Journal of Applied Statistics, Vol. 27, No. 7, 2000, 859–870

Problems of inference for Azzalini's skew-normal distribution

ARTHUR PEWSEY, *Departamento de Matemáticas, Universidad de Extremadura, Cáceres, Spain*

ABSTRACT *This paper considers various unresolved inference problems for the skew-normal distribution. We give reasons as to why the direct parameterization should not be used as a general basis for estimation, and consider method of moments and maximum likelihood estimation for the distribution's centred parameterization. Large sample theory results are given for the method of moments estimators, and numerical approaches for obtaining maximum likelihood estimates are discussed. Simulation is used to assess the performance of the two types of estimation. We also present procedures for testing for departures from the limiting folded normal distribution. Data on the percentage body fat of elite athletes are used to illustrate some of the issues raised.*

1 Introduction

Azzalini (1985) introduces the skew-normal class as one being able to reflect varying degrees of skewness, which is mathematically tractable and which includes the normal distribution as a special case. A random variable X has a (standard) skew-normal distribution with parameter λ, $X \sim SN(\lambda)$, if its density is of the form

$$\varphi(x;\lambda) = 2\phi(x)\Phi(\lambda x), \qquad -\infty < x < \infty, -\infty < \lambda < \infty,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and distribution functions, respectively. The case $\lambda = 0$ corresponds to the standard normal distribution. As $\lambda \to \infty$, the distribution tends to the standard folded normal distribution, i.e. that of $X = |Z|$, where $Z \sim N(0, 1)$. The distribution $SN(-\lambda)$ is the reflection of $SN(\lambda)$ in $x = 0$, and hence as $\lambda \to -\infty$ the limiting density is the standard negative folded normal distribution, i.e. that of $X = -|Z|$. The densities for certain relevant values

Correspondence: A. R. Pewsey, Departamento de Matemáticas, Escuela Politécnica, Universidad de Extremadura, 10071, Cáceres. Spain.

ISSN 0266-4763 print; 1360-0532 online/00/070859-12

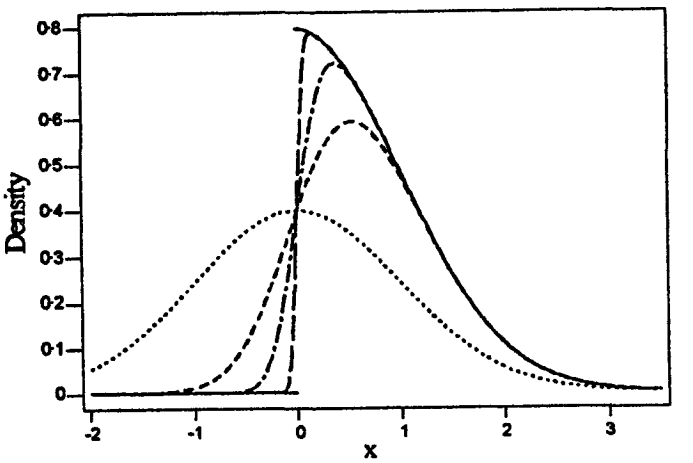

FIG. 1. Densities for various cases of the skew-normal distributions: -----, $SN(0) \equiv N(0,1)$; – – –, $SN(2)$; — — —, $SN(5)$; ———, $SN(20)$; ———, standard folded normal.

of λ are represented in Fig. 1. For $\lambda = 2$ the density is moderately asymmetric, whilst for $\lambda = 20$ the density differs little from the limiting standard folded normal distribution.

The class can be generalized by the inclusion of location and scale parameters which we identify here as ξ and η, respectively. Thus, if $X \sim SN(\lambda)$ then $Y = \xi + \eta X$ is a skew-normal random variable. Following Azzalini & Capitanio (1999) we refer to (ξ, η, λ) as the 'direct' parameterization and denote the distribution of Y by $SN_D(\xi, \eta, \lambda)$. As $\lambda \to \pm\infty$ the limiting distribution is the (negative) folded normal distribution with parameters ξ and η, defined by replacing X by a standard (negative) folded normal random variable in the definition of Y. Using the results from Azzalini (1985), the first three moments and variance of Y are:

$$\begin{aligned} E(Y) &= \xi + b\eta\delta, \; E(Y^2) = \xi^2 + 2b\xi\eta\delta + \eta^2 \\ E(Y^3) &= \xi^3 + 3b\xi^2\eta\delta + 3\xi\eta^2 + 3b\eta^3\delta - b\eta^3\delta^3 \\ \text{var}(Y) &= \eta^2(1 - b^2\delta^2) \end{aligned} \tag{1}$$

where $b = (2/\pi)^{\frac{1}{2}}$ and $\delta = \lambda/(1+\lambda^2)^{\frac{1}{2}} \in (-1, 1)$. The coefficient of skewness for Y is the same as that for X, namely

$$\gamma_1 = b\delta^3(2b^2 - 1)/(1 - b^2\delta^2)^{3/2} \in (-0.99527, 0.99527).$$

In this paper we consider various issues of inference for the skew-normal distribution. In Section 2, we explain why the direct parameterization should not be used as a general basis for estimation, and in Section 3 we consider the details of method of moments (MM) and maximum likelihood (ML) estimation under Azzalini's second, 'centred', parameterization.The results of a simulation study into the performance of these two forms of estimation are presented in Section 4. In Section 5 we consider limiting cases of the distribution and introduce new test procedures that may be applied in the search for more parsimonious models. The paper concludes with a consideration of some real data.

2 Rejection of the direct parameterization as a general basis for estimation

2.1 Method of moments

Let $Y_n = (y_1, \ldots, y_n)$ denote a random sample of n observations from a $SN_D(\xi, \eta, \lambda)$ distribution, with sample moments $m_1 = 0, m_2 = s^2, \ldots$ about the mean. We denote the moment estimates of ξ, η and δ by $\tilde{\xi}$, $\tilde{\eta}$, $\tilde{\delta}$, respectively. For simplicity, consider the standardized sample $Y_{nS} = (y_{S1}, \ldots, y_{Sn})$ where $y_{Si} = (y_i - \bar{y})/s$, $i = 1, \ldots, n$; this is a sample from a $SN_D(\xi_S, \eta_S, \lambda)$ distribution with $\xi_S = (\xi - \bar{y})/s$ and $\eta_S = \eta/s$. Equating the first three sample moments of the standardized data to their population counterparts from (1), the MM estimates of ξ_S, η_S and δ are

$$\tilde{\xi}_S = -cm_3^{1/3}/s, \times \tilde{\eta}_S = (1 + \tilde{\xi}_S^2)^{\frac{1}{2}} \times \text{ and } \tilde{\delta} = -\tilde{\xi}_S/b\tilde{\eta}_S, \tag{2}$$

where $c = \{2/(4-\pi)\}^{1/3}$. Then, $\tilde{\lambda} = \tilde{\delta}/(1-\tilde{\delta}^2)^{\frac{1}{2}}$, provided $|\tilde{\delta}| < 1$. Otherwise, $\tilde{\delta}$ is inadmissible and $\tilde{\lambda}$ is undefined. The estimates of ξ and η can be recovered using

$$\tilde{\xi} = \bar{y} + s\tilde{\xi}_S \quad \text{and} \quad \tilde{\eta} = s\tilde{\eta}_S. \tag{3}$$

Both the MM and ML estimates of the location and scale parameters for the normal distribution are, of course, $\bar{y}$ and s. Thus, if we use (2) and (3) to estimate the parameters of an assumed skew-normal distribution, when the data come from a normal population, we will tend to over-estimate η and over- or under-estimate ξ and δ, depending on the sign of $\tilde{\xi}_S$. In practice, the sampling distributions of $\tilde{\xi}$ and $\tilde{\delta}$ are bimodal for values of λ around zero. As n and $|\lambda|$ increase, the sampling distributions of all three estimators tend to unimodal distributions. Those of $\tilde{\xi}$ and $\tilde{\delta}$ for the normal case are bimodal whatever the sample size.

2.2 Maximum likelihood

Azzalini (1985) gives the Fisher information matrix for the direct parameterization and notes that it is singular for $\lambda = 0$. We trace this singularity to the parameter redundancy of the parameterization for the normal case, a fact easily identified using the results of Catchpole & Morgan (1997). They identify an exponential-family model as being parameter redundant if the mean can be expressed using a reduced number of parameters. From equation (1), $E(Y)$ is a function of all three parameters, whereas for the normal case it is just ξ. The singularity of the information matrix then follows from Remark 4 of Catchpole & Morgan (1997). According to Theorem 2 of the same paper, the likelihood surface for the normal case must contain a completely flat ridge. Thus, if we were to attempt to maximize the log-likelihood for this parameterization using numerical techniques, the results obtained could be highly misleading as, for this case, no unique solution exists. These consequences, and our findings for MM estimation, rule the direct parameterization out as a general basis from which to conduct estimation. They also shed a different light on previously published work concerning the estimation of the direct parameters of the skew-normal distribution; see for example Arnold *et al.*, 1993.

3 Estimation for the centred parameterization

Having identified the singularity problem associated with ML estimation for the direct parameterization, Azzalini (1985) introduces the 'centred' parameterization,

$(\mu, \chi, \sigma, \gamma_1)$. He defines a skew-normal random variable Y with $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2$ by

$$Y = \mu + \sigma\{X - E(X)\}/\{\text{var}(X)\}^{\frac{1}{2}}, \qquad -\infty < \mu < \infty, \sigma > 0,$$

where X is a $SN(\lambda)$ random variable. The parameter γ_1 is the coefficient of skewness of X, and hence also that of Y. We will denote the distribution of Y by $SN_C(\mu, \sigma, \gamma_1)$. Clearly, as $E(Y) = \mu$, this parameterization is not parameter redundant for the normal case.

3.1 Method of moments

MM estimation is trivial for the centred parameterization. As,

$$E(Y) = \mu, E(Y^2) = \mu^2 + \sigma^2 \quad \text{and } E(Y^3) = \mu^3 + 3\mu\sigma^2 + \sigma^3\gamma_1$$

the MM estimates are given by

$$\tilde{\mu} = \bar{y}, \tilde{\sigma} = s, \tilde{\gamma}_1 = g_1 = m_3/s^3.$$

The first two of these are also the MM estimates for the parameters of the normal, folded normal and negative folded normal distributions. Inadmissible values for $\tilde{\gamma}_1$, i.e. those for which $|\tilde{\gamma}_1| > 0.99527$, occur often for highly skewed populations. It is natural to interpret such estimates as indicating that a (negative) folded normal distribution is the underlying generating mechanism.

Considering the asymptotic distribution of the estimators, we follow established notation and let $\beta_2 = \mu_4/\mu_2^2$ and $\beta_4 = \mu_6/\mu_2^3$. In an extension of it we define $\beta_3 = \mu_5/\mu_2^{5/2}$. For $Y \sim SN_C(\mu, \sigma, \gamma_1)$,

$$\beta_2 = 3 + 2\tau^4(\pi - 3)$$

$$\beta_3 = 10\gamma_1 + \tau^5(3\pi^2 - 40\pi + 96)/4$$

$$\beta_4 = 15\{1 + \tau^4(2\pi - 6)\} - \tau^6(9\pi^2 - 80\pi + 160)/2$$

where $\tau = c\gamma_1^{1/3}$. Using the δ method (Rao, 1973, p. 388), we obtain:

$$\begin{aligned}
E(\tilde{\mu}) &= \mu, \; E(\tilde{\sigma}) = \sigma\{1 - (3 + \beta_2)/8n\} + O(n^{-3/2}), \\
E(\tilde{\gamma}_1) &= \gamma_1 + 3\{\gamma_1(7 + 5\beta_2) - 4\beta_3\}/8n + O(n^{-3/2}), \\
\text{var}(\tilde{\mu}) &= \sigma^2/n, \quad \text{var}(\tilde{\sigma}) = \sigma^2(\beta_2 - 1)/4n + O(n^{-3/2}), \\
\text{var}(\tilde{\gamma}_1) &= \{9 - 6\beta_2 - 3\gamma_1\beta_3 + \beta_4 + \gamma_1^2(35 + 9\beta_2)/4\}/n + O(n^{-3/2}), \qquad (4) \\
\text{cov}(\tilde{\mu}, \tilde{\sigma}) &= \sigma^2\gamma_1/2n + O(n^{-3/2}), \\
\text{cov}(\tilde{\mu}, \tilde{\gamma}_1) &= \sigma(\beta_2 - 3 - 3\gamma_1^2/2)/n + O(n^{-3/2}) \\
\text{cov}(\tilde{\sigma}, \tilde{\gamma}_1) &= \sigma\{2\beta_3 - \gamma_1(5 + 3\beta_2)\}/4n + O(n^{-3/2}).
\end{aligned}$$

Simulation confirms that all of these asymptotic results, apart from that for $\text{var}(\tilde{\gamma}_1)$, provide very good approximations, even for sample sizes as small as 20. Sample sizes of 50 or more are required before the expression for $\text{var}(\tilde{\gamma}_1)$ can reasonably be applied. In view of equation (4), the joint distribution of the estimators is asymptotically trivariate normal. However, it is known that the sampling distribu-

tion of $\hat{\gamma}_1$ tends to normality very slowly even for data from normal populations; see, for example, Pearson (1963) and D'Agostino (1970). For data from highly skewed populations the sampling distribution of $\hat{\gamma}_1$ is skewed, even for very large n. When asymptotic theory does not apply, inference can be based upon computer intensive methods such as Monte Carlo significance testing and the parametric bootstrap.

3.2 Maximum likelihood

Consider a random sample Y_n from the $SN_C(\mu, \sigma, \gamma_1)$ distribution and its standardized counterpart Y_{nS} from the $SN_C(\mu_S, \sigma_S, \gamma_1)$ distribution. The constraint

$$\hat{\eta} - \left\{\sum_{i=1}^{n} (y_i - \hat{\xi})^2/n\right\}^{\frac{1}{2}}$$

on the ML estimates for the direct parameterization leads to the constraint

$$\hat{\sigma}_S = \{\hat{\mu}_S^2(1 + \hat{t}^2) + 1\}^{\frac{1}{2}} - \hat{\mu}_S\hat{t}$$

where $\hat{t} = c\hat{\gamma}_1^{1/3}$, and the ML estimates, $\hat{\mu}_S$ and $\hat{\gamma}_1$, are those values that maximize the constrained log-likelihood

$$l(\mu_S, \gamma_1; Y_{nS}) = -n\log[\{\mu_S^2(1 + \tau^2) + 1\}^{\frac{1}{2}} - \mu_S\tau] - \frac{n}{2}\log(1 + \tau^2) + \sum_{i=1}^{n}\log\left\{\Phi\left(\frac{(y_{Si} - \mu_S)\tau}{\{\mu_S^2(1 + \tau^2) + 1\}^{\frac{1}{2}} - \mu_S\tau} + \tau^2\right)\right\} \qquad (5)$$

$-\infty < \mu_S < \infty$, $-0.99527 < \gamma_1 < 0.99527$. The ML estimates of μ and σ are given by $\hat{\mu} = \bar{y} + s\hat{\mu}_S$ and $\hat{\sigma} = s\hat{\sigma}_S$.

Unlike its MM counterpart, an ML estimate is always admissible. However, particularly for small n, and generally for samples drawn from populations with moderate to large values of $|\lambda|$, the global maximum of the log-likelihood can occur for a boundary value of γ_1. Generally, a boundary ML estimate indicates that the set of plausible models lies to only one side of the most likely one; see for example Lindsey (1996, p. 81). For the skew-normal distribution, then, a boundary estimate of γ_1 indicates that a folded normal distribution is the most likely generating mechanism. Subsequent inference for γ_1 is problematic as the conditions required for the attainment of the Crámer–Rao lower bound do not apply on the boundary of the parameter space. Depending on whether γ_1 is of interest, we might carry out ML-based inference for a folded normal distribution or apply the computer intensive methods referred to in Section 3.1. We see the re-estimation approach to dealing with boundary estimates of Azzalini & Capitanio (1999, p. 591) as one based on an interpretation of the likelihood which is difficult to defend on objective grounds.

Azzalini & Capitanio (1999) give a WWW address from which two S-PLUS routines for fitting the skew-normal distribution using maximum likelihood estimation can be obtained. The routine sn.em uses the EM algorithm to maximize the log-likelihood whilst sn.mle employs gradient based methods. Both routines use

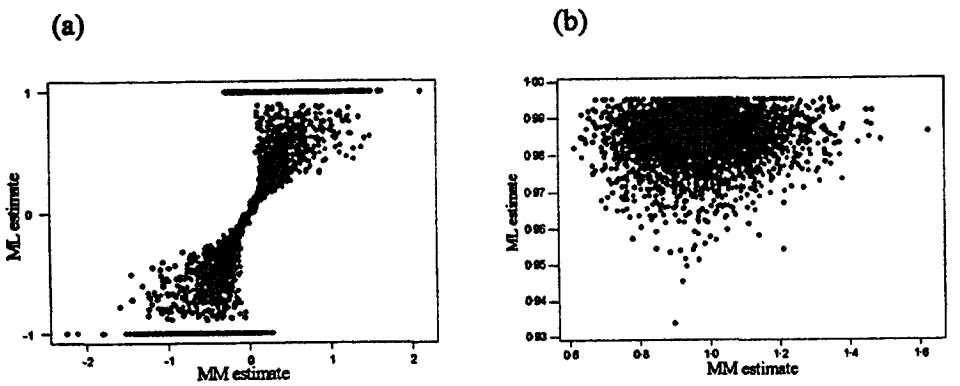

FIG. 2. Scatterplots of the ML versus the MM estimate of γ_1 for 5000 simulated samples of size: (a) 20 from the $N(0,1) \equiv SN_C(0,1,0)$ distribution; (b) 500 from the $SN_D(0, 1, 20) \equiv SN_C(0.7969, 0.6041, 0.9851)$ distribution.

the MM estimates as default starting values. We note that, whilst $\tilde{\mu}$ usually provides a good initial estimate of $\hat{\mu}$, in general $\tilde{\gamma}_1$ and $\hat{\gamma}_1$ are not strongly related, particularly if n is small or $|\gamma_1|$ is large. Another starting value needs to be used if $\tilde{\gamma}_1$ is inadmissible. Figure 2 illustrates the lack of any clear relation between $\hat{\gamma}_1$ and $\tilde{\gamma}_1$ using estimates obtained from simulated samples of size 20 from the $N(0,1)$ distribution and of size 500 from the $SN_D(0,1,20) \equiv SN_C(0.7969, 0.6041, 0.9851)$ distribution. The symbols forming the horizontal lines in both scatterplots correspond to boundary ML estimates.

Experience shows that if the MM estimates are used as starting values then both routines can converge to a local, rather than the global, maximum of the log-likelihood. It has long been known that log-likelihood surfaces can contain multiple maxima, the problem being most acute for small samples. Schemes for identifying the global maximum in such circumstances date back to the early computational work of Barnett (1966). A standard approach is to use a grid of starting values in an attempt to ensure that the true global maximum is identified. An optional argument of the routine sn.mle allows the user to specify their own starting values and so can be used to carry out a grid search. The routine sn.em does not have this argument and so its use is strictly limited. Moreover, it can be very slow to execute and, more problematically, is based upon optimization for the direct parameterization which we identified in Section 2.2 as being parameter redundant for the normal case.

Our approach to finding ML estimates is based upon the optimization of equation (5) using the simplex algorithm of Nelder & Mead (1965). We use the starting value $\tilde{\mu}_S = 0$ and a grid of starting values for γ_1 spanning its full range. So as to ensure that the true global maximum has been identified, the maximum value found during optimization is contrasted with the values of (5) on the γ_1 boundary, namely $-\frac{1}{2}n\log(1+y^2_{s(1)})$ and $-\frac{1}{2}n\log(1+y^2_{s(n)})$, where $y_{S(1)}$ and $y_{S(n)}$ denote the minimum and maximum values of the standardized data, respectively. A simulation study incorporating this approach showed that multiple maxima occur with greatest frequency for small samples from normal populations. The simulated sample in Table 1 provides a case in point. The log-likelihood associated with this sample has a local maximum at ($\mu = 0.01$, $\sigma = 1.31$, $\gamma_1 = 0.10$) and a global maximum at ($\mu = -0.03$, $\sigma = 1.36$, $\gamma_1 = 0.73$). Using the default starting values, both sn.mle

TABLE 1. Simulated sample of size 20 from the $SN_D(0,1,0) \equiv N(0,1)$ distribution

2.583	1.343	−1.107	−1.203	−0.984	1.359	−0.799	1.580	−0.540	−0.918
1.121	0.412	−0.763	1.333	−1.928	0.873	−1.076	1.139	0.050	−2.258

and sn.em converge to the local, rather than the global, maximum. For $n = 20$, approximately 5% of samples from normal populations have log-likelihood surfaces which contain multiple maxima within the parameter space, whilst for samples from skew-normal populations with $\lambda = 20$ the frequency is close to 2%. For sample sizes as large as 500, multiple maxima are generally rare for all but those samples from close to normal populations. Even for this limiting case, their frequency of occurrence is only around 0.4%.

4 Simulation study

In order to compare the performance of MM and ML estimation we conducted a simulation study. Samples of size $n = 20$, 50, 100, 200 and 500 were simulated from the $SN_D(0, 1, \lambda)$ distribution for $\lambda = 0$, 2, 5 and 20. For each (n, λ) combination, 5000 samples were simulated using the method of Henze (1986) in conjunction with the uniform random number generator of Wichman & Hill (1982), incorporating the amendment of McLeod (1985). The simplex algorithm used was an amendment of algorithm AS47 of O'Neill (1971), available from the HENSA archives. Programming was carried out in FORTRAN. As performance measures we used the mean value and mean squared error. For comparative purposes, the measures for $\tilde{\gamma}_1$ were calculated using truncation of inadmissible values to ± 0.99527. For the estimation of γ_1 we also recorded: the percentage of inadmissible MM estimates; the percentage of boundary ML estimates; and the percentage of samples for which the MM estimate was inadmissible and the ML estimate was a boundary estimate.

There was very little difference between the performance measures for the MM and ML estimates of μ and σ. Also, the sampling distributions of $\tilde{\mu}$ and $\hat{\mu}$, and those of $\tilde{\sigma}$ and $\hat{\sigma}$, were very similar. Even for n as small as 20, these distributions were all close to normal.

In Fig. 3 we present the sampling distributions of $\tilde{\gamma}_1$ and $\hat{\gamma}_1$ from the simulated samples of size 20 from the $SN_D(0,1,0)$ and $SN_D(0,1,20)$ distributions. Those given in Fig. 4 correspond to the samples of size 500 from the same two populations. The sampling distribution of $\hat{\gamma}_1$ in Fig. 3(b) is dominated by two spikes produced by boundary estimates. In comparison, that of $\tilde{\gamma}_1$ in Fig. 3(a) is close to normal. Comparing Fig. 4(a) with Fig. 4(b), we see that the sampling distribution of $\hat{\gamma}_1$ approaches that of $\tilde{\gamma}_1$ as n increases. The spikes around 0 in Figs. 3(b) and 4(b) are due to a compressing effect of the γ_1 scale. We note from Fig. 4(c) and (d) that the sampling distributions of $\tilde{\gamma}_1$ and $\hat{\gamma}_1$ are not normal for samples drawn from highly skewed populations, even for n as large as 500.

From Table 2, $\hat{\gamma}_1$ generally outperforms $\tilde{\gamma}_1$, although the performance of $\hat{\gamma}_1$ is inferior for close to symmetric populations, particularly for small samples. The large mean squared error for these cases is consistent with the content of Fig. 3(b). In general, the frequencies of inadmissible and boundary estimates of γ_1 diminish with increasing n and as $\lambda \to 0$. However, boundary ML estimates are still possible for n as large as 500, albeit for samples from highly asymmetric populations. The

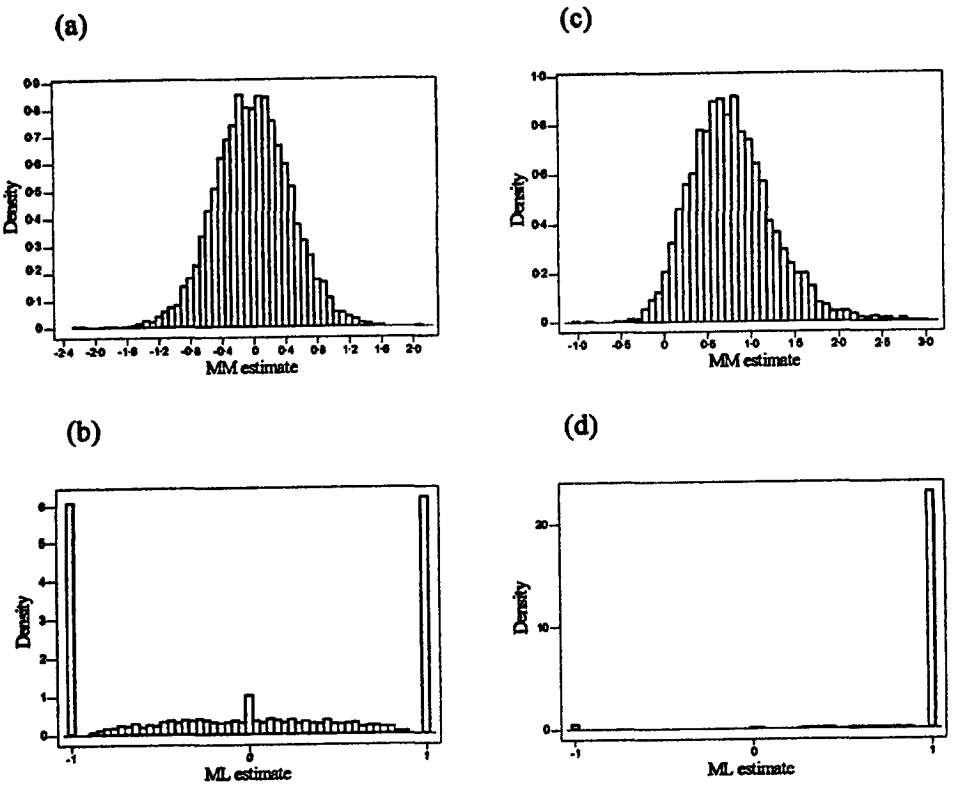

FIG. 3. Sampling distributions of the MM and ML estimates of γ_1 obtained from 5000 simulated samples of size 20: (a), (b) $SN_D(0,1,0) \equiv SN_C(0,1,0)$; (c), (d) $SN_D(0,1,20) \equiv SN_C(0.7969, 0.6041, 0.9851)$.

percentages appearing in the square brackets imply that there is little relation between the occurrence of inadmissible MM and boundary ML estimates of γ_1.

In many practical situations, inference will focus on μ and σ, with γ_1 being a nuisance parameter. Our results indicate that for these situations there is little or no benefit in using ML estimation, and the extreme simplicity of MM estimation strongly favours its adoption. Should a complete specification of the underlying distribution be required, then ML estimation is generally preferable. However, MM estimation performs better for small samples from close to symmetric populations.

5 Tests for limiting cases

The normal, folded normal and negative folded normal distributions warrant special attention as they can be specified in terms of two parameters rather than the three of the skew-normal class. Parsimony dictates that, for data displaying a high degree of asymmetry, or symmetry, we should investigate the appropriateness of the relevant limiting case.

Salvan (1986), has shown that $g_1 \equiv \tilde{\gamma}_1$ is the locally most powerful location and scale invariant statistic for testing for normality within the skew-normal class. Here we consider significance tests for departures from an underlying folded normal distribution. Those for its negative analogue are then obvious. Again, it is natural

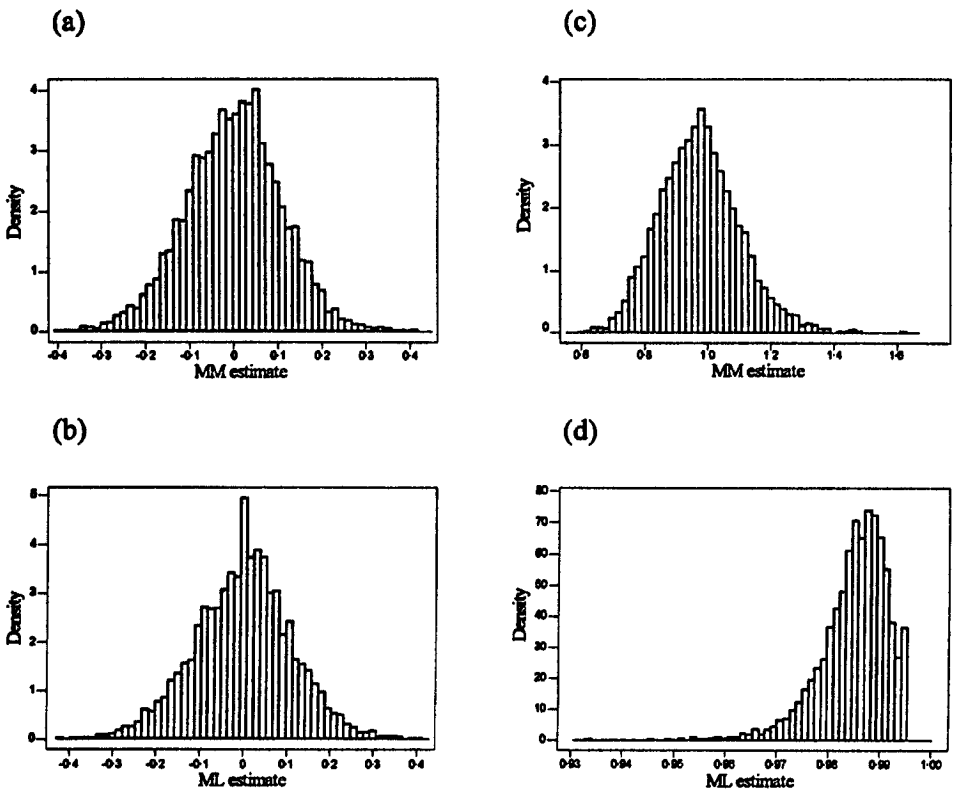

FIG. 4. Sampling distributions of the MM and ML estimates of γ_1 obtained from 5000 simulated samples of size 500: (a), (b) $SN_D(0,1,0) \equiv SN_C(0,1,0)$; (c), (d) $SN_D(0,1,20) \equiv SN_C(0.7969, 0.6041, 0.9851)$.

to consider $\hat{\gamma}_1$ as the basis for such a test procedure. From (4), the asymptotic distribution of $\hat{\gamma}_1$ for data from a folded normal distribution is normal with mean 0.99527 and variance $8.03572n^{-1}$. However, this asymptotic result should be used prudently as the sampling distribution of $\hat{\gamma}_1$ is positively skewed, even for very large samples. In the absence of a better approximation to this sampling distribution, we propose a computer intensive alternative for testing for departures from the folded normal distribution. As Barnard (1963) explains, a Monte Carlo based approach to significance testing is always available as long as data from the null model can be simulated. Here, such a test can be based on the rank of $\hat{\gamma}_1$ for the original data when ordered amongst the values of $\hat{\gamma}_1$ for samples of the same size simulated from the standard folded normal distribution.

Considering once more the percentages in square brackets in Table 2, and the content of Fig. 2(b), we note the following. If we use $\hat{\gamma}_1$ for testing for departures from a folded normal distribution, and base subsequent estimation upon the ML criterion, it is possible that the test might reject the null hypothesis and yet ML estimation could lead to the contradictory result of a boundary estimate for γ_1. Faced with this situation, the data analyst is advised to estimate the parameters using both methods and compare the resultant densities graphically to check for any gross disparities. Should we have reason to think that there is a threshold value associated with the variable of interest then this would favour the adoption of the

TABLE 2. Performance measures for MM and ML estimates of γ_1 from 5000 simulated samples of size n from the $SN_D(0,1,\lambda)$ distribution: mean; (mean squared error); {percentage of inadmissible or boundary estimates, respectively}; [percentage of samples for which MM estimate was inadmissible and ML estimate was a boundary estimate]

Sample size n	$\lambda=0$; $\gamma_1=0$		$\lambda=2$; $\gamma_1=0.4538$		$\lambda=5$; $\gamma_1=0.8510$		$\lambda=20$; $\gamma_1=0.9851$	
	MM	ML	MM	ML	MM	ML	MM	ML
20	−0.0015	0.0035	0.3171	0.4615	0.5806	0.7837	0.6769	0.9338
	(0.2064)	(0.5819)	(0.2081)	(0.4726)	(0.2045)	(0.2363)	(0.1947)	(0.0905)
	{3.74}	{49.16}	{8.86}	{54.74}	{21.70}	{77.24}	{29.40}	{95.08}
	[2.40]		[6.34]		[19.02]		[28.82]	
50	0.0014	−0.0003	0.3895	0.4464	0.7085	0.8448	0.7951	0.9811
	(0.1045)	(0.1812)	(0.1078)	(0.1351)	(0.0850)	(0.0405)	(0.0826)	(0.0049)
	{0.32}	{4.86}	{5.12}	{9.42}	{22.72}	{34.72}	{33.82}	{86.86}
	[0.12]		[1.14]		[10.02]		[30.26]	
100	−0.0002	−0.0012	0.4258	0.4467	0.7674	0.8474	0.8583	0.9851
	(0.0560)	(0.0731)	(0.0643)	(0.0576)	(0.0461)	(0.0138)	(0.0406)	(0.0007)
	{0.02}	{0.06}	{2.34}	{0.78}	{22.58}	{8.80}	{37.74}	{65.70}
	[0]		[0.02]		[2.42]		[25.34]	
200	−0.0013	−0.0015	0.4391	0.4473	0.8072	0.8499	0.8954	0.9857
	(0.0297)	(0.0339)	(0.0344)	(0.0266)	(0.0246)	(0.0054)	(0.0219)	(0.0002)
	{0}	{0}	{0.40}	{0}	{18.10}	{0.54}	{38.92}	{32.78}
	[0]		[0]		[0.22]		[13.08]	
500	−0.0002	−0.0004	0.4460	0.4505	0.8365	0.8494	0.9347	0.9852
	(0.0118)	(0.0122)	(0.0144)	(0.0104)	(0.0120)	(0.0019)	(0.0083)	(0.0000)
	{0}	{0}	{0.02}	{0}	{11.28}	{0}	{41.68}	{3.30}
	[0]		[0]		[0]		[1.26]	

ML estimates. If the contrary were true, we would side with those from MM estimation.

6 An example with real data

Often, the skewness of a sample distribution is a consequence of mixing data from two or more sub-populations. If this is the case, then a generally more informative analysis results from modelling the variable of interest within the various sub-samples rather than treating the data as a sample drawn from a single population. The data set we consider here provides a case in point. The data, taken from Cook & Weisberg (1994), consist of percentage body fat measurements made on 102 male and 100 female elite athletes representing ten different sports with highly disparate physiological demands. The athletes concerned trained at the Australian Institute of Sport. Figure 5 is a histogram of the data in which shading has been used to represent the density of male athletes in each class interval. The superimposed densities are those obtained from fitting the skew-normal distribution to the data treated as a single sample using MM and ML estimation. The corresponding estimates of the direct and centred parameters are:

$$\tilde{\xi}=6.04,\ \hat{\xi}=5.63;\ \tilde{\eta}=9.69,\ \hat{\eta}=10.00;\ \tilde{\lambda}=3.73,\ \hat{\lambda}=+\infty$$

$$\text{and } \tilde{\mu}=13.51,\ \hat{\mu}=13.62;\ \tilde{\sigma}=6.17,\ \hat{\sigma}=6.03;\ \tilde{\gamma}_1=0.76,\ \hat{\gamma}_1=0.99527.$$

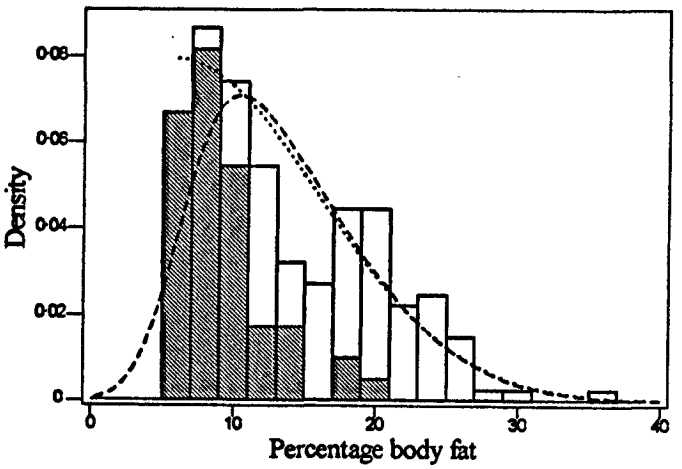

FIG. 5. Histogram of the percentage body fat data with superimposed skew-normal densities fitted via the method of moments (dashed curve) and maximum likelihood (dotted curve). Shaded rectangles represent the density for male athletes: non-shaded rectangles, the density for females.

We note the vast difference, both numerically and interpretationally, between $\tilde{\lambda}$ and $\hat{\lambda}$. The value for $\tilde{\lambda}$ of 3.73 would not lead us to suspect that a folded normal distribution was the underlying generating mechanism for the data, whereas the message from $\hat{\lambda}$ is clearly that it is. When we applied the Monte Carlo significance test of Section 5 with 4999 simulated samples we found that the value 0.76 corresponded to the 13th percentile of the sampling distribution of $\tilde{\gamma}_1$. Thus, according to the observed value of $\tilde{\gamma}_1$, the folded normal distribution is a possible, though unlikely, model for the data.

Whilst the forms taken by the two fitted densities in Fig. 5 are very similar across approximately 85% of the range of the data, the differences between their lower tails are important. It is known from the sports physiology literature that elite athletes seldom have less than 5% body fat. Generally, a minimum of between 3 and 4% body fat is necessary in order merely to survive. Actually, the minimum percentage body fat measurement within the sample is 5.63, which is the threshold value for ξ fitted under ML estimation. Thus, the MM solution ascribes non-zero probability to physically unattainable measurements whilst the ML solution appears to overestimate the threshold. Of course, we could have included the background information in the estimation process and fitted a folded normal distribution to the data with the value of ξ constrained to be some hard threshold value of, say, 3%.

Although we have modelled the data as a sample from a single population, the values for the male and female athletes are markedly different. The generally higher percentage body fat measurements for the female athletes are in keeping with results from research in sports physiology; see for example Wilmore & Costill (1994, Ch. 16). Obviously, a major part of the skewness within the data results from mixing the data for the two sexes. Splitting the data further by sport type provides greater insight into the factors influencing their values.

Acknowledgements

I would like to thank Adelchi Azzalini for supplying me with an extended version of Azzalini & Capitanio (1999) and stimulating communication via e-mail. Special

thanks go to Toby Lewis for introducing me to the skew-normal distribution and for his enthusiasm and support during the research which led to this paper. His comments, and those of a referee, improved the paper considerably.

REFERENCES

ARNOLD, B. C., BEAVER, R. J., GROENEVELD, R. A. & MEEKER, W. Q. (1993) The nontruncated marginal of a truncated bivariate normal distribution, *Psychometrika*, 58, pp. 471–488.

AZZALINI, A. (1985) A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, 12, pp. 171–178.

AZZALINI, A. & CAPITANIO, A. (1999) Statistical applications of the multivariate skew-normal distribution, *Journal of the Royal Statistical Society, Series B*, 61, pp. 579–602.

BARNARD, G. A. (1963) Comment on the paper by M. S. Bartlett, *Journal of the Royal Statistical Society, Series B*, 25, p. 294.

BARNETT, V. D. (1966) Evaluation of the maximum likelihood estimator where the likelihood equation has multiple roots, *Biometrika*, 53, pp. 151–165.

CATCHPOLE, E. A. & MORGAN, B. J. T. (1997) Detecting parameter redundancy, *Biometrika*, 84, pp. 187–196.

COOK, R. D. & WEISBERG, S. (1994) *An Introduction to Regression Graphics* (New York, Wiley).

D'AGOSTINO, R. B. (1970) Transformations to normality of the null distribution of g_1, *Biometrika*, 57, pp. 679–681.

HENZE, N. (1986) A probabilistic representation of the 'skew-normal' distribution, *Scandinavian Journal of Statistics*, 13, pp. 271–275.

LINDSEY, J. K. (1996) *Parametric Statistical Inference* (Oxford, Oxford University Press).

MCLEOD, A. I. (1985) A remark on AS183. An efficient and portable pseudo-random number generator, *Applied Statistics*, 34, pp. 198–200.

NELDER, J. A. & MEAD, R. (1965) A simplex method for function minimization, *Computer Journal*, 7, pp. 308–313.

O'NEILL, R. (1971) Algorithm AS47: function minimization using a simplex procedure, *Applied Statistics*, 20, pp. 338–345.

PEARSON, E. S. (1963) Some problems arising in approximating to probability distributions, using moments, *Biometrika*, 50, pp. 95–111.

RAO, C. R. (1973) *Linear Statistical Inference and its Applications*, 2nd Edn (New York, Wiley).

SALVAN, A. (1986) Locally most powerful invariant tests of normality, in: *Atti della XXXIII Riunione Scientifica della Società Italiana di Statistica*, 2, pp. 173–179 (Bari, Cacucci).

WICHMAN, B. A. & HILL, I. D. (1982) Algorithm AS 183: an efficient and portable pseudo-random number generator, *Applied Statistics*, 31, pp. 188–190.

WILMORE, J. H. & COSTILL, D. L. (1994) *Physiology of Sport and Exercise* (Champaign, Illinois: Human Kinetics).