Confidence Intervals for Seasonal Relative Risk with Null Boundary Values, and Application to Suicide Seasonality

Constantine E. Frangakis and Ravi Varadhan

Department of Biostatistics, Johns Hopkins University, 615 N.Wolfe Street, Baltimore, MD 21205.*

May 21, 2002

Abstract

In evaluating relative risk of seasonality, the null value is a boundary value. In this case, tests for the null hypothesis exist, but standard methods for confidence interevals are not appropriate. We provide a method for contructing confidence intervals under the circular normal model. The proposed confidence intervals are valid: (i) for all values of the underlying seasonal risk if the model is correct; and (ii) for the null boundary value of the seasonal risk, regardless of model assumptions and sample size. We apply our method to seasonal suicide data from a recent report.

Keywords: confidence interval, boundary value, seasonality, suicide.

In Petridou et al. (2002), the seasonality of suicides was evaluated, showing evidence of association with months of higher sunshine. The peer review of that paper had pointed out an interesting issue: how to construct a confidence interval (CI) for the relative risk (RR) of suicide (maximum over minimum frequency), because, here, the null value of RR (unity) is also a boundary of the possible values of RR. Although approaches exist for testing some boundary null hypotheses[2], we are unaware of an application in this area that had computed CIs when the null value is on the boundary. Therefore, in the original article, we first tested the null, finding strong evidence for seasonality for all countries, and then constructed the CIs using the maximum asymptotic margin of error across countries, in order to minimize potential false positive findings. Since then, we have developed and applied a more principled method to address the problem, which we briefly outline below. The results from the new method strengthen the substantive findings of the original paper.

*Address for correspondence:cfrangak@jhsph.edu.

Methods

Data and Model

The data are the numbers of suicides in each of the twelve months for twenty countries, for a period covering 4 to 24 of the most recent years. To demonstrate ideas, we focus on one such country, and let n_i denote the number of suicides for $i = 1, \dots, 12$ months. We assume that the counts $(n_1, \dots, n_{12})$ are from a multinomial distribution with probabilities of suicide $(p_1, \dots, p_{12})$, where p_i are proportional to the periodic factor $e^{k \cos(2\pi(i-\beta)/12)}$; e^{2k} is the ratio of maximum over minimum frequency of suicide, and β is the month of maximum frequency of suicides, where k is non-negative and β is an integer between 1 and 12. This model is a discrete version of the circular normal distribution for continuous data [3].

The model above has a few nice features. It allows for the null hypothesis of no seasonality of suicides, when $k = 0$, whereby all the probabilities, p_i, are equal to $\frac{1}{12}$. Furthermore, the above model will be more powerful in detecting true seasonality in suicides than a less structured model, e.g. the saturated model that fits eleven different probabilities. The model also possesses a useful property that allows us to construct CIs for the risk parameter, k, when standard methods are not generally appropriate, as discussed next.

Confidence Intervals when the Null Risk is on the Boundary

Suppose (k_0, β_0) are the true values of the parameters, (k, β), that generate the observed suicide counts $(n_1, \cdots, n_{12})$. We can write the likelihood function for any parameter values (k, β) as

$$\prod_{i=1}^{12} \{p_i(k,\beta)\}^{n_i}, \quad \text{where } p_i(k,\beta) = \frac{e^{k \cos(2\pi(i-\beta)/12)}}{\sum_{j=1}^{12} e^{k \cos(2\pi(j-\beta)/12)}}. \tag{1}$$

We wish to construct CIs for k_0 based on the maximum likelihood estimates (MLE), $\hat{k}$, and $\hat{\beta}$, which are the values of k and β that maximize the likelihood function in (1). However, because the null value, $k = 0$, is a boundary value, the usual theory for CIs of the form $\hat{k} \pm 1.96 se(\hat{k})$ obtained using a normal approximation of $\hat{k}$ is not applicable here. Moreover, although approaches exist for testing some boundary null hypotheses [2], we are unaware of a previous application in this area that has offered CIs when the null value is on the boundary. Here we propose an approach for valid CIs with our model based on first principles.

For this approach, it is important to consider, for every pair (k, β), the quantile function $q(k, \beta)$ defined so that $\Pr(|\hat{k} - k| \leq q(k, \beta) \mid k, \beta)$ is 95%. In words, the quantile $q(k, \beta)$ is the number q such that, when counts of total size n are generated repeatedly from model (1) with true parameter values (k, β), the fraction of times that $|\hat{k} - k|$ is less that q is 95%. For a fixed pair (k, β), we can calculate $q(k, \beta)$ as follows: (i) simulate a large number of sets of $(n_1, \ldots, n_{12})$ from model (1), each set totalling the fixed size n; (ii) for each set, calculate its MLE $\hat{k}$ and its distance $|\hat{k} - k|$ of $\hat{k}$ from k, and (iii) set $q(k, \beta)$ equal to the smallest number below which fall 95% of the calculated distances $|\hat{k} - k|$ across the simulations.

The quantile function gives an exact reference for how far the estimated risk parameter is expected to be from the true value. In particular, if for any month β we define the set $S(\hat{k}, \beta)$ of values of the risk parameter k to be:

$$S(\hat{k}, \beta) = \{k : |k - \hat{k}| \leq q(k, \beta)\},$$

then, by result of inversion of tests, we have that the range of values k defined by the set $S(\hat{k}, \beta_0)$ is an exact 95% CI for the true value k_0, whether or not k_0 lies on the boundary.

Generally, finding the CI for k_0 using $S(\hat{k}, \beta_0)$ is difficult because the true peak month β_0 is unknown. However, model (1) has the property that, for fixed risk parameter k, the quantile function $q(k, \beta)$ is not a function of β. The intuition for this is that, given the risk parameter k, the likely values of its estimator $\hat{k}$ are invariant of the month at which the peak occurs (a proof is given in the appendix). It follows that $S(\hat{k}, \beta_0) = S(\hat{k}, \hat{\beta})$, so that, instead of the unknown month β_0, we can simply use the MLE, $\hat{\beta}$, to calculate the set $S(\hat{k}, \hat{\beta})$, which will then be the exact 95% CI for k_0.

Algorithm

We have developed an algorithm to compute $S(\hat{k},\hat{\beta})$, the exact 95% CI for k_0, that involves three distinct computational tasks. The first is to compute the MLEs, $\hat{k},\hat{\beta}$. The second is to obtain the two roots of the equation, $f(k)=0$, where $f(k)=|k-\hat{k}|-q(k,\hat{\beta})$; the larger root is the upper endpoint of the 95% CI; the lower endpoint is the smaller root, provided that the root is positive, and is zero otherwise. In this task, the roots of the equation are obtained iteratively using the secant method. For the larger root, k_u, of the equation $f(k_u)=0$, the secant method starts with two points, $k_u^{(1)}$ and $k_u^{(2)}$, and proceeds iteratively to obtain a sequence $k_u^{(i)}, i=3,4,\cdots$, that converges to k_u. Here we choose $k_u^{(1)}$ and $k_u^{(2)}$ to be around the normal approximations $\hat{k}+1.96se(\hat{k})$ and $\hat{k}+1.96se(\hat{k})+\Delta$, respectively, where Δ is a small number. The smaller root, k_l, is obtained similarly where we choose the two starting points as, $k_l^{(1)}=\hat{k}-1.96se(\hat{k})$ and $k_l^{(2)}=k_l^{(1)}+\Delta$. A third task of our algorithm is required inside the secant method: because the quantile function, q, is not available in closed form, we compute q by simulation from the model at the values required by the iterations of the secant method. Although the parameter space for β is the integers, for ease of computation here we use the continuous scale.

In some simulations, we found that the endpoints of the exact 95% CI were relatively close to their starting normal approximations. However, as also discussed earlier, validity of the normal approximation cannot be guaranteed for arbitrary data with this model. So, use of the exact 95% CIs here is generally preferable.

Results

We applied our method to the suicide counts used in Petridou et al. (2002) across the twelve months, for each country. The sample counts overall months within a country ranged between 3,770 (Greece) and 222,227 (Japan), as summarized over recent years. For details, see Petridou et al. (2002).

The CIs for the relative risk RR computed with the new method are presented in Table 1. As expected, the point estimates of RR remain essentially the same as in Petridou et al. (2002). However, the exact CIs for RR computed with the new method are both, quite shorter than the conservative ones originally reported, and theoretically valid for the boundary constraint, a fact that further strengthens the findings of Petridou et al. (2002). Because we find strong evidence for seasonality ($k > 0$) with this method too, we also report the normal approximation to the CIs for the date of peak incidence of suicide, β_0 (in degrees).

Discussion

The method presented defines the relative risk in seasonality through a discrete version of the circular normal distribution, and provides CIs that are valid no matter the underline value of that risk. Moreover, because the circular normal model is correct under the null hypothesis of all seasons having the same risk, our CIs are guaranteed to have the right coverage under the null hypothesis, regardless of model or sample size assumptions.

The method we discussed, illustrated here for 12 seasons (months) can be used for arbitrary number of seasons when the researcher expects a single peak in the circular frequency of the studied event. Extensions of the method for CIs for multiple peaks, and for accounting for other covariates simultaneously, may be possible by introducing appropriate terms in the likelihood.

The computer program implementing the new method discussed here for obtaining CIs for a seasonal risk with general number of seasons is available from the web site
http://biosun01.biostat.jhsph.edu/∼cfrangak/papers/suicide/

Acknowledgments

The authors thank the Editor and three reviewers for constructive comments, Drs. D. Trichopoulos, E. Petridou, and K. Rothman for fruitful discussions, and the NIH grant 1R01EY014314-01 for partial support.

Appendix

To show that the quantile function $q(k, \beta)$ is free of β, it suffices to show that the distribution of the MLE $\hat{k}$ is the same when the data are generated by (k_0, β_0) or by $(k_0, \beta_0 + 1)$, i.e. that for any fixed k^*,

$$\text{pr}(\hat{k} = k^* | k_0, \beta_0) = \text{pr}(\hat{k} = k^* | k_0, \beta_0 + 1). \tag{2}$$

The key to showing (2) is to note that, for any 12-count $\underline{\text{n}} = (n_1, ..., n_{12})$ and (k, β), the model (1) is circularly invariant, that is

$$\prod_{i=1}^{12} \{p_i(k, \beta)\}^{n_i} = \prod_{i=1}^{12} \{p_i(k, \beta + 1)\}^{n_i^{(+1)}}, \tag{3}$$

where $n_i^{(+1)}$ $(i = 1, ..., 12)$ are the counts of the translated configuration $\underline{\text{n}}^{(+1)}$ defined as $(n_{12}, n_1, ..., n_{11})$. *This means that, given a fixed true value of the risk parameter k, an observation of a specific ordered pattern for the counts $\underline{\text{n}}$ when the true peak date is β is equally likely to an observation of that pattern rotated by an amount of time when the true peak date is also rotated by the same amount of time.* In particular, denote by N_{k^*} to be all configurations $\underline{\text{n}}$ that give $\hat{k}(\underline{\text{n}}) = k^*$. Then, the RHS and LHS of (2), respectively, become $\text{pr}(\underline{\text{n}} \in N_{k^*} | k_0, \beta_0 + 1)$ and $\text{pr}(\underline{\text{n}} \in N_{k^*} | k_0, \beta_0)$. But, by invariance (3), the latter equals $\text{pr}(\underline{\text{n}} \in N_{k^*}^{(+1)} | k_0, \beta_0 + 1)$, where $N_{k^*}^{(+1)}$ has the $\underline{\text{n}}^{(+1)}$ configurations of N_{k^*}. Therefore, to show (2), it suffices to show that $N_{k^*} = N_{k^*}^{(+1)}$, or, that if $\underline{\text{n}}$ belongs to N_{k^*} then it also belongs to $N_{k^*}^{(+1)}$ and vice versa. We show the first.

Take $\underline{\text{n}}$ in N_{k^*}, i.e., $\hat{k}(\underline{\text{n}}) = k^*$, and say the MLE of the date, $\hat{\beta}(\underline{\text{n}}) = \beta^*$. Now, letting $\underline{\text{n}}^{(-1)} = (n_2, ..., n_{12}, n_1)$, we have that:

$$\begin{aligned}
\max_{k,\beta} \prod_{i=1}^{12} \{p_i(k, \beta)\}^{n_i^{(-1)}} &= \max_{k,\beta} \prod_{i=1}^{12} \{p_i(k, \beta + 1)\}^{n_i} \quad &\text{(by invariance of model in (3))} \\
&= \prod_{i=1}^{12} \{p_i(k^*, \beta^*)\}^{n_i} \quad &\text{(by invariance property of MLEs)} \\
&= \prod_{i=1}^{12} \{p_i(k^*, \beta^* - 1)\}^{n_i^{(-1)}}, \quad &\text{(by invariance of model in (3))}
\end{aligned}$$

which shows that k^* is then also the MLE of k with data $\underline{\mathrm{n}}^{(-1)}$, so $\underline{\mathrm{n}}^{(-1)}$ belongs in N_{k^*}, and so $\underline{\mathrm{n}}^{(-1+1)} = \underline{\mathrm{n}}$ must also belong in $N_{k^*}^{(+1)}$, as well as in N_{k^*}. With a similar argument we can show that if $\underline{\mathrm{n}}$ is in $N_{k^*}^{(+1)}$ then $\underline{\mathrm{n}}$ must also be in N_{k^*}, which proves (2).

References

[1] Petridou E, Papadopoulos FC, Frangakis CE, Skalkidou A, Trichopoulos D. A role of sunshine in triggering of suicide. *Epidemiology* 2002; 13: 106–109.

[2] Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Statist Assoc* 1987; 82: 605-610.

[3] Gumbel EJ, Greenwood JA, Durand D. The circular normal distribution: theory and tables. *J Am Statist Assoc* 1953; 48: 131-152.

Country		Relative risk		Angle of Peak (in degrees)
		95% CI		95%CI
Austria	1.22	(1.17, 1.27)	150	(150, 180)
Canada	1.15	(1.11, 1.19)	180	(150, 180)
CzechRep	1.23	(1.21, 1.25)	150	(150,150)
Denmark	1.14	(1.10, 1.18)	150	(120, 150)
Finland	1.20	(1.14, 1.26)	180	(180, 210)
France	1.13	(1.11, 1.15)	150	(150, 150)
Germany	1.16	(1.14, 1.19)	150	(150, 150)
Greece	1.51	(1.38, 1.65)	150	(150, 180)
Hungary	1.45	(1.40, 1.50)	150	(150, 180)
Ireland	1.11	(1.02, 1.20)	150	(120, 210)
Japan	1.21	(1.20, 1.22)	150	(150, 150)
Mexico	1.16	(1.12,1.20)	150	(150, 180)
Netherlands	1.09	(1.04,1.14)	120	(90, 150)
Norway	1.11	(1.03, 1.20)	150	(90, 180)
Spain	1.25	(1.21, 1.28)	150	(120, 150)
Sweden	1.14	(1.09, 1.18)	150	(150, 180)
Switzerland	1.15	(1.10, 1.20)	150	(120, 150)
USA	1.09	(1.07,1.11)	180	(150, 180)
Australia	1.21	(1.15,1.27)	330	(330, 360)
New Zealand	1.14	(1.05, 1.23)	300	(270, 330)

Table 1: Results for relative risk of suicide (maximum over minimum frequency) within countries, using the new method applied to the data of Petridou et al. (2002). Angles are multiples of 30.