Remarks on the estimation of some variability measures

Marcella Polisicchio
Dipartimento di Metodi quantitativi per le Scienze economiche ed aziendali
Università degli Studi di Milano-Bicocca
marcella.polisicchio@unimib.it


1. Introduction

A location and a dispersion measure are usually given to start the analysis of a variate $X$ which is observed on a finite or infinite population. Sometimes the purpose of the application may suggest the use of a particular location and variability index. Focusing on the latter, the choice frequently falls on the standard deviation. This solution is often due either to the interesting properties of the square of the standard deviation, the variance, or to the hypothesis that $X$ is normally distributed. Very often only sample data selected from the population are available and the purpose is to estimate a variability measure. Hence it is necessary to propose an opportune estimator and to study its properties and the sample distribution. It seems reasonable to distinguish the situation where $X$ is a random variable (r.v.) with a known distribution function except for some parameters, from the non-parametric or the finite population case. In fact, in the first situation any dispersion measure is a function of the parameters of the distribution. Therefore using an opportune estimator of the parameters a good estimation of the variability measure can be obtained. The properties and the distribution of the dispersion measure estimator may be easily inferred from the characteristics of the estimator of the parameters. Instead in the non-parametric situation the properties of the estimator have to be verified directly. Hence for some estimators the obtained results are only approximate. Moreover for the finite population it is necessary to specify the sample design.

In this paper we intend to reflect on the choice of a variability measure and on its estimate by comparing their natural estimators and considering the conditions for the finite existence of the variance of these estimators.
2. Notation

In order to present an easy and complete treatment, it is assumed that 
\( a_1, a_2, \ldots, a_N \) are the values that the variate \( X \) takes on the \( N \) units of a finite population (f.p.). In all the other situations it is assumed, without losing in generality, that \( X \) is a continuous r.v. with probability density function \( f(x; \theta) \), where \( \theta \) is the vector of parameters. Obviously the dispersion of \( X \) may be measured in various ways, but we are interested only on the variability indexes that can be expressed as arithmetic mean – for a f.p. - or expectation value – for a continuous r.v. - of an opportune function of \( X \), say \( E[g(X)] \).

In particular, we are interested in the standard deviation of \( X \),

\[
\sigma = \left\{ E[ (X - \mu)^2 ] \right\}^{1/2},
\]

where \( \mu \) is the arithmetic mean or the expectation value of \( X \), \( \mu = E(X) \); in the mean absolute deviation from \( \mu \)

\[
S_\mu = E( |X - \mu| );
\]

and finally in the mean difference

\[
\Delta = E( |X - Y| ),
\]

where \( Y \) is a variate with the same distribution of \( X \) and independent from \( X \). In the following of the paper it is assumed that \( \sigma, S_\mu \) and \( \Delta \) are finite.

Let \( (X_1, \ldots, X_i, \ldots, X_n) \) be a simple random sample of size \( n \) \((n \geq 4)\) drawn from the variate \( X \) of the f.p. or from the continuous r.v. \( X \).

Various estimators of \( \sigma, S_\mu \) and \( \Delta \) can be defined. Successively it is necessary to analyze their dispersion from the measure that is intended to be estimated. In this paper we have decided to use unbiased estimators, even if this statement may be criticised (Frosini, 1986). On the other hand, we are not interested in the research of the best estimator in a wide class of estimators, but we intend to reflect on the choice of a sample dispersion measure.

3. General observations on the estimate of some variability measures

The statistical literature is rich in papers dedicated to the estimate of dispersion measures. A brief account of the history of sample measures of variability is
provided by H. A. David (1998). Many papers are referred to the normal distribution and only few works deal with the finite population.

At this point the estimate of the previous variability measures based on deviations from the mean $\mu$ can be taken into account. Moreover $\mu$ is assumed to be unknown.

It may be useful to introduce the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$  \hspace{1cm} (3.1)$$

where $\bar{X}$ is the sample mean. For this estimator it is well known that

$$E(S^2) = \sigma^2 \quad \text{for every } \sigma^2$$  \hspace{1cm} (3.2)$$

and

$$\text{Var}(S^2) = \frac{1}{n} \left( \mu_4 - \frac{n-3}{n-1} \sigma^4 \right) \quad \text{with } n>1$$  \hspace{1cm} (3.3)$$

where $\mu_4$ is the fourth central moment of $X$ (see, for example, Mood, Graybill, Boes, 1974, p. 229). Hence $S^2$ is an unbiased and mean square error consistent estimator of the variance $\sigma^2$. Yet it is opportune to point out that the variance of $S^2$ exists as finite only if the fourth central moment of the variate $X$ is finite. Naturally this condition is surely satisfied if we refer to a finite population, but if $X$ is a continuous r.v., it may have some infinite moments.

In order to estimate the standard deviation $\sigma$, the natural estimator

$$S_a = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2} = \sqrt{\frac{n-1}{n} S^2}$$  \hspace{1cm} (3.4)$$

may be used. It is possible to show that

$$E(S_a) \approx \sigma$$  \hspace{1cm} (3.5)$$

and

$$\text{Var}(S_a) \approx \frac{1}{4n} \left( \mu_4 - \frac{\sigma^4}{\sigma^2} \right)$$  \hspace{1cm} (3.6)$$

(see, for example, Cramér, 1946, p. 353). Hence $S_a$ is an unbiased estimator of $\sigma$ and its variance tends to zero when the sample size is increasing.

As for as the other two variability measures (2.2) and (2.3) are concerned, the estimation becomes a little more complicated because their analytic expressions present the absolute values. Moreover in the mean absolute deviation from $\mu$ the
unknown mean $\mu$ has to be estimated, just like in the estimation of the standard deviation $\sigma$. The natural estimator of $S_\mu$ is

$$K = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}| \quad (3.7)$$

The approximate expectation value and variance of $K$ are

$$E(K) \approx S_\mu \quad (3.8)$$

and

$$\text{Var}(K) \approx \frac{4}{n} \left[ \left( P(\mu) \right)^2 [\sigma^2 - T(\mu)] + [1 - P(\mu)]^2 T(\mu) - S_\mu^2 / 4 \right] \quad (3.9)$$

where

$$P(\mu) = \int_{-\infty}^{\mu} f(x) \, dx \quad \text{and} \quad T(\mu) = \int_{-\infty}^{\mu} (x - \mu)^2 \, f(x) \, dx \quad (3.10)$$

when $X$ is a continuous r.v., and

$$P(\mu) = \frac{\#\{a_i \leq \mu\}}{N} \quad \text{and} \quad T(\mu) = \frac{1}{N} \sum_{[a_i \leq \mu]} (a_i - \mu)^2 \quad (3.11)$$

when we are considering the variate $X$ in a f.p. This result is due to Gastwirth (1974) for a r.v. $X$, but it may be extended to the f.p. It is easy to show that for a symmetric r.v. $X$ being $P(\mu)=0.5$, the (3.9) is reduced to

$$\text{Var}(K) \approx \frac{1}{n} \left( \sigma^2 - S_\mu^2 \right) \quad (3.12)$$

(see, for example Stuart, Ord, 1994, p. 361). Hence $K$ is asymptotically an unbiased and mean square error consistent estimator of $S_\mu$.

Finally for the mean difference $\Delta$, the natural estimator is

$$\bar{\Delta} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| X_i - X_j \right| . \quad (3.13)$$

It is easy to show that
\[ E(\Delta) = \frac{n-1}{n} \Delta \quad \text{for every } \Delta, \quad (3.14) \]

while an unbiased estimator of \( \Delta \), is

\[ \hat{\Delta} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} |X_i - X_j| \quad (3.15) \]

The variance of \( \hat{\Delta} \) may be obtained with some tedious mathematical passages (see, Nair, 1936; Lomnicki, 1952; Michetti, Dall’Aglio, 1957), it is equal to

\[ \text{Var}(\hat{\Delta}) = \frac{4}{n(n-1)} \left[ \sigma^2 + (n-2)F \right] - \frac{2n-3}{2} \Delta^2 \quad (3.16) \]

for \( n \geq 4 \), where

\[ F = E(|X - Y| |X - Z|) \quad (3.17) \]

being \( Y \) and \( Z \) independent and identically distributed varieites like \( X \). It is obvious that the (3.16) tends to zero as \( n \) increases. The complex expression (3.16) and the difficulty in the evaluation of the functional \( F \) have determined a limiting use of the estimators (3.13) or (3.14). Therefore some interesting results regarding the calculation of \( F \) and the estimate of \( \text{Var}(\hat{\Delta}) \) are recently presented in Greselin, Polisicchio, Zenga (2004).

Thanks to all these characteristics, it seems reasonable to state that the mean difference is the dispersion measure, among those considered, with lower problems from the estimation point of view. In fact there exists an unbiased estimator of it, with variance depending on \( \sigma^2 \), \( F \) and the same \( \Delta \). Hence the variance of this estimator is finite even if, in the parent population, the moments of order higher than two are not finite. For the other two variability measures there are only approximated results. Besides, the variance of the estimator of the standard deviation is finite if, in the parent distribution \( X \), the fourth central moment is finite.

These remarks should be taken into account when economic or financial data are analysed. In fact such empirical distributions may be well fitted with some heavy tail distribution models. Therefore these fitted models may have an infinite fourth central moment.

In order to evaluate the behaviour of the estimator \( \hat{\Delta} \), in the following sections some sample variability measures are analysed in distribution models with all finite moments or with heavy tails.
4. Unbiased estimators of the standard deviation in the normal distribution

Let $X$ be a normal r.v. with probability density function

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$  \hspace{1cm} (4.1)

where $\mu$ and $\sigma$ are the unknown mean and standard deviation. The mean absolute deviation of $X$ is

$$S_\mu = \sigma \sqrt{\frac{2}{\pi}}$$  \hspace{1cm} (4.2)

and the mean difference is

$$\Delta = \frac{2\sigma}{\sqrt{\pi}}.$$  \hspace{1cm} (4.3)

In order to estimate the standard deviation $\sigma$, some unbiased estimators may be introduced using either the natural estimator $S_a$ or the relations (4.2) and (4.3). The specification of the density distribution, which the sample is drawn from, enables us to obtain exact results of the expectation value and of the variance of these estimators.

In relation to the natural estimator, it is convenient to rewrite the (3.4) as follows

$$S_a = \left[\frac{1}{n} \sigma^2 \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \right]^{1/2} = \frac{\sigma Y}{\sqrt{n}}$$  \hspace{1cm} (4.4)

where $Y = \left[\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \right]^{1/2}$.

Observing that the r.v. $X_i$ are independent with the same normal distribution, for $i=1,...,n$, we may deduce that $Y^2$ has a *chi-square* distribution with $n-1$ degree of freedom. Hence $Y$ has a *chi* distribution with $n-1$ degree of freedom. It is well known that

$$E(Y) = \sqrt{2} \frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)}$$
and

\[ \text{Var}(Y) = (n-1) - 2 \left[ \frac{\Gamma(n/2)}{\Gamma(1-1/n)} \right]^2, \]

(see, for example, Johnson, Kotz, Balakrishnan, 1994, p. 421). From the (4.4) it may be deduced

\[ E(S_a) = \frac{\sigma}{\sqrt{n}} E(Y) = \sigma \sqrt{\frac{2}{n}} \frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)} \]  

(4.5)

and

\[ \text{Var}(S_a) = \frac{\sigma^2}{n} \text{Var}(Y). \]  

(4.6)

The estimator \( S_a \) is biased for \( \sigma \), but it may be opportunely modified to obtain the following unbiased estimator

\[ S_{au} = \sqrt{\frac{n}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n/2)} S_a = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n/2)} \left[ \frac{1}{2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right]^{1/2}. \]  

(4.7)

The variance of this estimator is

\[ \text{Var}(S_{au}) = \frac{n}{2} \left[ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n/2)} \right]^2 \text{Var}(S_a) = \frac{\sigma^2}{2} \left[ \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n/2)} \right]^2 \text{Var}(Y) = \sigma^2 \left[ \frac{n-1}{2} \frac{\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma^2(n/2)} - 1 \right] \]  

(4.8)

(Holtzman, 1950; Cureton, 1968).

Another unbiased estimator of \( \sigma \) may be obtained from the relation (4.2) and from the estimation of the measure \( S_{\mu} \). In the normal distribution, it may be shown that the estimator \( K \) of \( S_{\mu} \), given by (3.7), has expectation value and variance respectively equal to

\[ E(K) = \sigma \sqrt{\frac{2}{\pi}} \left( \frac{1 - \frac{1}{n}}{n} \right) \]  

(4.9)

and
\[
\text{Var}(K) = \frac{2\sigma^2}{n\pi} \left(1 - \frac{1}{n} \right) \left[ \frac{\pi}{2} + \sqrt{n(n-2)} - n + \arcsen \left( \frac{1}{n-1} \right) \right]
\]  
(4.10)

(see, for example, Stuart, Ord, 1994, p. 362). Hence from the estimator \( K \) it may be obtained the following unbiased estimator of \( \sigma \)

\[
S_{Ku} = \sqrt{\frac{\pi}{2} \frac{n}{n-1} K} = \sqrt{\frac{\pi}{2} \frac{1}{n(n-1)} \sum_{i=1}^{n} |X_i - \bar{X}|}.
\]  
(4.11)

Using the (4.10) and (4.11) it may be shown that the variance of \( S_{Ku} \) is

\[
\text{Var}(S_{Ku}) = \frac{\pi}{2} \frac{n}{n-1} \text{Var}(K) = \frac{\sigma^2}{n} \left[ \frac{\pi}{2} + \sqrt{n(n-2)} - n + \arcsen \left( \frac{1}{n-1} \right) \right].
\]  
(4.12)

Finally another estimator of \( \sigma \) may be deduced analogously, using the relation (4.3) and estimating \( \Delta \). In the normal distribution, the estimator \( \hat{\Delta} \) given by (3.15) has the following expectation value and variance

\[
E(\hat{\Delta}) = \Delta = \frac{2\sigma}{\sqrt{\pi}}
\]  
(4.13)

\[
\text{Var}(\hat{\Delta}) = \frac{4\sigma^2}{n(n-1)} \left[ \frac{n+1}{3} + \frac{2}{\pi} \left( \sqrt{3} - 2 \right) \left( n + \sqrt{3} \right) \right]
\]  
(4.14)

(see, for example, Stuart, Ord, 1962, p. 363). Hence an unbiased estimator of \( \sigma \) is

\[
S_{\hat{\Delta}} = \sqrt{\frac{\pi}{2} \frac{\hat{\Delta}}{n(n-1)}} = \sqrt{\frac{\pi}{2} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|}
\]  
(4.15)

(see David, 1968). In fact

\[
E(S_{\hat{\Delta}}) = \sqrt{\frac{\pi}{2}} E(\hat{\Delta}) = \sigma \quad \text{for every} \quad \sigma
\]  
(4.16)

and

\[
\text{Var}(S_{\hat{\Delta}}) = \frac{\pi}{4} \text{Var}(\hat{\Delta}) = \frac{\sigma^2}{n(n-1)} \left[ \frac{n+1}{3} \frac{\pi}{2} + 2 \left( \sqrt{3} - 2 \right) \left( n + \sqrt{3} \right) \right].
\]  
(4.17)
In order to compare the previous three unbiased estimator of $\sigma$, the corresponding variance for some values of the sample size $n$ has been evaluated. The numerical results are presented in the following tables.

*Table 1: Values of the variance of some unbiased estimators of $\sigma$ as $n$ increases*

<table>
<thead>
<tr>
<th>Sample size $n$</th>
<th>$\text{Var}(S_{au})/\sigma^2$</th>
<th>$\text{Var}(S_{Ku})/\sigma^2$</th>
<th>$\text{Var}(S_{Au})/\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.570796</td>
<td>0.570796</td>
<td>0.570796</td>
</tr>
<tr>
<td>3</td>
<td>0.273240</td>
<td>0.275482</td>
<td>0.275482</td>
</tr>
<tr>
<td>4</td>
<td>0.177500</td>
<td>0.184765</td>
<td>0.180349</td>
</tr>
<tr>
<td>5</td>
<td>0.131768</td>
<td>0.139292</td>
<td>0.133775</td>
</tr>
<tr>
<td>6</td>
<td>0.104466</td>
<td>0.111856</td>
<td>0.106226</td>
</tr>
<tr>
<td>7</td>
<td>0.086498</td>
<td>0.093475</td>
<td>0.088050</td>
</tr>
<tr>
<td>8</td>
<td>0.073787</td>
<td>0.080293</td>
<td>0.075167</td>
</tr>
<tr>
<td>9</td>
<td>0.064324</td>
<td>0.070375</td>
<td>0.065566</td>
</tr>
<tr>
<td>10</td>
<td>0.057009</td>
<td>0.062641</td>
<td>0.058133</td>
</tr>
<tr>
<td>30</td>
<td>0.017387</td>
<td>0.019601</td>
<td>0.017768</td>
</tr>
<tr>
<td>50</td>
<td>0.010256</td>
<td>0.011620</td>
<td>0.010483</td>
</tr>
<tr>
<td>100</td>
<td>0.005056</td>
<td>0.005758</td>
<td>0.005177</td>
</tr>
</tbody>
</table>

*Table 2: Ratio between the variances of unbiased estimators of $\sigma$ as $n$ increases*

<table>
<thead>
<tr>
<th>Sample size $n$</th>
<th>$\text{Var}(S_{Ku})/\text{Var}(S_{au})$</th>
<th>$\text{Var}(S_{Au})/\text{Var}(S_{au})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1.008205</td>
<td>1.008205</td>
</tr>
<tr>
<td>4</td>
<td>1.040930</td>
<td>1.016051</td>
</tr>
<tr>
<td>5</td>
<td>1.057100</td>
<td>1.015231</td>
</tr>
<tr>
<td>6</td>
<td>1.070741</td>
<td>1.016848</td>
</tr>
<tr>
<td>7</td>
<td>1.080661</td>
<td>1.017943</td>
</tr>
<tr>
<td>8</td>
<td>1.088173</td>
<td>1.018702</td>
</tr>
<tr>
<td>9</td>
<td>1.094071</td>
<td>1.019309</td>
</tr>
<tr>
<td>10</td>
<td>1.098791</td>
<td>1.019716</td>
</tr>
<tr>
<td>30</td>
<td>1.127337</td>
<td>1.021913</td>
</tr>
<tr>
<td>50</td>
<td>1.132995</td>
<td>1.022133</td>
</tr>
<tr>
<td>100</td>
<td>1.137270</td>
<td>1.022516</td>
</tr>
</tbody>
</table>

Obviously for all these estimators the variance tends to zero as the sample size increases. The natural and unbiased estimator $S_{au}$ has the lower variance. This is an expected result, in fact in the normal distribution the conjoint sufficient statistics are $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$. These statistics only emerge in the estimator
Moreover it may be pointed out that the variances of \( S_{\kappa} \) and \( S_{\omega} \) are very close to the variance of \( S_{\omega} \) and that the variance of \( S_{\omega} \) is the nearer to that of \( S_{\omega} \).

Hence the estimator deduced by the sample mean difference has a good behaviour. This aspect turns more clear when we intend to estimate the mean difference \( \Delta \), given by (2.3) and (4.3), using the estimator \( \hat{\Delta} \) and the following estimator based on the sufficient statistics

\[
\hat{\Delta}_1 = \frac{2}{\sqrt{\pi}} S_a = \frac{2}{\sqrt{\pi}} \left[ \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right]^{1/2}.
\]  

(4.18)

It is easy to show that \( \hat{\Delta}_1 \) is biased for \( \Delta \), in fact

\[
E(\hat{\Delta}_1) = 2 \sqrt{\frac{2}{n\pi}} \frac{\Gamma(n/2)}{\Gamma(n-1)} \sigma
\]

(4.19)

Hence for a correct comparison with the unbiased estimator \( \hat{\Lambda} \), we need to evaluate the mean square error of \( \hat{\Delta}_1 \). It may be shown that

\[
\text{Var}(\hat{\Delta}_1) = \frac{4 \sigma^2}{\pi n} \left[ (n-1) - \frac{2 \Gamma^2(n/2)}{\Gamma^2 \left( \frac{n-1}{2} \right)} \right]
\]

(4.20)

and

\[
\text{MSE}(\hat{\Delta}_1) = \frac{4 \sigma^2}{n} \left[ 2 - \frac{1}{n} - 2 \sqrt{\frac{2}{\pi}} \frac{\Gamma(n/2)}{\Gamma \left( \frac{n-1}{2} \right)} \right].
\]

(4.21)

The table 3 presents the ratio between the variance of \( \hat{\Lambda} \) and the mean square error of \( \hat{\Delta}_1 \) as the sample size is increasing. These results highlight a very good behaviour of the estimator \( \hat{\Delta} \) in relation to the estimator based on the sufficient statistics. We may observe that the variance of \( \hat{\Delta} \) decreases more quickly than the mean square error of \( \hat{\Delta}_1 \) as the value of \( n \) is increasing. All these aspects should lead to some further meditation on the use of the sufficient statistics.
Table 3: Ratio between the variance of $\hat{\Delta}$ and the mean square error of $\hat{\Delta}_1$, as $n$ increases

<table>
<thead>
<tr>
<th>Sample size $n$</th>
<th>$Var(\hat{\Delta})/MSE(\hat{\Delta}_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.5358</td>
</tr>
<tr>
<td>3</td>
<td>1.2552</td>
</tr>
<tr>
<td>4</td>
<td>1.1724</td>
</tr>
<tr>
<td>5</td>
<td>1.1289</td>
</tr>
<tr>
<td>6</td>
<td>1.1056</td>
</tr>
<tr>
<td>7</td>
<td>1.0906</td>
</tr>
<tr>
<td>8</td>
<td>1.0801</td>
</tr>
<tr>
<td>9</td>
<td>1.0724</td>
</tr>
<tr>
<td>10</td>
<td>1.0665</td>
</tr>
<tr>
<td>30</td>
<td>1.0356</td>
</tr>
<tr>
<td>50</td>
<td>1.0302</td>
</tr>
<tr>
<td>100</td>
<td>1.0263</td>
</tr>
</tbody>
</table>

5. Estimation of some variability measures in Pareto distribution

Let $X$ be a Pareto distribution with probability density function

$$f(x) = \theta \theta x^{-\theta-1} \quad \text{for } x > x_0 > 0$$

(5.1)

and where $\theta > 0$.

In this model the considered dispersion measures are equal to

$$\sigma = \frac{x_0}{\theta - 1} \sqrt{\frac{\theta}{\theta - 2}}$$

(5.2)

$$S_\mu = \frac{2x_0(\theta - 1)^{-2}}{\theta^{\theta-1}}$$

(5.3)

$$\Delta = \frac{2x_0\theta}{(2\theta - 1)(\theta - 1)}$$

(5.4)

(see, Johnson, Kotz, Balekrishnan, 1994, p.577-578). In this section, applying the approximate results of the section 3, we intend to compare the natural estimators of the standard deviation, of the mean absolute deviation and of the mean difference. Hence, we have a different and more wide purpose than that one of the previous section. To evaluate the approximate variance of the sample standard
deviation we need the value of the fourth central moment. In Pareto distribution it is equal to

$$\mu_4 = \frac{3x_0^4 \theta (3 \theta^2 + 2)}{(\theta - 1)^4 (\theta - 2)(\theta - 3)(\theta - 4)} \quad \text{for } \theta > 4 \quad (5.5)$$

Hence the approximate expectation value and variance of the estimator $S_n$ are

$$E(S_n) \approx \sigma = \frac{x_0}{\theta - 1} \sqrt{\frac{\theta}{\theta - 2}} \quad \text{for } \theta > 2 \quad (5.6)$$

$$Var(S_n) \approx \frac{1}{4n} \left( \frac{\mu_4 - \sigma^4}{\sigma^2} \right) = \frac{x_0^2 [8 \theta^2 (\theta - 1) - 12 (\theta + 1)]}{4n(\theta - 1)^7 (\theta - 2)(\theta - 3)(\theta - 4)} \quad (5.7)$$

for $\theta > 4$.

The natural estimator $K$ of the variability measure $S_n$ is approximately unbiased. The variance of $K$ depends on the value of the two integrals given by (3.10). These integrals are equal to

$$P(\mu) = 1 - \left( \frac{x_0}{\mu} \right)^\theta \quad \text{and}$$

$$T(\mu) = \frac{x_0^2 \theta}{(\theta - 1)^3 (\theta - 2)} \left[ 1 - 2 \left( \frac{\theta - 1}{\theta} \right)^{\theta-1} \right] = \sigma^2 \left[ 1 - 2 \left( \frac{\theta - 1}{\theta} \right)^{\theta-1} \right] \quad \text{for } \theta > 2.$$

Hence by (3.9), the variance of $K$ is

$$Var(K) \approx \frac{4}{n} \frac{x_0^2 \theta}{(\theta - 1)^3 (\theta - 2)} \left( \frac{\theta - 1}{\theta} \right)^{\theta-1} \left[ 2 - 4 \left( \frac{\theta - 1}{\theta} \right)^\theta + \left( \frac{\theta - 1}{\theta} \right)^{\theta+1} \right] \frac{S_\mu^2}{n}$$

$$(5.8)$$

(Gastwirth, 1974).

Finally, the exact value of the variance of the natural and unbiased estimator $\hat{\lambda}$, given by (3.16), is known. Therefore, in order to realize a more adequate comparison with the variances of the other two dispersion estimators, we use the following approximate value
\[ \text{Var}(\hat{\Delta}) \approx \frac{4}{n} \left( F - \Delta^2 \right) \]  \hspace{1cm} (5.9)

In Pareto distribution, the functional \( F \) is

\[ F = \frac{x_0^2 \theta}{(\theta - 1)^2(\theta - 2)} + \frac{2x_0^2 \theta^2}{(\theta - 1)^2(2\theta - 1)(3\theta - 2)} = \frac{x_0^2 \theta \left( 8\theta^2 - 11\theta + 2 \right)}{(\theta - 1)^2(\theta - 2)(2\theta - 1)(3\theta - 2)} \]  \hspace{1cm} (5.10)

for \( \theta > 2 \) (see Michetti, Dall’Aglio, 1957, p. 244). Hence the approximate value of the variance of \( \hat{\Delta} \) is

\[ \text{Var}(\hat{\Delta}) \approx \frac{4}{n} \frac{x_0^2 \theta \left( \theta + 2 \right) \left( 4\theta^2 - 2\theta - 1 \right)}{(\theta - 1)^2(\theta - 2)(2\theta - 1)^2(3\theta - 2)} \]  \hspace{1cm} (5.11)

In order to compare the previous sample dispersion measures, the following table gives the values of the variances of those estimators for different values of \( \theta \) and for fixed \( x_0 = 1 \). Naturally, we have considered some values of \( \theta \) such that the fourth central moment is finite.

<table>
<thead>
<tr>
<th>Value of ( \theta )</th>
<th>( \text{Var}(S_n) )</th>
<th>( \text{Var}(K) )</th>
<th>( \text{Var}(\hat{\Delta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5</td>
<td>5.453/n</td>
<td>0.178/n</td>
<td>0.369/n</td>
</tr>
<tr>
<td>5.0</td>
<td>1.896/n</td>
<td>0.120/n</td>
<td>0.247/n</td>
</tr>
<tr>
<td>5.5</td>
<td>0.909/n</td>
<td>0.086/n</td>
<td>0.175/n</td>
</tr>
<tr>
<td>6.0</td>
<td>0.565/n</td>
<td>0.064/n</td>
<td>0.130/n</td>
</tr>
<tr>
<td>7.0</td>
<td>0.261/n</td>
<td>0.040/n</td>
<td>0.079/n</td>
</tr>
<tr>
<td>10.0</td>
<td>0.065/n</td>
<td>0.014/n</td>
<td>0.028/n</td>
</tr>
</tbody>
</table>

Among the considered estimators, the sample mean absolute deviation has the lower approximate variance. Moreover the sample mean difference has a variance very close to it.

### 6. Conclusions

A heavy tail distribution model fits many economic or financial empirical distributions. The problems connected with these models are represented by the possible presence of infinite moments starting from a certain order. The reality we are able to observe is finite and it may give only a finite estimate of any moments. It is sufficient to think Pareto distribution according to which the variance is infinite when \( \theta=2 \). Some economic phenomena may well be described by this
model and it may be possible that the observed data give an estimate of the parameter $\theta$ for which the model does not have a finite variance. In such a situation it seems more reasonable to apply a dispersion measure different from the standard deviation. Hence, keeping in mind these possibilities, this paper intends to highlight some remarks on the estimation of some variability measures. We focus on the estimation of the standard deviation, of the mean absolute difference and of the mean difference. In the non-parametric case, there are exact results in terms of expectation value and variance only for the sample mean difference. Moreover, the sample standard deviation has a finite variance only if in the population, which the sample is drawn from, the fourth central moment is finite. The variance of the sample mean absolute deviation is finite when in the parent population the variance and the mean absolute deviation are finite. While the variance of the sample mean difference is finite when the variance, the mean difference and the functional $F$, given by (3.17), are finite in the population. These aspects lead to the choice of the variability measure to synthesize the dispersion of the observed data when the sample is drawn from a r.v. that may have an infinite fourth central moment.

At the purpose to look into the behaviour of the sample mean difference we consider the normal distribution. We compare the variance of three unbiased estimators of $\sigma$. The estimator based on the sample standard deviation is the unique, among the considered ones, to be based on the sufficient statistics and it has the lower variance. Nevertheless the other two estimators have the variance close to the one of the previous estimator and the variance of the sample mean difference is the nearest. Finally we consider Pareto distribution. We compare the approximate variance of the natural estimators of different variability measures. In this case the estimator with a lower variance is the sample mean absolute deviation, but the sample mean difference presents a variance very close to it for different values of the parameter $\theta$. Moreover, for this model, it is possible to use the exact expression of the variance of the sample mean difference.

To conclude, Gini’s mean difference can be re-evaluated according to the existence of exact results in terms of the expectation value and the variance of the sample mean difference, to the good behaviour of this estimator in comparison with the sufficient statistics in the normal distribution and finally, to the consideration that many economic and financial empirical distributions may be well described by a heavy tail distribution model even without the finite fourth central moment.

References


