Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 14

5-1-2004

Kernel-Based Estimation of P(X less than Y)With Paired Data

Omar M. Eidous
Yarmouk University, Irbid, Jordan, omarm@yu.edu.jo

Ayman Baklizi
Yarmouk University,Irbid, Jordan, baklizi1@hotmail.com

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 143-148

1538 – 9472/04/$95.00

Kernel-Based Estimation of $\mathbf{P(X < Y)}$ With Paired Data

Omar Eidous
Department of Statistics
Irbid – Jordan

Ayman Baklizi
Department of Statistics
Irbid – Jordan

A point estimation of $P(X < Y)$ was considered. A nonparametric estimator for $P(X < Y)$ was developed using the kernel density estimator of the joint distribution of *X* and *Y*, may be dependent. The resulting estimator was found to be similar to the estimator based on the sign statistic, however it assigns smooth continuous scores to each pair of the observations rather than the zero or one scores of the sign statistic. The asymptotic equivalence of the sign statistic and the proposed estimator is shown and a simulation study is conducted to investigate the performance of the proposed estimator. Results indicate that the estimator has a good overall performance.

Key words: Kernel density estimation, stress-strength reliability, paired data

Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n independent pairs of possibly dependent jointly distributed random variables (X, Y). The aim is to estimate $p = P(X < Y)$ using this information. The problem of estimation the stress-strength reliability arises naturally in the context of mechanical reliability of a system with strength *X* and stress *Y*. The system fails any time its strength is exceeded by the stress applied to it. Another interpretation of p is that it measures the effect of the treatment when X is the response for a control group and Y refers to the treatment group. Other applications can be found in Johnson et. al. (1994) and the references therein.

Ayman Baklizi is an Assistant Professor of Applied Statistics. His research interests are in accelerated life tests and censored data. Email: baklizi1@hotmail.com. Omar Eidous is an Assistant Professor of Applied Statistics. His research interests are in line transect sampling and kernel methods. Email: omarm@yu.edu.jo

The sign statistic (Lehmann, 1975) is defined as $S = \sum_{i=1}^{n} \phi(X_i, Y_i)$ where $\phi(X_i, Y_i) = 1$ if $X_i < Y_i$ and 0 otherwise. It is readily seen that

$$E(S) = E\left(\sum_{i=1}^{n} \phi(X_i, Y_i)\right) = nP(X < Y),$$

therefore an unbiased estimator for p is given by

$$\hat{p}_1 = \frac{1}{n} E\left(\sum_{j=1}^{n} \phi(X_i, Y_i)\right).$$

Develop in this article is a new estimator for $P(X < Y)$ using kernel methods (Silverman, 1986). The kernel density estimators are used instead of the true unknown density and the estimator of p is introduced with some of its large sample properties. A simulation study was conducted to evaluate the performance of the proposed estimator and compare it with the estimator $\hat{p}_1$.

Methodology

The Kernel – Based Estimator

Let $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ be n independent pairs drawn from the distribution with joint probability density function $f(x,y)$. The desired parameter p to estimate is

$$p = P(X < Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(x,y)dxdy . \qquad (1)$$

In this article a nonparametric kernel method is used to estimate p. The kernel estimator of the two dimensional probability density function f at (x,y) is defined as (Scott, 1992).

$$\hat{f}(x,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)K\left(\frac{y-Y_i}{h_2}\right), \qquad (2)$$

where h_1 and h_2 are positive numbers control the smoothness of the fitted curve, usually called bandwidths or smoothing parameters. $K(u)$ is a kernel function which is a symmetric probability density. Comprehensive reviews of the kernel method are available in Silverman (1986); Scott (1992); Wand and Jones (1995). The proposed estimator of the parameter p is constituted by substituting formula (2) in (1) as an estimator for $f(x,y)$. The resulting estimator is of the form

$$\hat{p}_2 = \frac{1}{nh_1h_2}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{y} K\left(\frac{x-X_i}{h_1}\right)K\left(\frac{y-Y_i}{h_2}\right)dxdy . \qquad (3)$$

If the two random variables X and Y are defined on the positive real line, transform the positive data by taking logarithms of each observation as suggested by Silverman (1986).

To construct the kernel estimator $\hat{p}_2$ kernel function K and smoothing parameters h_1 and h_2 must be chosen. For example, the widely used criterion is to choose K, h_1 and h_2 that minimize the mean integral square error ($MISE$) of $\hat{f}(x,y)$. As many authors stated, there is very little to choose between the various kernels as they all contribute the similar amount to the $MISE$ (See Silverman, 1986 and Wand and Jones, 1995). The based-data formulas to choose h_1 and h_2 are given later in this paper.

Large Sample Properties

Consider the following transformation, $u = \dfrac{x-X_i}{h_1}$ and $v = \dfrac{y-Y_i}{h_2}$ it follows that

$$\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{(vh_2+Y_i-X_i)/h_1} K(u)K(v)dudv .$$

Since the kernel function K is a probability density function. When $h_1 \to 0$ and $h_2 \to 0$ such that $h_1 = O(h_2)$ as $n \to \infty$ the summand will be either zero or one depending on whether $X_i > Y_i$ or $X_i < Y_i$. Hence the integral approaches one if $X_i < Y_i$ and zero if $X_i > Y_i$. Thus the limiting value of $\hat{p}_2$ is $\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i, Y_i)$ which is the estimator based on the sign statistic. Consider,

$$\begin{aligned} E(\hat{p}_1) &= \frac{1}{n}\sum_{i=1}^{n} E(\phi(X_i, Y_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} P(X < Y) \\ &= P(X < Y). \end{aligned}$$

Also

$$\begin{aligned} \mathrm{var}(\hat{p}_1) &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(\phi(X_i, Y_i)) \\ &= \frac{1}{n}\mathrm{var}(\phi(X,Y)) \\ &= \frac{1}{n}p(1-p) . \end{aligned}$$

Because $\hat{p}_1$ and $\hat{p}_2$ are asymptotically equivalent, it follows that $\hat{p}_2$ is asymptotically unbiased. Because the variance of $\hat{p}_2$ tends to zero as $n \to \infty$, the proposed estimator is consistent. The limiting distributions of $\hat{p}_1$ and $\hat{p}_2$ are also the same.

Small Sample Performance of the Estimators

To implement the estimator $\hat{p}_2$ in practice the kernel function K and the smoothing parameters h_1 and h_2 need to be chosen. Consider the following two kernel functions. The standard normal kernel :

$$K_1(u) = (2\pi)^{-1/2} e^{-u^2/2}, \qquad -\infty < u < \infty$$

The rectangular kernel :

$$K_2(u) = \begin{cases} 1/2, & -1 < u < 1 \\ 0, & otherwise \end{cases},$$

Note here that if $u = (x - X_i)/h_1$ then

$$K_1\left(\frac{x - X_i}{h_1}\right) = (2\pi)^{-1/2} e^{-(x - X_i)^2/2h_1^2} \quad -\infty < x < \infty$$

and

$$K_2\left(\frac{x - X_i}{h_1}\right) = \begin{cases} 1/2, & -h_1 + X_i < x < h_1 + X_i \\ 0, & otherwise \end{cases}.$$

By adopting K_1 it is easy to show that (3) becomes

$$\hat{p}_2 = \frac{h_1}{nh_2\sqrt{2\pi}} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \Phi(t)\ e^{-(h_1 t + Y_i - X_i)^2/\ (2h_2^2)} dt \,. \qquad (4)$$

where $\Phi(t)$ is the normal distribution function at t. The estimator $\hat{p}_2$ which is given by equation (4) cannot be written in closed form, so derive the estimator $\hat{p}_2$ corresponding to K_2. If K_2 is adopted in (3) then we need to study six cases to find the double integral arises in (3). Let $a(i) = -h_1 + X_i$; $b(i) = h_1 + X_i$; $c(i) = -h_2 + Y_i$ and $d(i) = h_2 + Y_i$, where $i = 1,2,\cdots,n$. Notice that, $a(i) < b(i)$ and $c(i) < d(i)$. The proposed estimator is given by

$$\hat{p}_2 = \frac{1}{nh_1h_2} \sum_{i=1}^{n} Q_i \text{ , where}$$

$$Q_i = \int_{-\infty}^{\infty} \int_{-\infty}^{y} K\left(\frac{x - X_i}{h_1}\right) K\left(\frac{y - Y_j}{h_2}\right) dxdy \text{ , then}$$

Case1: If $d(i) < a(i)$ then $Q_i = 0$

Case2: If $b(i) < c(i)$ then $Q_i = h_1 h_2$

Case3: If $c(i) < a(i) < b(i) < d(i)$ then

$$Q_i = \int_{a(i)}^{b(i)} \int_{x}^{d(i)} \frac{1}{4} dydx$$

$$= \frac{1}{4}(b(i)d(i) - b(i)^2/2 - a(i)d(i) + a(i)^2/2)$$

Case4: If $a(i) < c(i) < d(i) < b(i)$ then

$$Q_i = \int_{c(i)}^{d(i)} \int_{a(i)}^{y} \frac{1}{4} dxdy$$

$$= \frac{1}{4}(d(i)^2/2 - a(i)d(i) - c(i)^2/2 + a(i)c(i)).$$

Case5: If $c(i) < a(i) < d(i) < b(i)$ then

$$Q_i = \int_{a(i)}^{d(i)} \int_{a(i)}^{y} \frac{1}{4} dxdy$$

$$= \frac{1}{4}(d(i)^2/2 + a(i)^2/2 - a(i)d(i)).$$

Case6: If $a(i) < c(i) < b(i) < d(i)$ then

$$Q_i = \int_{a(i)}^{c(i)} \int_{c(i)}^{d(i)} \frac{1}{4} dydx + \int_{c(i)}^{b(i)} \int_{x}^{d(i)} \frac{1}{4} dydx$$

$$= \frac{1}{4}\Big(-(c(i)^2 + b(i)^2)/2 + d(i)(b(i) - a(i)) + a(i)c(i)\Big).$$

On the other hand, the simulation results are depended on the following formulas to choose the smoothing parameters h_1 and h_2

which based on minimizing the asymptotic mean integral square error and by assuming the bivariate normal distribution and the rectangular kernel K_2 (Scott, 1992)

$$h_1 = 1.745(1-\rho^2)^{5/12}(1+\rho^2/2)^{-1/6}\sigma_1\ n^{-1/6}$$
$$h_2 = 1.745(1-\rho^2)^{5/12}(1+\rho^2/2)^{-1/6}\sigma_2\ n^{-1/6},$$

where σ_1, σ_2 and ρ are the standard deviations and the correlation coefficient respectively, they are estimated from the data. The performances of the sign estimator and the proposed estimator were investigated and compared. The criteria of the bias and mean squared error are used. The relative efficiency of the proposed estimator to the sign estimator is calculated as the ratio of mean squared errors.

A simulation study was conducted to investigate the performance of the estimators. The indices of our simulations are: $n = 10, 20, 40$ p : the true value of $p=p(X<Y)$ and is taken to be 0.1,0.3,…, 0.9.

The distribution from which the data are generated: two cases were considered;

1) The bivariate normal distribution $(X,Y) \sim BVN(0, \mu, 1, 1, \rho)$ where ρ is taken as -.8, -0.4, 0, 0.4, 0.8 and μ is chosen such that we get $p=p(X<Y)$ as chosen above.

2) The Gumbel bivariate exponential distribution (Johnson & Kotz, 1970) with probability density function

$$g(x,y) = e^{-(x+y)}\left(1+\alpha\left(2e^{-x}-1\right)\left(2e^{-y}-1\right)\right),$$
$$x>0, y>0.$$

The parameter α is chosen such that the correlation (r) between X and Y is -0.4, -0.2, 0, 0.2, 0.4. The variable X is transformed such that we get $p=p(X<Y)$ as chosen above.

For each combination of n and p, 1000 samples were generated for (X,Y). The estimators are calculated and the following quantities are obtained for both estimators:

The bias of the estimators, $Bias = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{p}_2^{(i)} - p\right)$.

Mean squared errors, $MSE_j = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{p}_j^{(i)} - p\right)^2$, j = 1, 2.

$Efficiency = \frac{MSE_1}{MSE_2}$. The results are presented in Tables 1 - 2.

Conclusion

Because the results for both kernels are similar, only the results for the uniform kernel are given. The results are presented in Table 1. In both cases of bivariate normal parent distribution and the bivariate exponential case, it is clear that the efficiency of the proposed estimator relative to the sign estimator is greater than one in all cases considered. Concerning the bias performance, it appears that the proposed estimator is almost unbiased. Overall it appears that the proposed estimator has a good performance, this performance may be improved when using more sophisticated types of kernels, bandwidth selection rules, and bias corrections.

References

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions: Volume 1*. New York: Wiley.

Lehmann, E. L. (1975) *Nonparametrics: Statistical methods based on ranks.* Holden Day: San Francisco.

Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.

Silverman, B. W. (1986). *Density estimation*. London: Chapman & Hall.

Wand, M., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Table 1: Mean Squared Errors and Efficiencies of the Estimators in the Bivariate Normal Case

n	ρ	p	Var	Bias	MSE	EFF	n	ρ	p	var	Bias	MSE	EFF
10	-0.80	0.10	0.009	0.003	0.009	1.052		0.40	0.10	0.005	0.006	0.004	1.128
	-0.80	0.30	0.021	0.001	0.021	1.027		0.40	0.30	0.011	0.003	0.009	1.129
	-0.80	0.50	0.025	-0.011	0.025	1.032		0.40	0.50	0.013	0.005	0.010	1.174
	-0.80	0.70	0.021	0.005	0.020	1.042		0.40	0.70	0.011	0.002	0.009	1.152
	-0.80	0.90	0.009	-0.001	0.009	1.045		0.40	0.90	0.005	0.000	0.004	1.128
	-0.40	0.10	0.009	0.001	0.008	1.132		0.80	0.10	0.005	0.002	0.004	1.114
	-0.40	0.30	0.021	-0.005	0.019	1.083		0.80	0.30	0.011	-0.003	0.009	1.109
	-0.40	0.50	0.025	-0.002	0.023	1.101		0.80	0.50	0.013	-0.004	0.011	1.102
	-0.40	0.70	0.021	0.003	0.019	1.099		0.80	0.70	0.011	0.000	0.009	1.080
	-0.40	0.90	0.009	0.002	0.008	1.120		0.80	0.90	0.005	0.000	0.004	1.175
	0.00	0.10	0.009	-0.002	0.007	1.188	40	-0.80	0.10	0.003	0.000	0.002	1.037
	0.00	0.30	0.021	0.000	0.019	1.135		-0.80	0.30	0.006	-0.001	0.005	1.023
	0.00	0.50	0.025	-0.002	0.023	1.132		-0.80	0.50	0.007	0.002	0.006	1.032
	0.00	0.70	0.021	0.005	0.018	1.126		-0.80	0.70	0.006	0.003	0.005	1.034
	0.00	0.90	0.009	-0.003	0.008	1.123		-0.80	0.90	0.003	0.000	0.002	1.030
	0.40	0.10	0.009	0.007	0.007	1.184		-0.40	0.10	0.003	0.001	0.002	1.074
	0.40	0.30	0.021	0.003	0.018	1.147		-0.40	0.30	0.006	0.006	0.005	1.079
	0.40	0.50	0.025	-0.002	0.020	1.129		-0.40	0.50	0.007	-0.003	0.006	1.109
	0.40	0.70	0.021	-0.004	0.019	1.178		-0.40	0.70	0.006	0.000	0.005	1.115
	0.40	0.90	0.009	-0.001	0.007	1.134		-0.40	0.90	0.003	-0.001	0.002	1.106
	0.80	0.10	0.009	0.002	0.007	1.154		0.00	0.10	0.003	0.002	0.002	1.112
	0.80	0.30	0.021	0.001	0.019	1.144		0.00	0.30	0.006	0.004	0.005	1.132
	0.80	0.50	0.025	0.002	0.023	1.147		0.00	0.50	0.007	0.000	0.005	1.104
	0.80	0.70	0.021	-0.004	0.019	1.106		0.00	0.70	0.006	0.001	0.004	1.124
	0.80	0.90	0.009	0.006	0.008	1.145		0.00	0.90	0.003	-0.004	0.002	1.164
20	-0.80	0.10	0.005	0.000	0.004	1.048		0.40	0.10	0.003	0.000	0.002	1.145
	-0.80	0.30	0.011	0.002	0.011	1.027		0.40	0.30	0.006	-0.002	0.004	1.144
	-0.80	0.50	0.013	0.004	0.012	1.030		0.40	0.50	0.007	-0.001	0.006	1.130
	-0.80	0.70	0.011	0.000	0.011	1.022		0.40	0.70	0.006	0.001	0.005	1.165
	-0.80	0.90	0.005	0.001	0.004	1.049		0.40	0.90	0.003	-0.003	0.002	1.161
	-0.40	0.10	0.005	-0.002	0.004	1.093		0.80	0.10	0.003	-0.001	0.002	1.093
	-0.40	0.30	0.011	0.000	0.009	1.097		0.80	0.30	0.006	0.000	0.005	1.090
	-0.40	0.50	0.013	0.000	0.011	1.092		0.80	0.50	0.007	-0.001	0.006	1.094
	-0.40	0.70	0.011	0.003	0.009	1.113		0.80	0.70	0.006	-0.004	0.005	1.082
	-0.40	0.90	0.005	0.000	0.004	1.111		0.80	0.90	0.003	-0.001	0.002	1.090
	0.00	0.10	0.005	0.001	0.004	1.178							
	0.00	0.30	0.011	0.004	0.009	1.148							
	0.00	0.50	0.013	-0.001	0.011	1.152							
	0.00	0.70	0.011	-0.003	0.009	1.110							
	0.00	0.90	0.005	-0.001	0.004	1.120							

Table 2: Mean Squared Errors and Efficiencies of the Estimators in the Bivariate Exponential Case

(r)	n	p	VAR	Bias	MSE	EFF	(r)	n	P	VAR	Bias	MSE	EFF
-0.4	10	0.1	0.009	-0.019	0.006	1.266	0.2	10	0.1	0.011	0.032	0.009	1.158
-0.4		0.3	0.021	-0.028	0.014	1.303	0.2		0.3	0.021	0.029	0.017	1.268
-0.4		0.5	0.025	0.004	0.019	1.320	0.2		0.5	0.024	-0.007	0.019	1.266
-0.4		0.7	0.021	0.029	0.016	1.266	0.2		0.7	0.022	-0.021	0.018	1.263
-0.4		0.9	0.009	0.017	0.006	1.202	0.2		0.9	0.010	-0.024	0.009	1.150
-0.4	20	0.1	0.005	-0.024	0.003	1.232	0.2	20	0.1	0.005	0.024	0.005	1.153
-0.4		0.3	0.011	-0.035	0.009	1.316	0.2		0.3	0.011	0.029	0.010	1.242
-0.4		0.5	0.013	-0.002	0.009	1.260	0.2		0.5	0.013	-0.003	0.010	1.328
-0.4		0.7	0.011	0.034	0.009	1.334	0.2		0.7	0.011	-0.020	0.009	1.206
-0.4		0.9	0.005	0.021	0.003	1.191	0.2		0.9	0.005	-0.030	0.005	1.119
-0.4	40	0.1	0.003	-0.024	0.002	1.292	0.2	40	0.1	0.003	0.025	0.003	1.056
-0.4		0.3	0.006	-0.036	0.005	1.319	0.2		0.3	0.006	0.024	0.005	1.186
-0.4		0.5	0.007	0.005	0.005	1.298	0.2		0.5	0.007	-0.003	0.005	1.300
-0.4		0.7	0.006	0.036	0.005	1.320	0.2		0.7	0.006	-0.021	0.004	1.183
-0.4		0.9	0.003	0.023	0.002	1.220	0.2		0.9	0.003	-0.025	0.003	1.066
-0.2	10	0.1	0.009	-0.006	0.007	1.190	0.4	10	0.1	0.013	0.046	0.012	1.119
-0.2		0.3	0.021	-0.018	0.016	1.277	0.4		0.3	0.024	0.041	0.019	1.244
-0.2		0.5	0.025	-0.005	0.019	1.326	0.4		0.5	0.025	0.001	0.019	1.294
-0.2		0.7	0.021	0.011	0.016	1.255	0.4		0.7	0.023	-0.037	0.019	1.232
-0.2		0.9	0.009	0.008	0.007	1.200	0.4		0.9	0.013	-0.047	0.011	1.135
-0.2	20	0.1	0.005	-0.010	0.003	1.243	0.4	20	0.1	0.005	0.050	0.008	1.074
-0.2		0.3	0.011	-0.010	0.008	1.275	0.4		0.3	0.011	0.043	0.011	1.161
-0.2		0.5	0.013	0.000	0.009	1.329	0.4		0.5	0.013	-0.001	0.010	1.364
-0.2		0.7	0.011	0.016	0.009	1.303	0.4		0.7	0.011	-0.040	0.010	1.204
-0.2		0.9	0.005	0.009	0.004	1.210	0.4		0.9	0.005	-0.046	0.007	1.098
-0.2	40	0.1	0.003	-0.010	0.002	1.262	0.4	40	0.1	0.003	0.048	0.005	0.986
-0.2		0.3	0.006	-0.013	0.004	1.288	0.4		0.3	0.006	0.034	0.006	1.135
-0.2		0.5	0.007	0.001	0.005	1.292	0.4		0.5	0.007	0.000	0.005	1.302
-0.2		0.7	0.006	0.016	0.004	1.258	0.4		0.7	0.006	-0.039	0.006	1.099
-0.2		0.9	0.003	0.010	0.002	1.241	0.4		0.9	0.003	-0.046	0.004	0.976
0.0	10	0.1	0.005	0.005	0.007	1.232							
0.0		0.3	0.011	0.006	0.017	1.275							
0.0		0.5	0.013	0.001	0.018	1.299							
0.0		0.7	0.011	-0.011	0.017	1.306							
0.0		0.9	0.005	-0.006	0.007	1.192							
0.0	20	0.1	0.005	0.006	0.004	1.225							
0.0		0.3	0.011	0.012	0.009	1.256							
0.0		0.5	0.013	-0.004	0.009	1.271							
0.0		0.7	0.011	-0.007	0.008	1.257							
0.0		0.9	0.005	-0.007	0.004	1.215							
0.0	40	0.1	0.003	0.005	0.002	1.179							
0.0		0.3	0.006	0.006	0.004	1.235							
0.0		0.5	0.007	-0.002	0.005	1.264							
0.0		0.7	0.006	-0.008	0.004	1.278							
0.0		0.9	0.003	-0.005	0.002	1.155							