Statistica & Applicazioni
Vol. V, n. 1, 2007

MINIMUM SAMPLE SIZES FOR CONFIDENCE INTERVALS FOR GINI'S MEAN DIFFERENCE: A NEW APPROACH FOR THEIR DETERMINATION

Francesca Greselin [1]

Walter Maffenini [1]

SUMMARY

The sample mean difference $\hat{\Delta}$ *is an unbiased estimator of Gini's mean difference* Δ*. It is well known that* $\hat{\Delta}$ *is asymptotically normally distributed (Hoeffding, 1948). In order to obtain confidence intervals for* Δ*,* $\hat{\Delta}$ *must be standardized and hence its variance* $Var(\hat{\Delta})$ *must be estimated. In this paper we study the effective coverage of the confidence intervals for* Δ*, when using a specific unbiased estimator* $\hat{Var}(\hat{\Delta})$ *for the variance of* $\hat{\Delta}$*, in a non-parametric framework. The empirical determination of the minimum sample size required to reach a good approximation of the nominal coverage is analyzed through a new approach. The reported results show that this threshold is critically related to the asymmetry and the tail heaviness in the underlying distribution.*

Keywords: *Gini's Mean Difference, asymptotic confidence interval, U-statistic, minimum sample size, heavy tails.*

1. INTRODUCTION

The mean difference, defined as the average of the absolute differences of all pairs of values in a population, is a measure of dispersion suggested in 1912 by C. Gini. Among variability measures, Gini's mean difference, denoted in the following by Δ, is not so widely employed as the standard deviation σ is. The broad diffusion of the standard deviation, in effect, is mainly due to some well-known interesting properties of its square, the variance σ^2, from both descriptive and inferential points of view.

The remarkable work due to Hoeffding (1948) opened a new perspective for the

---

[1] Quantitative Methods for Economics and Business Sciences, University of Milano-Bicocca, P.za dell'Ateneo Nuovo 1, 20126 Milan, Italy
(e-mail: francesca.greselin@unimib.it; walter.maffenini@unimib.it)

Although it is the result of a close collaboration, this paper was specifically elaborated as follows: Section 1, 2 and 6 are due to W. Maffenini, while the other Sections are due to F. Greselin.

mean difference Δ: the author derived the normal asymptotic distribution of its natural estimator $\hat{\Delta}$, under the sole condition of the existence of the second moment of the underlying distribution *F*. The variance of $\hat{\Delta}$ was first obtained by Nair (1936) by means of the order statistics; successively Lomnicki (1952) derived it in a more straightforward way. The variance of $\hat{\Delta}$ can be expressed as a function of the sample size *n* and of some regular functionals of *F*. Moreover an unbiased estimator $\hat{Var}(\hat{\Delta})$ of $Var(\hat{\Delta})$ is known (Hoeffding (1948), Cowell 1989 and, through a different approach, Zenga, Polisicchio and Greselin, 2004).

The literature about confidence intervals for concentration and inequality measures[1] is wide (see Mills and Zandvakili (1997), Sendler (1979), Xu (2000), among many others). The use of bootstrap resampling techniques - which estimate the standard error and the sample distribution of inequality estimators directly through the empirical distribution function of the sampled data - is often compared to some asymptotic methods (Biewen (2002)), where the variance of the indexes is estimated by the delta method, obtaining a strongly consistent estimator (see, among many others, Sendler (1979), Cowell and Flachaire (2002), Zitikis (2002)). To our best knowledge, in the wider context of variability indexes, Gini's mean difference has not received an analogous attention.

The specific contribution of the present work is hence to provide, in a non parametric framework and through an original methodology, the minimum samples sizes such that the coverage of the asymptotic confidence intervals for Δ is close to the nominal value. The paper is organized as follows: in section 2 some definitions and notations are introduced; in section 3 and 4 the methodology is presented in detail; section 4 deals also with the results of some simulation and their interpretation; section 5 shows some further analyses and finally section 6 concludes and points out some possible developments.

2. NOTATIONS AND DEFINITIONS

Let X be a continuous random variable (c.r.v.) with probability density function *f(x)*, for $x \in \mathbb{R}$. Gini's mean difference Δ is defined by:

$$\Delta = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x-y| f(x) f(y)\, dx\, dy \tag{1}$$

Let *D(x)* be the mean deviation of the c.r.v. *X* about *x*:

[1] As it is well known, concentration is a specific branch of variability, concerning transferable variables.

$$D(x) = \mathrm{E}\left[\left|x - X\right|\right] = \int_{-\infty}^{+\infty} \left|x - y\right| f(y)\, dy \tag{2}$$

It is easy to show that:

$$\Delta = \int_{-\infty}^{+\infty} D(x) f(x) dx \tag{3}$$

Moreover, the following functional, as a population characteristic:

$$\mathfrak{F} = \int D^2(x) f(x) dx. \tag{4}$$

will be useful for the expression of $Var(\hat{\Delta})$.

Let μ and σ^2 denote, respectively, the expectation and the variance of the c.r.v. X. In this paper, it is assumed that σ^2 is finite.

Denote by $(X_1,..,X_i,..,X_n)$ a random sample of size n ($n > 3$) from the c.r.v. X. Let $\hat{\Delta}$ denote the sample mean difference without repetition, unbiased estimator of Δ:

$$\hat{\Delta} = \frac{1}{n(n-1)} \sum_{\substack{i=1, \\ i \neq j}}^{n} \sum_{j=1}^{n} \left|X_i - X_j\right|. \tag{5}$$

The variance of $\hat{\Delta}$, when sampling from a c.r.v., is given by:

$$Var\left(\hat{\Delta}\right) = \frac{4}{n(n-1)} \left[\sigma^2 + (n-2)\mathfrak{F} - \frac{(2n-3)}{2}\Delta^2\right] \tag{7}$$

as derived in Nair (1936) and in Lomnicki (1952). $\hat{\Delta}$ is hence a mean squared error consistent estimator of Δ.

In (7) the variance of $\hat{\Delta}$ is expressed as a function of the sample size n, the variance σ^2, the functional $\mathfrak{F}$ and the square of Gini's mean difference. An unbiased estimator $\hat{Var}\left(\hat{\Delta}\right)$ for $Var\left(\hat{\Delta}\right)$ was proposed by Michetti and Dall'Aglio (1957) (successively Cowell (1989) and, recently, Zenga *et al.* (2004) derived it through a different methodology):

$$\hat{V}ar\left(\hat{\Delta}\right)=\frac{4}{(n-2)(n-3)}\left[S^2+(n-2)\hat{\mathfrak{F}}-\frac{(2n-3)}{2}\hat{\Delta}^2\right] \tag{8}$$

where S^2 and $\hat{\mathfrak{F}}$ are unbiased estimators of σ^2 and $\mathfrak{F}$, respectively given by:

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \tag{9}$$

and:

$$\hat{\mathfrak{F}}=\frac{1}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sum_{\substack{l=1\\ l\neq i,j}}^{n}\left|X_i-X_j\right|\left|X_i-X_l\right|. \tag{10}$$

3. CONFIDENCE INTERVALS FOR DELTA

Confidence intervals for Δ can be obtained by the sample statistic:

$$T=\frac{\left(\hat{\Delta}-\Delta\right)}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\Delta}\right)}}=\frac{\sqrt{n-2}\left(\hat{\Delta}-\Delta\right)}{2\sqrt{\frac{S^2}{(n-3)}+\frac{(n-2)}{(n-3)}\hat{\mathfrak{F}}-\frac{(n-1.5)}{(n-3)}\hat{\Delta}^2}}. \tag{11}$$

Like the variance σ^2, Gini's mean difference Δ and $\mathfrak{F}$ are regular functionals, hence the corresponding U-statistics, respectively given by S^2, $\hat{\Delta}$ and $\hat{\mathfrak{F}}$, converge almost surely to their mean values, by the U-statistics convergence theorem, generalizing the strong law of large numbers, in Lee (1990). This fact assures the strong convergence of $\hat{V}ar\left(\hat{\Delta}\right)$ to $Var(\hat{\Delta})$, so that, for Slustky's theorem, T has a standard normal asymptotic distribution.

4. A NEW METHODOLOGY FOR CONFIDENCE INTERVALS FOR DELTA

We are interested in distribution free confidence intervals for Δ, as an unbiased estimator $\hat{\Delta}$ for Δ, with normal asymptotic distribution, is available. Under these circumstances, the pivotal quantity method provides confidence intervals having – asymptotically – the chosen confidence level.

In applications, as a real sample has a finite size n, the issue of a possible difference between the nominal asymptotical confidence level (1-α) and the effective coverage p_n might be raised. Depending on how far the sample distribution of the statistic $\hat{\Delta}_n$ is from its asymptotic approximation, the confidence interval based on the normal distribution may give a coverage far from (often lower than) the nominal confidence level.

Hence, a methodology to evaluate the minimum sample size allowing a good approximation of the nominal confidence level seems to be very useful in applications. In this paper, simulations from different underlying distributions will be generated to explore whether the above discussed convergence is uniform or not.

The methodology can be described as follows:

- a sample size *n* is chosen;
- a series of *N* samples of size *n* is generated, from a given underlying distribution;
- for each sample, the (1-α)-confidence interval $\hat{\Delta} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\Delta})$ is computed, denoting by $\hat{\sigma}(\hat{\Delta})$ the square root of $\hat{Var}(\hat{\Delta})$. A point estimation $\hat{p}_n$ of the effective coverage p_n, defined as the ratio of the number of intervals containing Δ over *N,* is also evaluated.

In general, the effective unknown coverage p_n, estimated by $\hat{p}_n$, depends on the chosen sample size *n,* the (1-α)-confidence level and the underlying distribution. In any case, $\lim_{n\to\infty} p_n = (1\text{-}\alpha)$.

Relying on the assumption that (as in most applications) some coverage error with respect to the nominal confidence level 1-α is tolerable, hence, according to the needs of the researcher, an interval of values can be defined:

$$(1-\alpha-\gamma_1;\quad 1-\alpha+\gamma_2) \qquad \text{for } \gamma_1,\gamma_2 \in \mathbb{R}^+ \tag{12}$$

so that an effective coverage that falls in it can be considered as a satisfactory approximation of the nominal coverage 1-α.

The estimator $\hat{p}_n$, being defined as a sample frequency, is such that:

$$E(\hat{p}_n) = p_n \qquad \text{and} \qquad Var(\hat{p}_n) = p_n(1-p_n)/N\,. \tag{13}$$

When large values for *N* are chosen (for example *N*=5000), $\hat{p}_n$ can then be approximated by a normal distribution.

The issue is now to distinguish good values $\hat{p}_n$ from unacceptable ones. As the estimates $\hat{p}_n$ arise from simulations, they vary among different runs. Therefore a testing procedure is needed, in order to distinguish fluctuations of $\hat{p}_n$ due to sampling variability from substantial differences in estimation.

Our aim is to identify the sample size n assuring that the effective coverage is higher than the minimum acceptable value, so we want to test:

$H_0: \ p_n \geq 1-\alpha-\gamma_1$

against the alternative:

$H_1: \ p_n < 1-\alpha-\gamma_1$.

The choice of H_o and H_1 relates to the need of preventing undercoverage, due to some skewness observed in the sample mean difference distribution (Greselin and Zenga, 2006).

A critical region of size δ for the chosen set of hypotheses is given by:

$$\left\{ \hat{p}_n < c \,|\, c = 1-\alpha-\gamma_1 - \frac{1}{2N} - z_{1-\delta}\sqrt{(1-\alpha-\gamma_1)(\alpha+\gamma_1)/N} \right\} \qquad (14)$$

where the threshold c can be also obtained by the continuity correction (Newcombe, 1998).

From now on, after setting $\gamma_1 = k\alpha$ and choosing $k = 0.10$ and $\delta = 0.05$, the threshold c takes the values highlighted in Table 1, for some fixed values of the coefficient 1-α.

TABLE 1. *Critical thresholds c for the test of size δ to verify H_0 (where $\gamma_1 = k\alpha$)*

		1-α				
k	δ	0.9	0.925	0.95	0.975	0.99
0.05	0.10	0.8893	0.9163	0.9434	0.9708	0.9876
0.10	0.10	0.8842	0.9124	0.9408	0.9694	0.9870
0.15	0.10	0.8791	0.9086	0.9382	0.9681	0.9865
0.20	0.10	0.8740	0.9047	0.9356	0.9668	0.9859
0.05	0.05	0.8878	0.9149	0.9422	0.9699	0.9870
0.10	*0.05*	**0.8826**	**0.9110**	**0.9396**	**0.9686**	**0.9865**
0.15	0.05	0.8775	0.9071	0.9370	0.9673	0.9859
0.20	0.05	0.8723	0.9032	0.9344	0.9659	0.9854
0.05	0.01	0.8848	0.9123	0.9401	0.9684	0.9860
0.10	0.01	0.8796	0.9083	0.9374	0.9670	0.9855
0.15	0.01	0.8744	0.9044	0.9347	0.9657	0.9849
0.20	0.01	0.8692	0.9005	0.9321	0.9643	0.9844

The reported simulations were obtained by a software program written in C++, using pseudorandom numbers generated by the IMSL statistical library. For each continuous population, N = 5000 random samples, of different sample sizes from

n = 50 up to 2500, were generated. Simulations given by 5000 runs have been performed, to ensure a negligible Monte Carlo error. With reference to the s.q.m of the estimator $\hat{p}_n$, the approximation error is of order $N^{-1/2}$, so that, throughout all our simulations, the worst mean MC relative error resulted in estimating, say $\hat{\mathfrak{F}}$, is 1,7%, attained for the Pareto distribution with the heaviest tail.

The results of the obtained simulations are summarized in a separate table, for each chosen continuous population. Each row of a table shows the Monte Carlo coverages of confidence intervals for Δ and for μ, for a fixed sample size n and some selected confidence levels.

Our study deals with six underlying distributions: the normal, the uniform, the exponential, the lognormal, the Pareto, and the Dagum distribution. This choice is justified by their wide employment as models for real data and by the aim of considering different conditions of asymmetry and tail heaviness in the underlying distributions.

For the normal, uniform and exponential distribution, only one set of parameters is considered, as their shape indicators (i.e. the third and fourth standardized moments) do not depend on them. It is easy to show that any choice for their values would then yield the same results on the true (and estimated) coverages.

Along with the exponential, a second asymmetric case is represented by the lognormal distribution, whose density function is here recalled:

$$f(x) = \frac{1}{\delta\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}\left(\frac{\log x-\gamma}{\delta}\right)^2} \qquad \text{for } \delta > 0.$$

Three sets of parameters are examined in this case:

- the first set is obtained by considering a real economic distribution, i.e. the Italian income and expenses data (Banca d'Italia, 2002), giving γ = 10 and δ = 0.5.
- the second set, γ = 10 and δ = 0.35829, is chosen so that Gini's concentration ratio R is close to 0.2. This choice, along with the first and the third set (respectively giving R~0.3 and R~0.4) provides hence a wide view of real distributions.
- the third set is given by γ = 10 and δ = 0.74161.

The results obtained for the exponential and the lognormal distribution are then related to an increasing asymmetry, in some distributions owning all moments.

Many parametric distributions are heavy tailed, by which is meant that the upper tail decays like a power function. The Dagum and the Pareto distributions provide a framework in which the existence of moments, depending on the choice of the related parameters, can influence the coverage of confidence intervals for Δ.

The Dagum distribution has density function:

$$f(x) = \lambda\beta\theta\frac{x^{-(\theta+1)}}{\left[1+\lambda x^{-\theta}\right]^{(\beta+1)}} \qquad \text{for } x > 0, \quad \text{where} \quad \lambda, \beta, \theta > 0.$$

In a first case, the parameters are chosen by fixing a coefficient of variation close to 0.6 - as in real economic distributions - so that λ = 1, β = 1 and θ = 3.6. The second case is given by $\lambda = 1$, $\beta = 1$ and $\theta = 5$, to obtain a coefficient of variation $\sigma/\mu \sim 0.4$ (also in the range of values observed in real economic distributions) and a heavier right tail.

For the Pareto distribution, with density function:

$$f(x) = \theta x_0^{\theta} \; x^{-(\theta+1)} \quad \text{for } x \geq x_0 \text{, where } x_0 > 0 \text{ and } \theta > 0,$$

the parameters were set as follows: $x_0 = 1$ and $\theta = 3$, in a first case, and $x_0 = 1$ and $\theta = 4$ in a second set, respectively yelding $R = 0.2$ and $R = 0.33$, give the chance to evaluate how the existence of moments can influence the coverage.

4.1. Symmetric distributions: normal and uniform

Table 2 shows the estimated coverage of confidence intervals for Δ and for μ, based on $N = 5000$ random samples drawn from the normal distribution[2] $\mathcal{N}(\mu = 0.0; \sigma = 5.0)$.

TABLE 2. *Simulated percent coverage for different sample sizes n and confidence levels (1−α) (normal distribution)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for* μ				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	**0.8844**	0.9078	0.9346	0.9618	0.9756	**0.8962**	**0.9156**	**0.9488**	**0.9692**	0.9864
100	**0.8926**	0.9106	**0.9400**	0.9626	0.9862	**0.8918**	**0.9188**	**0.9410**	**0.9754**	**0.9894**
150	**0.8914**	**0.9232**	**0.9428**	**0.9706**	0.9844	**0.8966**	**0.9194**	**0.9534**	**0.9758**	**0.9910**
250	**0.8958**	**0.9220**	**0.9454**	**0.9704**	**0.9868**	**0.8918**	**0.9222**	**0.9476**	**0.9800**	**0.9896**
500	**0.8982**	**0.9220**	**0.9442**	**0.9736**	**0.9876**	**0.8998**	**0.9234**	**0.9522**	**0.9722**	**0.9890**
750	**0.8914**	**0.9232**	**0.9472**	**0.9736**	**0.9912**	**0.9078**	**0.9250**	**0.9530**	**0.9738**	**0.9902**
1000	**0.8960**	**0.9208**	**0.9514**	**0.9746**	**0.9900**	**0.9020**	**0.9234**	**0.9464**	**0.9742**	**0.9882**
1250	**0.9050**	**0.9274**	**0.9482**	**0.9756**	**0.9910**	**0.9060**	**0.9234**	**0.9482**	**0.9712**	**0.9894**
1500	**0.9008**	**0.9210**	**0.9470**	**0.9760**	**0.9866**	**0.9006**	**0.9216**	**0.9514**	**0.9792**	**0.9918**
1750	**0.8972**	**0.9264**	**0.9468**	**0.9744**	**0.9888**	**0.8988**	**0.9206**	**0.9468**	**0.9752**	**0.9924**
2000	**0.9060**	**0.9260**	**0.9486**	**0.9786**	**0.9904**	**0.8960**	**0.9266**	**0.9478**	**0.9768**	**0.9904**
2250	**0.9042**	**0.9226**	**0.9490**	**0.9704**	**0.9890**	**0.8954**	**0.9250**	**0.9498**	**0.9766**	**0.9892**
2500	**0.8988**	**0.9276**	**0.9484**	**0.9736**	**0.9908**	**0.9032**	**0.9160**	**0.9504**	**0.9786**	**0.9904**

In the tables presented in this Section, every simulated coverage is highlighted whenever it falls in the acceptance region for H_0, as described in Section 3.

Almost all the entries in Table 2 are highlighted: the coverage of confidence intervals based on $\overline{X}$ are hence satisfactory from a threshold of $n = 100$, while $n =$

[2] For comparability reasons, the proposed methodology is applied to all the considered distributions, although in the normal case exact confidence intervals based on the studentized mean are well known. However, notice that as n increases, the difference between the two approaches becomes negligible.

150 is suggested for a good coverage of intervals based on $\hat{\Delta}$.

Table 3 shows, in the case of sampling from the uniform distribution, the estimation of the effective coverage: the threshold of 150 is here confirmed, for both confidence intervals.

TABLE 3. *Simulated percent coverage for different sample sizes n and confidence levels (1−α) (uniform distribution)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for* μ				
1−α / *n*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
50	**0.8838**	0.9088	0.9394	0.9674	0.9830	**0.8942**	**0.9240**	**0.9454**	0.9664	0.9858
100	**0.8948**	**0.9128**	**0.9488**	0.9660	0.9854	**0.8958**	**0.9244**	**0.9450**	**0.9710**	0.9862
150	**0.8994**	**0.9294**	**0.9458**	**0.9730**	**0.9906**	**0.9024**	**0.9286**	**0.9512**	**0.9744**	**0.9868**
250	**0.9016**	**0.9222**	**0.9452**	**0.9742**	**0.9898**	**0.8960**	**0.9198**	**0.9466**	**0.9724**	**0.9922**
500	**0.8892**	**0.9266**	**0.9496**	**0.9724**	**0.9890**	**0.9050**	**0.9234**	**0.9468**	**0.9724**	**0.9906**
750	**0.8960**	**0.9224**	**0.9500**	**0.9720**	**0.9910**	**0.9010**	**0.9292**	**0.9478**	**0.9750**	**0.9904**
1000	**0.8924**	**0.9266**	**0.9488**	**0.9794**	**0.9888**	**0.8958**	**0.9296**	**0.9500**	**0.9784**	**0.9910**
1250	**0.8978**	**0.9230**	**0.9426**	**0.9776**	**0.9908**	**0.9012**	**0.9260**	**0.9542**	**0.9752**	**0.9914**
1500	**0.9034**	**0.9204**	**0.9542**	**0.9760**	**0.9916**	**0.8958**	**0.9274**	**0.9478**	**0.9754**	**0.9892**
1750	**0.9034**	**0.9256**	**0.9466**	**0.9726**	**0.9896**	**0.9026**	**0.9248**	**0.9460**	**0.9752**	**0.9904**
2000	**0.8972**	**0.9200**	**0.9536**	**0.9746**	**0.9904**	**0.9056**	**0.9272**	**0.9492**	**0.9766**	**0.9908**
2250	**0.8974**	**0.9290**	**0.9482**	**0.9728**	**0.9912**	**0.8994**	**0.9214**	**0.9518**	**0.9744**	**0.9916**
2500	**0.9048**	**0.9230**	**0.9542**	**0.9738**	**0.9884**	**0.9018**	**0.9244**	**0.9508**	**0.9752**	**0.9900**

4.2. Asymmetric distributions: exponential and lognormal

The estimated coverages in Table 4 were obtained through simulations on the exponential distribution with $\lambda = 0.2$.

TABLE 4. *Simulated percent coverage for different sample sizes n and confidence levels (1-α) (exponential distribution)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for* μ				
1−α / *n*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
50	0.8512	0.8742	0.9028	0.9284	0.9508	0.8792	**0.9122**	0.9314	0.9548	0.9772
100	0.8720	0.8926	0.9282	0.9514	0.9684	**0.8858**	**0.9150**	**0.9446**	0.9670	0.9814
250	**0.8836**	**0.9118**	**0.9402**	0.9652	0.9780	**0.8912**	**0.9206**	**0.9436**	**0.9698**	0.9854
500	**0.8910**	**0.9176**	**0.9436**	**0.9706**	0.9838	**0.9006**	**0.9254**	**0.9454**	**0.9728**	**0.9892**
750	**0.8986**	**0.9216**	**0.9490**	0.9664	**0.9876**	**0.8944**	**0.9262**	**0.9488**	**0.9702**	**0.9900**
1000	**0.8996**	**0.9200**	**0.9442**	**0.9732**	**0.9848**	**0.9052**	**0.9192**	**0.9458**	**0.9750**	**0.9866**
1250	**0.8978**	**0.9228**	**0.9458**	**0.9744**	**0.9894**	**0.9004**	**0.9246**	**0.9508**	**0.9732**	**0.9892**
1500	**0.8924**	**0.9216**	**0.9460**	**0.9722**	**0.9864**	**0.8918**	**0.9270**	**0.9524**	**0.9704**	**0.9892**
1750	**0.8966**	**0.9256**	**0.9488**	**0.9718**	**0.9892**	**0.8998**	**0.9224**	**0.9502**	**0.9712**	**0.9880**
2000	**0.9042**	**0.9236**	**0.9476**	**0.9752**	**0.9882**	**0.9014**	**0.9306**	**0.9486**	**0.9756**	**0.9900**
2250	**0.8936**	**0.9240**	**0.9486**	**0.9782**	**0.9874**	**0.8940**	**0.9236**	**0.9518**	**0.9792**	**0.9870**
2500	**0.8970**	**0.9278**	**0.9482**	**0.9712**	**0.9890**	**0.8932**	**0.9306**	**0.9480**	**0.9716**	**0.9888**

For the exponential distribution, the minimum sample size for the Gini mean difference confidence interval is of 250 units at the lowest nominal confidence level 0.9, and it raises to 750 units for higher confidence levels. The threshold sample size for the mean confidence interval is also increasing from 100 to 500 units as the nominal confidence level ranges from 0.9 to 0.99.

Tables 5, 6 and 7 (below) report the results obtained for another asymmetric distribution still possessing all moments, the lognormal distribution. Recall that, as described in the previous Section, three sets of parameters deserve here our attention.

For these three cases, as the parameter δ ranges from δ= 0.35829 to δ= 0.5 and to δ= 0.74161, the coefficient of skewness, given by the third standardized moment $\beta_1 = \mu_3 / \sigma^3$, increases from $\beta_1 = 0.6555$ to $\beta_1 = 0.9658$ and to $\beta_1 = 1.6545$. As asymmetry increases, larger samples are hence required for a satisfactory approximation of the nominal coverage.

This happens also in the interval estimation of the mean μ where, for example, a sample size of n = 500 is not enough to get good confidence intervals with (1-α) = 0.99.

TABLE 5. *Simulated percent coverage for different sample sizes n and confidence levels (1-α) (lognormal distribution γ=10, δ=0.35829)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for μ*				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	0.8630	0.8816	0.9170	0.9422	0.9640	0.8748	**0.9122**	**0.9456**	**0.9692**	0.9840
100	0.8762	0.9058	0.9314	0.9584	0.9766	**0.8932**	**0.9198**	**0.9474**	**0.9706**	**0.9874**
250	**0.8852**	**0.9158**	0.9338	0.9680	0.9832	**0.9010**	**0.9266**	**0.9424**	**0.9742**	**0.9878**
500	**0.8952**	**0.9124**	**0.9462**	**0.9710**	0.9814	**0.8928**	**0.9240**	**0.9508**	**0.9764**	**0.9892**
750	**0.9010**	**0.9272**	**0.9466**	**0.9760**	0.9842	**0.9044**	**0.9318**	**0.9452**	**0.9736**	**0.9888**
1000	**0.8996**	**0.9206**	**0.9492**	**0.9714**	**0.9896**	**0.8972**	**0.9222**	**0.9446**	**0.9778**	**0.9922**
1250	**0.8994**	**0.9234**	**0.9512**	**0.9740**	**0.9908**	**0.9084**	**0.9194**	**0.9514**	**0.9766**	**0.9912**
1500	**0.8980**	**0.9236**	**0.9438**	**0.9728**	**0.9880**	**0.8990**	**0.9290**	**0.9510**	**0.9756**	**0.9878**
1750	**0.8992**	**0.9206**	**0.9504**	**0.9748**	**0.9872**	**0.9036**	**0.9258**	**0.9532**	**0.9736**	**0.9904**
2000	**0.8984**	**0.9266**	**0.9516**	**0.9782**	**0.9888**	**0.9006**	**0.9312**	**0.9414**	**0.9762**	**0.9920**
2250	**0.8998**	**0.9256**	**0.9442**	0.9680	**0.9904**	**0.9024**	**0.9242**	**0.9432**	**0.9744**	**0.9886**
2500	**0.8966**	**0.9264**	**0.9500**	**0.9760**	**0.9912**	**0.8992**	**0.9234**	**0.9476**	**0.9726**	**0.9892**

Generally speaking, in all results shown in these two subsections, the confidence intervals for Δ have an acceptable coverage almost for every sample size leading to good results for μ. When this general rule does not hold, however, the maximum relative deviation between the effective coverage of intervals for Δ and the threshold c of the acceptance region is less than 2.8%.

***TABLE 6.** Simulated percent coverage for different sample sizes n and confidence levels (1-α) (lognormal distribution γ=10, δ=0.5)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for* μ				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	0.8482	0.8658	0.9014	0.9318	0.9550	0.8716	0.9060	0.9374	0.9634	0.9786
100	0.8698	0.8988	0.9224	0.9524	0.9700	**0.8904**	**0.9172**	**0.9462**	0.9678	0.9838
250	0.8816	**0.9116**	0.9302	0.9640	0.9798	**0.8966**	**0.9262**	**0.9416**	**0.9752**	**0.9866**
500	**0.8918**	**0.9130**	**0.9440**	**0.9686**	0.9796	**0.8928**	**0.9212**	**0.9502**	**0.9760**	**0.9896**
750	**0.8974**	**0.9270**	**0.9440**	**0.9738**	0.9826	**0.9040**	**0.9346**	**0.9428**	**0.9736**	**0.9882**
1000	**0.8986**	**0.9214**	**0.9462**	**0.9708**	**0.9892**	**0.8984**	**0.9230**	**0.9454**	**0.9774**	**0.9920**
1250	**0.8982**	**0.9188**	**0.9486**	**0.9708**	**0.9898**	**0.9056**	**0.9154**	**0.9532**	**0.9738**	**0.9908**
1500	**0.8962**	**0.9254**	**0.9446**	**0.9710**	**0.9874**	**0.8982**	**0.9286**	**0.9512**	**0.9750**	0.9862
1750	**0.8996**	**0.9218**	**0.9506**	**0.9772**	**0.9872**	**0.9024**	**0.9242**	**0.9524**	**0.9738**	**0.9902**
2000	**0.9010**	**0.9290**	**0.9488**	**0.9768**	**0.9876**	**0.8976**	**0.9322**	0.9386	**0.9778**	**0.9914**
2250	**0.8986**	**0.9276**	**0.9432**	0.9682	**0.9892**	**0.9036**	**0.9222**	**0.9444**	**0.9716**	**0.9884**
2500	**0.8958**	**0.9242**	**0.9498**	**0.9738**	**0.9914**	**0.9014**	**0.9242**	**0.9512**	**0.9732**	**0.9888**

***TABLE 7.** Simulated percent coverage for different sample sizes n and confidence levels (1-α) (lognormal distribution γ=10, δ=0.74161)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for* μ				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	0.8198	0.8446	0.8762	0.9096	0.9328	0.8588	0.8930	0.9214	0.9486	0.9664
100	0.8564	0.8854	0.9050	0.9354	0.9558	0.8814	**0.9124**	0.9366	0.9596	0.9784
250	0.8798	0.9046	0.9206	0.9544	0.9722	**0.8934**	**0.9218**	0.9358	**0.9692**	0.9840
500	**0.8828**	0.9062	0.9346	0.9646	0.9762	**0.8880**	**0.9162**	**0.9480**	**0.9752**	0.9854
750	**0.8960**	**0.9198**	0.9382	0.9682	0.9804	**0.9048**	**0.9294**	**0.9396**	**0.9736**	**0.9874**
1000	**0.8948**	**0.9204**	**0.9436**	0.9674	0.9860	**0.8990**	**0.9216**	**0.9436**	**0.9758**	**0.9910**
1250	**0.8962**	**0.9178**	**0.9448**	0.9684	**0.9870**	**0.9010**	**0.9206**	**0.9492**	**0.9736**	**0.9894**
1500	**0.8918**	**0.9196**	**0.9414**	**0.9704**	0.9844	**0.9000**	**0.9224**	**0.9488**	**0.9714**	0.9856
1750	**0.8986**	**0.9208**	**0.9486**	**0.9754**	0.9862	**0.8992**	**0.9270**	**0.9516**	**0.9738**	**0.9886**
2000	**0.8956**	**0.9316**	**0.9454**	**0.9762**	**0.9868**	**0.8994**	**0.9326**	**0.9406**	**0.9762**	**0.9906**
2250	**0.8986**	**0.9260**	**0.9424**	0.9670	**0.9872**	**0.9004**	**0.9252**	**0.9448**	**0.9706**	**0.9894**
2500	**0.8984**	**0.9204**	**0.9454**	**0.9722**	**0.9898**	**0.8984**	**0.9252**	**0.9498**	**0.9718**	**0.9892**

4.3. Heavy-tailed distributions: Dagum and Pareto

This last subsection shows some results obtained for distributions that do not possess all moments. As above-mentioned, two cases for the Dagum and two cases for the Pareto distribution are considered, in Tables 8-11.

As the number of existing moments decreases, particularly for high values of the confidence level, the estimated coverage moves away from the related acceptance thresholds. This fact is true even for very large sample sizes, for both confidence intervals for Δ and for μ in the Dagum and in the Pareto distribution as well.

***TABLE 8.** Simulated percent coverage for different sample sizes n and confidence levels (1−α) (Dagum distribution θ=5.0, λ=1.0, β=1.0)*

	coverage of confidence intervals for Δ					*coverage of confidence intervals for μ*				
1−α / *n*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
50	0.8398	0.8690	0.8804	0.9228	0.9408	0.8800	0.9004	0.9356	0.9614	0.9786
100	0.8668	0.8882	0.9038	0.9400	0.9564	**0.8930**	**0.9194**	0.9338	**0.9688**	0.9796
250	0.8820	0.9070	0.9382	0.9568	0.9752	**0.9032**	**0.9220**	**0.9480**	**0.9692**	**0.9872**
500	**0.8856**	**0.9130**	0.9336	0.9606	0.9844	**0.8908**	**0.9168**	0.9388	**0.9756**	**0.9886**
750	**0.8912**	**0.9188**	**0.9408**	0.9656	0.9816	**0.8956**	**0.9216**	**0.9566**	**0.9758**	**0.9902**
1000	**0.8924**	**0.9178**	**0.9498**	**0.9700**	0.9854	**0.9032**	**0.9216**	**0.9488**	**0.9778**	**0.9876**
1250	**0.8962**	**0.9184**	**0.9414**	**0.9720**	**0.9868**	**0.9010**	**0.9188**	**0.9510**	**0.9756**	**0.9884**
1500	**0.8918**	**0.9250**	**0.9470**	**0.9692**	**0.9870**	**0.8978**	**0.9226**	**0.9494**	**0.9744**	**0.9884**
1750	**0.8926**	**0.9210**	**0.9472**	**0.9690**	**0.9872**	**0.9014**	**0.9248**	**0.9510**	**0.9774**	**0.9898**
2000	**0.8958**	**0.9250**	**0.9436**	**0.9714**	**0.9864**	**0.9046**	**0.9286**	**0.9466**	**0.9758**	**0.9876**
2250	**0.8924**	**0.9178**	**0.9494**	**0.9772**	**0.9882**	**0.9106**	**0.9188**	**0.9480**	**0.9818**	**0.9912**
2500	**0.8990**	**0.9210**	**0.9422**	**0.9680**	**0.9882**	**0.9002**	**0.9252**	**0.9474**	**0.9756**	**0.9886**

Table 9, 10 and 11 show the worst cases of coverage. The situation is somewhat better for confidence intervals for μ. Indeed, when the confidence intervals for μ have an estimated coverage leading to the acceptance of H_0, the maximum relative deviation between the effective coverage of intervals for Δ and the threshold for acceptance is less than 2.5% for the Dagum (1;1;5) case, 2.9% for the Dagum (1;1;3.6) case, 2.2% for the Pareto (1;4) case and 2.5% for the Pareto (1;3) case.

***TABLE 9.** Simulated percent coverage for different sample sizes n and confidence levels (1-α) (Dagum distribution θ=3.6, λ=1.0, β=1.0)*

	coverage of the confidence intervals for Δ					*coverage of the confidence intervals for μ*				
1−α / *n*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
50	0.8184	0.8376	0.8622	0.8860	0.9152	0.8764	0.8934	0.9216	0.9488	0.9676
100	0.8450	0.8602	0.8868	0.9166	0.9438	0.8818	0.9096	0.9316	0.9578	0.9752
250	0.8632	0.8910	0.9128	0.9448	0.9652	**0.8844**	**0.9150**	**0.9400**	0.9658	0.9844
500	0.8780	0.9070	0.9328	0.9592	0.9766	**0.8856**	**0.9178**	**0.9426**	**0.9706**	**0.9866**
750	0.8792	**0.9110**	0.9344	0.9544	0.9748	**0.8994**	**0.9174**	**0.9432**	**0.9708**	**0.9874**
1000	**0.8854**	**0.9142**	0.9334	0.9602	0.9802	**0.8992**	**0.9222**	**0.9450**	0.9676	**0.9880**
1250	**0.8938**	**0.9144**	0.9344	0.9654	0.9806	**0.8980**	**0.9212**	**0.9452**	**0.9726**	0.9864
1500	**0.8892**	**0.9160**	**0.9420**	0.9640	0.9844	**0.8974**	**0.9210**	**0.9440**	**0.9710**	**0.9890**
1750	**0.8870**	**0.9204**	0.9378	0.9662	0.9846	**0.9016**	**0.9230**	**0.9498**	**0.9740**	**0.9882**
2000	**0.8952**	**0.9190**	**0.9412**	**0.9714**	0.9850	**0.9060**	**0.9212**	**0.9504**	**0.9760**	0.9864
2250	**0.8968**	**0.9256**	0.9350	0.9674	0.9836	**0.9046**	**0.9330**	**0.9456**	**0.9722**	**0.9906**
2500	**0.8908**	**0.9188**	**0.9442**	0.9654	0.9844	**0.8954**	**0.9214**	**0.9490**	**0.9704**	**0.9878**

TABLE 10. *Simulated percent coverage for different sample sizes n and confidence levels (1-α) (Pareto distribution x_0=1.0, θ=4.0)*

	coverage of confidence intervals for Δ					coverage of confidence intervals for μ				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	0.7978	0.8294	0.8508	0.8790	0.9092	0.8436	0.8762	0.8916	0.9292	0.9484
100	0.8404	0.8578	0.8810	0.9106	0.9358	0.8686	0.8990	0.9188	0.9452	0.9626
250	0.8632	0.8844	0.9068	0.9438	0.9596	**0.8858**	0.9062	0.9314	0.9638	0.9772
500	**0.8854**	0.9076	0.9218	0.9542	0.9724	**0.8932**	**0.9142**	0.9392	0.9684	0.9814
750	**0.8864**	0.9088	0.9320	0.9568	0.9762	**0.8892**	**0.9218**	**0.9418**	**0.9678**	0.9814
1000	**0.8862**	0.9106	0.9372	0.9614	0.9764	**0.8894**	**0.9194**	**0.9456**	**0.9662**	0.9832
1250	**0.8944**	**0.9202**	0.9380	0.9610	0.9806	**0.8986**	**0.9292**	**0.9448**	**0.9690**	**0.9866**
1500	**0.8858**	**0.9166**	**0.9424**	0.9664	0.9798	**0.8966**	**0.9208**	**0.9454**	**0.9704**	0.9852
1750	**0.8926**	**0.9206**	**0.9368**	0.9646	0.9796	**0.8912**	**0.9256**	**0.9434**	**0.9700**	0.9848
2000	**0.8924**	**0.9202**	**0.9468**	0.9672	0.9830	**0.8972**	**0.9268**	**0.9536**	**0.9708**	**0.9874**
2250	**0.8974**	**0.9214**	**0.9412**	0.9652	0.9860	**0.9020**	**0.9238**	**0.9468**	**0.9684**	**0.9896**
2500	**0.8986**	**0.9212**	**0.9436**	**0.9694**	0.9844	**0.8978**	**0.9254**	**0.9470**	**0.9710**	**0.9892**

TABLE 11. *Simulated percent coverage for different sample sizes n and confidence levels (1-α) (Pareto distribution x_0=1.0, θ=3.0)*

	coverage of confidence intervals for Δ					coverage of confidence intervals for μ				
1−α	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*	*0.9*	*0.925*	*0.95*	*0.975*	*0.99*
n										
50	0.7846	0.8012	0.8294	0.8558	0.8802	0.8362	0.8508	0.8812	0.9108	0.9340
100	0.8198	0.8404	0.8658	0.8850	0.9240	0.8536	0.8776	0.9042	0.9244	0.9514
250	0.8496	0.8648	0.8992	0.9224	0.9504	0.8778	0.8936	0.9212	0.9438	0.9696
500	0.8602	0.8880	0.9218	0.9440	0.9610	**0.8828**	0.9068	0.9352	0.9586	0.9738
750	0.8680	0.9008	0.9178	0.9500	0.9670	**0.8906**	0.9076	0.9332	0.9622	0.9784
1000	**0.8836**	0.9028	0.9268	0.9476	0.9686	**0.8934**	**0.9122**	0.9386	0.9602	0.9794
1250	0.8756	0.9078	0.9266	0.9526	0.9762	**0.8938**	**0.9156**	0.9360	0.9624	0.9836
1500	0.8742	0.9090	0.9294	0.9546	0.9790	**0.8850**	**0.9172**	**0.9404**	0.9640	**0.9880**
1750	**0.8844**	**0.9152**	0.9344	0.9604	0.9768	**0.8934**	**0.9202**	**0.9436**	0.9674	0.9836
2000	**0.8968**	0.9088	0.9330	0.9590	0.9776	**0.8964**	**0.9202**	**0.9414**	0.9662	0.9854
2250	0.8816	**0.9152**	0.9294	0.9600	0.9772	**0.8882**	**0.9206**	**0.9380**	**0.9694**	0.9850
2500	**0.8862**	**0.9148**	0.9324	0.9640	0.9792	**0.8904**	**0.9194**	0.9376	**0.9696**	0.9846

In order to get an overall view of the simulation results, the following plot of coverages, presented in Figure 1 below, can be helpful. It clearly depicts that the convergence of the studentized sample mean difference T to the normal distribution is not uniform.

If the value 0.933 is considered as a satisfactory approximation of 0.95 it may be argued that a minimum threshold of n = 1500 allows us to use the asymptotic confidence intervals for Δ.

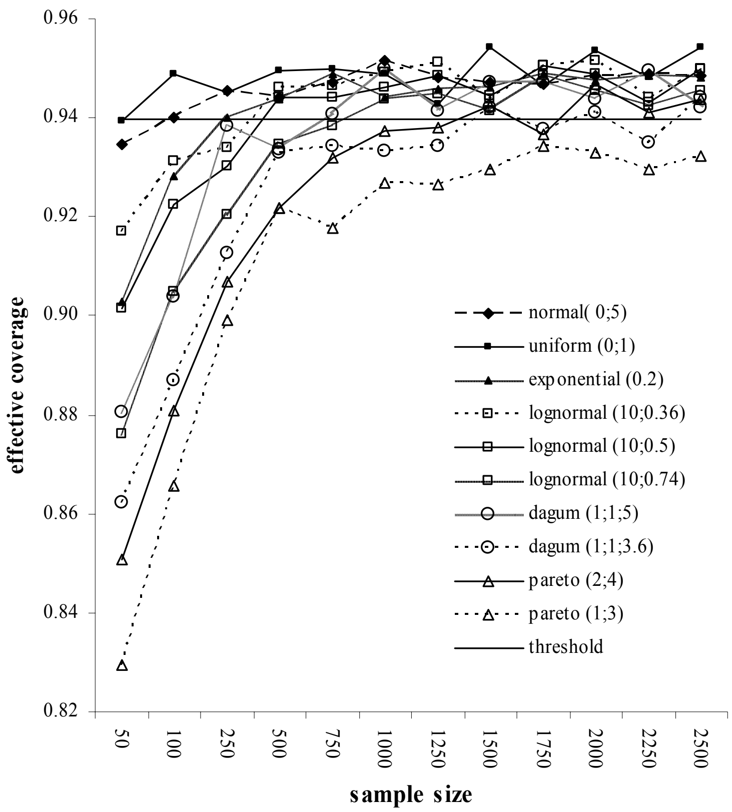

FIGURE 1. *Effective coverage of asymptotic confidence intervals for* Δ, *with (1-* α*)=0.95, compared to the threshold established by the test on* H_0

Analogously, as it is well known, also the convergence of the sample mean assured by the central limit theorem is not uniform, as shown by Figure 2.

A minimum value of n=1000 for the sample size guarantees an effective coverage of 0.935, for all underlying distributions. Whenever this coverage is accepted as a satisfactory approximation of the nominal value 0.95, n = 1000 could then be considered as a suitable threshold for broad application purposes, for confidence intervals for the mean.

With reference to the whole set of results, as it is well known in the literature, a reduction of the required sample size is observed if the confidence level decreases. Indeed, higher confidence levels are more sensitive to the exact nature of the tails of the sample distribution of the used estimator.

In social sciences, where the sample sizes are usually high, nearly around the thousands (see, for instance and among many others, the income surveys of the Luxembourg Income Study and the Banca d'Italia surveys), asymptotic confidence intervals provides hence reasonable inference. Conversely, where the sample size is lower than the proposed thresholds, bootstrap inference is to be advised, despite its computational complexity.

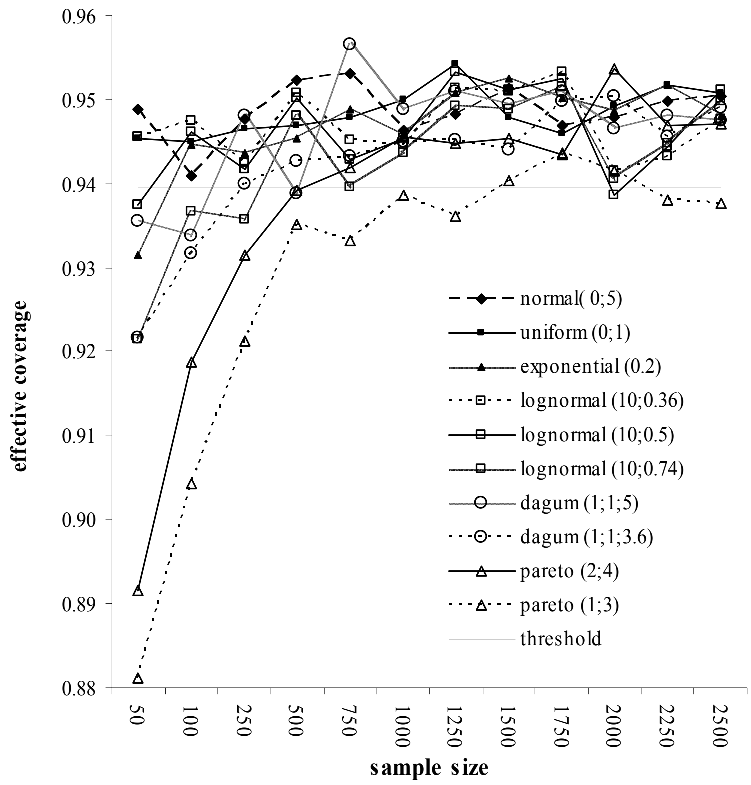

FIGURE 2. *Effective coverage of asymptotic confidence intervals for the mean, with (1-α)=0.95, and the threshold established by the test on H_0*

5. A FURTHER ANALYSIS: LEFT AND RIGHT COVERAGE ERRORS

In order to better understand the behaviour of the effective coverage, it might be interesting to evaluate also the left coverage and the right coverage, in all the considered distributions. To give a rough idea of the behaviour of the tails of the estimator T in (11), only the nominal confidence 0.95 is here considered. All results have hence to be compared with the value $\alpha/2$=0.025, that is the asymptotic proportion of intervals $\hat{\Delta} \pm z_{1-\alpha/2}\hat{\sigma}(\hat{\Delta})$ not covering the true value of Δ. To this attempt, for each confidence interval which fails to contain the true value Δ, it was checked whether Δ lies on its left or on its right side. The proportion of intervals satisfying this condition provides an estimate for the left (respectively right) rejection probability (denoted by LRP, and, respectively, RRP).

Notice that the true value of Gini's mean difference lies on the left side of the confidence interval if and only if a large value of the sample statistic has occurred. Therefore, the LRP actually describes the right tail of the sample distribution, whereas the RRP, conversely, refers to the left tail.

TABLE 13. *Simulated LRP for different sample sizes n and for different underlying distributions, with confidence level 0.95*

n	*normal*	*uniform*	*expo nential*	*lognorm 10;0.36)*	*lognorm (10;0.5)*	*lognorm (10;0.74)*	*Dagum (1;1;5)*	*Dagum (1;1;3.6)*	*Pareto (1;4)*	*Pareto (1;3)*
50	0.0172	0.0332	0.0086	0.0118	0.0098	0.0062	0.0066	0.0028	0.0042	0.0016
100	0.0164	0.0290	0.0100	0.0128	0.0110	0.0062	0.0076	0.0040	0.0042	0.0016
250	0.0202	0.0310	0.0096	0.0154	0.0124	0.0078	0.0048	0.0066	0.0020	0.0026
500	0.0256	0.0262	0.0142	0.0154	0.0136	0.0106	0.0110	0.0062	0.0078	0.0034
750	0.0214	0.0272	0.0142	0.0202	0.0190	0.0146	0.0148	0.0078	0.0090	0.0058
1000	0.0234	0.0248	0.0194	0.0184	0.0178	0.0136	0.0120	0.0106	0.0088	0.0072
1250	0.0208	0.0288	0.0192	0.0170	0.0160	0.0130	0.0142	0.0106	0.0096	0.0094
1500	0.0236	0.0224	0.0208	0.0210	0.0186	0.0144	0.0134	0.0108	0.0104	0.0068
1750	0.0230	0.0264	0.0190	0.0158	0.0158	0.0138	0.0142	0.0102	0.0142	0.0090
2000	0.0254	0.0240	0.0228	0.0220	0.0210	0.0176	0.0166	0.0134	0.0136	0.0112
2250	0.0220	0.0250	0.0212	0.0242	0.0230	0.0184	0.0160	0.0150	0.0138	0.0082
2500	0.0228	0.0252	0.0228	0.0194	0.0176	0.0168	0.0146	0.0138	0.0128	0.0096

Notice that nearly all simulated LRPs are lower than 0.025 and that they improve as *n* increases. Furthermore, as asymmetry increases, the LRP goes far from its asymptotical value; as the tail heaviness increases LRP dramatically decreases, that is the right tail of the sample mean difference becomes thinner.

By left coverage error we mean the difference between the LRP and its asymptotic value 0.025. Figure 3 offers a plot of the left coverage error for 0.95 nominal confidence.

The most relevant result is that the left coverage error is very large in moderate samples and decreases very slowly as the sample size increases.

Since the LRP is mostly lower than its asymptotic value, the RRP is expected to be higher than 0.025, as Table 14 shows.

It is noticeable that RRP behaves worse in the normal distribution than in the uniform, the latter being the distribution offering the best estimation of the RRP, for all the values of the sample size *n*. Furthermore, data from the lognormal distribution show that, as asymmetry increases, the RRP increases as well, even reaching values very far from 0.025. In addition, as the tail heaviness grows (see the last four columns of Table 14) the RRP increases dramatically, emphasizing the shape of the right tail of the studentized $\hat{\Delta}$.

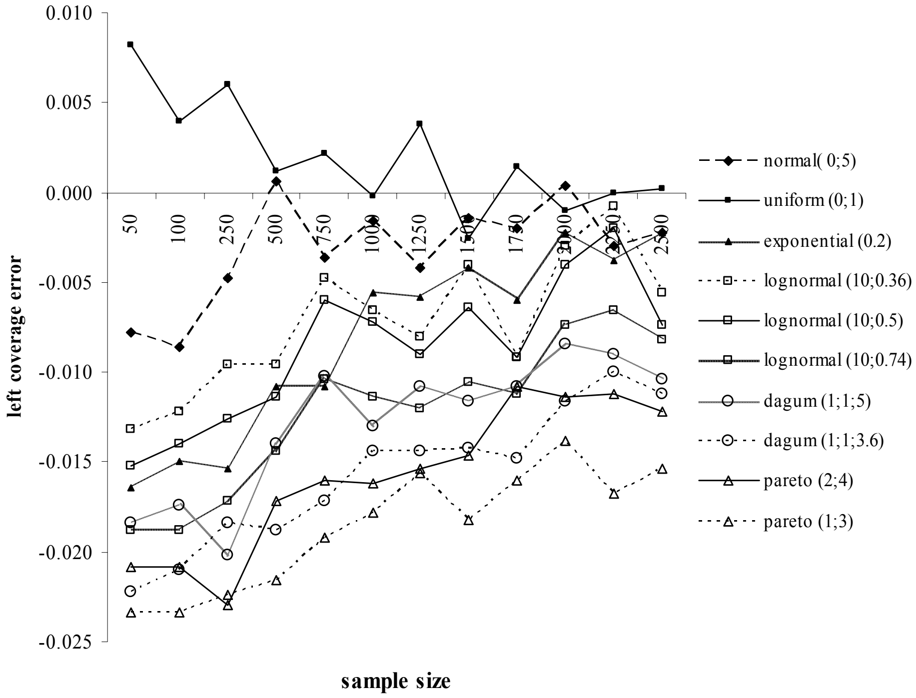

FIGURE 3. *Left coverage error for confidence level (1-α)=0.95*

Figure 3 offers a plot of the left coverage error (i.e. the difference between the RRP and its asymptotic value 0.025) for 0.95 nominal confidence. All left coverage error converges to zero as the sample size increases.Almost all of them are negative, only samples drawn from the uniform give positive values. A deeper understanding of this behaviour could be obtained by means of the exact distribution of the sample mean difference that, in this particular case, is known (see Ali 1969).

TABLE 14. *Simulated RRP for different sample sizes n and for different underlying distributions, with confidence level 0.95*

n	*normal*	*uniform*	*expo nential*	*logn. 10;0.36)*	*logn. (10;0.5)*	*logn. (10;0.74)*	*Dagum (1;1;5)*	*Dagum (1;1;3.6)*	*Pareto (1;4)*	*Pareto (1;3)*
50	0.0482	0.0274	0.0886	0.0712	0.0888	0.1176	0.1130	0.1350	0.1450	0.1690
100	0.0436	0.0222	0.0618	0.0558	0.0666	0.0888	0.0886	0.1092	0.1148	0.1326
250	0.0344	0.0238	0.0502	0.0508	0.0574	0.0716	0.0570	0.0806	0.0912	0.0982
500	0.0302	0.0242	0.0422	0.0384	0.0424	0.0548	0.0554	0.0610	0.0704	0.0748
750	0.0314	0.0228	0.0368	0.0332	0.0370	0.0472	0.0444	0.0578	0.0590	0.0764
1000	0.0252	0.0264	0.0364	0.0324	0.0360	0.0428	0.0382	0.0560	0.0540	0.0660
1250	0.0310	0.0286	0.0350	0.0318	0.0354	0.0422	0.0444	0.0550	0.0524	0.0640
1500	0.0294	0.0234	0.0332	0.0352	0.0368	0.0442	0.0396	0.0472	0.0472	0.0638
1750	0.0302	0.0270	0.0322	0.0338	0.0336	0.0376	0.0386	0.0520	0.0490	0.0566
2000	0.0260	0.0224	0.0296	0.0264	0.0302	0.0370	0.0398	0.0454	0.0396	0.0558
2250	0.0290	0.0268	0.0302	0.0316	0.0338	0.0392	0.0346	0.0500	0.0450	0.0624
2500	0.0288	0.0206	0.0290	0.0306	0.0326	0.0378	0.0432	0.0420	0.0436	0.0580

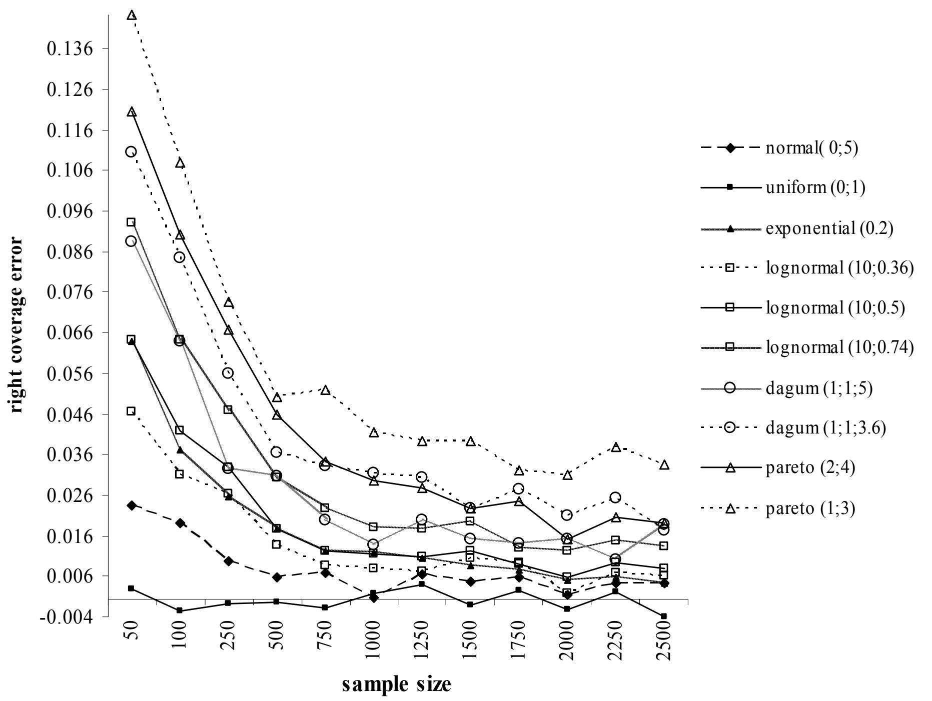

***FIGURE 4**. Right coverage error for confidence level (1-α)=0.95*

It is worthwhile to notice that the scale in the vertical axis of Figure 4 is 5 times bigger than in Figure 3, this fact indicating the higher relevance of the right coverage error with respect to the left one.

The knowledge of the LRP and the RRP gives more information on the distribution of the studentized sample Gini's mean difference: it has always some asymmetry and a long right tail when sampling from heavy-tailed distributions. This behaviour has a great influence on the sample sizes required to attaining a good coverage for confidence intervals for Δ.

6. CONCLUDING REMARKS

The simulation results presented in this paper are encouraging. The outlined methodology to determine the effective coverage of confidence intervals for Δ seems really useful. It gives a good approximation nearly in all cases in which the confidence intervals for μ are valid as well. The minimum sample sizes obtained throughout this work can be helpful from an operational point of view. As expected, in accordance with the literature, the effective coverages depends upon the population distribution and sometimes inaccuracies may risk to be quite serious on small sample sizes if one just relies upon the asymptotic approximation. In fact, in some of the considered cases, the minimal threshold indicated by the reported simulations is rather high, but similar results hold also for the confidence intervals for μ, whenever the variance of the underlying population is unknown. This fact is not a relevant issue in social sciences, where data usually consists of large samples drawn from populations.

A careful inspection of the simulated results allows us to observe three main factors affecting the effective percent coverage of confidence intervals for Δ:

- the existence of the r-th moments of the distribution, with r >2,
- the degree of asymmetry of the underlying distribution, and
- the value of the nominal risk α.

The simulations from the Pareto and the Dagum distribution show that heavier tails produce a considerable decrease in coverages. According to the simulated data, furthermore, the approximation for the two considered continuous symmetric distributions (the uniform and the normal distributions, for which the minimum sample size is n = 150 for 1- α = 0.975) is better than the approximation obtained for the asymmetric ones. For the exponential and the lognormal distribution with slight asymmetry, the threshold is n = 500, while for the lognormal distribution with higher asymmetry n = 2000 is needed, and finally for the Pareto and the Dagum distribution, in case of heavy tails, n > 2500. In particular, the analysis carried out on the lognormal distribution, with different sets of the parameters, indicates that increasing asymmetry in the underlying distribution has a worse effect on the coverage.

Finally, as it is well known in literature, the coverage, in all cases, has a different behaviour for different values of the nominal risk α, so that the minimum sample sizes required for high values of α are always lower than those needed for lower values of the nominal risk: the approximation is worse in the tails when α is small.

Remarkably, the simulated coverage is always lower than the nominal one, in all considered cases. This fact can be ascribed (as also LRP and RRP errors denote) to some asymmetry or/and tail heaviness in the distribution of $\hat{\Delta}$, which tend to vanish for increasing values of the sample size n, as $\hat{\Delta}$ approaches the asymptotic normal distribution. More investigation can be done by producing further simulations from other continuous distributions. We hope that the new proposed methodology for determining minimum sample sizes for confidence intervals may be useful also in other contexts.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous referees for their valuable comments and suggestions. The work is partially supported by F.A.R. 2005 - Università degli Studi di Milano-Bicocca.

RIASSUNTO

La differenza media campionaria $\hat{\Delta}$ *è uno stimatore non distorto per la differenza media di Gini. ed ha distribuzione asintoticamente normale (Hoeffding, 1948). In questo lavoro, utilizzando uno specifico stimatore della varianza della differenza media campionaria si studiano gli intervalli di confidenza per* Δ*, in ambito non-parametrico. Le coperture effettive sono state stimate mediante simulazioni ottenute da una varietà di modelli di fenomeni reali, considerando diversi set di parametri per ciascuno di essi. La determinazione empirica delle soglie minime necessarie ad assicurare una buona approssimazione della copertura nominale è ottenuta applicando un test d'ipotesi a dati ottenuti per via simulativa. I risultati raggiunti mostrano che tali soglie campionarie sono criticamente legate alla presenza di asimmetria e di code pesanti nella distribuzione da cui provengono i campioni. Esse sono comunque inferiori a quelle usualmente reperibili negli studi delle scienze sociali.*

REFERENCES

Ali, M.M. (1969). Distribution of linear combination of order statistics from rectangular population. *Bull.Inst. Statist. Res*., tr. **3**, 1-21.

Banca d'Italia (2004). I bilanci delle famiglie italiane nel 2002. Supplemento al *Bollettino Statistico*. **XIV**, n.1-2.

Biewen, M. (2002). Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics*, **108**, 317-342.

Cowell F.A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics*, **42**, 27-41.

Cowell, F.A. and Flachaire E. (2002). Sensitivity of Inequality Measures to Extreme values. *Working paper 2002.40 EUREQua*, University Paris I Panthéon-Sorbonne.

Fishman, G.S. (1996). *Monte Carlo: Concepts, algorithms and Applications.* Springer Series in Operation Research. Springer. New York

Greselin F. and Zenga M. (2006). Convergence of the Sample Mean Difference to the normal distribution: simulation results. *Statistica & Applicazioni*, **4**, special issue n°1, 99-122.

Hoeffding W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, **19**, No.3, 293-325.

Lee A.J. (1990), *U-Statistics*, *Theory and Practice*. Marcel Dekker, Inc. New York and Basel.

Lomnicki Z.A. (1952). The standard error of Gini's mean difference. *The Annals of Mathematical Statistics*, **23**, No.4, 635-637.

Michetti B. and Dall'aglio G. (1957). La differenza semplice media. *Statistica*, **17**, 159-254.

Mills J.A. and Zandvakili S. (1997). Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics*, **12**, 133-150.

Nair U.S. (1936). The standard error of Gini's mean difference. *Biometrika*, **28**, 428-436.

Newcombe R.G. (1998). Two sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine,* **17**, 857-872.

Sendler W. (1979). On Statistical Inference in Concentration Measurement. *Metrika*, **26**, 109-122.

Xu K. (2000). Inference for Generalized Gini Indices using Iterated-Boostrap Method. *Journal of Business & Economic Statistics*, **18**, 223-227.

Zenga M., Polisicchio M., and Greselin, F. (2004). The variance of Gini's mean difference and its estimators. *Statistica*, **3**, 455-475.

Zitikis R. (2002). Large sample estimation of a family of economic inequality indices. *Pak.J. Statist*, S.Ejaz Ahmed Special Issue, **18**, 225-248.