

Prior predictive p -values for assessing the Goodness-of-fit of the Skew Normal model.

Stefano Cabras and Maria Eugenia Castellanos*

Abstract We review a series of goodness-of-fit tests for the family of Skew Normal models when all parameters are unknown. As the null sample distributions of the considered test statistics depend only on the asymmetry parameter, we use a default, but proper, prior on skewness parameter to integrate it out. We thus obtain prior predictive distributions of the test statistics in which the prior predictive p -value, advocated by G. Box, is used to check the fit of the Skew Normal model. Goodness-of-fit tests, here proposed, depend only on the sample size and exhibit full agreement between nominal and actual size. We show also that they have good power against local alternative models which also account for asymmetry in the data. We exemplify how to use these tests in some anthropometric measures using the GOF_{SN} library in R.

Key words: Empirical distribution function test, Jeffreys' prior, prior predictive distribution, testing statistical hypotheses.

1 Introduction

In this paper we consider the univariate Skew Normal model ($\mathcal{SN}$ hereafter) described in [1]. One of the main features of the $\mathcal{SN}$ distribution is the ability to model the skewness through a shape parameter, λ . This model exhibits remarkable properties in terms of mathematical tractability and it has been applied, as a reference model, in several scientific disciplines as, for example: economics, epidemiol-

Stefano Cabras

Department of Mathematics and Informatics - University of Cagliari, Via Ospedale, 72 09124-Cagliari (Italy), e-mail: s.cabras@unica.it

Maria Eugenia Castellanos

Department of Statistics and O.R. - Rey Juan Carlos University, C/Tulipan, sn 28938-Mostoles (Spain), e-mail: maria.castellanos@urjc.es

ogy and geology (i.e.: [4, 7, 14, 16]). In order to use a parametric model to make inference it is first necessary to analyze its fit on observed data. In this context, we approach model checking as a preliminary analysis in that if the assumed model is compatible with observed data, then the full (and difficult) process of model elaboration and model selection (or averaging) can be avoided. Works on goodness of fit (GOF hereafter) for $\mathcal{SN}$ model can be found in: [22] where normality is tested within the $\mathcal{SN}$ class by showing that the coefficient of skewness is the locally most powerful location and scale invariant statistic; [21] used the same test statistics for testing departures from the half-normal distribution; [15] studied the Kolmogorov-Smirnov test to assess GOF of $\mathcal{SN}$ model assuming all parameters known; where instead [20] derived the distribution of several test statistics when all unknown parameters are estimated with the maximum likelihood estimators (MLE). This last work reports tables with critical values of GOF test for fixed values of λ and different sample sizes, n , obtaining an agreement between nominal and actual size of the test only for large n ($n > 400$). [12] proposed a test based on the Anderson-Darling test statistic (AD statistic, in the sequel) to analyze the fit of the $\mathcal{SN}$ model. In order to avoid the dependence of the approximated quantiles from the λ parameter, in [12] it is considered a polynomial regression on the quantiles of the AD statistic as function of the skewness parameter. Unfortunately, in such GOF test, there is a lack of agreement between nominal and actual size of the test, at least, for the analyzed sample sizes ($n = 50$ and $n = 100$, see Table 4 and 5 in [12]).

In this paper we follow, instead, the Bayesian methodology described in [10]. Such GOF tests are based on the prior predictive p -value of [9] and later extensively discussed in [5, 6]. We describe the sampling distribution under the $\mathcal{SN}$ model, (null distribution in the sequel), for some empirical distribution function test statistics (EDF here after). This is achieved by integrating out λ with respect to the Jeffreys' prior distribution discussed in [18]. In this way we can assess the compatibility of whole $\mathcal{SN}$ class with data, moreover, the null distribution does not depend on unknown parameters and the resulting GOF tests have full agreement between nominal and actual size for $n \geq 3$. The full agreement is assessed by the fact that the sampling null distribution of associated p -values are uniform. The key point is that marginal prior predictive distributions of the used EDF test statistics:

- i) do not depend on location and scale parameters when they are efficiently estimated in the sense of ([13]) and
- ii) the Jeffreys' prior on λ is a proper distribution for the $\mathcal{SN}$ model. Therefore, the prior predictive distribution can be used to calculate the p -value, while this is not possible in general.

The rest of the paper is organized as follows: Section 2 describes GOF test statistics and derives their approximate null distributions. Section 4 illustrates the power of the tests against local alternatives. Application to two data sets is illustrated in Section 5. Further remarks are contained in Section 6.

2 Goodness-of-fit tests

A random variable X is distributed as $\mathcal{SN}$ if X has density

$$f(x | \xi, \eta, \lambda) = \frac{2}{\eta} \phi\left(\frac{x - \xi}{\eta}\right) \Phi\left(\lambda \frac{x - \xi}{\eta}\right), \quad -\infty < x < \infty,$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively density and cumulative distribution function (cdf) of the standard normal distribution. Parameters $-\infty < \xi < \infty$ and $0 < \eta < \infty$ are location and scale, respectively. Parameter $-\infty < \lambda < \infty$ regulates the skewness. The $\mathcal{SN}$ is exactly the normal model for $\lambda = 0$, while as $\lambda \rightarrow \pm\infty$ the limiting distribution is the half normal model. It is interesting to note that the coefficient of skewness, γ , depends only on λ and varies in the interval $(-0.99527, 0.99527)$ ([21]). This fact may cause incompatibility with data in which the skewness is larger.

As reported in [18], maximum likelihood estimators (MLE) of $\mathcal{SN}$ parameters may be infinite with high probability for small n and large λ . It has been shown that the Fisher information matrix, using the above parametrization, called “direct” parametrization, is singular at $\lambda = 0$ (see [1]). Finally, another problem with this model is that the profile-likelihood function for λ may have different stationary points, with one at $\lambda = 0$ independently of the observed sample ([18]).

Papers [2], [21] and [18] analyze different estimation techniques in order to mitigate or avoid the mentioned problems. For our purposes, it is interesting to note that singularity of the Fisher information matrix can be avoided by working with the central parametrization ([1]), denoted by $\mathcal{SN}_{CP}(\mu, \sigma, \gamma)$, where μ , σ and γ represent mean, standard deviation and coefficient of skewness, respectively (see equation (1) in [21]).

In order to mitigate the irregularities of the likelihood function, [18] proposes to use the reference prior ([8]) on λ (or equivalently on γ). The reference prior equals the Jeffreys’ prior for the case of a scalar parameter. The scalar parameter is the case for our GOF tests as it will be illustrated in the next section.

For constructing a GOF test we need three elements:

- 1) a diagnostic test statistic, T , to quantify incompatibility with data, using the convention that large observed values, t_{obs} , indicate incompatibility;
- 2) a completely specified distribution for T , $h(t)$, under the null sampling model;
- 3) a way to measure the conflict between t_{obs} and $h(t)$ such as the p -value used here.

All these three elements are described in the sequel, while a full description can be retrieved in [10].

EDF statistics measure the difference between the empirical cumulative distribution function, F_n , and a theoretical one, F , that, in our case, is an element of the $\mathcal{SN}$ class. We fix F using the following estimates: MLE of γ , $\hat{\gamma}$, sample mean, $\tilde{\mu}$, and sample standard deviation, $\tilde{\sigma}$, for μ and σ respectively. Therefore, the theoretical distribution used is

$$F \equiv \mathcal{SN}_{CP}(\tilde{\mu}, \tilde{\sigma}, \hat{\gamma}).$$

As it is commented in [2], even when using the central parameterization for moderate sample sizes, there are still problems with the MLE estimator. In fact, the estimation of λ can occur in the frontier of λ 's support, that is $\hat{\lambda} \rightarrow \pm\infty$ ($\hat{\gamma} \rightarrow \pm 0.99527$). For the samples in which this occurs we adopt the strategy mentioned in [2]: when the maximization of λ is diverging, we choose the smallest $\hat{\lambda}$ whose likelihood is not significantly smaller than the likelihood with $\lambda = \infty$ according to a likelihood ratio test.

The EDF statistics we report, in this paper, are the Kolmogorov-Smirnov, D ,

$$\begin{aligned} D^+ &= \sup_x \{F_n(x) - F(x)\}; & D^- &= \sup_x \{F(x) - F_n(x)\}; \\ D &= \max\{D^+, D^-\}; \end{aligned}$$

and the Cramér-von Mises, W^2 , statistics

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 dF(x);$$

Other EDF statistics, more than those mentioned here, can be found in [10]. However, the GOF tests, here mentioned, exhibit the largest power in the analyzed local alternative models.

The corresponding sample version of both statistics are obtained using the following formulas:

$$\begin{aligned} D^+ &= \max_i \left(\frac{i}{n} - z_i \right); & D^- &= \max_i \left(z_i - \frac{i-1}{n} \right); \\ D &= \max\{D^+, D^-\}; \\ W^2 &= \sum_{i=1}^n \left(z_i - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}; \end{aligned} \tag{1}$$

where $z_i = F(x_{(i)})$, $\bar{z} = \sum_{i=1}^n z_i/n$ and $x_{(i)}$ denotes the i order statistics ($x_{(i)} < x_{(i+1)}$, $i = 1, \dots, n$).

When location-scale parameters, as μ and σ in our case, are unknown and replaced with efficient estimators in the sense of [13], such as $\tilde{\mu}$ and $\tilde{\sigma}$, then the null sampling distribution $h(t)$ of the EDF statistics does not depend on the true values of μ and σ ([23], Section 4.3.2).

However, $h(t)$ still depends on γ since it is a shape parameter. In the frequentist paradigm, $h(t)$ is usually obtained by plugging in a point estimator of the unknown parameter (i.e. $\hat{\gamma}$), obtaining the corresponding plug-in p -value, p_{plug} . It is known ([5, 6]) that this type of p -value is conservative in small samples because of the double use of the data: first to estimate $h(t)$ and second to calculate the p -value using again the data through the observed statistics, t_{obs} . In [20] p_{plug} is used, and because of the double use of data, it results in less power than the proposal reported in this work.

Our proposal is to use, as null sampling distribution $h(t)$, the marginal prior predictive distribution of T obtained by integrating γ with respect to the Jeffreys' prior. In general, it is not possible to use this procedure because non-informative priors, as the Jeffreys' prior or the reference prior, are usually improper. In the case of the $\mathcal{SN}$ model, the Jeffreys' prior, that equals the reference prior, is a proper density ([18]).

The Jeffreys' prior for γ is given by

$$\pi(\gamma) \propto \pi(\lambda) \left| \frac{d\lambda}{d\gamma} \right|, \quad \pi(\lambda) = K^{-1} \sqrt{\int_{-\infty}^{\infty} 2z^2 \phi(z) \frac{\phi^2(\lambda z)}{\Phi(\lambda z)} dz}, \quad (2)$$

where the constant K is obtained numerically using adaptive quadrature integration for the integrals with respect to z and λ .

The marginal prior predictive distribution of T , $m(t)$, is defined by averaging in the conditional distribution $h(t | \gamma)$ with respect to the prior $\pi(\gamma)$:

$$m(t) = \int h(t | \gamma) \pi(\gamma) d\gamma, \quad (3)$$

Prior $\pi(\lambda)$ plays an important rule in this model checking procedure and it shows useful properties for GOF:

- i) $m(t)$ always exists;
- ii) it is invariant under reparameterization allowing to easily switch between central and direct parametrization.

For any statistic T we have approximated $h(t)$ for a particular n (in our case this distribution only depend on n) using the following algorithm:

Algorithm 1-

1. Draw $\gamma^{(1, \dots, m, \dots, M)} \sim \pi(\gamma)$, for $M = 10^6$;
2. for each $\gamma^{(m)}$ generate a random sample of size n from the $\mathcal{SN}$ model with $\mu = 0$, $\sigma = 1$ ($h(t)$ does not depend on μ, σ) and $\gamma = \gamma^{(m)}$;
3. for the m -th sample calculate the EDF statistic on the random sample following the definition in (1).

Critical quantiles of the prior predictive distribution (3) approximated by algorithm 1 for W^2 and $\sqrt{n}D$ statistics with different n appear in Table 1. Critical quantiles for other EDF statistics can be found in [10].

3 Size of the proposal tests

In order to assess the GOF of the $\mathcal{SN}$ model for a data set, we can use the critical values reported in Table 1 or we can calculate the prior predictive p -value associated with the observed data. The prior predictive p -value is defined as

Table 1 Critical quantiles of $m(t)(=h(t))$ for W^2 and $\sqrt{n}D$ test statistics. (Source: [10])

Probability	Sample size (n)										
	5	10	20	30	50	100	150	200	300	400	500
W^2											
0.100	0.073	0.079	0.087	0.092	0.095	0.095	0.095	0.096	0.096	0.096	0.096
0.050	0.089	0.098	0.106	0.113	0.118	0.120	0.120	0.121	0.120	0.121	0.120
0.025	0.105	0.118	0.128	0.134	0.142	0.147	0.147	0.147	0.147	0.148	0.147
0.010	0.128	0.148	0.161	0.166	0.176	0.187	0.186	0.187	0.186	0.186	0.183
$\sqrt{n}D$											
0.100	0.608	0.679	0.723	0.745	0.762	0.775	0.781	0.783	0.787	0.790	0.789
0.050	0.658	0.734	0.784	0.812	0.831	0.847	0.850	0.858	0.857	0.863	0.860
0.025	0.702	0.785	0.845	0.870	0.892	0.915	0.922	0.926	0.923	0.930	0.930
0.010	0.753	0.853	0.915	0.945	0.971	1.000	1.010	1.013	1.021	1.018	1.015

$$Pr^{m(t)}(T \geq t_{obs}). \quad (4)$$

To calculate (4) we use the approximated distribution obtained by algorithm 1 for a sample size n .

We have simulated data for different values of n from the $\mathcal{SN}$ model with $\mu = 0$ $\sigma = 1$ and γ values generated from prior in (2). For each sample size, n , we have simulated 1000 data sets, and for each data set (4) has been approximated. Figure 1 shows that prior predictive p -values are uniformly distributed for $n = 3$. The uniformity of the prior predictive p -value assures that the size of the GOF tests equals the nominal one.

4 A power study under local alternatives

We study the power of GOF tests, here proposed, against alternative distributions to the $\mathcal{SN}$ that account for asymmetry in the data. Figure 2 reports the power for a test of size 95% against the following models:

Lognormal. $Lognormal(\mu = 0, \sigma)$ with location μ and shape parameter σ (see [17], p. 208). We consider σ equals 0.3 as small values of σ makes Log-normal distribution similar to $\mathcal{SN}$.

Gamma. $Gamma(\alpha, \beta)$ with shape α and rate β . Large values of α makes this model similar to $\mathcal{SN}$. For seek of comparison with [20] we consider these special cases: exponential distribution and the χ^2 distribution with 4 degrees of freedom.

Skew- T . $ST(\mu = 0, \sigma = 1, \gamma = 0.5, df)$ where μ , σ and γ are location, scale and shape parameter respectively, and df indicates the degrees of freedom. This model, introduced by [3], encompasses the $\mathcal{SN}$ one through df parameter: $ST(\mu, \sigma, \gamma, df) \rightarrow \mathcal{SN}(\mu, \sigma, \gamma)$ as $df \rightarrow \infty$. We consider Skew- T model with 1, 3 and 10 df .

Weibull. *Weibull*($c = 2, \alpha = 1$) where c and α are shape and scale parameters, respectively (see [17], p. 641). Large values of c makes the Weibull model similar to the $\mathcal{SN}$.

Statistic W^2 is the most powerful test, also it results the most powerful between all the EDF test considered in [10]. When the alternative models are closed to the $\mathcal{SN}$ one, the power decreases, it is the case for the $\mathcal{ST}$ with $df \geq 10$, *Gamma* model with $\alpha \geq 4$, *Weibull* model and for the Log-Normal model. The power of the tests presented in this paper are, in general, greater than the best tests reported in previous works [20] and [12] as it is pointed out in [10].

5 Application with the GOF SN library

We illustrate the proposed GOF procedure on two datasets showing how to use the GOF SN library (<http://cran.r-project.org/web/packages/GOF SN/GOF SN.pdf>) from R to make the tests.

First we consider the well known Frontier dataset used in [1] to show a case when MLE of λ does not exist. The moment estimators of μ and σ are respectively 0.88 and 0.76, while as estimation of γ we take the limit value 0.99527. Substituting the parameters by these estimations and computing the statistics as mentioned in

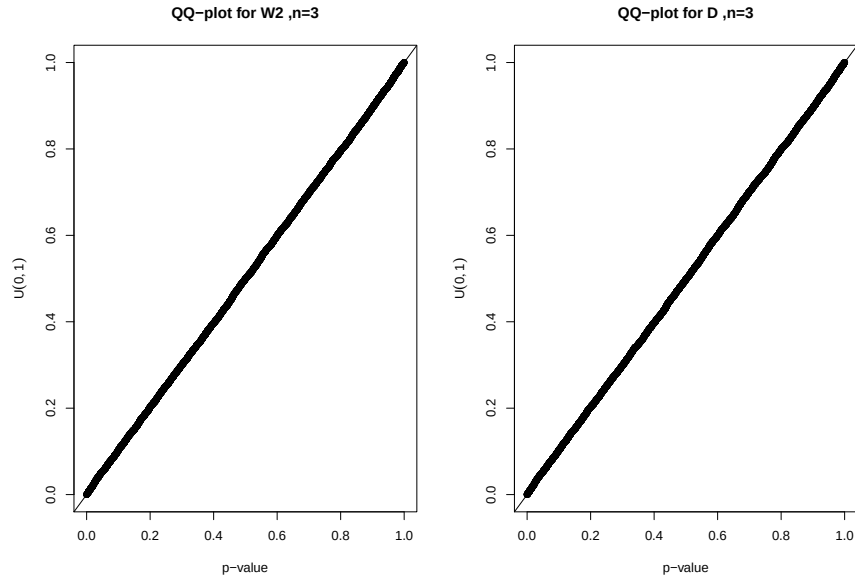


Fig. 1 Quantile-Quantile plot of p -value distribution (1000 samples) under the null model $SN(x | \lambda), \lambda \sim \pi(\lambda)$ against $U(0, 1)$ for $n = 3$. and different test statistics. (Source: [10])

equations (1), we get $W^2 = 0.025$ and $\sqrt{n}D = 0.407$, both values are well below the quantile 90% showing that the Frontier data are compatible with the $\mathcal{SN}$ model.

These calculations are automatically done using the following code in R:

```
>library(GOFSN)
>data(frontier)
>W2(frontier)
>D(frontier)
```

It is possible make directly the test using the commands:

```
>W2.sn(frontier)
>D.sn(frontier)
```

The second dataset is a subset of metric traits from an observational study conducted for anthropologists on around 100 male and 100 female individuals from an homogeneous population, 20 to 25 years old people in Sardinia ([11, 19]). The aim

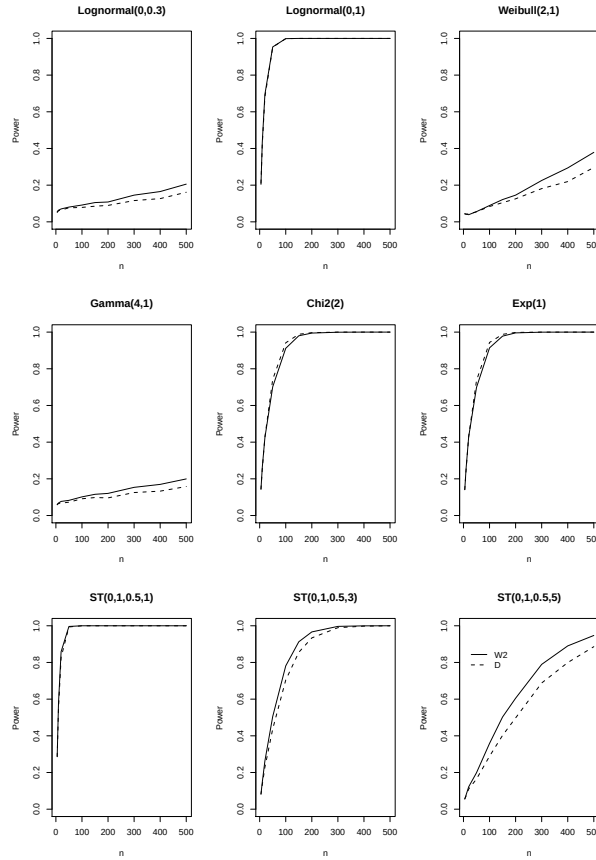


Fig. 2 Power functions for the alternative models. (Source: [10])

of the study is to analyze the sexual dimorphism through the comparison of asymmetry in males and females metric traits. $\mathcal{SN}$ seems a flexible model to fit each metric trait, but before using this model to make inference, we have to study the GOF of $\mathcal{SN}$ in each metric trait. We consider three metric traits: stature, triceps skin fold and mid-thigh skin fold. Figure 3 shows a kernel density estimator of each metric trait for each sex. It can be seen that most of the distributions are asymmetric.

Table 2 shows parameter estimations for each variable and sex, the value of W^2 (the most powerful EDF statistic), the 95% quantile for W^2 for $n = 100$, and the conclusion about the fitting of the $\mathcal{SN}$. Table 2 shows that $\mathcal{SN}$ results to be compatible for all measurements but the triceps and mid-thigh skin folds in males. The bad fit of the $\mathcal{SN}$ model in these data set could be related with the limited skewness allowed by the $\mathcal{SN}$, about 0.99527. To avoid this, in [11, 19] the $\mathcal{SN}$ and Skew-t models have been compared through Bayes factors selecting the best model for each metric trait.

6 Remarks

In this paper we review the GOF approach for the $\mathcal{SN}$ discussed in [10]. This approach proposes several GOF tests which depend only on sample size. This proce-

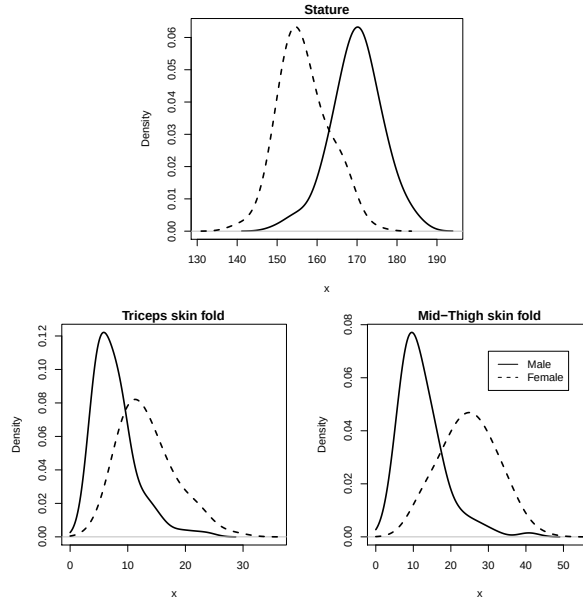


Fig. 3 Kernel density estimation for several metric traits in male and female populations.

Table 2 Estimations and test statistic W^2 for metric traits in Sardinian young people.

Variables	Sex	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$	W^2	0.95-Q	$\mathcal{SN}$ fit
Stature	F	156.7	6.14	0.41	0.06	0.12	yes
	M	169.9	6.36	-0.23	0.08	0.12	yes
Triceps Skin.	F	13.33	4.78	0.83	0.02	0.12	yes
	M	7.76	3.76	0.98	0.21	0.12	no
Mid-thigh Skin.	F	24.65	7.33	0.02	0.03	0.12	yes
	M	12.35	6.09	0.95	0.19	0.12	no

ture, based on $m(t)$, fully agrees with the Neymann hypothesis testing, as rejection regions are fixed prior to observe the data. The tests are based on the Jeffreys' prior (2) which is used to eliminate the nuisance parameter γ present in the distribution of the test statistics. This may be viewed in contrast to the methodology in [12], where the dependence from γ is avoided using a post-processing of the quantiles of the null distribution of test statistics. Other GOF tests, that we used as benchmarks, are less powerful ([20, 12]) because they are based on a double use of the data: determination of $h(t)$, and calculus of p -value on observed t_{obs} . Another open field of research could be the assessing of the GOF for the Skew- T model. There, degrees of freedom represents another nuisance parameter for the null distribution of the test statistic.

Acknowledgements This work has been partially supported by M.I.U.R. of Italy, by the Spanish Ministry of Science and Technology, under grants MTM2007-61554 and TIN2008-0679-C04-0 and by grant S2009/esp-1594 from the Madrid Region of Spain. Part of the work was done while the second author was visiting the Department of Mathematics and Informatics, University of Cagliari, Italy.

References

1. Azzalini A.: A class of distribution which includes the normal ones. *Scand. J. Stat.* **12**, 171-178 (1985)
2. Azzalini A., Capitanio A.: Statistical applications of the multivariate skew normal distribution. *J. Roy. Statist. Soc. Ser. B* **61**, 579-602 (1999)
3. Azzalini A., Capitanio A.: Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J. Roy. Statist. Soc. Ser. B* **65**, 367-389 (2003)
4. Ball L., Mankiw N.G.: Relative price changes as aggregate supply shocks. *Quarterly J. Economics* **CX**, 161-193 (1995)
5. Bayarri M.J., Berger J.: Quantifying surprise in the data and model verification. In: Bernardo J, Berger J, Dawid A, Smith A (eds) *Bayesian Statistics*, Oxford University Press, London, **6**, 53-82 (1999)
6. Bayarri M.J., Berger J.: P-values for composite null models. *Journal of the American Statistical Association*, **95**, 1127-1142 (2000)
7. Bazan J.L., Branco M.D., Bolfarine H.: A skew item response model. *Bayesian Analysis* **1**, 861-892 (2006)
8. Berger J., Bernardo J.M.: On the development of reference priors. In: Berger JO, Bernardo J, Dawid A, Smith A (eds) *Bayesian Statistics*, Oxford University Press, London, **4**, 35-60 (1992)

9. Box G.: Sampling and bayes inference in scientific modeling and robustness. *J. Roy. Statist. Soc. Ser. A*, **143**, 383-430 (1980)
10. Cabras S., Castellanos M.E.: Default bayesian goodness-of-fit tests for the skew-normal model. *Journal of Applied Statistics* **36**, 223-232 (2009)
11. Cabras S., Castellanos M.E.: Comparing skewness in two populations using bayesian inference on skew normal and skew t-models. Tech. rep., Department of Statistics Rey Juan Carlos University (Spain), **5/12** (ISSN: 1699-6895)
12. Dalla Valle A.: A test for the hypothesis of skew-normality in a population. *J. Statist. Comput. Simul.* **77**, 63-77 (2007)
13. Durbin J.: Weak convergence of the sample distribution function when parameters are estimated. *Ann. Stat.* **1**, 279-290 (1973)
14. Ghosh P., Branco M.D., Chakraborty H.: Bivariate random effect model using skew-normal distribution with application to hivrna. *Statistics in Medicine* **26**, 1255-1267 (2006)
15. Gupta A., Chen T. Goodness-of-fit tests for the skew-normal distribution. *Commun. Stat. Simul. Comp.*, **30**, 907-930 (2001)
16. Gupta R.C., Brown N.: Reliability studies of the skew-normal distribution and its application to a strength-stress model. *Commun Statist - Theory and Methods* **30**, 2427-2455 (2001)
17. Johnson K.S., Balakrishnan N.: Continuous Univariate Distributions. John Wiley & Sons, New York (1994)
18. Liseo B., Loperfido N.: A note on reference priors for the scalar skew-normal distribution. *J. Stat. Plan. Infer.*, **136**, 373-389 (2006)
19. Marini E., Racugno W., Cabras S., Castellanos M.E., Buffa R., Rebato E., Salces I., Borgognini Tarli S.: Dimorphic skewed distributions of anthropometric variables. In: 1st Congress of Italian Evolutionary Biologists, Ferrara-Italy (2005)
20. Mateu-Figueras G., Puig P., Pewsey A.: Goodness-of-fit tests for the skew-normal distribution when the parameters are estimated from the data. *Commun. Statist. - Theory and Methods*, **36**, 9 (2003)
21. Pewsey A.: Problems of inference for azzalini's skew-normal distribution. *J. Applied Statistics*, **27**, 1184-1190 (2000)
22. Salvan A.: A locally most powerful invariant tests of normality. In: Atti della XXXIII Riunione Scientifica della Società Italiana di Statistica, **2**, 173-179 (1986)
23. Stephens M.A.: Goodness-of-fit Techniques, Marcel Dekker, NY (1986)