2011

Approximating Confidence Intervals About Discrete Time Survival/Cumulative Incidence Estimates Using the Delta Method

Alexis Dinno
Portland State University, alexis.dinno@pdx.edu

Jong-Sung Kim
Portland State University

Citation Details

Dinno, Alexis and Kim, Jong-Sung, "Approximating Confidence Intervals About Discrete Time Survival/Cumulative Incidence Estimates Using the Delta Method" (2011). *Community Health Faculty Publications and Presentations*. 37.
https://pdxscholar.library.pdx.edu/commhealth_fac/37

Approximating Confidence Intervals About Discrete Time Survival/Cumulative Incidence Estimates Using the Delta Method

For some of our recent work, see:
http://stats.stackexchange.com/questions/114246/how-to-analytically-estimate-cis-on-the-survival-function-s-t-in-a-logit-h

Alexis Dinno[1] and Jong-Sung Kim[2]

[1]Portland State University, School of Community Health, Portland, OR, USA. [2]Portland State University, Department of Mathematics and Statistics, Portland, OR, USA.

1 Introducing discrete time event history models

Event history analysis answers whether and when an event will happen in a population at risk, and goes by different names such as 'survival analysis' or 'failure time models.' A good reference for this material is found in Singer, J. D. and Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.*

The basic discrete time event history model

$$\begin{aligned} h_t &= \frac{1}{1+e^{-(\alpha_1 d_1+\cdots+\alpha_T d_T+\beta_1 X_1+\cdots+\beta_p X_p)}} \\ \ln\left(\frac{h_t}{1-h_t}\right) &= \alpha_1 d_1+\cdots+\alpha_T d_T+\beta_1 X_1+\cdots+\beta_p X_p \\ \text{logit}\left(h_t\right) &= \alpha_1 d_1+\cdots+\alpha_T d_T+\beta_1 X_1+\cdots+\beta_p X_p \end{aligned}$$

Time in discrete time event history analysis is represented discretely. In the above there are T discrete time periods represented, and p predictor variables. These model conditional log-odds of event occurrence as a linear function of time t (specified by α_t) and the predictors (the Xs) multiplied by their slope parameters (the βs).

Data and representation in discrete time event history models

Discrete time event history models employ *person-period* data formats wherein observations in each discrete period are nested within individuals. Event occurrence is represented as a nominal variable coded as: no event = 1; and event = 1. *Post-event observations are right censored*, and right censoring is modeled with the *absence* of observations corresponding to the censored periods. Excerpted are data for T=12 with two predictors, sex and support:

ID	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	female	support	Y
01	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
02	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
02	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
02	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
03	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
03	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
03	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
04	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
04	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1

(Adapted from Singer, J. and Willett, J. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, 18(2):155–195.)

Discrete time conditional hazard and survival functions

The discrete time hazard function h_t describes the risk of event occurrence at time t conditional on the predictors for a randomly selected individual who has not experienced an event before time t, the estimated discrete time conditional hazard is thus:

$$\hat{h}_t = \frac{1}{1+e^{-(\hat{\alpha}_t+\beta_1 X_1+\cdots+\beta_p X_p)}}$$

The survival function S_t describes the proportion of the study population for whom an event has not occurred by time t conditional on the predictors since the study's beginning of time, and is thus estimated the product of the complements of the discrete time conditional hazards up to time t as:

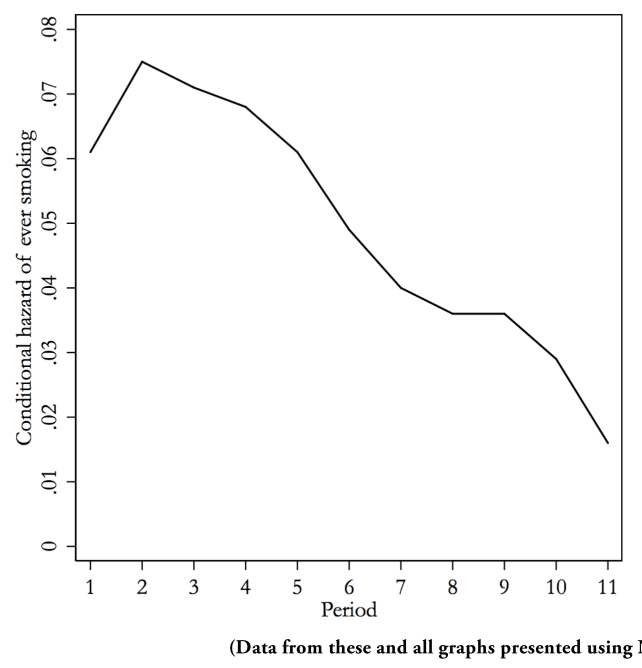
$$\hat{S}_t = \prod_{i=1}^{t}\left(1-\hat{h}_t\right)$$

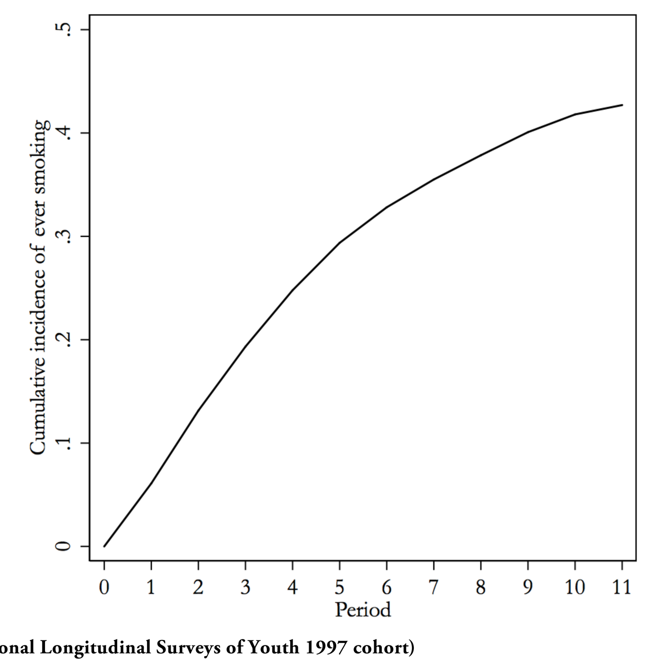
S_t can be reframed in terms of cumulative incidence, which is the proportion of the study population for whom an event has occurred by time t conditional on the predictors since the study's beginning of time. Cumulative incidence is simply $1 - S_t$, and the variance of S_t and $1 - S_t$ are identical.

The term 'cumulative incidence' is sometimes used with a different meaning in competing risks event history models. The definition used here is consistent with how epidemiologists conceive of risk: the proportion experiencing an event in a specific period of time.

Baseline (unconditional) models of ever smoking a single cigarette

Results from event history models are communicated graphically (e.g. using 'hazard curves' and 'survival curves').





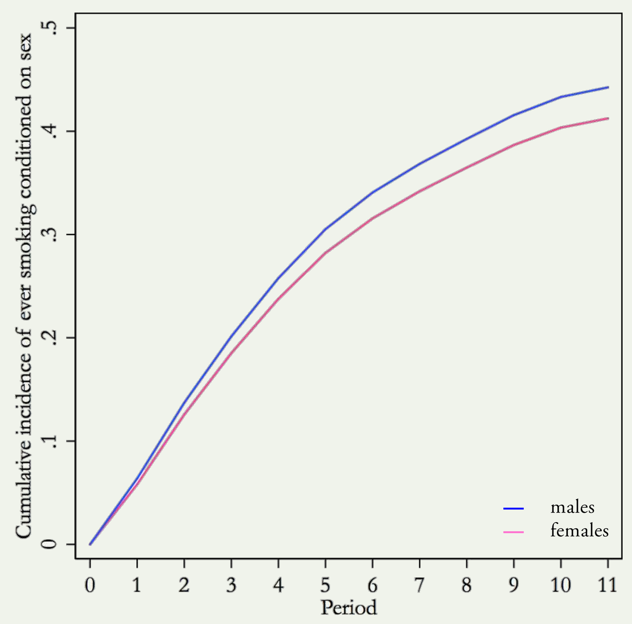
(Data from these and all graphs presented using National Longitudinal Surveys of Youth 1997 cohort)

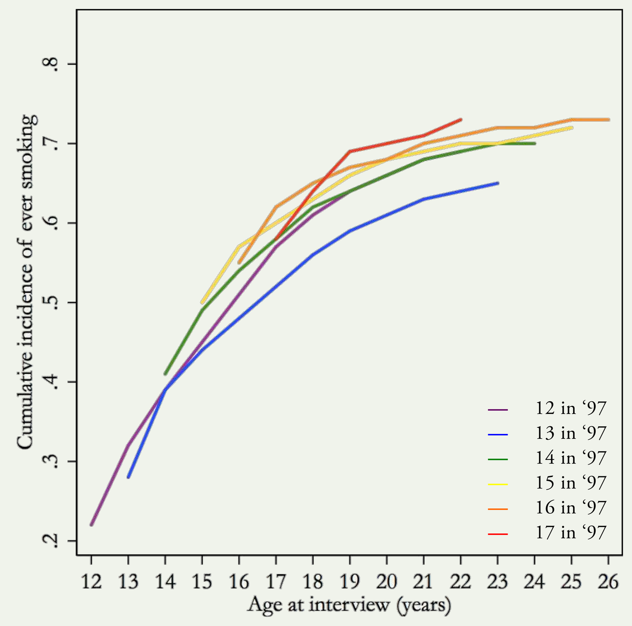
Portland State UNIVERSITY

2 The problem: no confidence intervals

To effectively interpret and communicate the results of discrete time event history models we would like to be able to infer whether the differences between different groups are meaningful. Consider the below graphs. Does males' smoking initiation differ significantly from females'? Does the 13 in 1997 cohort differ significantly from others?

Conditioning smoking initiation on sex, and (separately) on cohort





Confidence intervals about survival function and cumulative incidence function values, and the corresponding curves, would facilitate rapid 'eyeball' hypothesis testing over the duration of study time, between arbitrary numbers of groups. Two caveats: 1) the usual concerns about multiple comparisons still apply; and 2) employing confidence intervals for visual tests of significance require modification to correspond to t-tests of mean difference—typically in the form of a scalar adjustment to the value of $z_{\alpha/2}$.

A naïve approach to estimating $\sigma_{\hat{h}_t}$ and $\sigma_{\hat{S}_t}$

Typically discrete time event history models are estimated using maximum likelihood techniques, such as those available in the logistic regression packages for major statistical software packages. These packages estimate the standard errors of the parameter estimates (i.e. the $\sigma_{\hat{\beta}}$ terms), and β confidence intervals can be normally approximated with the familiar:

$$\text{CI}_{\hat{\beta}} = \hat{\beta} \pm z_{\frac{\alpha}{2}}\sigma_{\hat{\beta}}$$

One approach would perform the same transform on the $\sigma_{\hat{\beta}}$ terms as is used with the $\hat{\beta}$ terms to estimate $\hat{h}_t$, $\hat{S}_t$ or $1-\hat{S}_t$ in order to estimate confidence intervals for $\sigma_{\hat{h}_t}$, $\sigma_{\hat{S}_t}$ or (remember $\sigma_{\hat{S}_t} = \sigma_{1-\hat{S}_t}$). However this approach produces inaccurate results because $\sigma^2_{g(\theta)}=g(\sigma^2_\theta)$ only when $g(\bullet)$ is ~~a linear~~ an identity function—which is the case for neither logit nor product functions. But normally approximation of confidence intervals for $\hat{h}_t$, $\hat{S}_t$ or $1-\hat{S}_t$ with any accuracy requires $\sigma_{\hat{h}_t}$ or $\sigma_{\hat{S}_t}$ for which there are no commonly accepted estimators.

$$\text{CI}_{\hat{h}_t} = \hat{h}_t \pm z_{\frac{\alpha}{2}}\sigma_{\hat{h}_t} \qquad \text{CI}_{\hat{S}_t} = \hat{S}_t \pm z_{\frac{\alpha}{2}}\sigma_{\hat{S}_t} \qquad \text{CI}_{1-\hat{S}_t} = 1-\hat{S}_t \pm z_{\frac{\alpha}{2}}\sigma_{\hat{S}_t}$$

Approximate confidence intervals using the univariate delta method

Fortunately, $\sigma_{\hat{h}_t}$ and $\sigma_{\hat{S}_t}$ can be approximated using the delta method[1] with a first-order Taylor series expansion. Here the univariate case (e.g. for unconditional models):

$$\begin{aligned} g\left(X\right) &\approx g\left(\mu\right)+g'\left(\mu\right)\left(X-\mu\right) \\ \sigma^2_{g(X)} &= \left[g'\left(\mu\right)\right]^2\sigma^2_X \end{aligned}$$

If $g(X)=e^X/(1+e^X)$ (the anti-logit, or logistic function), then we obtain $\sigma^2_{\hat{h}_t}$ as:

$$\begin{aligned} \sigma^2_{\hat{h}_t} &= \sigma^2_{\hat{\alpha}_t}\left(\frac{d}{d\alpha_t}\frac{e^{\alpha_t}}{1+e^{\alpha_t}}\bigg|_{\alpha_t=\hat{\alpha}_t}\right)^2 \\ &= \sigma^2_{\hat{\alpha}_t}\left(\frac{e^{\hat{\alpha}_t}\left(1+e^{\hat{\alpha}_t}\right)-e^{\hat{\alpha}_t}e^{\hat{\alpha}_t}}{\left(1+e^{\hat{\alpha}_t}\right)^2}\right) \\ &= \sigma^2_{\hat{\alpha}_t}\left(\frac{e^{\hat{\alpha}_t}}{\left(1+e^{\hat{\alpha}_t}\right)^2}\right)^2 \end{aligned}$$

$\ln(\hat{S}_t)$ is more tractable than $\hat{S}_t$, and if the function $g(X)=1-X$, we first estimate $\sigma^2_{\ln(\hat{S}_t)}$:

$$\begin{aligned} \sigma^2_{\ln(\hat{S}_t)} &= \sigma^2_{\sum_{i=1}^{t}\ln\left(1-\hat{h}_i\right)} \\ &= \sum_{i=1}^{t}\sigma^2_{\ln\left(1-\hat{h}_i\right)} \quad \text{(by independence)} \\ &= \sum_{i=1}^{t}\sigma^2_{\hat{h}_i}\left[\frac{d}{dh_i}\ln\left(1-h_i\right)\bigg|_{h_i=\hat{h}_i}\right]^2 \\ &= \sum_{i=1}^{t}\sigma^2_{\hat{h}_i}\left(\frac{1}{\hat{h}_i-1}\right)^2 \\ &= \sum_{i=1}^{t}\sigma^2_{\hat{\alpha}_i}\left(\frac{e^{\hat{\alpha}_i}}{1+e^{\hat{\alpha}_i}}\right)^2 \end{aligned}$$

Finally, if $g(X)=e^X$, substitute for $\hat{h}_t$ and $\hat{S}_t$ to obtain $\sigma^2_{\hat{S}_t}$ using only $\sigma_{\hat{\alpha}}$ and $\hat{\alpha}$.

$$\begin{aligned} \sigma^2_{\hat{S}_t} &= \sigma^2_{\ln(\hat{S}_t)}\left(\frac{d}{d\ln\left(S_t\right)}e^{\ln(S_t)}\bigg|_{S_t=\hat{S}_t}\right)^2 \\ &= \sigma^2_{\ln(\hat{S}_t)}\left(e^{\ln\left(\hat{S}_t\right)}\right)^2 \\ &= \sigma^2_{\ln(\hat{S}_t)}\hat{S}_t^2 \\ &= \left[\sum_{i=1}^{t}\sigma^2_{\hat{\alpha}_i}\left(\frac{e^{\hat{\alpha}_i}}{1+e^{\hat{\alpha}_i}}\right)^2\right]\left[\prod_{i=1}^{t}\left(\frac{1}{1+e^{\hat{\alpha}_i}}\right)\right]^2 \end{aligned}$$

[1] See Oehlert G. W. (1992) A note on the delta method. *American Statistician* 46(1):27–29.

Approximation for conditional models

The delta method be extended to a multivariate case (1st-order Taylor series expansion):

$$\begin{aligned} g(\mathbf{Z}) &\approx g(\mu)+\mathbf{V}(\mathbf{Z}-\mu) \\ \sigma^2_{g(\mathbf{Z})} &= \mathbf{V}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \end{aligned}$$

where: $\mathbf{Z}$ is a 1 by p + 1 row vector of d_t, X_{1t},…, X_{pt}, $\mathbf{V}$ is a is a 1 by p + 1 vector of partial derivatives of $\mathbf{Z}$ evaluated at the p + 1 by 1 mean vector μ, and $\mathbf{\Sigma}$ is the p + 1 by p + 1 variance-covariance matrix of $\mathbf{Z}$.

$$\begin{aligned} \hat{h}_t &= \frac{e^{\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t}}{1+e^{\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t}} \\ \mathbf{V}_t &= \left[\begin{array}{cccc} \frac{\partial h_t}{\partial\alpha_t} & \frac{\partial h_t}{\partial\beta_1} & \ldots & \frac{\partial h_t}{\partial\beta_p} \end{array}\right]\Bigg|_{\alpha_t=\hat{\alpha}_t,\beta_1=\hat{\beta}_1,\ldots,\beta_p=\hat{\beta}_p} \\ &= \frac{e^{\left(\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t\right)}}{\left(1+e^{\left(\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t\right)}\right)^2}\left[\begin{array}{cccc} 1 & X_{1t} & \ldots & X_{pt} \end{array}\right] \\ \sigma^2_{\hat{h}_t} &= \mathbf{V}_t\mathbf{\Sigma}_t\mathbf{V}_t^{\mathrm{T}} \\ &= \left[\frac{e^{\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t}}{\left(1+e^{\hat{\alpha}_t+\hat{\mathbf{B}}\mathbf{X}_t}\right)^2}\right]^2\mathbf{Z}_t\mathbf{\Sigma}_t\mathbf{Z}_t^{\mathrm{T}} \end{aligned}$$

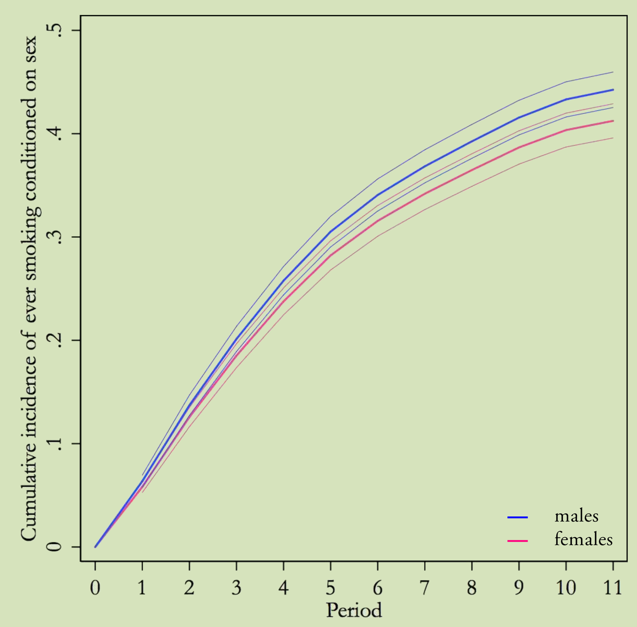
Again assuming independence for $\sigma^2_{\ln(\hat{S}_t)}$, the form of $\sigma^2_{\hat{S}_t}$ follows:

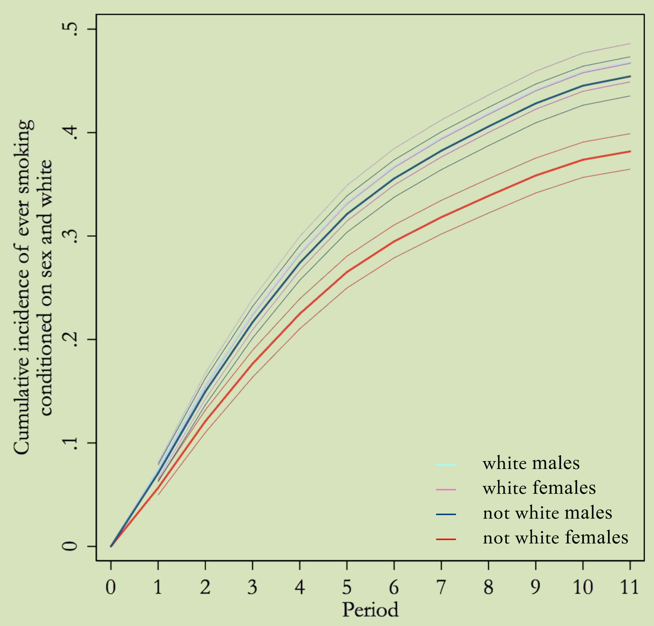
$$\begin{aligned} \sigma^2_{\hat{S}_t} &= \left[\sum_{i=1}^{t}\sigma^2_{\hat{h}_i}\left(\frac{1}{\hat{h}_i-1}\right)^2\right]\hat{S}_t^2 \\ &= \left[\sum_{i=1}^{t}\mathbf{V}_i\mathbf{\Sigma}_i\mathbf{V}_i^{\mathrm{T}}\left(\frac{1}{\hat{h}_i-1}\right)^2\right]\hat{S}_t^2 \\ &= \left[\sum_{i=1}^{t}\mathbf{Z}_i\mathbf{\Sigma}_i\mathbf{Z}_i^{\mathrm{T}}\left(\frac{e^{\hat{\alpha}_i+\hat{\mathbf{B}}\mathbf{X}_i}}{1+e^{\hat{\alpha}_i+\hat{\mathbf{B}}\mathbf{X}_i}}\right)^2\right]\left[\prod_{i=1}^{t}\left(\frac{1}{1+e^{\hat{\alpha}_i+\hat{\mathbf{B}}\mathbf{X}_i}}\right)\right]^2, \end{aligned}$$

where $\hat{\mathbf{B}}$ is a row vector of estimated parameters $\hat{\beta}_1$, …, $\hat{\beta}_p$ and $\mathbf{X}_i$ is a column vector of observed variables X_{1i}, …, X_{pi} indexed by i = 1, …, t.

3 Visual inference for discrete time survival curves

Adding 95% confidence intervals to the model adjusting for sex (below left) permits inference that cumulative smoking initiation for females is significantly lower than for males after about the fourth period.[3] Additional adjustment for white/not white plus an interaction term (below right) reveals that this difference is explained by not-white females' cumulative incidence of initiation, which is significantly lower than all other groups at all periods; the overlap of confidence bounds with each of the cumulative incidence curves of the remaining three groups implies that they do not differ significantly. (Our examples are pedagogical, and neglect the nuance of serious analysis.)





Inferences may also be made with conditional hazard curves using these methods. The quantities $\sigma_{\hat{h}_t}$ and $\sigma_{\hat{S}_t}$ may also be used in numerical inference using t-tests.

Planned methodological development

- Derivation and application for models using complementary log-log links under an assumption of proportional hazards, rather than the assumption of proportional odds of the logistic regression model presented here.
- Derivation and application for multilevel discrete-time event history models
- To be implemented in free software (in R , & in the **dthaz** package for Stata)

Future application to disparities in smoking initiation & progression

National Longitudinal Surveys of Youth 1997 data include detailed annual self reports of 30-day smoking behavior; include a wealth of socio-demographic information, including characteristics of the parenting and household environment; span the ages at which the vast majority of regular smokers initiate the and progress to their established smoking careers; include state and local geocodes permitting linkage to state and local tobacco control policies.

3 For a current description of visual hypothesis testing using confidence intervals, see Afshartous D, Preston R. Confidence intervals for dependent data: Equating non-overlap with statistical significance. *Computational Statistics & Data Analysis* 2010;54(10):2296–2305.