*, 1934-1943.

Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy, 2nd Edition*. New York, NY: Springer.