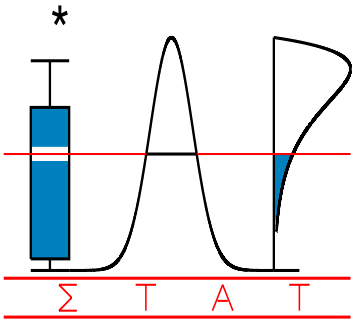
T E C H N I C A L

R E P O R T

0454

ESTIMATING KENDALL'S TAU FOR BIVARIATE INTERVAL CENSORED DATA WITH A SMOOTH ESTIMATE OF THE DENSITY.

E. LESAFFRE and K. BOGAERTS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Estimating Kendall's tau for bivariate interval censored data with a smooth estimate of the density.

Emmanuel Lesaffre[1] and Kris Bogaerts

Katholieke Universiteit Leuven, Biostatistical Centre, Leuven, Belgium

Abstract

Measures of association for bivariate interval censored data have not yet been studied extensively. Betensky and Finkelstein[3] proposed to calculate Kendall's coefficient of concordance using a multiple imputation technique. However, this method is quite computer intensive.
Our approach is based on two steps. First, we fit a bivariate smooth estimate of the density of log-event times on a fixed grid. The smoothing technique is based on a mixture of Gaussian densities with weights determined by a penalized likelihood approach. In a second step we plug the expression of the smoothed density into the population's version of Kendall's tau, which becomes a weighted sum of constants calculated from the grid.
The performance of our method is illustrated by a simulation study and is applied to tooth emergence data of 7 permanent teeth measured on 4468 children from the Signal-Tandmobiel®study.

Key words: Bivariate survival; interval-censored; Kendall's tau.

1 Introduction

Measures of association are well studied and often applied to data that are completely observed. Some measures have been extended to right censored data (e.g. [16], [8], [7]). However for interval censored data, association measures have not yet been studied extensively.
In absence of censoring, Kendall's tau (τ) is based on scores assigned to each pair of bivariate observations, say $(X_1, Y_1), (X_2, Y_2)$ that measure the concordance between the two observations. More specifically, (X_1, Y_1) and (X_2, Y_2) are said to be concordant if $X_1 > X_2$ and $Y_1 > Y_2$ or if $X_1 < X_2$ and $Y_1 < Y_2$ and they are discordant if $X_1 > X_2$ and $Y_1 < Y_2$ or if $X_1 < X_2$ and $Y_1 > Y_2$. Concordant pairs are assigned a score of 1, discordant pairs are assigned a score of -1, and pairs in which there is equality among either variable are assigned a score of 0. Kendall's tau is calculated as the average of these scores over all pairs of observations and in this way it estimates the difference between the probability of concordance and the probability of discordance. In the presence of censoring, things are more complicated. Oakes[16] proposed to estimate Kendall's tau for bivariate right censored data by assigning zero to pairs of observations that cannot be compared. Following Oakes's approach, Betensky of Finkelstein[3] suggested to calculate Kendall's tau in the presence of interval censoring using a multiple imputation strategy. However, this method is quite computer intensive for large data sets.
Our method is based on 2 steps. First we approximate the bivariate density of the log

[1]Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium, E-mail: Emmanuel.Lesaffre@med.kuleuven.ac.be

of the event times by a smoothing technique. The smoothing technique is an extension of the approach used by Ghidey et al.[10] for smoothing the random effects distribution in a linear mixed model and by Komárek et al.[13] for smoothing the distribution of the error term in an accelerated failure time model. More specifically the smooth density is a mixture of Gaussian densities fixed on a bivariate grid with weights determined by a penalized likelihood approach. In the second step, the estimated smoothed bivariate cumulative distribution function $\widehat{F}$ can be plugged into the expression $\tau = 4\int F dF - 1$, which is the population's version of Kendall's tau.
The next section describes the smoothing procedure and its mathematical properties. The calculation and properties of Kendall's tau are described in Section 3. The results of the simulations are presented in Section 4. The application to tooth emergence data of the Signal Tandmobiel®study is described in Section 5. In Section 6 our approach is critically examined.

2 Smooth estimate of the bivariate density

2.1 Smoothing Method

A detailed description of the smoothing method can be found in Bogaerts and Lesaffre[4]. Briefly, let (T_1, T_2) represent a positive valued bivariate random vector with density f. Let T_1 and T_2 be interval censored in the rectangle $(t_{1l}, t_{1r}] \times (t_{2l}, t_{2r}]$ by an independent censoring process. We also include here the special cases, i.e. left ($t_l = 0$) and right censoring ($t_r = \infty$). The smoothing procedure is an extension of the approach of Ghidey et al.[10] and Komárek et al.[13]. The bivariate density of $\log(T_1)$ and $\log(T_2)$ is modelled as a weighted sum of bivariate normal distributions with zero correlation over a (fixed) fine grid of size $k_1 \times k_2$ with means equal to the gridpoints of the grid and variances equal but fixed. Thus, we assume that

$$\begin{pmatrix} \log(T_1) \\ \log(T_2) \end{pmatrix} \sim \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} c_{ij}\mathcal{N}(\mu_{\mathbf{ij}}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}) \tag{1}$$

where $c_{ij} > 0, \forall i, j$ and $\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} c_{ij} = 1$. The aim is to estimate the weights $c_{ij} (i = 1, \ldots, k_1, j = 1, \ldots, k_2)$. Note that this involves a constrained maximum likelihood procedure in the $k_1 \times k_2$ weight parameters. Unconstrained maximum likelihood estimation is obtained by introducing parameters a_{ij} as

$$c_{ij} = \frac{e^{a_{ij}}}{\sum_{i,j} e^{a_{ij}}},$$

with, say $a_{11} = 0$ to ensure identifiability.
Following the work of Eilers and Marx [9] a penalty term is used to smooth our approximation to the true density f. The penalty equals to

$$p = \frac{\lambda_1}{2}\sum_{i,j}(\nabla_i^k a_{ij})^2 + \frac{\lambda_2}{2}\sum_{i,j}(\nabla_j^k a_{ij})^2 \tag{2}$$

where λ_1 (> 0) and λ_2 (> 0) are "smoothing" parameters and ∇^k is the k^{th} order difference operator.
Given λ_1, λ_2, let l_n denote the loglikelihood for a sample of size n and p the penalty defined in (2). Maximizing the penalized loglikelihood $l_{P,n} = l_n - p$ with respect to $\mathbf{a} = (a_{11}, \ldots, a_{k_1 k_2})^T$, yields estimates $\widehat{a}_{ij} (i = 1, \ldots, k_1, j = 1, \ldots, k_2)$.
The parameters λ_1 and λ_2 are assumed to be given, they determine the smoothness of the density, i.e. the larger the smoother the density will be. The optimum λ_1 and λ_2 correspond to a minimum Akaike's Information Criterium (AIC) ([1]) defined as $AIC = -2 \times$log-likelihood$+2 \times$"effective degrees of freedom". The "effective degrees of freedom" can be determined as follows[11]:

$$df \;=\; trace\left[H_{L_P}^{-1} H_L\right]$$

where $H_L = -\frac{\partial^2 l_n}{\partial \mathbf{a} \partial \mathbf{a}^T}$ and $H_{L_P} = -\frac{\partial^2 l_{P,n}}{\partial \mathbf{a} \partial \mathbf{a}^T}$. The optimum λ_1 and λ_2 can be found by a grid search or a parabolic interpolation search.

2.2 Statistical Properties

2.2.1 Consistency

When parameters $\mathbf{a}_0$ exist such that the truth can be written like (1) with a grid and variances equal to the one to analyze the data, it can be shown that the parameters are consistently estimated[10]. So, $\widehat{\mathbf{a}}_n \to \mathbf{a}_0$. However, in general the true distribution can not be written as a weighted sum of normal distributions with means at a pre-specified grid. Using White's theory[20], one can show that all the parameter estimates asymptotically converge to a function that minimizes the Kullback-Leibler distance. From limited simulations (results are not reported) we observed that the Kullback-Leibler distance goes to zero for a fine enough grid and a large sample size.

2.2.2 Asymptotical Normality

Under the same conditions as in 2.2.1 it can also be shown that the estimated parameters are asymptotically normally distributed. Namely, $\sqrt{n}(\widehat{\mathbf{a}}_n - \mathbf{a}_0) \to \mathcal{N}(\mathbf{0}, \Sigma)$ where Σ can be consistently estimated by

$$n H_{L_P}^{-1}(\widehat{\mathbf{a}}_n) H_L^{-1}(\widehat{\mathbf{a}}_n) H_{L_P}^{-1}(\widehat{\mathbf{a}}_n),$$

where $H_L = -\frac{\partial^2 l_n}{\partial \mathbf{a} \partial \mathbf{a}^T}$ and $H_{L_P} = -\frac{\partial^2 l_{P,n}}{\partial \mathbf{a} \partial \mathbf{a}^T}$.

3 Kendall's tau

The association between two survival times can be expressed by Kendall's tau[12], which is equal to

$$\tau \;=\; 4 \cdot \int F \mathrm{d}F - 1, \tag{3}$$

where F is the bivariate cumulative distribution function of f. Our approach consists in replacing F by the cumulative distribution of the bivariate smoothed function in expression (3). This leads to the following expression for the estimate of τ (see Appendix):

$$\widehat{\tau} \;=\; 4 \cdot \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \sum_{k=1}^{k_1} \sum_{l=1}^{k_2} \widehat{c}_{ij} \widehat{c}_{kl} \Phi\left(\frac{\mu_{1,i} - \mu_{1,k}}{\sqrt{2}\sigma_1}\right) \Phi\left(\frac{\mu_{2,j} - \mu_{2,l}}{\sqrt{2}\sigma_2}\right) - 1$$

where $\widehat{c}_{ij}$ and $\widehat{c}_{kl}$ are the estimated coefficients and Φ denotes the univariate cumulative standard normal distribution. Clearly, given the coefficients $\widehat{c}_{ij}$, the calculation of $\widehat{\tau}$ is readily done.
Based on the variance-covariance matrix of $\widehat{\mathbf{a}}$ (see Section 2.2.2) and using the delta method, one can easily derive the variance and also a (95%) confidence interval for $\widehat{\tau}$. Further, for $\widehat{\tau_1}$ and $\widehat{\tau_2}$ estimated for two independent groups of subjects a two-sample Z-test can be derived to test $H_0 : \tau_1 = \tau_2$ (see Appendix for details). A SAS macro (version 8.2) has been written to estimate $\widehat{\tau}$ for interval censored data and can be downloaded from http://www.med.kuleuven.ac.be/biostat/research/software.htm.

4 Simulation Study

For the simulation study, independent failure times were simulated from a bivariate log-normal distribution (scenario 1, $\tau = 0$). In addition, failure times were simulated from 5 different scenario's (scenario 2 to 6) with a given τ different from zero: 1) a bivariate log-normal distribution ($\tau = 0.41$), 2) an equal mixture of two bivariate log-normal distributions with the same variance ($\tau = 0.63$), 3) an equal mixture of two bivariate log-normal distributions with the different variances ($\tau = 0.49$), 4) an unequal mixture of two bivariate log-normal distributions with the same variance and with two modes for both marginals ($\tau = 0.54$) and 5) an unequal mixture of two bivariate log-normal distributions with the same variance but with only one mode for one marginal and two modes in the other marginal ($\tau = 0.26$).
Two different independent censoring schemes were applied to the (uncensored) data: 1) about 10% left, 70% interval and 20% right censoring and 2) about 10% left, 50% interval and 40% right censoring. This was done by generating 6 visit times and a drop out process (both independently of the failure times).
The sample sizes were 100 and 500. Two gridsizes were examined i.e. 10×10 and 20×20, but the 10×10 grid was not always satisfactory and is therefore not further considered here. For each setting 1000 simulations were performed. Both smoothing parameters were assumed to be equal to each other. A grid search with 10 values ranging from 0.001 to 500 was performed in order to choose the smoothing parameters. Third order differences were used in the penalty. The variances were set to the square of 2/3 of the gridsize (see [4]).
For each choice of the smoothing parameters, Kendall's tau and its corresponding variance were calculated using the method. As a benchmark, Kendall's tau was also estimated for the uncensored failure times using the standard expression.
For scenario 1 we investigated the type I error for testing $H_0 : \tau = 0$. Our results showed that the probability of the type I error ranged from 5.5% to 7.2% and approached the nominal level when the sample size increases.
Table 1 displays the mean difference with corresponding standard error between Kendall's tau calculated using our method and the true Kendall's tau from the distribution from which data was simulated. Table 2 displays the mean difference with corresponding standard error between Kendall's tau calculated on the censored (using our method) and uncensored observations (using the standard formula). For both tables, very small mean differences were observed and the mean difference decreased with increasing sample size for all simulation settings.

		Censoring 70% interval 10% left 20% right				Censoring 50% interval 10% left 40% right			
		N=100		N=500		N=100		N=500	
scenario	τ	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
1	0	-0.004	0.0027	0.000	0.0012	-0.005	0.0027	-0.001	0.0012
2	0.41	0.004	0.0023	0.004	0.0010	-0.003	0.0022	0.003	0.0009
3	0.63	0.002	0.0014	0.006	0.0006	-0.005	0.0014	-0.001	0.0006
4	0.49	0.015	0.0027	0.013	0.0013	-0.023	0.0019	-0.012	0.0009
5	0.54	-0.012	0.0017	-0.005	0.0007	-0.016	0.0018	-0.006	0.0008
6	0.26	0.002	0.0023	0.002	0.0010	0.003	0.0028	0.001	0.0013

Table 1: Simulation study: Mean difference with standard error (s.e.) between Kendall's tau calculated using our method with a 20×20 grid and the true Kendall's tau from the distribution from which data was simulated.

		Censoring 70% interval 10% left 20% right				Censoring 50% interval 10% left 40% right			
		N=100		N=500		N=100		N=500	
scenario	τ	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
1	0	0.001	0.0018	0.001	0.0008	-0.000	0.0017	0.000	0.0007
2	0.41	0.004	0.0017	0.001	0.0007	-0.003	0.0014	0.001	0.0006
3	0.63	0.001	0.0011	0.005	0.0005	-0.006	0.0010	-0.002	0.0004
4	0.49	0.014	0.0022	0.012	0.0011	-0.025	0.0012	-0.013	0.0006
5	0.54	-0.016	0.0010	-0.008	0.0004	-0.020	0.0012	-0.009	0.0005
6	0.26	-0.002	0.0011	-0.001	0.0005	0.001	0.0019	0.006	0.0010

Table 2: Simulation study: Mean difference with standard error (s.e.) between Kendall's tau calculated from the censored and the uncensored observations for the simulation study using a 20×20 grid.

Tooth number	Tooth name	Boys Median (years)	Boys % censoring left	Boys % censoring interval	Boys % censoring right	Girls Median (years)	Girls % censoring left	Girls % censoring interval	Girls % censoring right
11	Central incisor	7.08	49	45	6	6.85	62	34	4
12	Lateral incisor	8.25	9	77	14	7.84	21	68	11
13	Canine	11.53	0	39	61	10.91	0	56	44
14	First premolar	10.73	1	56	43	10.31	0	68	32
15	Second premolar	11.62	1	37	62	11.26	0	47	53
16	First molar	6.31	83	15	2	6.14	89	10	1
17	Second molar	12.27	0	19	81	11.95	0	29	71

Table 3: Signal-Tandmobiel®study: Median emergence times and censoring distribution for the teeth of the right side of the upper jaw for boys and girls.

5 Application to Signal Tandmobiel®Study

The emergence age of a tooth is the chronological age of a child at which that tooth appears in the mouth. Not only the timing, but also the association pattern of (permanent) tooth emergence is of interest to dentists.
The Signal-Tandmobiel®study is a prospective longitudinal survey, which collected dental and oral health behaviour data from a representative sample (N=4468) of Flemish children born in 1989. An elaborate description of the Signal-Tandmobiel®project can be found in Vanobbergen et al.[18]. The children were examined annually on pre-scheduled visits (from the age of 7 to the age of 12) by 16 trained dentist-examiners in a mobile dental clinic on the school premises. Tooth emergence was recorded at each examination by direct inspection. Each permanent tooth was scored according to its clinical eruption stage (adapted from Carvalho et al.[6]). However, for the present analysis, the status of tooth eruption was dichotomized: not emerged versus emerged. As the children were examined annually, the emergence times are interval-censored. Since a tooth can emerge before the first or after the last visit also left and right censored emergence times are encountered. Based on data obtained from the Signal-Tandmobiel®study emergence times of 28 permanent teeth were determined for Flemish children from 7 to 12 years of age[14]. In Europe, the teeth are numbered with a two digit number as follows: the first digit represents the quadrant numbered from 1 to 4 (the upper right quadrant is "1", upper left "2", lower left " 3" and lower right "4"), the second digit refers to the place within the quadrant starting from the midline towards the back of the mouth. The last molar (tooth 18, a wisdom tooth) emerges (if it emerges) at the age of 17 years or later. Since its emergence time could not be recorded in our study we discarded that tooth here. Table 3 displays the median emergence times and the censoring distribution for teeth 11 to 17 for for the 2315 boys and 2153 girls of the Signal-Tandmobiel®study, separately. The median emergence times were estimated by fitting a log-logistic model to the data.

As an illustration we measured the association between the emergence times by means of Kendall's tau for each pair of the first quadrant. A 20×20 grid and a third order difference penalty was applied. The results are presented in Table 4. The highest association for both boys and girls was observed between the two incisors (teeth 11 and 12) and the two premolars (teeth 14 and 15). The lowest association was 0.28, between the second premolar and first molar for boys. From Figure 1 the following trend can be observed for boys: the closer the median emergence times, the higher the correlation. A similar

	11	12	13	14	15	16	17
11	1	0.53	0.45	0.38	0.42	0.43	0.42
		(0.50-0.56)	(0.40-0.50)	(0.34-0.41)	(0.38-0.46)	(0.36-0.50)	(0.35-0.48)
12	0.52	1	0.46	0.32	0.35	0.33	0.35
	(0.48-0.55)		(0.43-0.50)	(0.29-0.35)	(0.31-0.39)	(0.25-0.41)	(0.30-0.41)
13	0.43	0.46	1	0.49	0.49	0.39	0.35
	(0.39-0.47)	(0.43-0.49)		(0.45-0.52)	(0.41-0.57)	(0.31-0.47)	(0.27-0.44)
14	0.36	0.35	0.48	1	0.57	0.33	0.38
	(0.32-0.40)	(0.32-0.39)	(0.45-0.51)		(0.53-0.60)	(0.24-0.41)	(0.29-0.47)
15	0.35	0.34	0.42	0.56	1	0.28	0.35
	(0.31-0.40)	(0.31-0.38)	(0.38-0.47)	(0.52-0.60)		(0.23-0.33)	(0.26-0.43)
16	0.40	0.36	0.42	0.33	0.34	1	0.35
	(0.24-0.57)	(0.25-0.48)	(0.32-0.52)	(0.15-0.51)	(0.25-0.42)		(0.19-0.50)
17	0.36	0.32	0.39	0.38	0.44	0.48	1
	(0.30-0.41)	(0.27-0.38)	(0.32-0.45)	(0.33-0.43)	(0.38-0.50)	(0.39-0.57)	

Table 4: Signal-Tandmobiel®study: Kendall's tau with a 95% Confidence Interval between brackets for the teeth of the right side of the upper jaw. Results for boys and girls are presented in the upper and lower part, respectively.

pattern is found for the girls. This relates to the two emergence phases that are observed in Table 3. Namely, there is an early emergence phase for the first molar and the two incisors around 7 years and a later emergence phase for the canine, the two pre-molars and the second molar around 11 and 12 year. Although the emergence times of girls are significantly earlier than those of boys[14], no significant difference in association could be shown between boys and girls using a two-sample Z-test. The width of the confidence interval is apparently related with the proportion of left, right or interval censored data. Namely, the larger the proportion of left or right censored data, the wider the confidence interval is. This can be explained by the fact that an interval censored observation contains more information about the event time than a left or right censored observation.

Parner et al.[17] fitted a bivariate normal distribution to tooth emergence data of Danish children born in 1978. More than 12000 children were analyzed for both boys and girls. The children were examined annually from 3 to 18 years old. They reported Pearson correlations for all pairs of teeth. For a bivariate normal distribution there exists a relation between Kendall's tau and Pearson's correlation (ρ), namely $\tau = 2sin^{-1}(\rho)/\pi$. When transforming the Pearson correlations reported by Parner et al.[17] to Kendall's tau's, we found for most teeth similar results. Though in our study the correlations were somewhat lower. However, with the exception for 4 and 6 associations for girls and boys respectively, the estimates of Parner et al.[17] felt always within our 95% confidence intervals. Several reasons can explain this discrepancy: the use of another population or a possibly bad fitting bivariate normal distribution. As reported by Leroy et al.[14], emergence standards should be derived from the population in which they are to be applied, as factors related to emergence may vary considerably.

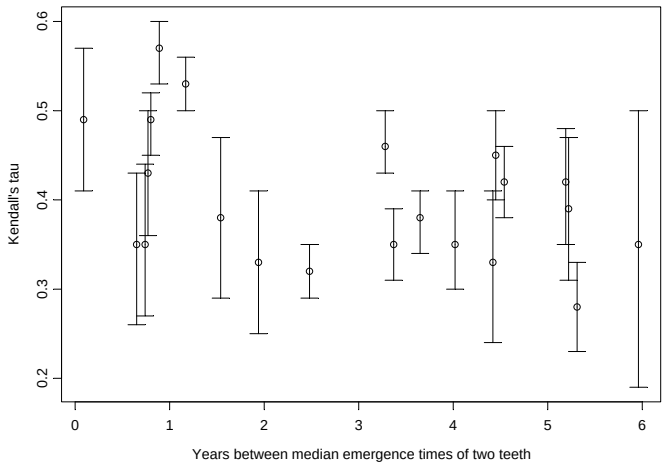


Figure 1: Signal-Tandmobiel®study: Kendall's tau with a 95% Confidence Interval versus the time between median emergence times of two teeth of the right side of the upper jaw for boys.

6 Discussion

It is important to note that our smoothing method is not a classical mixture problem. Indeed, only the mixing weights are estimated because the means and variances of the bivariate standard normal densities are fixed.

The penalty was defined in the parameters a_{ij}. On first sight, it seems more natural to define the penalty in the parameters c_{ij}. However, the computation in c_{ij} implied a significantly higher computation time with more numerical instability. The same was true when the penalty was expressed in terms of c_{ij} but the computations were done in a_{ij}.

If $\lambda_1 \to \infty$ and $\lambda_2 \to \infty$, the penalty becomes more important. If the order of the penalty is k, then $k^2 - 1$ well chosen conditions specify the limit distribution of the parameters $\mathbf{c}$ uniquely as the grid grows finer and wider. For $k = 1$, the penalty implies that all a_{ij} are equal to each other and that the limit distribution will resemble an uniform distribution. For $k = 2$, the penalty implies that the a_{ij}'s lie on a straight line for a specific i or j. Three well chosen conditions will fix the limit distribution. For $k = 3$, the penalty implies that the a_{ij}'s lie on a quadratic curve for a specific i or j. For 8 well chosen conditions the limit distribution will be member of the Bhattacharryya's distribution, the family of all bivariate densities with normal conditionals [2]. This family includes (among others) the bivariate normal distributions. For our simulations and the application a third order difference operator was used. In a limited simulation study, a second order difference showed also to provide adequate results.

In all our settings, a 20×20 grid provided good results. In practice, one can fit several increasing grid sizes to the data. If the results remain similar, this would indicate that a good fit is obtained. Given the grid, calculating Kendall's tau using our macro is done in a fairly automated way. On a Pentium IV 2GHz, calculation of Kendall's tau for a data set of size 100 and 500 in our simulations took on average 2 and 3 minutes respectively.

The method of Betensky of Finkelstein[3] starts with modelling the bivariate survivor function, this may be done in a parametric or non-parametric way. The parametric approach has the obvious drawback that choosing the correct distribution is hard especially with interval censored observations. For the non-parametric fit, there are two drawbacks. First,

the non-parametric maximum likelihood estimate (NPMLE) is not necessarily unique for interval censored data. Betensky of Finkelstein[3] do not describe how this affects their estimator. Secondly, although some recent progress in the computation of the NPMLE (e.g. [5], [15]) has been made, the estimation of the NPMLE is still quite computationally intensive for large data sets. This implies for the analysis of emergence times of the Signal-Tandmobiel®study that the calculation of the NPMLE for a pair of teeth is impossible with the current computing power due to an excessive large number of regions of possible support. The procedure of Betensky of Finkelstein[3] can therefore even not be performed with the NPMLE as starting point. Also we are not aware of a program that is currently available to fit the method of Betensky of Finkelstein[3].
Further, for right censored data, Wang and Wells[19] reported that the estimator of Oakes[16] is not consistent when the true value of τ is not equal to zero. The bias even increases as the degree of dependence increases. As the estimator of Betensky of Finkelstein[3] is based on Oakes's approach, it is likely that their estimator is also biased when the true value of τ is not equal to zero. In their simulations, Betensky of Finkelstein[3] only examine a situation where the true τ equals 0.224. For this setting, the mean bias was limited to 0.01. Situations with a true higher association were not examined. In our simulations, we obtained good results for all examined true τ's, i.e. up to 0.65.
Often the problem of interval censored data is overcome by approximating the event time by the midpoint of the interval. When applying this technique to the Signal-Tandmobiel®data quite large differences (up to 0.27) in the estimate of Kendall's tau were observed. Similar results were found by Parner et al.[17] who reported a bias in Pearson's correlation from 0.09 to 0.57. Therefore when trying to estimate an association measure on bivariate interval censored data, an adequate technique should be used.
In conclusion, we provide a relative easy method for estimating a measure of association for bivariate interval censored data. It performs well for both examined censoring schemes (up to 40% right censoring) and a fine and wide enough grid must be taken. A grid of size 20×20 was sufficient for all our simulations.
Finally, one can also derive an estimate for Spearman's correlation using our technique, namely

$$\widehat{\rho} \;=\; 12 \cdot \sum_i \sum_j \sum_k \sum_l \sum_p \sum_q \widehat{c}_{ij}\widehat{c}_{kl}\widehat{c}_{pq}\Phi\left(\frac{\mu_{1,i}-\mu_{1,p}}{\sqrt{2}\sigma_1}\right)\Phi\left(\frac{\mu_{2,j}-\mu_{2,q}}{\sqrt{2}\sigma_2}\right) - 3.$$

Details are given in the Appendix.

Acknowledgements

Both authors acknowledge support from the Interuniversity Attraction Poles Program P5/24 - Belgian State - Federal Office for Scientific, Technical and Cultural Affairs. The Signal-Tandmobiel®project comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Working Group Oral Health Promotion and Prevention, Flemish Dental Association; Dental School, University Ghent), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, University Leuven), K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

Appendix

Calculation of Kendall's Tau in terms of the coefficients c

The population measure of Kendall's tau for a cumulative distribution function F is defined as $\tau = 4 \cdot \int\int F(x,y)\mathrm{d}F(x,y) - 1$. Denote by $\Phi_2(\boldsymbol{\mu_{ij}}, \Sigma)$ the cumulative bivariate normal distribution with mean $\boldsymbol{\mu_{ij}} = (\mu_{1,i}, \mu_{2,j})$ and variance-covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ and let $\phi_2(\boldsymbol{\mu_{ij}}, \Sigma)$ denote the corresponding density. When we replace F by the cumulative distribution of our smooth estimate in the expression of τ, so $\widehat{F} = \sum_i \sum_j \widehat{c}_{ij} \Phi_2(\boldsymbol{\mu_{ij}}, \Sigma)$, then we obtain

$$\begin{aligned} \widehat{\tau} &= 4 \cdot \int\int \sum_i \sum_j \widehat{c}_{ij} \Phi_2(\boldsymbol{\mu_{ij}}, \Sigma) \cdot \sum_k \sum_l \widehat{c}_{kl} \phi_2(\boldsymbol{\mu_{kl}}, \Sigma)\mathrm{d}x\mathrm{d}y - 1 \\ &= 4 \cdot \sum_i \sum_j \sum_k \sum_l \widehat{c}_{ij}\widehat{c}_{kl} \int\int \Phi_2(\boldsymbol{\mu_{ij}}, \Sigma) \cdot \phi_2(\boldsymbol{\mu_{kl}}, \Sigma)\mathrm{d}x\mathrm{d}y - 1 \end{aligned}$$

These integrals can be simplified using the fact that in the variance-covariance matrix Σ zero correlation is assumed. Thus

$$\begin{aligned} &\int\int \Phi_2(\boldsymbol{\mu_{ij}}, \Sigma) \cdot \phi_2(\boldsymbol{\mu_{kl}}, \Sigma)\mathrm{d}x\mathrm{d}y = \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\cdot\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{z-\mu_{1,i}}{\sigma_1}\right)^2\right]\mathrm{d}z \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_{1,k}}{\sigma_1}\right)^2\right]\mathrm{d}x \times \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\cdot\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{z-\mu_{2,j}}{\sigma_2}\right)^2\right]\mathrm{d}z \cdot \frac{1}{\sqrt{2\pi}\cdot\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_{2,l}}{\sigma_2}\right)^2\right]\mathrm{d}y \end{aligned}$$

Using transformations these integrals can be converted to integrals of bivariate standard normal densities with zero correlation, i.e.

$$\begin{aligned} &\int_{-\infty}^{\infty}\int_{-\infty}^{\frac{t\sigma_1+\mu_{1,i}-\mu_{1,k}}{\sigma_1}} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}z^2\right]\mathrm{d}z \cdot \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}t^2\right]\mathrm{d}t \times \\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\frac{t\sigma_2+\mu_{2,j}-\mu_{2,l}}{\sigma_2}} \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}z^2\right]\mathrm{d}z \cdot \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}t^2\right]\mathrm{d}t \end{aligned}$$

By taking advantage of the symmetry of the bivariate standard normal distribution, we can rewrite this product of bivariate integrals as a product of univariate cumulative standard normal distributions.
Hence, finally

$$\widehat{\tau} = 4 \cdot \sum_i \sum_j \sum_k \sum_l \widehat{c}_{ij}\widehat{c}_{kl} \Phi\left(\frac{\mu_{1,i}-\mu_{1,k}}{\sqrt{2}\sigma_1}\right)\Phi\left(\frac{\mu_{2,j}-\mu_{2,l}}{\sqrt{2}\sigma_2}\right) - 1.$$

Comparing Kendall's tau between two independent groups

Assume we have two independent groups. Let τ_1 and τ_2 denote the true Kendall's tau's in both groups. For a large enough number of observations in both groups (n_1 and n_2),

we have that $\widehat{\tau}_i \sim \mathcal{N}(\tau_i, \sigma_i^2/n_i)$ for i=1,2. Therefore we can test $H_0 : \tau_1 = \tau_2$ by a simple two-sample Z-test. Namely $Z = (\widehat{\tau}_1 - \widehat{\tau}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Calculation of Spearman's correlation in terms of the coefficients c

The population measure of Spearman's correlation is defined as
$\rho = 12 \cdot \int\int F_1(x)F_2(y)\mathrm{d}F(x,y) - 3$ where $F(x,y)$ is the bivariate cumulative distribution function and $F_1(x)$ and $F_2(y)$ are the corresponding univariate marginal distributions. Using the same arguments as for the derivation of Kendall's tau, one can derive that ρ can be estimated by

$$\widehat{\rho} \;=\; 12 \cdot \sum_i \sum_j \sum_k \sum_l \sum_p \sum_q \widehat{c}_{ij}\widehat{c}_{kl}\widehat{c}_{pq}\Phi(\frac{\mu_{1,i} - \mu_{1,p}}{\sqrt{2}\sigma_1})\Phi(\frac{\mu_{2,j} - \mu_{2,q}}{\sqrt{2}\sigma_2}) - 3.$$

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

[2] Arnold, B., Castillo, E. and Sarabia, J. (2001). Conditionally specified distributions: an introduction. *Statistical Science* **16**, 249–274.

[3] Betensky, R. and Finkelstein, D. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine* **18**, 3101–3109.

[4] Bogaerts, K. and Lesaffre, E. (2003). A smooth estimate of the bivariate survival density in the presence of left, right and interval censored data. In *Proceedings of the Joint Statistical Meetings, Biometrics Section [CD-ROM]*, pages 633–639, Alexandria, VA: American Statistical Association.

[5] Bogaerts, K. and Lesaffre, E. (2004). A new fast algorithm to find the regions of possible support for bivariate interval censored data. *Journal of Computational and Graphical Statistics* **13**, 330–340.

[6] Carvalho, J., Ekstrand, K. and Thylstrup, A. (1989). Dental plaque and caries on occlusal surfaces of first permanent molars in relation to stage of eruption. *J Dent Res* **68**, 773–779.

[7] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–152.

[8] Dabrowska, D. (1986). Rank tests for independence for bivariate censored data. *The Annals of Statistics* **14**, 250–264.

[9] Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties and discussion. *Statistical Science* **11**, 89–121.

[10] Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics* **60**, 945–953.

[11] Gray, R. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942–951.

[12] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer Verlag New York.

[13] Komárek, A., Lesaffre, E. and Hilton, J. (2003). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. In Verbeke, G., Molenberghs, G., Aerts, M. and Fieuws, S., editors, *Proceedings of the 18th International Workshop on Statistical Modelling*, pages 233–238, Leuven, Belgium.

[14] Leroy, R., Bogaerts, K., Lesaffre, E. and Declerck, D. (2003). The emergence of permanent teeth in Flemish children. *Community Dent Oral Epidemiol* **31**, 30–39.

[15] Maathuis, M. (2004). Reduction algorithm for the npmle for the distribution function of bivariate interval censored data. *Accepted in JCGS* .

[16] Oakes, D. (1982). A concordance test for independence in the presence of censoring. *Biometrics* **38**, 451–455.

[17] Parner, E., Heidmann, J., Kjaer, I., Vaeth, M. and Poulsen, S. (2002). Biological interpretation of the correlation of emergence times of permanent teeth. *Journal of Dental Research* **81**, 451–454.

[18] Vanobbergen, J., Martens, L., Lesaffre, E. and Declerck, D. (2000). The Signal-Tandmobiel® project, a longitudinal intervention heath promotion study in Flanders (Belgium): baseline and first year results. *Eur J Paediatr Dent* **2**, 87–96.

[19] Wang, W. and Wells, M. (2000). Estimation of Kendall's tau under censoring. *Statistica Sinica* **10**, 1199–1215.

[20] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.