Journal of Applied Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Large sample confidence intervals for the skewness parameter of the skew-normal distribution based on Fisher's transformation

Valentina Mameli a, Monica Musio a, Erik Sauleau b & Annibale Biggeri c

a Dipartimento di Matematica, Università di Cagliari, Italy
b Laboratoire de Biostatistique, Faculté de Médecine, Université de Strasbourg, France
c Dipartimento di Statistica, Università di Firenze, Italy

Available online: 13 Mar 2012

To cite this article: Valentina Mameli, Monica Musio, Erik Sauleau & Annibale Biggeri (2012): Large sample confidence intervals for the skewness parameter of the skew-normal distribution based on Fisher's transformation, Journal of Applied Statistics, DOI:10.1080/02664763.2012.668177

To link to this article: http://dx.doi.org/10.1080/02664763.2012.668177

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings,
demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Large sample confidence intervals for the skewness parameter of the skew-normal distribution based on Fisher’s transformation

Valentina Mamelia, Monica Musioa, Erik Sauleaub and Annibale Biggeric

Dipartimento di Matematica, Università di Cagliari, Italy; Laboratoire de Biostatistique, Faculté de Médecine, Université de Strasbourg, France; Dipartimento di Statistica, Università di Firenze, Italy

(Received 25 February 2011; final version received 16 February 2012)

The skew-normal model is a class of distributions that extends the Gaussian family by including a skewness parameter. This model presents some inferential problems linked to the estimation of the skewness parameter. In particular its maximum likelihood estimator can be infinite especially for moderate sample sizes and is not clear how to calculate confidence intervals for this parameter. In this work, we show how these inferential problems can be solved if we are interested in the distribution of extreme statistics of two random variables with joint normal distribution. Such situations are not uncommon in applications, especially in medical and environmental contexts, where it can be relevant to estimate the distribution of extreme statistics. A theoretical result, found by Loperfido [7], proves that such extreme statistics have a skew-normal distribution with skewness parameter that can be expressed as a function of the correlation coefficient between the two initial variables. It is then possible, using some theoretical results involving the correlation coefficient, to find approximate confidence intervals for the parameter of skewness. These theoretical intervals are then compared with parametric bootstrap intervals by means of a simulation study. Two applications are given using real data.

Keywords: skew-normal distribution; skewness parameter; Fisher transformation; maximum likelihood estimator; bootstrap intervals; PM10 concentration; creatinine clearance

1. Introduction

The skew-normal (SN) distribution, introduced by Azzalini [1] in 1985, refers to a parametric class of probability distributions that includes the standard normal as a special case. A random variable \( Z \) is said to be skew-normal with parameter \( \lambda \), if its density function is

\[
\varphi(z; \lambda) = 2\phi(z)\Phi(\lambda z) \quad \text{with} \quad \lambda, \ z \in \mathbb{R},
\]

where \( \phi(z) \) and \( \Phi(z) \) are the density and distribution functions of the standard normal distribution, respectively.
where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian density and cumulative distribution function, respectively. The parameter $\lambda$ controls skewness, which is positive when $\lambda > 0$ and negative when $\lambda < 0$. The standard normal distribution can be obtained by setting $\lambda = 0$. We shall write $Z \sim SN(\lambda)$ to denote a random variable with density (1). The literature related to the skew-normal distribution has grown rapidly in recent years, because of the nice properties of this distribution. In particular, we recall the following properties:

- As $\lambda \to \infty$, $\varphi(z; \lambda)$ tends to the half-normal density.
- If $Z$ is a SN($\lambda$) random variable, then $-Z$ is a SN($-\lambda$) random variable.
- If $Z \sim SN(\lambda)$, then $Z^2 \sim \chi^2(1)$, i.e. a chi-squared random variable with one degree of freedom.

The class of SN distributions can be generalized by the inclusion of location and scale parameters which we denote by $\xi$ and $\psi > 0$. Thus if $Z \sim SN(\lambda)$, then $W = \xi + \psi Z$ is a skew-normal variable with parameters $(\xi, \psi, \lambda)$, i.e. its density function is given by:

$$
\varphi(w; \lambda) = \frac{2}{\psi} \phi\left(\frac{w - \xi}{\psi}\right) \Phi\left(\frac{\lambda (w - \xi)}{\psi}\right).
$$

Despite its nice properties, inferential problems arise in the estimation of the skewness parameter. More specifically, its maximum likelihood estimator (MLE) can take infinite values with positive probability, especially for small or moderate sample sizes. In addition, it is not clear how to calculate confidence intervals for this parameter. Furthermore, the method of moments can give even worse results. Several solutions have been proposed to solve these problems, using numerical approximation methods, in both a classical and a Bayesian approach.

Azzalini and Capitanio [2], suggested stopping the log-likelihood maximization procedure when the log-likelihood function reaches a value not significantly lower than the maximum. Another solution was proposed by Sartori [10], who developed a method based on a second-order modification of the likelihood equation that never produces boundary estimates. Another contribution to the problem can be found in Liseo and Loperfido [6], who use a Bayesian approach which modifies the likelihood function with a Jeffreys priors for the skewness parameter. They also prove that such prior is proper.

In this paper, we present a solution to these estimation problems in a particular situation. We are interested in the distribution of the maximum or minimum of two random variables which have a bivariate normal distribution. Order statistics of correlated normal variables appear in statistical applications. In a number of situations, especially in medical and the environmental contexts, even if observations are taken in pairs, interest centres on the maximum or minimum value of the observations. The distribution of the minimum of two standardized correlated normal variables was found by Roberts [9] in the context of twin studies. The resulting distribution is now recognized as the skew-normal. Loperfido [7] shows that the minimum or maximum of two random variables with same mean and variance, whose distribution is jointly normal, is skew-normal with skewness parameter that can be expressed as a function of the correlation coefficient between the two initial variables. In this specific case, we use the MLE of the correlation coefficient $\rho$ between the two initial variables to find the MLE of the corresponding skewness parameter of the skew-normal. Using the Fisher transformation [4,5], we approximate the distribution of the skewness parameter $\lambda$ and we are able to test hypotheses and to compute confidence intervals for $\lambda$. The Fisher transformation is accurate for small $\rho$ and large values of the sample size $n$. We conduct a simulation study to compare such theoretical confidence intervals with the corresponding confidence intervals computed using the parametric bootstrap, for different values of $n$ and $\rho$. 
Finally we present two applications of such method using real data. In the first we are interested in the natural logarithm of PM$_{10}$ concentration. The fine particles PM$_{10}$ in the atmosphere are pollutants that cause serious health effects, so it is important to keep their concentration under control. Keeping PM$_{10}$ to acceptable levels, could prevent hospitalizations for respiratory and cardiovascular causes and acute bronchitis and asthma among children. It becomes relevant to study the distribution of the natural logarithm of the highest concentrations of such fine particles. We analyse data from a monitoring survey carried out daily between 2003 and 2005 in two different areas of the town of Cagliari in Italy.

In the second example, we consider a medical data set concerning the follow-up of patients who had surgical operation for a renal cancer. Depending on the tumour (size, extension, etc.), the operation can be a total or a partial nephrectomy. The removal of part of a kidney or a whole kidney causes a decrease in the filtering function performed by the kidney. But this glomerular filtration rate recovers progressively through an increase in the capacity of the remaining kidney (and also, if the nephrectomy was only partial, of the remaining part of the operated kidney). Full renal capabilities are observed at around 1–3 post-operative months, depending on the surgical procedure, the pre-operative renal function, the patient, etc. When the patient has a medical examination, for example, six months after the operation, the maximum of the glomerular function rate between its value at 1 month and its value at 3 months can be considered as a value of the new capacity of the patient’s renal function.

The paper is organized as follows: in Section 2, after recalling a theoretical result concerning the SN distribution, we use Fisher’s transformation to construct test and confidence intervals for the skewness parameter. Section 3 is devoted to the construction of confidence intervals for the skewness parameter using the parametric bootstrap. In Section 4, we summarize the results of a simulation study that we conducted to compare confidence intervals constructed using both methods. In Section 5, we apply the proposed methodology to construct approximate (ACIs) and bootstrap confidence intervals (BCIs) for the skewness parameter using data from the above-mentioned monitoring survey in Cagliari and from the follow-up data set of patients operated for a renal cancer in Strasbourg (France). We finish with a discussion in Section 6.

2. Approximate confidence intervals for skewness parameter based on Fisher’s transformation

We denote a random vector $(X, Y)$ having a bivariate normal distribution by $(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, where $\rho$ is the correlation coefficient. Its density is then

$$f(x,y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left[ \frac{(x - \mu_x)^2}{\sigma_x^2} - \frac{2\rho(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y} + \frac{(y - \mu_y)^2}{\sigma_y^2} \right] \right\}.$$  

Loperfido [8] has shown that any weighted average of the extremes of an exchangeable and bivariate normal random vector is skew-normal. More specifically, the following result holds:

**Theorem 2.1** Let $X$ and $Y$ be two random variables whose joint distribution is bivariate normal with $\mu_X = \mu_Y = \xi$, $\sigma_X^2 = \sigma_Y^2 = \psi^2$ and $\text{Cov} (X, Y) = \rho \psi^2$. Then for any two constants $h$ and $k \neq -h$ the distribution of

$$h \min(X, Y) + k \max(X, Y)$$
Then the random set

\[ C(R) = \left[ \exp \left( \frac{-z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \frac{1+R}{1-R} \right) , \exp \left( \frac{z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \frac{1+R}{1-R} \right) \right] \]

is an approximate 1 - \( \alpha \) confidence interval for \( \lambda \).

A proof of Theorem 2.1 can be found in [7] for \( \xi = 0 \) and \( \psi = 1 \), and in the more general case in [8].

Suppose now that we are interested in constructing a confidence intervals for \( \lambda = \sqrt{(1 - \rho)/(1 + \rho)} \). We know that the coefficient of correlation \( \rho \) is the sample correlation coefficient

\[ R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}. \]

Using the invariance property of MLE, we know that the statistic \( \hat{\lambda} = \sqrt{(1 - R)/(1 + R)} \) is the MLE of \( \lambda \).

It is customary to base tests concerning \( \rho \) on the statistic \( \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right) \). This is the Fisher transformation of \( R \) (see, for instance, [4,5]). It can be shown that the distribution of this statistic, for \( n > 50 \), is approximately normal with mean \( \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) \) and variance \( 1/(n-3) \). Then the variable

\[ Z = \frac{(1/2) \ln \left( \frac{1+R}{1-R} \right) - (1/2) \ln \left( \frac{1+\rho}{1-\rho} \right)}{1/\sqrt{n-3}} \]

(2)

has approximately standard normal distribution. Using the above approximation we can calculate 1 - \( \alpha \) confidence intervals for the parameter \( \lambda = \sqrt{(1 - \rho)/(1 + \rho)} \). We have

\[ P \left( -z_{\alpha/2} \leq \frac{(1/2) \ln \left( \frac{1+R}{1-R} \right) - (1/2) \ln \left( \frac{1+\rho}{1-\rho} \right)}{1/\sqrt{n-3}} \leq z_{\alpha/2} \right) \approx 1 - \alpha \]

which is equivalent to

\[ P \left( \exp \left( \frac{-z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \frac{1+R}{1-R} \right) \leq \lambda \leq \exp \left( \frac{z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \frac{1+R}{1-R} \right) \right) \approx 1 - \alpha. \]

Then the random set

In particular, for the choice \( k = 0 \) and \( h = 1 \), we find that the distribution of \min(X, Y) is

\[ SN \left[ \xi, \psi, \lambda = -\sqrt{\frac{1-\rho}{1+\rho}} \right] \]

while, for \( h = 0 \) and \( k = 1 \), we have that the distribution of \max(X, Y) is

\[ SN \left[ \xi, \psi, \lambda = \sqrt{\frac{1-\rho}{1+\rho}} \right]. \]
This approximation can also be used to test hypotheses concerning \( \lambda = \sqrt{(1-\rho)/(1+\rho)} \).

If we are interested in testing

\[
H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0,
\]

we find that an appropriate critical region of size \( \alpha \) for testing the null hypothesis against the alternative is \(|Z| \geq z_{\alpha/2}\), where \(Z\) is defined as in Equation (2) and \(z_{\alpha/2}\) is defined by \(P(Z \geq z_{\alpha/2}) = \alpha/2\). Then we can write the rejection region as

\[
\left\{ r : \left| \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right) + \ln \lambda_0 \right| \geq z_{\alpha/2} \right\}.
\]

The same procedure can be applied to compute the confidence intervals of level \(1 - \alpha\) and the critical region for the hypothesis (4) for \( \lambda = -\sqrt{(1-\rho)/(1+\rho)} \). For instance, an approximate confidence interval (ACI) is

\[
C(R) = \left[ \exp \left( \frac{z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right) \right), \exp \left( \frac{-z_{\alpha/2}}{\sqrt{n-3}} - \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right) \right) \right].
\]

Using Theorem 2.1 and these procedures, we can compute confidence intervals and critical regions for the unknown skewness parameter \( \lambda \) when the other unknown parameters (means and variances) of the random variables \(X\) and \(Y\) are estimated by the corresponding MLEs. Note that the length of the \(1 - \alpha\) confidence interval (3)

\[
L(R, n) = \exp \left( \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right) \right) \left[ \exp \left( \frac{z_{\alpha/2}}{\sqrt{n-3}} \right) - \exp \left( \frac{-z_{\alpha/2}}{\sqrt{n-3}} \right) \right]
\]

is a decreasing function of \(R\) for fixed \(n\) and a decreasing function of \(n\) for fixed values of \(R\). We expect to have shorter intervals for \(R\) closed to 1 and for large samples.

### 3. Parametric BCIs

In this section, we use the parametric bootstrap method for constructing confidence intervals (see [3]). This method relies on resampling with replacement from an estimated parametric model and calculating the required statistic from these repeated samples. The values of the statistic from the repeated sampling can then be used to generate standard errors and confidence intervals for the statistic of interest.

In our specific case we consider a random sample of \(n\) pairs \(Z = (X, Y)\) where \((X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)\). For the parametric bootstrap, instead of estimating the theoretical distribution function \(F\) by the empirical distribution function, we estimate the five parameters of the bivariate normal by the corresponding MLEs. We denote the bivariate normal distribution with these values for the parameters by \(\hat{F}_{\text{norm}}\). Suppose that our functional of interest is \(\Theta = \Theta(F)\), which we estimate by the statistic \(\hat{\Theta} = \hat{\Theta}(Z_1, \ldots, Z_n)\). In order to construct a confidence interval for \(\Theta\) we introduce the bootstrap random variables \(Z_1^*, Z_2^*, \ldots, Z_n^*\) i.i.d with distribution \(\hat{F}_{\text{norm}}\).

Then we generate \(B\) bootstrap samples from \(Z_1^*, Z_2^*, \ldots, Z_n^*\), denoted by \(z^{a1}, z^{a2}, \ldots, z^{aB}\), and for each we compute the bootstrap replication \(\hat{\Theta}^*(b) = \hat{\Theta}(z^{ab}), b = 1, \ldots, B\). Let \(\hat{\Theta}_B^{* \text{(1-\alpha)}}\) be the 100 \(\cdot\ \alpha\)th empirical percentile of the \(\hat{\Theta}^*(b)\) values, that is, the \(B \cdot \alpha\)th value in the ordered list of the \(B\) replications of \(\hat{\Theta}^*\). Likewise, let \(\hat{\Theta}_B^{* \text{up}}\) be the 100 \(\cdot\ (1-\alpha)\)th empirical percentile. The approximate \(1 - 2\alpha\) percentile interval is:

\[
[\hat{\Theta}_{% \text{lo}}, \hat{\Theta}_{% \text{up}}] \approx [\hat{\Theta}_B^{* \text{(1-\alpha)}}(\alpha), \hat{\Theta}_B^{* \text{(1-\alpha)}}(1-\alpha)].
\]

In our case \(\Theta = \lambda = \pm \sqrt{(1-\rho)/(1+\rho)}\) and \(\hat{\Theta} = \hat{\lambda} = \pm \sqrt{(1-R)/(1+R)}\).
4. Simulation study

Typically, studies of the comparative performance of confidence intervals rely on simulations. In this section, we have performed a small simulation study to compare coverage probability and expected length of the ACI and BCI methods, for constructing confidence intervals for \( \lambda = \sqrt{(1 - \rho)/(1 + \rho)} \) (of course a similar study can be provided for \( \lambda = -\sqrt{(1 - \rho)/(1 + \rho)} \)). Samples of size \( n = 15, 30, 40, 50, 80, 100, 500, 1000 \) were simulated from the bivariate normal distribution \( N_2(0, 0, 1, 1, \rho) \) for the values \( \rho = -0.9, -0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8, 0.9 \) of the correlation coefficient. For each sample size \( n \) and each value of \( \rho \), we generate 1000 ACI and BCI intervals for the parameter \( \lambda \) and then we compute AVL and AVU, the average lower and upper confidence bounds, CP, the actual coverage probability of the two-sided confidence intervals (obtained as the ratio of the number of intervals containing the true values over the total number of simulations) and EL, the estimate of the expected length. Then for these values of \( n \) and \( \rho \), the bootstrap distribution of \( \hat{\lambda}^* = \sqrt{(1 - R)/(1 + R)} \) was calculated, based on \( B = 1000 \) bootstrap replications. Partial results of the simulation study are reported in Tables 1 and 2 and are summarized in Figure 1.

A confidence interval with a narrower expected length implies a more accurate estimate of the parameter and thus is always preferred to a longer one. The actual coverage probability should be near the nominal coverage 0.95. As Tables 1, 2 and other data not presented indicate, for \( n \geq 50 \), the simulation study gives similar results for the two methods and for different values of \( \rho \), both in terms of coverage probability and expected length. For small and moderate sample sizes the ACI has actual coverage probabilities reasonably close to the nominal value of 0.95. In contrast, the intervals based on the BCI method have poor coverage when \( n \) is small or moderate. We note that, in general, the bootstrap method has a coverage rate slightly less than 95%. As expected, with larger sample sizes the confidence intervals become narrower.

For both methods, the expected length becomes larger for negative values of \( \rho \). This behaviour is particularly evident when \( \rho \) is closed to \(-1\) and \( n \) small or moderate. This is not surprising and is in agreement with other results in literature. In fact, as \( \rho \to -1, \lambda \to \infty \) and estimation problems

<table>
<thead>
<tr>
<th>( \rho = 0.5, \lambda = 0.5774 )</th>
<th>n</th>
<th>Method</th>
<th>AVL</th>
<th>AVU</th>
<th>EL</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>ACI</td>
<td>0.3337</td>
<td>1.0347</td>
<td>0.7010</td>
<td>0.9390</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.3267</td>
<td>1.0322</td>
<td>0.7055</td>
<td>0.9180</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>ACI</td>
<td>0.3963</td>
<td>0.8427</td>
<td>0.4464</td>
<td>0.944</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.3965</td>
<td>0.8293</td>
<td>0.4328</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>ACI</td>
<td>0.4196</td>
<td>0.7992</td>
<td>0.3797</td>
<td>0.952</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.4176</td>
<td>0.7934</td>
<td>0.3758</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>ACI</td>
<td>0.4339</td>
<td>0.7687</td>
<td>0.3347</td>
<td>0.9470</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.4335</td>
<td>0.7619</td>
<td>0.3283</td>
<td>0.9380</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>ACI</td>
<td>0.4646</td>
<td>0.7263</td>
<td>0.2617</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.4639</td>
<td>0.7234</td>
<td>0.2595</td>
<td>0.948</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>ACI</td>
<td>0.4732</td>
<td>0.7045</td>
<td>0.2313</td>
<td>0.9540</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.4725</td>
<td>0.7017</td>
<td>0.2292</td>
<td>0.9410</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>ACI</td>
<td>0.5286</td>
<td>0.6302</td>
<td>0.1016</td>
<td>0.9470</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.5282</td>
<td>0.6298</td>
<td>0.1017</td>
<td>0.9500</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>ACI</td>
<td>0.5424</td>
<td>0.6141</td>
<td>0.0717</td>
<td>0.9480</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>0.5422</td>
<td>0.6140</td>
<td>0.0718</td>
<td>0.9470</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Results of the simulations with $\rho = -0.5$: AVL (average of the lower confidence bound), AVU (average of the upper confidence bound), EL (average of the length) and CP (coverage probability).

<table>
<thead>
<tr>
<th>$n$</th>
<th>Method</th>
<th>AVL</th>
<th>AVU</th>
<th>EL</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>ACI</td>
<td>1.0574</td>
<td>3.2786</td>
<td>2.2212</td>
<td>0.9490</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.1078</td>
<td>3.4013</td>
<td>2.2935</td>
<td>0.9150</td>
</tr>
<tr>
<td>30</td>
<td>ACI</td>
<td>1.2247</td>
<td>2.6041</td>
<td>1.3794</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.2495</td>
<td>2.6192</td>
<td>1.3697</td>
<td>0.922</td>
</tr>
<tr>
<td>40</td>
<td>ACI</td>
<td>1.2877</td>
<td>2.4530</td>
<td>1.1653</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.3036</td>
<td>2.4604</td>
<td>1.1567</td>
<td>0.929</td>
</tr>
<tr>
<td>50</td>
<td>ACI</td>
<td>1.3202</td>
<td>2.3386</td>
<td>1.0184</td>
<td>0.9570</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.3360</td>
<td>2.3421</td>
<td>1.0061</td>
<td>0.9370</td>
</tr>
<tr>
<td>80</td>
<td>ACI</td>
<td>1.4027</td>
<td>2.1927</td>
<td>0.7899</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.4115</td>
<td>2.1967</td>
<td>0.7852</td>
<td>0.941</td>
</tr>
<tr>
<td>100</td>
<td>ACI</td>
<td>1.4265</td>
<td>2.1238</td>
<td>0.6973</td>
<td>0.9480</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.4332</td>
<td>2.1260</td>
<td>0.6928</td>
<td>0.9440</td>
</tr>
<tr>
<td>500</td>
<td>ACI</td>
<td>1.5850</td>
<td>1.8897</td>
<td>0.3047</td>
<td>0.9560</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.5855</td>
<td>1.8912</td>
<td>0.3058</td>
<td>0.9560</td>
</tr>
<tr>
<td>1000</td>
<td>ACI</td>
<td>1.6299</td>
<td>1.8453</td>
<td>0.2154</td>
<td>0.9530</td>
</tr>
<tr>
<td></td>
<td>BCI</td>
<td>1.6301</td>
<td>1.8463</td>
<td>0.2161</td>
<td>0.9540</td>
</tr>
</tbody>
</table>

Figure 1. Results of the simulation study.
can arise. As expected, the simulation study confirms that, for all sample sizes, the length of the interval decreases as $\rho$ increases.

5. Examples

To highlight the applicability of the method presented in Section 2 we consider two situations of different nature.

5.1 $PM_{10}$ concentrations

In environmental or epidemiological studies it is relevant to estimate the distribution of extreme statistics. If you are monitoring the pollution in different areas of a region or a town it is important to model appropriately the order statistic maximum and/or minimum or the range. In this example, we analyse data from $PM_{10}$ concentrations recorded daily between the 1 December 2003 and the 1 February 2005 in two different stations in Cagliari, Italy. After removing missing values, from each station we have 111 observations. Our interest rests on the natural logarithm of the maximum value of $PM_{10}$ concentrations in the two stations. We assume that their joint distribution is bivariate normal. We standardized the variables using the MLEs of the unknown means and standard deviations. We are interested in the distribution of the maximum of such standardized random variables. The conditions of Theorem 2.1 are satisfied. Then we know that our extreme statistic has a SN distribution with location parameter equal to 0, scale parameter equal to 1 and skewness parameter $\lambda$ equal to $-\sqrt{(1 - \rho)/(1 + \rho)}$. To evaluate a confidence interval for $\lambda$ we apply the procedure described in Section 2. In order to check the fit of the bivariate normal distribution to the data we use the Shapiro–Wilk multivariate normality test (see [11]). This is based on the Shapiro–Wilk statistic defined as the ratio of two estimates of the variance of a normal distribution based on a random sample of ordered $n$ observations $y_1 \leq y_2 \leq \cdots \leq y_n$. Analytically, $W = (\sum a_i y_i)^2 / \sum (y_i - \bar{y})^2$ where $a = (a_1, \ldots, a_n)^T$ is such that $(n - 1)^{1/2} \sum a_i y_i$ is the best unbiased estimate of the standard deviation of the $y_i$ assuming normality. The observed value is $W = 0.9823$ and the corresponding $p$-value is 0.1474. The estimated value for $R$ is 0.5145. In Table 3 (left size) are reported the estimated value of $\lambda$ together with an approximate 95% confidence interval. This confidence interval is then compared with the bootstrap interval constructed as described in Section 3. ACI provides slightly better results than BCI (assuming that $\hat{\lambda}$ is the true value of $\lambda$).

5.2 Creatinine clearance

In our second example, we consider a data set concerning the follow-up of 145 patients who had operation for renal cancer, in the University hospital of Strasbourg. The follow-up consists of several medical examinations (1, 3, 6, 12 and 24 months after the operation) with blood tests

<table>
<thead>
<tr>
<th>Method</th>
<th>CI</th>
<th>Length</th>
<th>CI</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACI</td>
<td>(0.4688, 0.6836)</td>
<td>0.2148</td>
<td>(0.1655, 0.2300)</td>
<td>0.0645</td>
</tr>
<tr>
<td>BCI</td>
<td>(0.4401, 0.6701)</td>
<td>0.23</td>
<td>(0.1607, 0.2352)</td>
<td>0.07454</td>
</tr>
</tbody>
</table>
and further investigations. Glomerular filtration rate is a measure of renal function using the flow rate of filtered fluid through the kidney. Creatinine clearance rate is the volume of blood plasma that is cleared of creatinine per unit time and is a common measure for approximating the glomerular filtration rate. When the patient has his medical consultation 6 months after operation, the maximum of the creatinine clearance rate between the value at 1 month and the value at 3 months can be considered as the value of his new renal function after recovery.

Statistical tests confirm that the two measures (creatinine clearance at 1 month and at 3 months) have the same mean and the same variance. The Shapiro–Wilks test cannot reject the hypothesis that the distribution is joint normal ($W = 0.9893$ and $p$-value is 0.3302). The estimated correlation coefficient is $R = 0.9266$. Table 3 (right side) summarizes the point estimate of $\lambda$ and its confidence interval at the 95% level using Fisher’s transformation and the bootstrap technique. As previously, the length of this interval is narrower with the theoretical approximate method than with the bootstrap.

6. Discussion and conclusions

We show that, in the specific case described in the paper, the problem of finding the maximum likelihood estimate of the skewness parameter, which in general is not an easy task, can be solved easily using the Fisher Transformation if the parameters of the bivariate normal are supposed to be all unknown and we fix both means and standard deviations equal to their MLEs. The main goal of this paper is to compute confidence intervals for the skewness parameter $\lambda$. It is well known that the Fisher Transformation is adequate for $n > 50$ and that this approximation is more accurate, for small $n$, when $\rho$ is close to zero [5]. This begs the question as to how well the ACI method perform when $|\rho|$ is close to 1 and the sample size $n$ is small or moderate. To investigate this dependence we conduct a simulation study to compare the ACI with another procedure to construct confidence intervals, the percentile parametric BCI method. Comparison of the performance of the confidence intervals is conducted in terms of their: (1) coverage probability and (2) length. The simulation study reveals that ACI performs better in terms of coverage probability. We see that, for the most part, actual coverage levels vary but the ACI coverage is little larger and closer to the nominal coverage. The differences between the two methods are particularly important for small and moderate sample sizes. For example, for $n = 15$ and $\rho = -0.2$ the 95% confidence intervals cover the true value only the 91% of times if we use the BCI and the 94.2% using the ACI. The two methods are comparable in terms of expected length. Results provided allow us to adopt the ACI procedure to compute confidence intervals for $\lambda$. The approximation used is good enough to lead us always to prefer the ACI method. Results are not satisfactory in terms of expected width when $n$ is small and $\rho$ is closed to $-1$.

The main advantages of the ACI method are that it is based on a theoretical approximation, and it gives rapid solutions, even for very large sample sizes $n$, whereas the percentile bootstrap can take hours. Results of both examples are in agreement with the findings of the simulation study. The approximate method proposed here could also be applied to other types of data, for instance, to data coming from double measurements with the same instrument, such as spirometry. Another potential application is epidemiological studies on twins [9].

Acknowledgements

We are grateful to the editor and the referees for helpful and constructive comments. Authors would like to thank Prof. Herv Lang and Dr. Charles Thomas from the University hospital of Strasbourg, for providing the creatine clearance data set. This work was supported by the project start up giovani ricercatori of the University of Cagliari, and partially to the visiting professor program of ‘Regione Autonoma della Sardegna’. 

Downloaded by [Universita Degli Studi di Cagliari], [Monica Musio] at 02:00 14 March 2012
References