

Comparison of Skewness in two populations ^(*)

Confronto di asimmetria in due popolazioni

Stefano Cabras¹, Maria Eugenia Castellanos²,

¹ Dipartimento di Matematica, Università di Cagliari
e-mail: s.cabras@unica.it

² Departamento de Estadística e Investigación Operativa, Universidad Rey Juan Carlos

Riassunto: L'obiettivo di questo lavoro è il confronto dell'asimmetria in due popolazioni indipendenti utilizzando le classi di distribuzioni normale e t asimmetriche. I parametri sono stimati in un approccio bayesiano con distribuzioni a priori da default poichè l'elicitazione diretta sui parametri di asimmetria risulta difficoltosa per l'esperto. Dopo aver studiato la compatibilità dei modelli con i dati osservati, si costruisce il fattore di Bayes per il confronto tra i due modelli. Sulla base del modello scelto si ottiene una approssimazione della distribuzione a posteriori dell'asimmetria nelle due popolazioni. Si considera un esempio applicativo a dati di natura antropometrica.

Keywords: goodness-of-fit, Jeffrey's prior, skew normal, skew t .

1. Introduction

In this paper we compare skewness of two independent populations. Population skewness is used in a variety of applications among them we concentrate in analyzing differences in asymmetry in male and female body measurements to study sexual dimorphism (cfr. Cabras and Castellanos, 2005). We consider two independent not necessarily symmetric populations X and Y representing sexes, where the first three centered moments, μ_j , $j = 1, 2, 3$ exist. This paper focuses on inference about the coefficients of asymmetry, γ_X and γ_Y , defined as $\gamma = \mu_3/\mu_2^{3/2}$. The methods for testing differences between γ_X and γ_Y reported in the literature usually rely on the normality assumption (cfr. Sokal and Rohlf, 1994). In Joanes and Gill (1998) different sample coefficients are evaluated under non symmetric populations showing that the bias increases with population skewness. The non parametric approach to this problem using methods, such as bootstrap, does not give satisfactory results for small samples and large asymmetry (cfr. Cabras and Castellanos, 2005). For non-symmetric populations, we believe that inference on γ_X and γ_Y may be improved using suitable parametric models. The problem of evaluating the difference between γ_X and γ_Y may be satisfactorily addressed using skew normal and skew t models (cfr. Azzalini, 1985, Azzalini and Capitanio, 2003) in a Bayesian framework. We determine the overlap of posterior credible intervals for γ_X and γ_Y , at a given level α . Bayes Factors are discussed and used to choose between the two models.

(*) Work partially supported by M.I.U.R. (Italy) and Spanish Ministry of Science and Technology and ERDF, under grants MTM2004-03290, and TSI2004-06801-C04-01

2. Methods

To model anthropometric measurements we use skew normal, $\mathcal{SN}$, and skew t , $\mathcal{ST}$, distributions. The actual names of these distributions were given by Azzalini in Azzalini (1985) and Azzalini and Capitanio (2003). The density of $\mathcal{SN}$ is

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ are respectively density and cumulative distribution function (cdf) of the standard normal distribution. Space parameter is $-\infty < \xi < \infty$ and $0 < \eta < \infty$, where ξ, η are respectively location and scale parameters and $-\infty < \lambda < \infty$ is the shape parameter. The $\mathcal{ST}$ distribution has density for $-\infty < x < \infty$,

$$f(x | \zeta, \omega, \lambda, \nu) = \frac{2}{\omega} t_\nu\left(\frac{x - \zeta}{\omega}\right) T_{\nu+1}\left(\lambda \frac{x - \zeta}{\omega} \left(\frac{\nu + 1}{(x - \zeta)^2/\omega^2 + \nu}\right)^{1/2}\right),$$

where $t_\nu(x)$ is the density function of a standard scalar t distribution with ν degrees of freedom and $T_{\nu+1}$ denotes the cdf of a standard scalar t distribution with $\nu + 1$ degrees of freedom. Again, $-\infty < \zeta < \infty$, $0 < \omega < \infty$, and $-\infty < \lambda < \infty$ are location, scale and shape parameters, respectively, where $\nu > 0$ represents the degrees of freedom. $\mathcal{SN}$ model is nested in $\mathcal{ST}$ ($\mathcal{SN} \equiv \mathcal{ST}$ for $\nu \rightarrow \infty$) which accounts for heavy tails and larger skewness than $\mathcal{SN}$. In Liseo and Loperfido (2006) are reviewed a variety of estimation techniques.

2.1. Prior and posterior distributions

For both $\mathcal{SN}$ and $\mathcal{ST}$ models default Bayesian inference is considered for location, scale and shape parameters, while an empirical Bayes approach is used for the degrees of freedom for the $\mathcal{ST}$ model. As noted in Geweke (1993), it is important to use a proper prior distribution on ν since an improper one essentially implicates the $\mathcal{SN}$ model, due to the fact that the prior mass on $\nu \rightarrow \infty$ far exceeds the mass for any finite ν .

2.1.1. Skew normal model

The paper of Liseo and Loperfido (2006) introduces default Bayesian inference, we use this prior construction technique, concretely, the prior we consider is:

$$\pi(\xi, \eta, \lambda) \propto \pi(\xi, \eta) \pi(\lambda) = \frac{1}{\eta} \sqrt{\int_{-\infty}^{\infty} 2z^2 \phi(z) \frac{\phi^2(\lambda z)}{\Phi(\lambda z)} dz}.$$

Given a random sample $\mathbf{x} = (x_1, \dots, x_n)$, the posterior density is

$$\pi(\xi, \eta, \lambda | \mathbf{x}) \propto \pi(\xi, \eta, \lambda) \eta^{-n} \prod_{i=1}^n \phi\left(\frac{x_i - \xi}{\eta}\right) \Phi\left(\lambda \frac{x_i - \xi}{\eta}\right).$$

Liseo and Loperfido (2006) showed that this posterior is proper when there are at least two different observations, we use MCMC simulation techniques to approximate it.

2.1.2. Skew t model

To construct the prior for the $\mathcal{ST}$ model, we rely on the following way to define the $\mathcal{ST}$ model: $X = \zeta + \theta^{-1/2}Z$, where $Z \sim \mathcal{SN}(0, \omega, \lambda)$ and $\theta \sim \chi_\nu^2/\nu$, independent of Z .

The shape parameter λ is the shape parameter of the random variable Z , therefore the same prior for λ is used, as in the $\mathcal{SN}$ model. We assume for ζ and ω the usual default prior for location and scale parameters, $\pi(\zeta, \omega) \propto 1/\omega$. Then the joint prior becomes

$$\pi(\zeta, \omega, \lambda, \nu) \propto \pi(\zeta, \omega)\pi(\lambda)\pi(\nu) = \frac{1}{\omega} \pi^{EB}(\nu) \sqrt{\int_{-\infty}^{\infty} 2z^2 \phi(z) \frac{\phi^2(\lambda z)}{\Phi(\lambda z)} dz},$$

where $\pi^{EB}(\nu)$ is a gamma distribution with the MLE of ν , $\hat{\nu}$, as mean, and variance, $\sigma_{\hat{\nu}}^2$ equals the standard error of $\hat{\nu}$. The corresponding posterior distribution is proper Cabras and Castellanos (2005).

Again $\pi(\zeta, \omega, \theta, \lambda, \nu \mid \mathbf{x})$ cannot be obtained analytically and MCMC algorithms are used. Details about the algorithms used appear in Cabras and Castellanos (2005).

3. Application

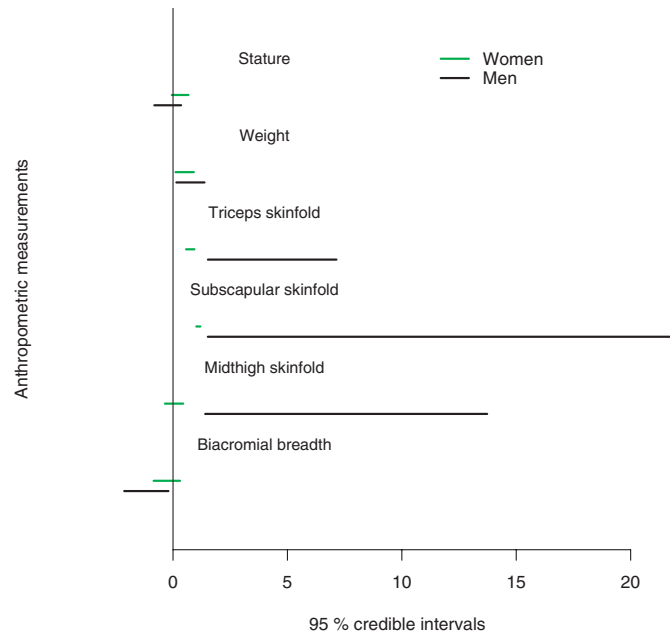
The aim in this application is to make inference about differences in male and female body measurements, as such differences may be indicative of sexual dimorphism. The sample considered here consists of six anthropometric measurements taken in Sardinian adults (Italy). To choose between the $\mathcal{SN}$ or $\mathcal{ST}$ model we consider the Bayes factor. Given the data, $\mathbf{x}$, the Bayes factor is defined as the ratio of the predictive densities for the data; $B = m(\mathbf{x} \mid \mathcal{SN})/m(\mathbf{x} \mid \mathcal{ST})$. In general, when improper priors are used under different models, Bayes factors does not exist. However, in this case the $\mathcal{SN}$ model may be thought of as a particular case of $\mathcal{ST}$ with $\nu = M \subseteq \mathbb{R}^+$, where M is sufficiently large, moreover for the particular prior distributions we use, Bayes Factors are correctly defined (see Kass and Raftery (1995) for an extensive discussion of similar cases). Predictive densities $m(\mathbf{x} \mid \mathcal{SN})$ and $m(\mathbf{x} \mid \mathcal{ST})$ are approximated using the output of the MCMC algorithms with the method described in Chib and Jeliazkov (2001). Table 1 shows the observed Bayes factors. Values of $B < 1$ indicate that $\mathcal{ST}$ is preferred to $\mathcal{SN}$.

Table 1: Observed Bayes factors, B , of $\mathcal{SN}$ against $\mathcal{ST}$.

	Stature	Weight	Triceps	Subscapular	Midthigh	Biacromial
Men	0.594	0.34	0.033	0.013	0.323	0.033
Women	1.335	1.402	0.707	0.79	1.065	0.059

Figure 1 contains posterior credible intervals of γ_X and γ_Y (obtained from the corresponding model, $\mathcal{ST}$ if $B < 1$ and $\mathcal{SN}$ in other case) which show that some measurements are clearly asymmetric, such as *weight*, and almost all *skinfold* measurements, (except *midthigh skinfold* in women). Skinfold measures clearly exhibit differences in asymmetry. This feature is also found in other populations as shown in Cabras and Castellanos (2005).

Figure 1: 95% credible intervals for γ_X and γ_Y of anthropometric measurements in the Sardinian population, which are showed in the vertical dimension along with the corresponding names.



References

- Azzalini A. (1985) A class of distribution which includes the normal ones, *Scandinavian Journal of Statistics*, 12, 171–178.
- Azzalini A., Capitanio A. (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society, B*, 65, 367–389.
- Cabras S., Castellanos M.E. (2005) Comparing the skewness in two populations using Bayesian inference on skew normal and skew t models, *Working Paper, Dep. Stat., Rey Juan Carlos University*, <http://bayes.escet.urjc.es/publicaciones/wp05-11.pdf>.
- Chib S., Jeliazkov I. (2001) Marginal likelihood from the Metropolis Hastings output, *Journal of the American Statistical Association*, 96, 270–281.
- Geweke J. (1993) Bayesian Treatment of the Independent Student-t Linear Model, *Journal of Applied Econometrics*, 8 Supplemental, 19–40.
- Joanes D. N., Gill C. A. (1998) Comparing Measures of Sample Skewness and Kurtosis, *The Statistician*, 47, 183–189.
- Kass R., Raftery A. (1995) Bayes factors, *Journal of the American Statistical Association*, 90, 773–795.
- Liseo B., Loperfido N. (2006) A note on reference priors for the scalar skew-normal distribution, *Journal of Statistical Planning and Inference*, 136, 373–389.
- Sokal R.R., Rohlf F.J. (1994) *Biometry (3rd ed.)*, Freeman W.H., New York.