JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

MARC G. GENTON
YANYUAN MA
EMANUEL PARZEN

Discussion of « Sur une limitation très générale de la dispersion de la médiane » by M. Fréchet

Journal de la société française de statistique, tome 147, n° 2 (2006), p. 51-60

<http://www.numdam.org/item?id=JSFS_2006__147_2_51_0>

DISCUSSION OF «SUR UNE LIMITATION TRÈS GÉNÉRALE DE LA DISPERSION DE LA MÉDIANE» BY M. FRÉCHET

Marc G. GENTON [*,‡], Yanyuan MA [†,‡], and Emanuel PARZEN [‡]

This is the first time we had an opportunity to read this paper by M. Fréchet and we enjoyed doing that very much. In brief, M. Fréchet demonstrates that measured with the classical sample variance estimator, the variability of the sample mean is smaller than the variability of the sample median in fairly general classes of distributions for the sample. However, he points out the fact that measured with other variability estimators, for instance such as the semi interquartile range, this order may reverse. M. Fréchet then presents an ingenious experiment that illustrates this fact.

Our discussion is centered around three main themes. The first theme concentrates on the definition of quantiles for discrete random variables and samples with ties. The second theme is devoted to the analysis of M. Fréchet's experiment through computer simulations. The third theme is concerned with the exact distribution of the median and the semi interquartile range for correlated samples, possibly from heavy-tailed distributions.

1. Sample quantiles for discrete random variables and samples with ties

Following Parzen (2004), we introduce various concepts defining the sample versions of the probability distribution function $F(x) = P(X \leqslant x)$ and the quantile function $Q(u) = F^{-1}(u)$, $0 \leqslant u \leqslant 1$, of a random variable X. We define the probability mass function $p(x) = P(X = x)$, the probability density function $f(x) = F'(x)$, and the mid-distribution function $F^{\text{mid}}(x) = F(x) - 0.5p(x)$. Note when the random variable X is continuous, $p(x) \equiv 0$ and thus $F^{\text{mid}}(x) = F(x)$. The mid-distribution plays a crucial role in defining sample quantiles for samples with ties as shown below.

The sample distribution function of a sample $X_1, \ldots, X_n$ is defined as $\tilde{F}(x) = \tilde{P}(X \leqslant x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x)$ where $I(\cdot)$ denotes the indicator function.

* Department of Econometrics, University of Geneva, Bd du Pont-d'Arve 40, CH-1211 Geneva 4, Switzerland. E-mail : Marc.Genton@metri.unige.ch

† Group of Statistics, University of Neuchâtel, Pierre à Mazel 7, CH-2000 Neuchâtel, Switzerland.

‡ Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. E-mail : {genton, ma, eparzen}@stat.tamu.edu

Consequently, the sample quantile function, denoted by $\tilde{Q}(u) = \tilde{F}^{-1}(u)$, is defined to be a piecewise constant function such that $\tilde{Q}(u) = X_{j:n}$ for $(j-1)/n < u \leqslant j/n$, where $X_{1:n} \leqslant \ldots \leqslant X_{n:n}$ are the order statistics of the sample. Similarly, we define the sample probability mass function $\tilde{p}(x) = \tilde{P}(X = x)$ and the sample mid-distribution function $\tilde{F}^{\text{mid}}(x) = \tilde{F}(x) - 0.5\tilde{p}(x)$.

The sample median $\tilde{Q}(0.5)$, based on the sample quantile function defined above, unfortunately does not agree with the usual definition : $X_{m+1:n}$ if $n = 2m+1$ and $0.5(X_{m:n} + X_{m+1:n})$ if $n = 2m$. This motivates the definition of the continuous sample quantile function $\tilde{Q}^{\text{c}}(u)$ as being piecewise linear and connecting the values $\tilde{Q}^{\text{c}}((j-0.5)/n) = X_{j:n}$ for samples with distinct values, i.e.

$$\tilde{Q}^{\text{c}}(u) = (0.5 - nu + j)X_{j:n} + (0.5 + nu - j)X_{j+1:n}, \tag{1}$$

for $(j-0.5)/n < u \leqslant (j+0.5)/n$. It can be verified that this definition agrees with the usual definition of the sample median, whereas many computer programs use ad hoc sample quantile definitions that do not.

The extension of these concepts to samples with ties is based on the mid-distribution function. Denote the distinct values in the sample by $x_1, \ldots, x_d$. Then, the continuous sample quantile function $\tilde{Q}^{\text{c}}(u)$ for samples with ties is defined as being piecewise linear and connecting the values $\tilde{Q}^{\text{c}}(\tilde{F}^{\text{mid}}(x_j)) = x_j$, which can be viewed as a definition of fractional order statistics. Although it is a bit tedious to write down the analytic form of $\tilde{Q}^{\text{c}}$, the concept is very simple to grasp. As an illustrative example, consider a sample of size $n = 5$ of binary data $0, 1, 1, 1, 1$. The usual definition yields a sample median of 1. However, the previous definition for samples with ties yields a sample median of $4/5$, which is also the empirical proportion of 1's, that is, $\tilde{p}(1)$. This answer appears to be much more natural.

In his experiment, M. Fréchet uses neither the continuous sample quantile function $\tilde{Q}^{\text{c}}(u)$ defined above, nor the definition for samples with ties. Nevertheless, the samples he considered do involve ties. For example on p. 76, M. Fréchet reports the sample medians of eight different series as being 3, 2, 1, -1, 0, -3, 0, 1. Based on the definition above for samples with ties, we find instead the sample median values to be $2.6, 2, 1, -0.25, 0.5, -3.25, 1/3, 7/6$. Table 1 describes the steps associated with the computation of the sample median of the first series reported by M. Fréchet. A linear interpolation between (2, 9/24) and (3, 14/24) yields a sample median of 2.6.

The semi interquartile range used by M. Fréchet as a measure of variability is half the difference between the upper sample quartile and the lower sample quartile, i.e., $(\tilde{Q}(0.75) - \tilde{Q}(0.25))/2$. Obviously, the computation of the semi interquartile range is also influenced by ties in the sample and will yield different results from the traditional definition of the semi interquartile range. For example, treating $\pm\alpha$ as specific observations, the lower sample quartiles of the eight series are $-2, -1.25, -7/3, -3, -2.5, -(\alpha+4)/2, -2.5, -2/3$ and the upper sample quartiles are $3.8, 3.5, 2.8, (9+\alpha)/4, 3.5, 1, 3+\alpha/4, 2.2$. Hence, the semi interquartile ranges are $2.9, 2.38, 77/30, (21+\alpha)/8, 3, (\alpha+6)/4, (22+$

TABLE 1. — Computation of the sample median of the first series.

X	$-\alpha$	-4	-3	-2	-1	$+1$	$+2$	$+3$	$+4$	$+\alpha$
Series 1	0	2	1	0	1	0	1	4	1	2
$\tilde{F}(x)$	0	$\frac{2}{12}$	$\frac{3}{12}$	$\frac{3}{12}$	$\frac{4}{12}$	$\frac{4}{12}$	$\frac{5}{12}$	$\frac{9}{12}$	$\frac{10}{12}$	1
$\tilde{p}(x)$	0	$\frac{2}{12}$	$\frac{1}{12}$	0	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{4}{12}$	$\frac{1}{12}$	$\frac{2}{12}$
$\tilde{F}^{\text{mid}}(x)$	0	$\frac{2}{24}$	$\frac{5}{24}$	$\frac{6}{24}$	$\frac{7}{24}$	$\frac{8}{24}$	$\frac{9}{24}$	$\frac{14}{24}$	$\frac{19}{24}$	$\frac{22}{24}$

$\alpha)/8, 43/30$. We return to the effect of neglecting ties in sample quantile computations on M. Fréchet's experiment in the next section.

2. M. Fréchet's experiment in the computer age

The experiment described by M. Fréchet is based on samples of size $n = 12$ and 8 replicates which allowed him to perform computations "by hand". We take advantage of modern computational capabilities provided by computers to reanalyze and extend M. Fréchet's experiment with larger sample sizes and 500 simulation replicates. Throughout we used the statistical software R for our implementation, noting that sample quantiles for samples with ties are not part of the software and needed to be specifically programmed for implementation.

First, we investigate the difference between the use of the classical median and the median for ties in M. Fréchet's experiment. Figure 1 depicts histograms of the median of samples from the discrete uniform distribution used by M. Fréchet, with $\alpha = 5$ and sizes $n = 12$ (original experiment; top panels) and $n = 200$ (bottom panels). The left panels use the classical sample median whereas the right panels use the sample median for ties. When $n = 12$, the samples do not have many ties, so the difference between the two median estimators is hardly noticeable. However, when $n = 200$, the samples have many ties and the difference becomes very obvious. The classical median often takes the values -1 or $+1$, and is occasionally 0. Intuitively, the reason is the following. The definition of median used by M. Fréchet, in the presence of ties, is equivalent to the following procedure : 1) draw observations to form a sample; 2) perturb the observations with the same tied value by adding small different noises, for example, 1, 1, 1, 1, will be perturbed to $1, 1 + \epsilon, 1 + 2\epsilon, 1 + 3\epsilon$, for sufficiently small ϵ (say $\epsilon = 1/(4n)$, note that one observation of each set of ties should be left unperturbed); 3) calculate the median of the perturbed sample, which does not contain any ties; 4) "unperturb" the resulting median. For large sample size (in our case, $n = 200$), it is very likely that each possible tied value will have *approximately* the same number of observations in the sample, hence the median will almost always be among

-1, 0 and $+1$. Only when exactly $n/2$ observations in the sample are positive, the sample median will be 0. Otherwise, the sample median is typically -1 or $+1$. Hence, the fundamental reason that the classical sample median does not treat ties correctly is that it does not really recognize ties. Indeed, it essentially considers tied observations as different observations that happen to be extremely close. On the contrary, the median for ties recognizes the special meaning of ties, by reporting the proportion of certain tied observations in a sample. Moreover, unlike the classical median, the asymptotic distribution of the sample median for ties seems to be of normal type. We have not found a formal proof of the asymptotic normality of the sample median for ties in the literature, but of course it is closely linked to the case of continuous variables. Indeed, it is well-known that if $X_1, \ldots, X_n$ is a random sample, independently and identically distributed according to a distribution $F(x)$ with continuous density $f(x)$, mean μ, finite variance σ^2, and such that $f(Q(p)) > 0$, then

$$\sqrt{n}(\tilde{Q}(p) - Q(p)) \to N(0, p(1-p)/f(Q(p))^2), \tag{2}$$

in distribution when $n \to \infty$. We conjecture a similar result holds for discrete distributions, such as the one used by M. Fréchet in his experiment, as long as the quantiles defined for samples with ties are adopted. A rigorous proof is beyond the scope of this discussion though.

Next, we extend the previous simulations by letting α vary from 5 to 60, thus creating heavier tails in the discrete uniform distribution. We are interested in studying the relation between α and the dispersion reduction studied by M. Fréchet, $\sqrt{n}S_{D_n}/S_X$. We experiment with three different descriptions for the location of the data, i.e. we let D_n be the sample mean, the classical sample median and the sample median for ties, respectively. In terms of the dispersion measure, we experiment with two different criteria, where S is sample standard deviation and sample semi interquartile range, respectively. Note that there are two different approaches in forming the semi interquartile range, resulting from two different ways of calculating the lower and upper quartiles. In our experiment, we form the semi interquartile range using classical quartile calculations when D_n is sample mean and classical median, while using sample quartile with ties when D_n is the sample median with ties. The result of these experiments is presented in the top panels of Figure 2.

Agreeing with M. Fréchet's conclusion, the dispersion reduction of the sample mean based on the semi interquartile range is unbounded in α, whereas all other dispersion reductions are bounded. We can specifically calculate that under sample standard deviation, the dispersion reduction is 1 for the sample mean and is $\sqrt{n}/\sqrt{6+\alpha^2/5}$ for the classical median. Similarly, under the classical semi interquartile range, the dispersion reduction is $0.1927\sqrt{6+\alpha^2/5}$ for the sample mean and $\sqrt{n}/3.5$ for the classical sample median. Specific calculations for the sample median with ties are not available at this moment since the asymptotic behavior for sample quantiles with ties has not been established rigorously. Note that the dispersion reduction of the medians based on the standard deviation tends to zero as $\alpha \to +\infty$. This is an indication of

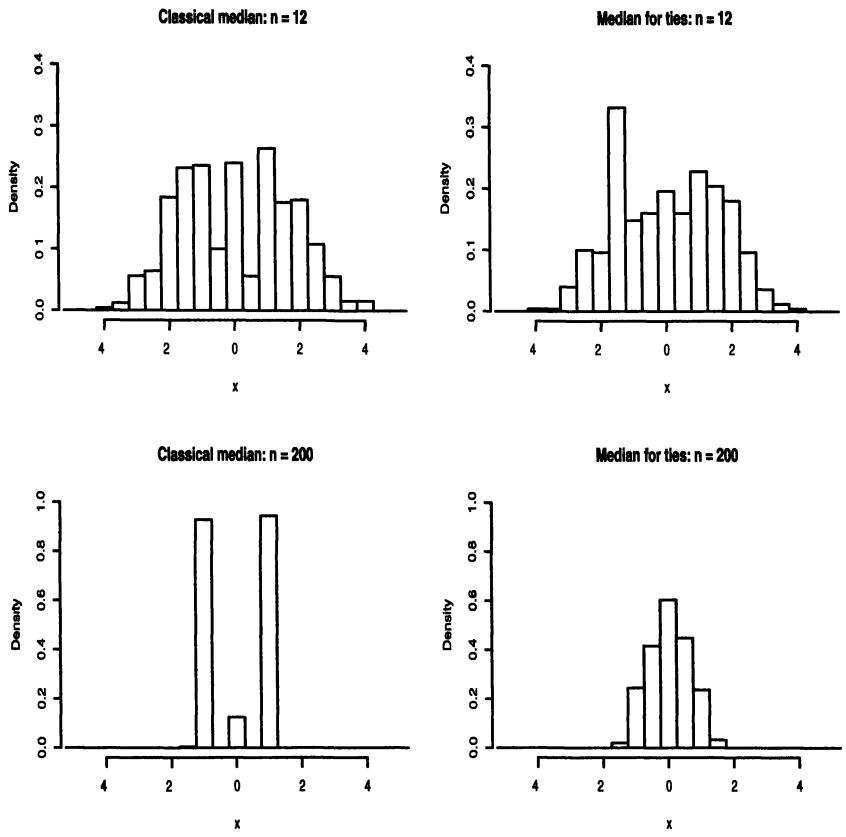

FIG 1. — Simulations of M. Fréchet's experiment based on 500 replicates with sample size $n = 12$ (original experiment; top panels) and $n = 200$ (bottom panels) : histograms comparing the classical sample median (left panels) and the sample median for ties (right panels).

the robustness of the median compared to the sample mean when $\alpha \to +\infty$, that is, when the tails of the distribution of the sample become heavier.

We are also interested in the performance of the dispersion reduction when the sample size increases. In the bottom panels of Figure 2, we present the dispersion reduction for different sample sizes, with $\alpha = 5$ fixed. We can see that, contrary to the conclusion of M. Fréchet, the dispersion reduction for the classical sample median increases as n increases, under both the sample standard deviation measure and the sample semi interquartile range measure. In fact, from our calculation in the previous paragraph, it can be easily verified that they increase as a function $h(n) = c\sqrt{n}$ for $c = 0.3015$ and $c = 0.2857$, respectively. Only when the sample median with ties is implemented, will the dispersion reductions be bounded asymptotically. On the other hand, the dispersion reductions for the sample mean remain approximately at 1 under

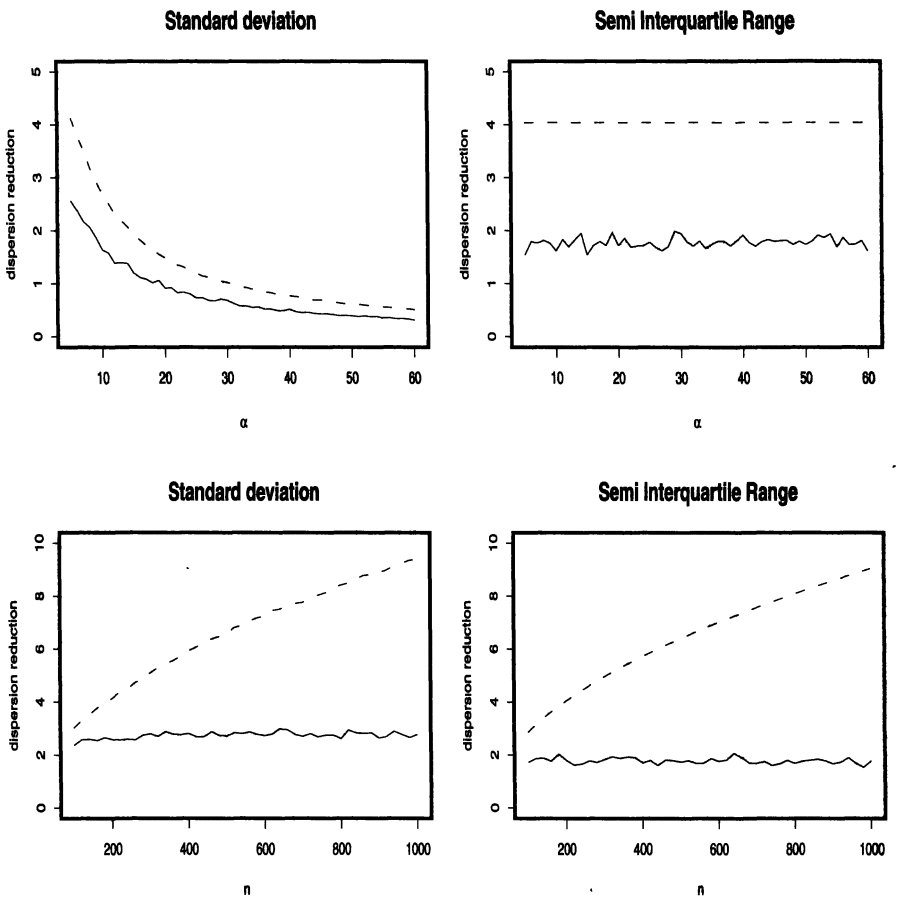

FIG 2. — Simulation analysis of the dispersion reduction as a function of α (top panels) and a function of n (bottom panels). In the left panels, the dispersion reduction is measured with sample standard deviation, in the right panels, it is measured with sample semi interquartile range. Experiment with sample mean (dotted lines), classical sample median (dashed lines) and sample median for ties (solid lines) are presented. All results are based on 500 replicates. In the top panels, $n = 200$, in the bottom panels, $\alpha = 5$.

the sample standard deviation measure and at 0.6391 under the sample semi interquartile range measure.

3. Exact distribution of the median and semi interquartile range

The sample median and semi interquartile range described by M. Fréchet are special linear combinations of order statistics, also called L-statistics, and their exact distribution is known in case of independent and identically distributed random variables. Recently, Arellano-Valle and Genton (2006a,b) have shown that the exact distribution of L-statistics from absolutely continuous dependent random variables is closely related to the so-called fundamental skew distributions introduced by Arellano-Valle and Genton (2005). Specifically, let $\mathbf{X} \overset{d}{=} (\mathbf{Y}|\mathbf{Z} \geqslant \mathbf{0})$, where $\mathbf{Y} \in \mathbb{R}^n$ is a random vector with probability density function $f_{\mathbf{Y}}$, $\mathbf{Z} \in \mathbb{R}^m$ is a random vector, and the notation $\mathbf{Z} \geqslant \mathbf{0}$ is meant component-wise. Then, following Arellano-Valle and Genton (2005), $\mathbf{X}$ has an n-dimensional fundamental skew (FUS) distribution with probability density function given by :

$$f_{\mathbf{X}}(\mathbf{x}) = K_m^{-1} f_{\mathbf{Y}}(\mathbf{x}) Q_m(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \tag{3}$$

where $Q_m(\mathbf{x}) = P(\mathbf{Z} \geqslant \mathbf{0}|\mathbf{Y} = \mathbf{x})$ and $K_m = \mathrm{E}(Q_m(\mathbf{Y})) = P(\mathbf{Z} \geqslant \mathbf{0})$ is a normalizing constant. When $f_{\mathbf{Y}}$ is a symmetric probability density function (i.e. $f_{\mathbf{Y}}(-\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y})$ for all $\mathbf{y} \in \mathbb{R}^n$), (3) defines the fundamental skew-symmetric ($FUSS$) class of distributions. Because we consider $\mathbf{Y}$ conditionally on $\mathbf{Z} \geqslant \mathbf{0}$, this selection mechanism induces skewness in the probability density function $f_{\mathbf{X}}$ through the term Q_m. Arellano-Valle, Branco, and Genton (2006) present a unified view on skewed distributions resulting from selections. The book edited by Genton (2004) describes further properties and applications of these distributions.

We briefly investigate the change of shape of the exact distribution of the sample median and semi interquartile range as a function of the correlation between the observations and the heaviness of the tails of the original data. For simplicity, we focus on the case of exchangeable absolutely continuous random variables, that is, the correlation between any two observations is the same $\rho \in [0, 1)$, although general dependence structures can be handled as well. Let $\mathbf{X} = (X_1, \ldots, X_n)^T \sim EC_n(\mu \mathbf{1}_n, \sigma^2(1-\rho)\Omega_n, \varphi)$, $\mu \in \mathbb{R}$, $\sigma > 0$, be an exchangeable elliptically contoured absolutely continuous random vector with $\Omega_n = I_n + \frac{\rho}{1-\rho}\mathbf{1}_n\mathbf{1}_n^T$ and characteristic generator φ. The probability density function of $\mathbf{X}$ is denoted by $f_n(\mathbf{x}; \mu\mathbf{1}_n, \sigma^2(1-\rho)\Omega_n, h^{(n)})$ with density generator $h^{(n)}$, and its cumulative distribution function by $F_n(\mathbf{x}; \mu\mathbf{1}_n, \sigma^2(1-\rho)\Omega_n, \varphi)$, for $\mathbf{x} \in \mathbb{R}^n$. Denote by $\mathbf{X}_{(n)} = (X_{1:n}, \ldots, X_{n:n})^T$ the vector of order statistics induced by $\mathbf{X}$. Then Arellano-Valle and Genton (2006b) have shown that for any matrix $L \in \mathbb{R}^{p \times n}$ of rank p ($1 \leqslant p \leqslant n$), the probability density function $f_{L\mathbf{X}_{(n)}}$ of $L\mathbf{X}_{(n)}$ is :

$$\begin{aligned} &f_{L\mathbf{X}_{(n)}}(\mathbf{y}) = n! f_p(\mathbf{y}; \mu L\mathbf{1}_n, \sigma^2(1-\rho)L\Omega_n L^T, h^{(p)}) \times \\ &F_{n-1}(\Delta L^T (L\Omega_n L^T)^{-1}\mathbf{u}; \mathbf{0}, \Delta\{I_n - L^T(L\Omega_n L^T)^{-1}L\}\Delta^T, \varphi_{q(\mathbf{y})}), \quad \mathbf{y} \in \mathcal{S}_p, \end{aligned} \tag{4}$$

where $\Delta \in \mathbb{R}^{(n-1)\times n}$ is a difference matrix such that $\Delta\mathbf{X} = (X_2 - X_1, X_3 - X_2, \ldots, X_n - X_{n-1})^T$, $q(\mathbf{y}) = \mathbf{u}^T(L\Omega_n L^T)^{-1}\mathbf{u}$, with $\mathbf{u} = (\mathbf{y} - \mu L\mathbf{1}_n)/(\sigma\sqrt{1-\rho})$, and the region of support defined on $\mathbb{R}^p$ is $\mathcal{S}_p = \{\mathbf{y} = L\mathbf{x};\ \Delta\mathbf{x} \geqslant \mathbf{0},\ \mathbf{x} \in \mathbb{R}^n\}$.

The median and semi interquartile range each correspond to the case $p = 1$, that is, univariate L-statistics of the form $L\mathbf{X}_{(n)} = \sum_{i=1}^n a_i X_{i:n}$, $a_i \in \mathbb{R}$. By (4), their densities are of the form

$$f_{L\mathbf{X}_{(n)}}(y) = n! f_1(y; \eta, \tau^2, h^{(1)}) F_{n-1}(z\Delta\mathbf{a}_\gamma; \mathbf{0}, \Delta\{I_n - \mathbf{a}_\gamma\mathbf{a}_\gamma^T\}\Delta^T, h_{z^2}^{(n-1)}),\ y \in \mathbb{R}, \tag{5}$$

where $\eta = \mu\sum_{i=1}^n a_i$, $\tau^2 = \sigma^2(1-\rho)\{\sum_{i=1}^n a_i^2 + \gamma(\sum_{i=1}^n a_i)^2\}$, $z = (y-\eta)/\tau$, $\mathbf{a} = (a_1, \ldots, a_n)^T$, and $\mathbf{a}_\gamma = \mathbf{a}/\{\sum_{i=1}^n a_i^2 + \gamma(\sum_{i=1}^n a_i)^2\}^{1/2}$, with $\gamma = \rho/(1-\rho)$. Moreover, those cases with $L\mathbf{1}_n = \sum_{i=1}^n a_i = 0$, for example such as the semi interquartile range, yield additional simplifications.

As an illustration, consider an exchangeable sample of size $n = 8$ with a multivariate distribution with correlation $\rho = 0, 0.1, \ldots, 0.9$ of two types : Normal, and Student t with 3 degrees of freedom. Then, the median corresponds to $\mathbf{a} = (0,0,0,1/2,1/2,0,0,0)^T$ whereas the semi interquartile range to $\mathbf{a} = (0,-1/4,-1/4,0,0,1/4,1/4,0)^T$ according to formula (1). Figures 3 and 4 depict the exact density of the median and semi interquartile range in that setting (top panel : Normal; bottom panel : Student t) based on (5). The bold curve is the density for $\rho = 0$. The dashed curve is the marginal density of the sample.

From Figure 3, we note that the exact density of the median is symmetric and has increasingly heavier tails when either the correlation ρ increases or the heaviness of the tails of the sample's distribution increases. When $\rho \to 1$, the exact density of the median tends to the marginal density of the sample.

From Figure 4, we note that the exact density of the semi interquartile range is skewed, even more so when the heaviness of the tails of the sample's distribution increases. When the correlation ρ increases, the spread of the distribution decreases.

References

ARELLANO-VALLE R. B., BRANCO M. D. and GENTON M. G. (2006). A unified view on skewed distributions arising from selections. *The Canadian Journal of Statistics*, 34, 581-601.

ARELLANO-VALLE R. B. and GENTON M. G. (2005). On fundamental skew distributions. *Journal of Multivariate Analysis*, 96, 93-116.

ARELLANO-VALLE R. B. and GENTON M. G. (2006a). On the exact distribution of the maximum of absolutely continuous dependent random variables. *Statistics & Probability Letters*, in press.

ARELLANO-VALLE R. B. and GENTON M. G. (2006b). On the exact distribution of linear combinations of order statistics from dependent random variables. *Journal of Multivariate Analysis*, in press.

GENTON M. G. (2004). *Skew-Elliptical Distributions and Their Applications : A Journey Beyond Normality.* Edited Volume, Chapman & Hall/CRC, Boca Raton, FL, 416 pp.

PARZEN E. (2004). Quantile probability and statistical data modeling. *Statistical Science*, 19, 652–662.

R Development Core Team (2004). 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`

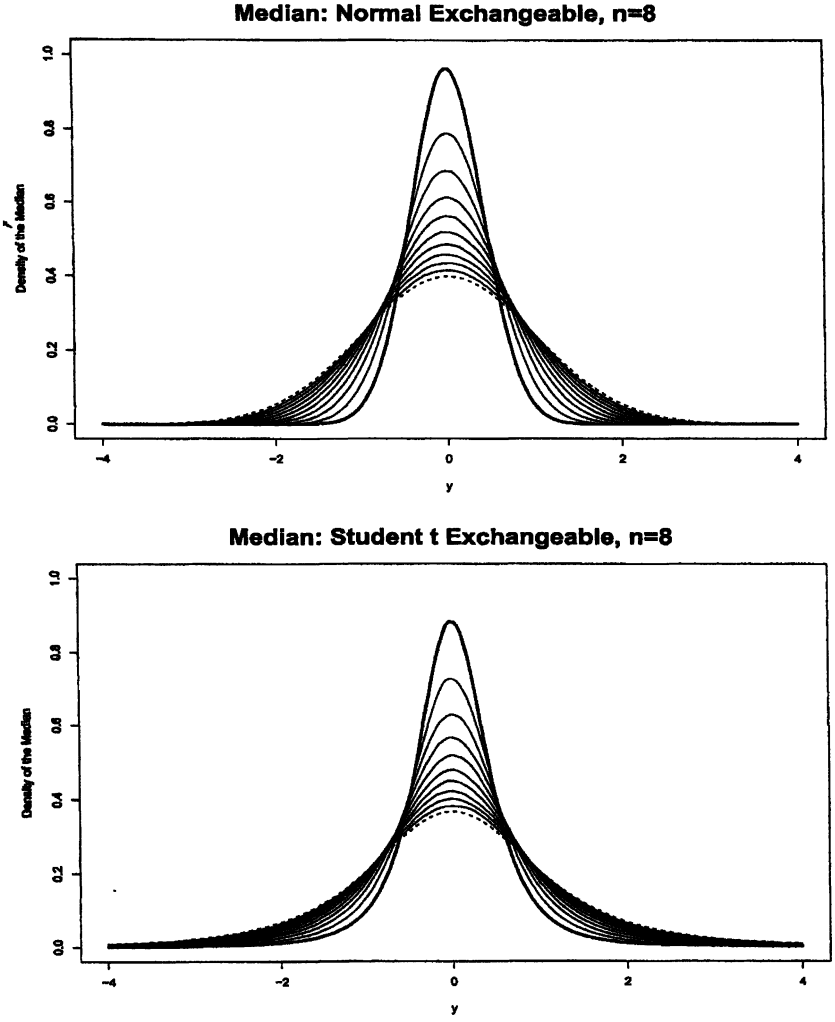

FIG 3. — Exact density of the median for an exchangeable sample of size $n = 8$ with a multivariate distribution with correlation $\rho = 0, 0.1, \ldots, 0.9$: Normal (top panel); Student t with 3 degrees of freedom (bottom panel). The bold curve is the density for $\rho = 0$. The dashed curve is the marginal density of the sample.

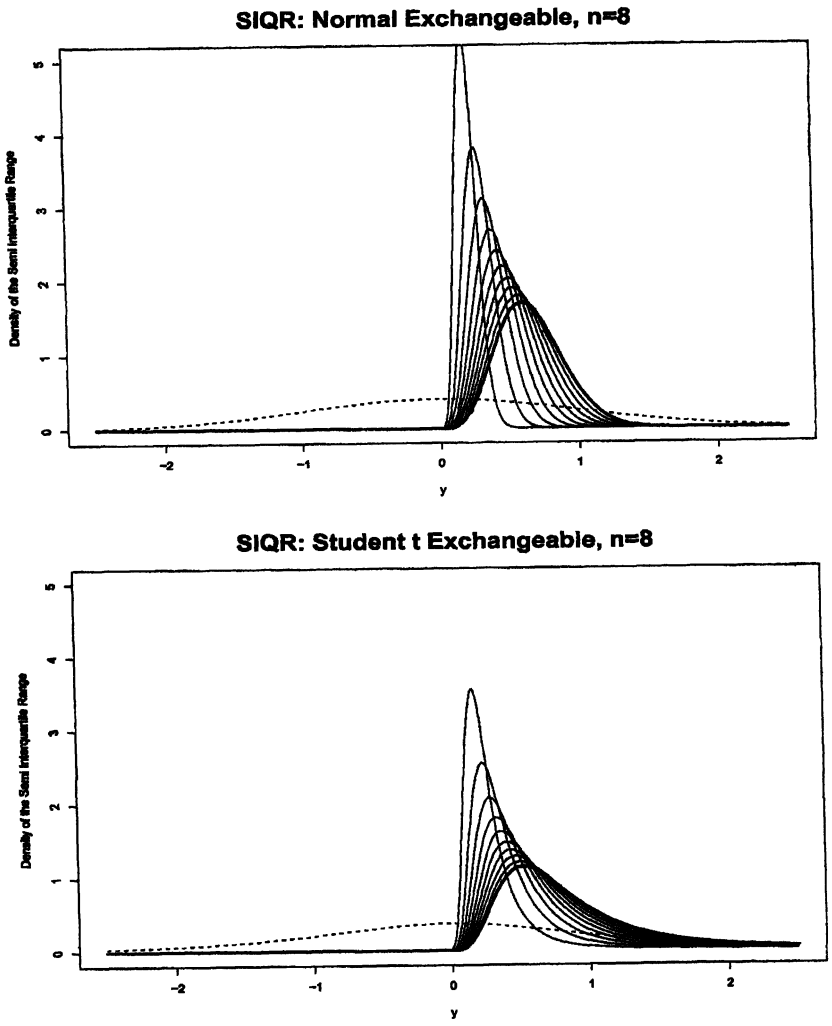

FIG 4. — Exact density of the semi interquartile range (SIQR) for an exchangeable sample of size $n = 8$ with a multivariate distribution with correlation $\rho = 0, 0.1, \ldots, 0.9$: Normal (top panel); Student t with 3 degrees of freedom (bottom panel). The bold curve is the density for $\rho = 0$. The dashed curve is the marginal density of the sample.