Available online at www.sciencedirect.com

ScienceDirect

C. R. Acad. Sci. Paris, Ser. I 344 (2007) 509–514

COMPTES RENDUS MATHEMATIQUE

http://france.elsevier.com/direct/CRASS1/

Statistics/Probability Theory

A test for the equality of marginal distributions

Vilijandas Bagdonavičius [a], Ruta Levuliene [a], Mikhail Nikulin [b]

[a] *University of Vilnius, 24, Naugarduko, 01513 Vilnius, Lithuania*
[b] *Université Victor-Segalen Bordeaux 2, BP 26, 146, rue Léo-Saignat, 33076 Bordeaux cedex, France*

Received 21 March 2006; accepted after revision 14 February 2007

Available online 19 April 2007

Presented by Paul Deheuvels

Abstract

We present a test for the equality of marginals of bi-dimensional distribution functions under censoring. The asymptotic power of the test under approaching alternatives and the simulation analysis for finite samples are done. The test is more powerful than classical tests in situations where the marginals differ in shape parameters. ***To cite this article: V. Bagdonavičius et al., C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Résumé

Un test pour l'égalité des répartitions des marginales. On propose un test pour l'égalité des répartitions des marginales pour des données complètes ou censurées à droite. On obtient une estimation de la puissance asymptotique pour des alternatives approchées. On étudie par simulation des propriétés du test pour des échantillons finis. Le test est plus puissant que les tests classiques quand les marginales diffèrent en paramètre de forme. ***Pour citer cet article : V. Bagdonavičius et al., C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Version française abrégée

Soient $X = (X_1, X_2)$ un vecteur aléatoire de la fonction de répartition $\tilde{F}$, $\tilde{f}$ la densité, et F_1, F_2 les fonctions de répartition marginales. On considère l'hypothèse classique $H_0 : F_1 = F_2$.

On suppose que, pour le j-ième objet, les vecteurs $(X_{1j}, X_{2j}, \delta_{1j}, \delta_{2j})$ sont observés, $X_{ij} = T_{ij} \wedge C_{ij}$, $\delta_{ij} = \mathbf{1}_{\{T_{ij} < C_{ij}\}}$, où (T_{1j}, T_{2j}) sont des vecteurs aléatoires indépendants bidimensionels de la fonction de survie $\tilde{S}$ $(j = 1, \ldots, n)$ et (C_{1j}, C_{2j}) sont les moments des censures, indépendants de (T_{1j}, T_{2j}) et identiquement distribués, de la fonction de survie $\tilde{G}$.

En utilisant les notations (2)–(6), la statistique du test proposé, a la forme (7). La loi limite de la statistique est la loi du chi-deux à deux degrés de liberté. La statistique est obtenue en modifiant les statistiques pour une classe des alternatives assez générales. La puissance asymptotique pour des alternatives approchées s'écrit en termes de loi non-centrale du chi-deux (la loi limite est donnée par (8)).

E-mail addresses: vilijandas.bagdonavicius@mif.vu.lt (V. Bagdonavičius), nikou@sm.u-bordeaux2.fr (M. Nikulin).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.03.001

On a étudié par simulation des propriétés du test pour des échantillons finis quand les lois marginales sont Weibull, loglogistique et lognormale. Le test est plus puissant que les tests classiques quand les marginales diffèrent en paramètre de forme.

1. Introduction

Let $X = (X_1, X_2)$ be a random vector with dependent components, the joint distribution function $\tilde{F}$ and the joint probability density $\tilde{f}$ with the marginal absolutely continuous distribution functions F_1, F_2. So we exclude continuous failure times where $X_1 = X_2$ with positive probability.

Consider the classical hypothesis $H_0 : F_1 = F_2$ of the equality of the marginal distributions of paired samples. The classical non-parametric tests are the sign test and the Wilcoxon signed-rank test. In situations where crossings of distribution functions are possible these tests are not powerful. It is explained by the fact that they are based on the number (sign test) or the sums of ranks (Wilcoxon signed-rank test) of the objects corresponding to the positive differences between the coordinates of the observed two-dimensional random vectors. If there are no crossings then it is rather probable that most differences between coordinates take the same sign but in the case of crossings the numbers of positive and negative values may be similar as in the case of equality of distributions.

Examples of crossings of distribution functions are numerous, see in Bagdonavicius and Nikulin [1]. For example, if the marginals have the cumulative distribution functions of the form:

$$F_i(x) = F\big((x/\theta_i)^{\nu_i}\big) \quad (i = 1, 2), \tag{1}$$

where F is the c.d.f. of the standard Weibull (exponential), loglogistic, lognormal law then the marginal c.d.f. F_1 and F_2 cross if $\nu_1 \neq \nu_2$.

The classical tests cannot be used in the case of censored samples.

Parametric tests of the equality of marginal distributions were considered in Cantor and Knapp [3], Owen et al. [7], Owen [8].

2. Alternatives

Sklar [10] introduced a representation of m-dimensional distribution function as a composition of a distribution function concentrated in the unit cube $[0, 1]^m$ and marginal distribution functions. The analogous representation holds also for common survival function: in the case $m = 2$ it has the form:

$$\tilde{S}(x_1, x_2) = P(X_1 > x_1, X_2 > x_2) = C_\alpha\big(S_1(x_1), S_2(x_m)\big),$$

where the function C_α is parametrised by the association parameter α. The function C_α is called the survival copula. The joint behaviour of a random vector with continuous marginals F_i can be characterised uniquely by its associated copula.

We note here that more about the properties and different applications of copulas in survival analysis can be seen in a recent paper of Georges et al. [5], and applications of copulas in reliability can found, for example, in Bagdonavicius and Nikulin [2].

Suppose that the marginals F_i and the copula C_α are unknown. Denote by $S_i = 1 - F_i$, f_i, $\Lambda_i = -\ln S_i$, and $\lambda_i = f_i/S_i$ the survival function, probability density function, cumulative hazard, and hazard rate of X_i, respectively $(i = 1, 2)$.

One of the possible ways of test construction could be to use the following simple idea: suppose that the differences between the marginal distributions are defined by the changing shape and scale model (1) with completely unknown F, to write some modification of parametric score functions to this semiparametric situation and to consider the distribution of these functions in the case of the equality of marginals. Unfortunately, such modifications require estimators of the derivatives of baseline density function and are not useful in practice. We use another way: formulate rather general semiparametric alternatives which give the possibility to obtain tractable modifications of parametric score functions to the semiparametric case and obtain the limit distribution. Naturally, the power of the obtained test is investigated not only in the case of formulated alternatives but also in the case of the changing shape and scale alternatives (1). Without loss of generality we can suppose that the random variables X_i are non-negative (we always can consider the variables $\exp\{X_i\}$ if it is not so).

Consider the following alternative to hypothesis (1):

$$H_\theta:\quad \lambda_1(x)=\lambda(x),\quad \lambda_2(x)=e^{\beta}\{1+e^{\beta+\gamma}\Lambda(x)\}^{e^{-\gamma}-1}\lambda(x),$$

where $\lambda=\lambda_1$, $\Lambda=\Lambda_1$, $\theta=(\beta,\gamma)$, $\beta^2+\gamma^2\neq 0$, the unknown copula C_α being the same as under H_0. The alternative contains a number of different possibilities: if $\beta\gamma>0$ then the marginal hazard rates and the cumulative distribution functions cross in $(0,\infty)$; if $\beta\gamma<0$ then the ratios of marginal hazard rates are monotone without crossings neither of the hazard rates nor the survival functions in $(0,\infty)$; if $\gamma=0$ then the ratio of hazard rates is constant.

3. The data

Suppose that (T_{1j},T_{2j}) are n independent two-dimensional vectors, each having joint survival function $\tilde S$ $(j=1,\dots,n)$. In situations when T_{ij} can be interpreted as failure times, data may be censored. For example T_{ij} may be the failure time of the ith component of the jth system with two dependent components or time to some event of the jth object in the ith experiment. So we suppose that right censoring is possible. Let (C_{1j},C_{2j}) be the censoring times, independent of the failure times (T_{1j},T_{2j}) and identically distributed with joint survival function $\tilde G$ and with marginal survival functions G_i, $i=1,2$.

Then, for subject j, the vectors $(X_{1j},X_{2j},\delta_{1j},\delta_{2j})$ are observed; here $X_{ij}=T_{ij}\wedge C_{ij}$, $\delta_{ij}=\mathbf{1}_{\{T_{ij}<C_{ij}\}}$.

Set

$$N_{ij}(x)=\mathbf{1}_{\{T_{ij}\leqslant x,T_{ij}<C_{ij}\}},\quad Y_{ij}(x)=\mathbf{1}_{\{X_{ij}\geqslant x\}},\quad N(x)=N_1(x)+N_2(x),\quad Y(x)=Y_1(x)+Y_2(x),$$

$$N_i(x)=\sum_{j=1}^n N_{ij}(x),\quad Y_i(x)=\sum_{j=1}^n Y_{ij}(x),\quad M_i(x)=N_i(x)-\int_0^x Y_i(u)\,d\Lambda_i(u).\tag{2}$$

4. The test

Suppose at first that λ is known. Then under H_θ the parameter θ can be estimated by the method of maximum likelihood using only the data corresponding to the second component. The score functions are:

$$U_1(\theta,\Lambda)=\int_0^\tau\left[1+\left(e^{-\gamma}-1\right)\frac{e^{\beta+\gamma}\Lambda(x)}{1+e^{\beta+\gamma}\Lambda(x)}\right]\left(dN_2(x)-Y_2(x)e^{\beta}\{1+e^{\beta+\gamma}\Lambda(t)\}^{e^{-\gamma}-1}\,d\Lambda(x)\right),$$

$$U_2(\theta,\Lambda)=\int_0^\tau\left[-e^{-\gamma}\ln\left(1+e^{\beta+\gamma}\Lambda(t)\right)+\left(e^{-\gamma}-1\right)\frac{e^{\beta+\gamma}\Lambda(x)}{1+e^{\beta+\gamma}\Lambda(x)}\right]$$
$$\times\left(dN_2(x)-Y_2(x)e^{\beta}\{1+e^{\beta+\gamma}\Lambda(t)\}^{e^{-\gamma}-1}\,d\Lambda(x)\right),$$

$\tau<\infty$ being the maximum follow-up time such that $P(X_{ij}\geqslant\tau)>0$ for all i,j.

Suppose now that Λ is unknown. To obtain the test statistic we replace the parameter $\theta=(\beta,\gamma)$ by its value under H_0, i.e. by $(0,0)$ and the function Λ by its non-parametric estimator (also under H_0) from all data:

$$\hat\Lambda(x)=\int_0^x\frac{dN(u)}{Y(u)}.\tag{3}$$

Set $\hat U_k=U_k(0,0,\hat\Lambda)$. For $k=1,2$, we have:

$$\hat U_k=(-1)^k\left(\int_0^\tau\frac{Y_2(u)}{Y(u)}\ln^{k-1}\left(1+\hat\Lambda(u-)\right)dN_1(u)-\int_0^\tau\frac{Y_1(u)}{Y(u)}\ln^{k-1}\left(1+\hat\Lambda(u-)\right)dN_2(u)\right)$$
$$=(-1)^k\left(\int_0^\tau\frac{Y_2(u)}{Y(u)}\ln^{k-1}\left(1+\hat\Lambda(u-)\right)dM_1(u)-\int_0^\tau\frac{Y_1(u)}{Y(u)}\ln^{k-1}\left(1+\hat\Lambda(u-)\right)dM_2(u)\right).\tag{4}$$

Under H_0 the survival functions coincide, i.e. $S_1 = S_2 =: S$, and the means Y_i/n converge to $y_i = G_i S$ uniformly on $[0, \tau]$ as $n \to \infty$. Set $y = y_1 + y_2$. Under H_0 the two-dimensional stochastic process $(n^{-1/2}M_1, n^{-1/2}M_2)$ converges to a Gaussian process $(\tilde{M}_1, \tilde{M}_2)$ with the mean $(0, 0)$ and the covariances

$$\mathbf{cov}\big(\tilde{M}_i(x_1), \tilde{M}_i(x_2)\big) = \int_0^{x_1 \wedge x_2} y_i(u)\,\mathrm{d}\Lambda(u),$$

$$\mathbf{cov}\big(\tilde{M}_1(x_1), \tilde{M}_2(x_2)\big) = \int_0^{x_1}\int_0^{x_2} \tilde{G}(u, v)\big\{\tilde{S}(\mathrm{d}u, \mathrm{d}v) + \tilde{S}(u, \mathrm{d}v)\,\mathrm{d}\Lambda(u) + \tilde{S}(\mathrm{d}u, v)\,\mathrm{d}\Lambda(v) + \tilde{S}(u, v)\,\mathrm{d}\Lambda(u)\,\mathrm{d}\Lambda(v)\big\}$$

(see Prentice and Cai [9]).

Proposition. *The random vector* $(n^{-1/2}\hat{U}_1, n^{-1/2}\hat{U}_2)$ *converges in distribution to the normal random vector* (V_1, V_2) *with the mean* $(0, 0)$ *and the covariance matrix* $\Sigma = \|\sigma_{kl}\|$, *where*

$$\sigma_{kl} = \mathrm{cov}(V_1, V_2) = (-1)^{k+l}\Bigg(\int_0^{\tau} \frac{y_1(u)y_2(u)}{y(u)} \ln^{k+l-2}\big(1 + \Lambda(u)\big)\,\mathrm{d}\Lambda(u) - \int_0^{\tau}\int_0^{\tau} \frac{y_2(u)y_1(v)}{y(u)y(v)}$$
$$\times\, r_{kl}\big(\Lambda(u), \Lambda(v)\big)\tilde{G}(u, v)\big\{\tilde{S}(\mathrm{d}u, \mathrm{d}v) + \tilde{S}(u, \mathrm{d}v)\,\mathrm{d}\Lambda(u) + \tilde{S}(\mathrm{d}u, v)\,\mathrm{d}\Lambda(v) + \tilde{S}(u, v)\,\mathrm{d}\Lambda(u)\,\mathrm{d}\Lambda(v)\big\}\Bigg),$$

here

$$r_{kl}(x, y) = \ln^{k-1}(1 + x)\ln^{l-1}(1 + y) + \ln^{l-1}(1 + x)\ln^{k-1}(1 + y): \quad r_{11}(x, y) = 2, \tag{5}$$
$$r_{12}(x, y) = r_{21}(x, y) = \ln(1 + x) + \ln(1 + y), \qquad r_{22}(x, y) = 2\ln(1 + x)\ln(1 + y).$$

The covariances are consistently estimated by:

$$\hat{\sigma}_{kl} = \frac{(-1)^{k+l}}{n}\Bigg(\int_0^{\tau} \frac{Y_1(u)Y_2(u)}{Y(u)} \ln^{k+l-2}\big(1 + \hat{\Lambda}(u-)\big)\,\mathrm{d}\hat{\Lambda}(u) - \sum_{i=1}^{n}\int_0^{\tau}\int_0^{\tau} \frac{Y_2(u)Y_1(v)}{Y(u)Y(v)} r_{kl}\big(\hat{\Lambda}(u-), \hat{\Lambda}(v-)\big)$$
$$\times \big\{\mathrm{d}N_{1i}(u)\,\mathrm{d}N_{2i}(v) - Y_{1i}(u)\,\mathrm{d}N_{2i}(v)\,\mathrm{d}\hat{\Lambda}(u) - Y_{2i}(v)\,\mathrm{d}N_{1i}(u)\,\mathrm{d}\hat{\Lambda}(v) + Y_{1i}(u)Y_{2i}(v)\,\mathrm{d}\hat{\Lambda}(u)\,\mathrm{d}\hat{\Lambda}(v)\big\}\Bigg). \tag{6}$$

Set $\tilde{\sigma}_{kl} = n\hat{\sigma}_{kl}$, $\tilde{\Sigma} = (\tilde{\sigma}_{kl})$. The limit distribution of the test statistic,

$$X^2 = (\hat{U}_1, \hat{U}_2)\tilde{\Sigma}^{-1}(\hat{U}_1, \hat{U}_2)^{\mathrm{T}}, \tag{7}$$

is chi-square with two degrees of freedom.

The hypothesis is rejected with the approximate significance level α if $X^2 > \chi^2_{1-\alpha}(2)$.

5. The asymptotic power of the test under approaching alternatives

Under the sequence of approaching alternatives,

$$H_n:\quad \lambda_1(x) = \lambda(x), \quad \lambda_2(x) = \mathrm{e}^{c_1/\sqrt{n}}\big\{1 + \mathrm{e}^{(c_1+c_2)/\sqrt{n}}\Lambda(x)\big\}^{\mathrm{e}^{-c_2/\sqrt{n}}-1}\lambda(x),$$

the random vector $(n^{-1/2}\hat{U}_1(x), n^{-1/2}\hat{U}_2(x))$ converges to a normal random vector $(V_1 + \mu_1, V_2 + \mu_2)$, where

$$\mu_k = (-1)^{k-1}\int_0^{\tau} \frac{y_1(u)y_2(u)}{y(u)} \ln^{k-1}\big(1 + \Lambda(u)\big)\big(c_1 - c_2\ln\big(1 + \Lambda(u)\big)\big)\,\mathrm{d}\Lambda(u),$$

so the limit distribution of the statistic X^2 is non-central chi-square with two degrees of freedom and the non-centrality parameter δ:

$$X^2 \to \chi^2(2, \delta), \quad \delta = \mu^{\mathrm{T}} \Sigma^{-1} \mu, \quad \mu = (\mu_1, \mu_2)^{\mathrm{T}}. \tag{8}$$

So the power of the test under the approaching alternatives is written in terms of non-central chi-square distribution. More about chi-squared testing one can see, for example, in Greenwood and Nikulin [6].

6. Simulation study

We studied the significance level and the power of the tests for finite samples when the marginals have the following distributions:

(1) Weibull: $S_1(x) = \mathrm{e}^{-x}$, $S_2(x) = \mathrm{e}^{-(x/\theta)^\nu}$;
(2) loglogistic: $S_1(x) = 1/(1+x)$, $S_2(x) = 1/(1+(x/\theta)^\nu)$;
(3) lognormal: $S_1(x) = 1 - \Phi(\ln x)$, $S_2(x) = 1 - \Phi(\ln(x/\theta)^\nu)$,

and the Clayton copula model holds:

$$C_\alpha(u, v) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}, \quad \alpha > 0$$

Table 1
Finite sample first type error rate ($\alpha = 0.5$, $N = 5000$ simulations, uncensored)

n	Weibull	Loglogistic	Lognormal
50	0.0410	0.0414	0.0368
100	0.0474	0.0434	0.0454

Table 2
Power of the tests ($N = 1000$ simulations)

Law	θ	n	$\alpha = 0.5$			$\alpha = 4$		
			New	Sign	Wilcoxon signed-rank	New	Sign	Wilcoxon signed-rank
Weibull	0.8	50	0.218	0.117	0.131	0.670	0.463	0.583
		100	0.429	0.217	0.277	0.968	0.798	0.873
	0.7	50	0.499	0.278	0.289	0.961	0.785	0.822
		100	0.865	0.552	0.505	1.000	0.982	0.985
	0.6	50	0.801	0.453	0.381			
		100	0.988	0.793	0.664			
	0.5	50	0.960	0.649	0.492			
		100	1.000	0.917	0.734			
Loglogistic	0.8	50	0.148	0.052	0.050	0.383	0.042	0.062
		100	0.329	0.080	0.058	0.877	0.069	0.082
	0.7	50	0.381	0.065	0.054	0.837	0.091	0.076
		100	0.714	0.135	0.064	1.000	0.142	0.084
	0.6	50	0.687	0.114	0.059			
		100	0.949	0.185	0.055			
	0.5	100	0.894	0.140	0.058			
		100	0.999	0.256	0.068			
Lognormal	0.8	50	0.228	0.105	0.097	0.604	0.221	0.181
		100	0.486	0.197	0.149	0.956	0.463	0.307
	0.7	50	0.529	0.181	0.133	0.957	0.428	0.245
		100	0.868	0.394	0.187	1.000	0.750	0.417
	0.6	50	0.834	0.316	0.151			
		100	0.986	0.540	0.230			
	0.5	50	0.969	0.403	0.166			
		100	1.000	0.695	0.235			

(see Clayton and Cuzick [4], Bagdonavicius and Nikulin [2]). Sample sizes $n = 50$ and $n = 100$ were considered. The values of the first type error rate with nominal level 0.05 are given in Table 1.

The power of the test for slightly correlated data ($\alpha = 0.5$ which corresponds to the Kendal's correlation coefficient $\tau_K = 0.2$) and strongly correlated data ($\alpha = 4$, $\tau_K = 0.67$) is given in Table 2. The proposed test is considerably more powerful than the sign and Wilcoxon signed-rank tests.

References

[1] V. Bagdonavicius, M. Nikulin, Accelerated Life Models, Chapman and Hall/CRC, Boca Raton, 2002.

[2] V. Bagdonavicius, M. Nikulin, Semiparametric Models in Accelerated Life Testing, Queen's Papers in Pure and Applied Mathematics, vol. 98, Kingston, Ontario, Canada, 1995.

[3] A. Cantor, R. Knapp, A test of the equality of survival distributions based on paired observations from conditionally exponential distributions, IEEE Trans. Reliability 34 (1985) 342–346.

[4] D. Clayton, J. Cuzick, EM algorithm for Cow's regression model using GLIM, J. Roy. Statist. Soc. Ser. C 34 (2) (1985) 148–156.

[5] P. Georges, A.-G. Lamy, E. Nicolas, G. Quibel, T. Roncalli, Multivariate survival modelling: a unified approach with copulas, Groupe de Recherch Operationelle, Crédit Lyonnais, 2001.

[6] P.E. Greenwood, M. Nikulin, A Guide to Chi-Squared Testing, John Wiley and Sons, New York, 1996.

[7] W. Owen, D. Sinha, M. Capozzoli, A paired-data analysis for a lifetime distribution, Amer. Statist. 54 (2000) 252–256.

[8] W. Owen, A power analysis of tests for paired lifetime data, Lifetime Data Anal. 11 (2005) 233–243.

[9] R. Prentice, J. Cai, Covariance and survivor function estimation using censored multivariate failure time data, Biometrika 79 (1992) 495–512.

[10] A. Sklar, Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8 (1959) 229–231.