An Approximation to the Small Sample Distribution of the Trimmed Mean for Gaussian Mixture Models

Alfonso García-Pérez*

Departamento de Estadística I. O. y C. N.,
Universidad Nacional de Educación a Distancia (UNED),
Paseo Senda del Rey 9, 28040-Madrid, Spain
agar-per@ccia.uned.es

Abstract. The α-trimmed mean, a statistic commonly used in robustness studies, has an intractable small sample distribution. For this reason, an asymptotic normal distribution or a Student t distribution are commonly used as approximations when the sample size is small. In this article we obtain an approximation for the small sample distribution of the α-trimmed mean, based on the von Mises expansion of a functional, which is valid for the case in which the observations come from a Gaussian Mixture Model.

Keywords: Robustness, α-trimmed mean, von Mises expansion.

1 Introduction

The α-trimmed mean is a very popular robust statistic used for location problems. If we trim the $100 \cdot \alpha\%$ of the smallest and the $100 \cdot \alpha\%$ of the largest ordered sample data $X_{(i)}$, the symmetrically α-trimmed mean is defined by

$$\overline{X}_\alpha = \frac{1}{n-2k}\left(X_{(k+1)} + ... + X_{(n-k)}\right)$$

where $k = [n\alpha]$ if $[\,.\,]$ stands for the integer part.

Its exact distribution is intractable (see for instance [13] pp. 31). Its large-sample approximation is asymptotically normal under some conditions although more complicated than for other L-estimates; see for instance [12] pp. 361, [13] pp. 31, [1] or [15].

When the sample size is small and the data are normally distributed, a Student's t distribution is used as an approximation for the standardized trimmed mean; see for instance [14] pp. 105 or pp. 156-157, or [16]. In fact, if it is

$$W_i = \begin{cases} X_{(k+1)} \ , & X_i \leq X_{(k+1)} \\ X_i \quad , X_{(k+1)} < & X_i \leq X_{(n-k)} \\ X_{(n-k)} \ , & X_i \geq X_{(n-k)} \end{cases}$$

* The author is very grateful to two anonymous referees for their comments. This work was partially supported by the Grant MTM2012-33740.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*, Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_14

and $\overline{x}_{\alpha}^{W}$ is the α-Winsorized mean

$$\overline{x}_{\alpha}^{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$$

being also the α-Winsorized quasi-variance

$$S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(W_i - \overline{x}_{\alpha}^{W}\right)^2$$

then it is

$$\frac{\overline{X}_{\alpha} - \mu_{\alpha}}{\sqrt{\widehat{V}(\overline{X}_{\alpha})}} = \frac{(1-2\alpha)\sqrt{n}\,(\overline{X}_{\alpha} - \mu_{\alpha})}{S_W} \approx t_{n-2k-1}$$

where

$$\mu_{\alpha} = \frac{1}{1-2\,\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(p)\,dp = \frac{1}{1-2\,\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} y\,dF(y)$$

is the functional associated with the trimmed mean $\overline{X}_{\alpha}$.

Nevertheless, if the data are supposed to come from a normal distribution (i.e., no contamination is assumed) the trimmed mean is not really needed.

There are some Edgeworth expansions used as approximations, [10], but it is well known that these approximations are accurate only in the center of the distribution and not in the tails where they can even be negative.

The only accurate approximations for the distribution of $\overline{X}_{\alpha}$, when the sample size is small and the distribution not normal, are the saddlepoint approximations given in [11] or [2], although these are almost impossible to apply and the elements involved in them, difficult to interpret.

In some articles, [3], [4], [5], [6], [7], [8] and [9], a linear approximation, based on a von Mises expansion plus an iterative procedure, was used to obtain accurate approximations of some classical statistics when the underlying model is close to the normal distribution. In these articles a saddlepoint approximation was used in the computation of the Tail Area Influence Function (TAIF) that appears in the von Mises expansion. But, in two recent articles, [8] and [9], a new expression to compute exactly the TAIF was obtained, formula that can be used in the von Mises expansion instead of the saddlepoint approximation.

We shall use the von Mises expansion in combination with the exact expression of the TAIF, to obtain an accurate approximation to the small sample distribution of the trimmed mean when the underlying model is close to the normal.

2 Definitions and Computations

Although the random variables X_i in the sample $(X_1, ..., X_n)$ are independent and identically distributed (iid), in this section we shall consider statistics

(e.g., the trimmed mean) for which it could be $T_n(X_1 + c, X_2, ..., X_n) \neq T_n(X_1, X_2 + c, ..., X_n)$ for a constant c. For this reason, in the following we shall consider statistics $T_n(X_1, ..., X_n)$ based on independent but not necessarily identically distributed univariate random variables X_i, being $X_i \equiv G_i$, $i = 1, ..., n$ ($X \equiv H$ stands for "X is distributed as H"), statistics that, in the case of a hypothesis testing problem, will reject the null hypothesis (usually about a parameter $\theta \in \Theta$) for large values of T_n, although the results can easily be extended to other situations.

Under very general conditions (Section 2 in [17]) we can use the first-order von Mises expansion (see Corollary 2 in [9]) to compute the tail probability functional under a model $\mathbf{F} = (F_1, ..., F_n)$ as

$$P_{\mathbf{F}}\{T_n(X_1, X_2, ..., X_n) > t\} = P_{F_1,...,F_n}\{T_n(X_1, X_2, ..., X_n) > t\} =$$

$$= P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i\,(x; t; T_n, \mathbf{G})\; dF_i(x) + Rem$$

where TAIF_i is the i-th Partial Tail Area Influence Function of T_n at $\mathbf{G} = (G_1, ..., G_n)$ with relation to G_i, $i = 1, ..., n$, defined in [9] by

$$\mathrm{TAIF}_i(x; t; T_n, \mathbf{G}) = \frac{\partial}{\partial \epsilon} P_{G_i^{\epsilon,x}}\{T_n(X_1, ..., X_n) > t\}\Big|_{\epsilon=0}$$

in those $x \in \mathcal{X}$ where the right hand side exists, being $G_i^{\epsilon,x} = (1 - \epsilon)G_i + \epsilon\,\delta_x$, $i = 1, ..., n$, and δ_x the probability measure which assigns mass 1 at the point $x \in \mathcal{X} \subset \mathbb{R}$.

In the computation of the TAIF_i only G_i is contaminated; the other distributions remain fixed, $i = 1, ..., n$.

Here we assume this situation and also that the X_i's are univariate although an extension to multivariate case would be straightforward (see [9]).

The remainder term

$$Rem = \frac{1}{2} \int \int T^{(2)}_{\mathbf{G_F}}(x_1, x_2)\, d[\mathbf{F}(x_1) - \mathbf{G}(x_1)]\, d[\mathbf{F}(x_2) - \mathbf{G}(x_2)]$$

is small if distributions $\mathbf{F}$ and $\mathbf{G}$ are close. ($T^{(2)}_{\mathbf{G_F}}$ is the *second derivative* of the tail probability functional at the mixture distribution $\mathbf{G_F} = (1 - \lambda)\mathbf{G} + \lambda\mathbf{F}$, for some $\lambda \in [0, 1]$.)

Hence, if $\mathbf{F}$ and $\mathbf{G}$ are close enough, we can write, using the exact expression for the TAIF_i obtained in [9]

$$P_{\mathbf{F}}\{T_n(X_1, X_2, ..., X_n) > t\} \simeq P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \sum_{i=1}^{n} \int_{\mathcal{X}} \mathrm{TAIF}_i\,(x; t; T_n, \mathbf{G})\, dF_i(x) \tag{1}$$

$$= (1 - n) P_{\mathbf{G}}\{T_n(X_1, X_2, ..., X_n) > t\} + \int_{\mathcal{X}} P_{G_2,...,G_n}\{T_n(x, X_2, ..., X_n) > t\}\, dF_1(x) +$$

$$+\int_{\mathcal{X}} P_{G_1,G_3,\ldots,G_n}\{T_n(X_1,x,\ldots,X_n)>t\}\,dF_2(x)+\cdots$$

$$+\int_{\mathcal{X}} P_{G_1,\ldots G_{n-1}}\{T_n(X_1,\ldots,X_{n-1},x)>t\}\,dF_n(x) \tag{2}$$

that allows an approximation of the tail probability $P_{\mathbf{F}}\{T_n > t\}$ under models $(F_1,\ldots,F_n)$, knowing the value of this tail probability under near models $(G_1,\ldots,G_n)$.

In order to value the influence of outliers, we shall consider as model $\mathbf{F} = (1-\epsilon)\mathbf{G} + \epsilon\mathbf{G}_s$ where $\mathbf{G}_s$ is a shift version of $\mathbf{G}$ and $\epsilon \in [0, 0.5]$ a parameter which measures the contamination.

Namely, if $\mathbf{G}$ are location families with a common location parameter θ_0, we shall suppose that $\mathbf{G}_s$ have a common location parameter $\theta > \theta_0$.

In this case, we shall have, for instance, in the last integral (2), if $t = t_n$ is a possible value of T_n and φ the random function (test or critical function in a hypothesis testing problem)

$$\varphi(x_1,\ldots,x_n) = \begin{cases} 1 & if \quad T_n(x_1,x_2,\ldots,x_n) > t_n \\ \\ 0 & if \quad T_n(x_1,x_2,\ldots,x_n) \le t_n \end{cases}$$

that

$$\int_{\mathcal{X}} P_{G_{1;\theta_0},\ldots G_{n-1;\theta_0}}\{T_n(X_1,\ldots,X_{n-1},x)>t_n\}\,dF_n(x)$$

$$=(1-\epsilon)\int_{\mathcal{X}} P_{G_{1;\theta_0},\ldots G_{n-1;\theta_0}}\{T_n(X_1,\ldots,X_{n-1},x)>t_n\}\,dG_{n;\theta_0}(x)$$

$$+\epsilon\int_{\mathcal{X}} P_{G_{1;\theta_0},\ldots G_{n-1;\theta_0}}\{T_n(X_1,\ldots,X_{n-1},x)>t_n\}\,dG_{n;\theta}(x)$$

$$=(1-\epsilon)\int_{\mathcal{X}}\left[\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\varphi(x_1,\ldots,x_{n-1},x)\,dG_{1;\theta_0}(x_1)\cdots dG_{n-1;\theta_0}(x_{n-1})\right]dG_{n;\theta_0}(x)$$

$$+\epsilon\int_{\mathcal{X}}\left[\int_{\mathcal{X}}\cdots\int_{\mathcal{X}}\varphi(x_1,\ldots,x_{n-1},x)\,dG_{1;\theta_0}(x_1)\cdots dG_{n-1;\theta_0}(x_{n-1})\right]dG_{n;\theta}(x)$$

$$=(1-\epsilon)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,\ldots,X_n)>t_n\}+\epsilon P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,\ldots,X_n+(\theta-\theta_0))>t_n\}$$

moving the shift parameter in the last integral with a simple change of variable. Hence, if $\mathbf{F} = (1-\epsilon)\mathbf{G}_{\theta_0} + \epsilon\mathbf{G}_\theta$

$$P_{\mathbf{F}}\{T_n(X_1,X_2,\ldots,X_n)>t_n\} \simeq (1-\epsilon\,n)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,\ldots,X_n)>t_n\}+$$
$$+\epsilon\left(P_{\mathbf{G}_{\theta_0}}\{T_n(X_1+(\theta-\theta_0),X_2,\ldots,X_n)>t_n\}+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2+(\theta-\theta_0),\ldots,X_n)>t_n\}\right.$$

$$+\cdots+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n+(\theta-\theta_0))>t_n\}\big)$$

$$=(1-\epsilon\, n)P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2,...,x_n)\}+\epsilon$$

$$\big(P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1-(\theta-\theta_0),x_2,...,x_n)\}$$

$$+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2-(\theta-\theta_0),...,x_n)\}$$

$$+\cdots+P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,x_2,...,x_n-(\theta-\theta_0))\}\big)\,.$$

3 Iterative Procedure

The previous approximation is accurate if $\mathbf{F}=(1-\epsilon)\mathbf{G}_{\theta_0}+\epsilon\mathbf{G}_{\theta}$ is close to $\mathbf{G}_{\theta_0}$, i.e., if ϵ is small and/or θ is close to θ_0. Nevertheless, in some situations, ϵ is not small or θ is far from θ_0. In these cases we can use an alternative iterative procedure considering intermediate distributions between $\mathbf{G}_{\theta_0}$ and $\mathbf{F}=\mathbf{G}_{(1-\epsilon)\theta_0+\epsilon\theta}$; namely, distributions $\mathbf{F}_j=(F_{1;\theta_j},...,F_{n;\theta_j})=(F_{1j},...,F_{nj})=\mathbf{G}_{\theta_0+(\theta-\theta_0)\epsilon j/(k+1)}$, $j=1,...,k+1$, where $\mathbf{F}_0=\mathbf{G}_{\theta_0}=(G_{1;\theta_0},...,G_{n;\theta_0})$ and $\mathbf{F}_{k+1}=\mathbf{F}=\mathbf{G}_{\theta_0+(\theta-\theta_0)\epsilon}$. With k iterations, equation (1) now becomes

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n\}+$$

$$+\sum_{j=1}^{k+1}\int_{\mathcal{X}}\sum_{i=1}^{n}\mathrm{TAIF}_i\,(x;t_n;T_n,\mathbf{F}_{j-1})\;dF_{ij}(x)$$

Moreover, since

$$\mathrm{TAIF}_i\,(x;t_n;T_n,\mathbf{F}_{j-1})=$$

$$=P_{(F_{1,j-1},...,F_{i-1,j-1},F_{i+1,j-1},...,F_{n,j-1})}\{T_n(X_1,...,X_{i-1},x,X_{i+1},...,X_n)>t_n\}$$

$$-P_{\mathbf{F}_{j-1}}\{T_n(X_1,...,X_n)>t_n\}$$

if we consider again a location family as underlying distribution, i.e., that $\mathbf{F}_j=(F_{1;\theta_j},...,F_{n;\theta_j})=(F_{1j},...,F_{nj})$ is a location family with location parameter $\theta_j=\theta_0+\epsilon j(\theta-\theta_0)/(k+1)$ and the random function φ, we can move again the shift parameter in the distribution to the random variable with a change of variable, obtaining

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n\}+$$

$$+\sum_{j=1}^{k+1}[P_{\mathbf{G}_0}\{T_n(X_1+c_{2j},X_2+c_{1j},...,X_n+c_{1j})>t_n\}$$

$$+P_{\mathbf{G}_0}\{T_n(X_1+c_{1j},X_2+c_{2j},X_3+c_{1j},...,X_n+c_{1j})>t_n\}$$

$$+...+P_{\mathbf{G}_0}\{T_n(X_1+c_{1j},...,X_{n-1}+c_{1j},X_n+c_{2j})>t_n\}$$

$$-nP_{\mathbf{G}_0}\{T_n(X_1+c_{1j},...,X_n+c_{1j})>t_n\}]$$

where $c_{1j}=\epsilon(j-1)(\theta-\theta_0)/(k+1)$ and $c_{2j}=\epsilon j(\theta-\theta_0)/(k+1)$. Hence,

$$P_{\mathbf{F}}\{T_n(X_1,X_2,...,X_n)>t_n\}\simeq P_{\mathbf{G}_{\theta_0}}\{T_n(X_1,X_2,...,X_n)>t_n(x_1,...,x_n)\}$$

$$+\sum_{j=1}^{k+1}[P_{\mathbf{G}_0}\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{2j},x_2-c_{1j},...,x_n-c_{1j})\}$$

$$+P_{\mathbf{G}_0}\{T_n(X_1,X_2,...,X_n)>t_n(X_1-c_{1j},x_2-c_{2j},...,x_n-c_{1j})\}$$

$$+...+P_{\mathbf{G}_0}\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{1j},...,,x_n-c_{2j})\}$$

$$-nP_{\mathbf{G}_0}\{T_n(X_1,X_2,...,X_n)>t_n(x_1-c_{1j},...,x_n-c_{1j})\}].$$

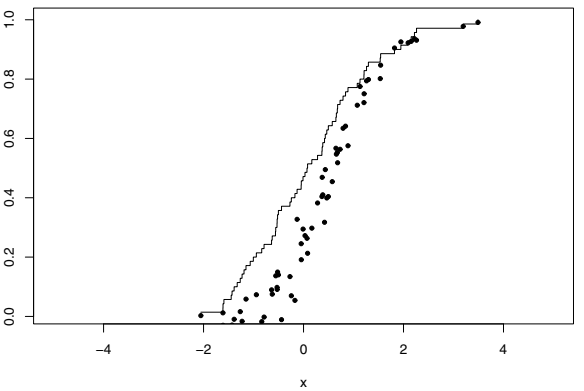

Fig. 1. Simulated (solid line) and von Mises approximation given by (3) (dotted) distributions of T_n with a $N((1-\epsilon)\theta_0+\epsilon\theta,1)$ model

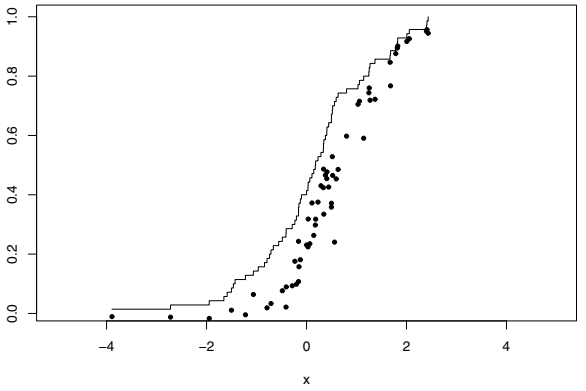

Fig. 2. Simulated (solid line) and von Mises approximation given by (3) (dotted) distributions of T_n with a $(1-\epsilon)N(\theta_0,1)+\epsilon N(\theta,1)$ model

If we considere now the standardized trimmed mean

$$T_n = \frac{\overline{X}_\alpha - \mu_\alpha}{\sqrt{\widehat{V}(\overline{X}_\alpha)}} = \frac{(1-2\alpha)\sqrt{n}\,(\overline{X}_\alpha - \mu_\alpha)}{S_W} \approx t_{n-2k-1}$$

and, as $\mathbf{G}_0$, standard normal distributions, for which we know that $T_n \approx t_{n-2k-1}$ we have

$$P_{G_{(1-\epsilon)\theta_0+\epsilon\theta}}\{T_n(X_1,X_2,...,X_n) > t_n\} \simeq P\{W > t_n(x_1,...,x_n)\}$$
$$+\sum_{j=1}^{k+1}[P\{W > t_n(x_1-c_{2j},x_2-c_{1j},...,x_n-c_{1j})\}+P\{W>t_n(x_1-c_{1j},x_2-c_{2j},...,x_n-c_{1j})\}$$
$$+...+P\{W > t_n(x_1-c_{1j},...,x_n-c_{2j})\}-nP\{W > t_n(x_1-c_{1j},...,x_n-c_{1j})\}] \quad (3)$$

where W is a random variable with a Student's t distribution with $n-2k-1$ degrees of freedom. Hence, with this approximation, we transfer computations under the Gaussian Mixture Model $\mathbf{F}$ to computations of a Student's t distribution.

4 Simulations

If we consider a $N((1-\epsilon)\theta_0+\epsilon\theta,1)$ model as distribution G in the von Mises approximation (3), we observe in Fig. 1 that this approximation (dotted) is accurate considering $n=10,\theta_0=0,\epsilon=0.05,\alpha=0.1,\theta=1$, only with $k=20$ iterations and a simulation of $B=70$ replications in the computations of the simulated distribution of T_n.

In Fig. 2 we see that, even in the case that the underlying model is a $(1-\epsilon)N(\theta_0,1)+\epsilon N(\theta,1)$, the approximation is also accurate with the same values in the parameters as before.

References

1. Bickel, P.J.: On Some Robust Estimates of Location. Ann. Math. Statist. 36, 847–858 (1965)
2. Easton, G.S., Ronchetti, E.: General Saddlepoint Approximations with Applications to L Statistics. J. Amer. Statist. Assoc. 81, 420–430 (1986)
3. García-Pérez, A.: Von Mises Approximation of the Critical Value of a Test. Test 12, 385–411 (2003)
4. García-Pérez, A.: Chi-Square Tests under Models Close to the Normal Distribution. Metrika 63, 343–354 (2006)
5. García-Pérez, A.: t-tests with Models Close to the Normal Distribution. In: Balakrishnan, N., Castillo, E., Sarabia, J.M. (eds.) Advances in Distribution Theory, Order Statistics, and Inference, pp. 363–379. Birkhäuser-Springer, Boston (2006)
6. García-Pérez, A.: Approximations for F-tests which are Ratios of Sums of Squares of Independent Variables with a Model Close to the Normal. Test 17, 350–369 (2008)
7. García-Pérez, A.: Hotelling's T^2-Test with Multivariate Normal Mixture Populations: Approximations and Robustness. In: Pardo, L., Balakrishnan, N., Gil, M.Á. (eds.) Modern Mathematical Tools and Techniques in Capturing Complexity. Understanding Complex Systems, vol. 9, pp. 437–452. Springer, Heidelberg (2011)
8. García-Pérez, A.: Another Look at the Tail Area Influence Function. Metrika 73, 77–92 (2011)
9. García-Pérez, A.: A Linear Approximation to the Power Function of a Test. Metrika 75, 855–875 (2012)
10. Hall, P., Padmanabhan, A.P.: On the Bootstrap and the Trimmed Mean. J. Multivariate Anal. 41, 132–153 (1992)
11. Helmers, R., Jing, B.-Y., Qin, G., Zhou, W.: Saddlepoint Approximations to the Trimmed Mean. Bernoulli 10, 465–501 (2004)
12. Lehmann, E.L.: Theory of Point Estimation. John Wiley and Sons (1983)
13. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. John Wiley and Sons (2006)
14. Staudte, R.G., Sheather, A.J.: Robust Estimation and Testing. John Wiley and Sons (1990)
15. Stigler, S.M.: The Asymptotic Distribution of the Trimmed Mean. Ann. Stat. 1, 472–477 (1973)
16. Tukey, J.W., McLaughlin, D.H.: Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: trimming/winsorization 1. Sankhya A 25, 331–352 (1963)
17. Withers, C.S.: Expansions for the Distribution and Quantiles of a Regular Functional of the Empirical Distribution with Applications to Nonparametric Confidence Intervals. Ann. Stat. 11, 577–587 (1983)