Studying seasonality by using sine and cosine functions in regression analysis

A M Stolwijk, H Straatman, G A Zielhuis

Abstract

***Study objective*—A statistical test that allows for adjustment of confounding can be helpful for the study of seasonal patterns. The aim of this article is to supply a detailed description of such a method. An example of its application is given.**

***Design*—A statistical test is presented that retains the information on the connection of time periods by describing the seasonal pattern as one sine and one cosine function. Such functions can be included into a regression model. The resulting form of the seasonal pattern follows a cosine function with variable amplitude and shift.**

***Main results*—The test is shown to be applicable to test for seasonality. Not only one cosine function per time period, but also a mixture of cosine functions can be used to describe the seasonal pattern. Adjustment for confounding effects is possible.**

***Conclusions*—This method for studying seasonal patterns can be applied easily in a regression model. Adjusted prevalences and odds ratios can be calculated.**

(*J Epidemiol Community Health* 1999;**53**:235–238)

Many studies have been published that concerned seasonal variation, for instance in births,[1] early pregnancy loss,[2] and in congenital malformations.[3] Whether a seasonal pattern exists can be studied in several ways. In this article we demonstrate a method that allows for adjustment of confounding. The first part is a general approach of studying seasonal patterns. In the second part we show a more detailed description of the method by means of mathematical functions. Subsequently, we give an application of this method using fictitious data. As an example we use the study of seasonality in anencephaly frequency at birth.

Studying seasonality, a general approach

To study the seasonal variation in congenital malformations, data analysis can be performed in successive stages. The first step is to calculate and plot the prevalence of malformations at birth per month. Then confidence intervals surrounding the monthly prevalences can be calculated and added to the figure. From this information, it can be inferred whether there are differences in malformations per month and whether there are differences between the months. In the same way, clusters of months can be formed and compared with other clusters of months. If confounding is presumed to occur, the next step is to adjust for such confounding effects. One way to perform this is by means of stratification so that insight can be gained into whether prevalences differ between months or clusters of months after adjustment for confounding. In this phase, problems may occur if several confounding factors are present. Adjustment for their effects simultaneously by means of stratification will often lead to small numbers of observations per month and thus to imprecise estimations of the prevalences. None the less, these preliminary phases of analysis will provide the first indications of whether there is a specific seasonal pattern in malformations. Rough evidence of such a pattern warrants a statistical test. In addition, a method is necessary that allows for adjustment of the effects of several confounders simultaneously.

We focus on the question of whether there is a seasonal pattern in malformations during the course of a year, without paying attention to changes between years. To test for seasonality, a χ^2 test can be used to detect any departure from a uniform distribution. A more specific test should take into account the connection between time periods such as months or weeks. The method of Edwards[4] tests whether frequencies follow a sine function over 12 months. Also adaptations of the Edwards' test are suitable, for instance the one of Cave and Freedman[5] to test a bimodal seasonal pattern over 12 months, of Walter and Elwood,[6] which can be used in the case of unequal populations at risk, of Roger[7] for small sample sizes, and of Jones *et al*[8] for an arbitrary shape of the seasonal effect. The non-parametric Hewitt's test[9] or its adaptation for other than six month periods by Rogerson[10] can also be applied, but they are less powerful than parametric tests. A Kolmogorov-Smirnov type statistic of Freedman[11] has a better power than the χ^2 test and the Hewitt's test in samples of moderate size. None of these tests allows for adjustment of confounding effects, except the method of Jones *et al*.[8] Some of them, including the latter, require special software. Moreover, the test of Jones *et al* can only be used for—usually rare—events that follow a Poisson distribution. Therefore another test that allows for adjustment of confounding and that can be performed by widely available statistical computer programs is warranted.

In epidemiological practice, multivariate analysis techniques are commonly used to adjust for confounding. Linear regression analysis is often performed if the dependent variable has a normal distribution. In studies on seasonality in malformations, the dependent

Department of Epidemiology, University of Nijmegen, the Netherlands

Correspondence to: Annette M Stolwijk, Department of Epidemiology, University of Nijmegen, PO Box 9101, NL-6500 HB Nijmegen, the Netherlands.

Accepted for publication 3 September 1998

variable is likely to be dichotomous, for example anencephaly is either present or it is not. In such a case, logistic regression analysis can be used.

To test whether congenital malformations are seasonally distributed, one sine and one cosine function can be introduced into the regression model. This results in a pattern following a cosine function with variable amplitude and shift. Depending on the hypothesis being tested, the period of the cosine function can be one year, half a year or shorter. The maximum likelihood method estimates the regression coefficients for the best fitting regression line. The amplitude and the amount of shift of the cosine function can be calculated from the regression coefficients. For each time period, the probability of a malformation and the odds ratios can be calculated using the logistic regression model. Such a method was applied before, for instance by Bound *et al*[3] and Woodhouse *et al*,[12] but not fully explicated. Therefore this paper will give a detailed description and an example. For more advanced models in time series we refer to Fahrmeir and Tutz.[13]

Detailed description of the method

A linear regression model can be developed to analyse seasonality in congenital malformations. Generally speaking, such a model will have the following form:

$$y = \beta_0 + \beta_{season} \times season + \beta_{C_1} \times C_1 + \ldots . + \beta_{C_N} \times C_N$$

where β_0 is the intercept and C indicates a confounder; y is a continuous variable related to the presence of a congenital malformation and is normally distributed or can be transformed into such a distribution. An example of such a variable is the level of α fetoprotein. In many studies on congenital malformations, the outcome parameter is defined as the probability of malformation. This probability can be modelled in a logistic regression model such as:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{season} \times season + \beta_{C_1} \times C_1 + \ldots . + \beta_{C_N} \times C_N$$

where P is the probability of a malformation, for instance the probability of anencephaly. To define the variable "season" in these models, it is hypothesised that the seasonal pattern under study follows a cosine function with variable amplitude and horizontal shift. In this cosine function, two periods must be defined: (a) the time period that defines the measure of malformation, for example, "month" in "the probability of an anencephalus birth per month" and (b) the period described by one cosine function. As an example we take "month" as the time period under study, and "one year" as the period of the cosine function. The cosine function can be described as:

$$f(t) = \alpha \times \cos\left[\left(\frac{2\pi t}{T}\right) - \theta\right] \quad (1)$$

KEY POINTS
- A statistical test is presented to study seasonal patterns.
- This method can be applied in a regression model.
- Adjustment of confounding is possible.

T= number of time periods described by one cosine function over (0, 2π) (for example, T = 12 months);

t = time period (for example, for January: t = 1, for February: t = 2, etc);

α = amplitude, > 0;

θ = horizontal shift of the cosine function (in radials).

As θ is unknown, transformation of this cosine function is required before the regression analysis can be performed. Therefore the following formula is included into a regression model:

$$f(t) = \beta_1 \times \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times \cos\left(\frac{2\pi t}{T}\right) \quad (2)$$

Including this formula into the logistic regression model results in:

$$\ln\left(\frac{P_t}{1-P_t}\right) = \beta_0 + \beta_1 \times \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times \cos\left(\frac{2\pi t}{T}\right) + \beta_{C_1} \times C_1 + \ldots . + \beta_{C_N} \times C_N$$

Then the probability of malformation in each time period can be calculated by:

$$P_t = \frac{e^{\beta_0 + \beta_1 \times \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times \cos\left(\frac{2\pi t}{T}\right) + \beta_{C_1} \times C_1 + \ldots . + \beta_{C_N} \times C_N}}{1 + e^{\beta_0 + \beta_1 \times \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times \cos\left(\frac{2\pi t}{T}\right) + \beta_{C_1} \times C_1 + \ldots . + \beta_{C_N} \times C_N}}$$

With this information the best fitting seasonal pattern can be plotted. To describe the function of this plot by means of the cosine function presented in formula (1), the amplitude and shift have to be calculated using formula (2). The amplitude is:

$$\alpha = \sqrt{\beta_1^2 + \beta_2^2}$$

and two extreme values in $(0,T)$ can be found at the solutions of:

$$\tan\left(\frac{2\pi t}{T}\right) = \frac{\beta_1}{\beta_2}$$

By solving:

$$t = \arctan\left(\frac{\beta_1}{\beta_2}\right) \times \frac{T}{2\pi}$$

one value t is retrieved. If $\beta_1/\beta_2 > 0$, then $t > 0$ and indicates the first extreme; the other extreme value is found at $t + T/2$. If $\beta_1/\beta_2 \leqslant 0$, then $t \leqslant 0$; the extreme values are found at $t + T/2$ and at $t + T$. If $\beta_1 > 0$, the first extreme is a maximum and the second a minimum; if $\beta_1 \leqslant$

Table 1 Numbers of anencephalus cases and total births, fictitiously divided into data from boys and girls

	Data from Walter and Elwood[6]			*Fictitious data*					
	Total			*Boys*			*Girls*		
Month of birth	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*	*Anencephalus cases*	*Total births*	*Prevalence (per 100 000)*
January	468	340 797	137	463	252 695	183	5	88 102	5.68
February	399	318 319	125	392	215 431	182	7	102 888	6.80
March	471	363 626	130	459	205 341	224	12	158 285	7.58
April	437	359 689	121	417	156 571	266	20	203 118	9.85
May	376	373 878	101	347	120 846	287	29	253 032	11.46
June	410	361 290	113	375	93 400	401	35	267 890	13.07
July	399	368 867	108	364	95 359	382	35	273 508	12.80
August	472	358 531	132	444	115 886	383	28	242 645	11.54
September	418	363 551	115	398	158 252	251	20	205 299	9.74
October	448	352 173	127	436	198 874	219	12	153 299	7.83
November	409	331 964	123	402	224 665	179	7	107 299	6.52
December	397	336 894	118	392	249 801	157	5	87 093	5.74

Table 2 Results of the models and accompanying tests for seasonality, with degrees of freedom (df) and p values

Test		*Maximum*	*Minimum*	*Test statistic*	*Df*	*p Value*	*Deviance*	*Df*	*p Value*
A	Edwards, 12 month period: Frequencies proportional to $1 + \alpha \cos(\Theta_i - \Theta^*)$	Dec	Jun	0.80†	2	<0.7			
B	Walter and Elwood, 12 month period: $m_i[1 + \alpha \cos(\Theta_i - \Theta^*)]$, m_i = number of births per month	Dec	Jun	12.48†	2	<0.005			
C	Logistic, cosine function with 12 month period: $\ln[P/(1-P)] = -6.718 + 0.009822 \times \sin(2\pi t/12) + 0.06929 \times \cos(2\pi t/12)$	Dec	Jun	12.41‡	2	0.002	23.93	9	0.004
D	Logistic, cosine function with 6 month period: $\ln[P/(1-P)] = -6.719 + 0.03647 \times \sin(2\pi t/6) - 0.04525 \times \cos(2\pi t/6)$	Feb, Aug	May, Nov	8.55‡	2	0.01	27.80	9	0.001
E	Logistic, mix of two cosine functions with 12 and 6 month periods: $\ln[P/(1-P)] = -6.719 + 0.009869 \times \sin(2\pi t/12) + 0.06985 \times \cos(2\pi t/12) + 0.03661 \times \sin(2\pi t/6) - 0.04511 \times \cos(2\pi t/6)$	Feb	Jun	8.54‡§	2	0.01	15.39	7	0.03

†Information from Walter and Elwood.[6] ‡Likelihood ratio test result $[-2\ln(L_1/L_2)]$. §Test of model E versus model C.

0 it is the other way around. The maximum extreme (t_{max}) indicates the shift θ in formula (1), which can be calculated by: $2\pi t_{max}/T$ radials.

Application to data

As an example we applied this method for studying seasonality to data from anencephalus births and total births described by Walter and Elwood[6] (table 1). In their article they presented the results of several tests of seasonality: the method of Edwards using only case frequencies or using adjusted frequencies, their own method assuming months of equal length or exact month lengths and Hewitt's non-parametric test. They found that neither Edwards's test using frequencies nor Hewitt's test detected a seasonal pattern. The other three methods did find a seasonal pattern with the maximum prevalence of anencephaly in late December.

We used sine and cosine functions in a logistic regression analysis to test for seasonality in the prevalence of anencephaly. Firstly, we tested whether there was a seasonal pattern with one maximum level and one minimum level per year—that is, a cosine function with a period of 12 months. Secondly, we studied whether the seasonal pattern in the prevalence of anencephaly was better described by a cosine functions with a period of six months (that is, T = 6), or, thirdly, by a mix of one cosine function with a period of 12 months and one with a period of six months. In figure 1 the cosine functions are shown. In table 2 the logistic regression models and their results are presented: regression coefficients, maximum and minimum levels, likelihood ratio test results, and deviances.

Firstly, the test of a seasonal pattern with one maximum per year. The time period is the "month" and the period of the cosine function is "one year". We also found a seasonal pattern. The maximum prevalence was observed in December (t_{max} = 0.27) and the minimum in June (t_{min} = 0.27 + T/2 = 6.27), in agreement with the results described by Walter and Edwards. The pattern could be described by a

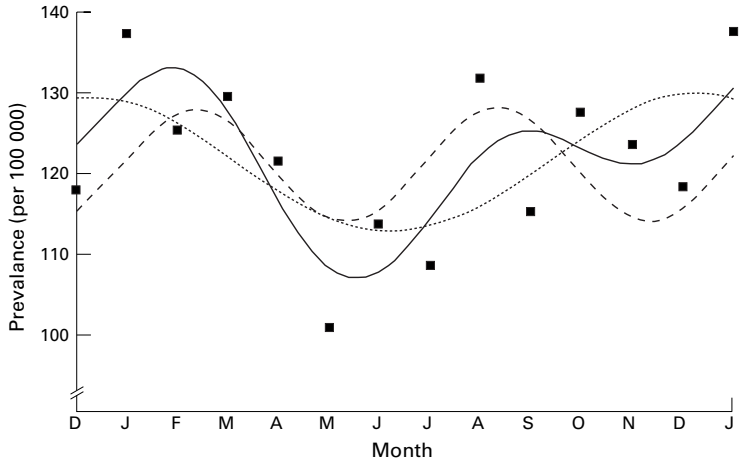

Figure 1 Probability of a child with anencephaly (P), expressed as the prevalence of anencephaly per 100 000 births. Square symbols, prevalence; dotted line, cosine function with period of 12 months; dashed line, cosine function with period of six months; solid line, mix of two cosine functions (see table 2 for formulas).

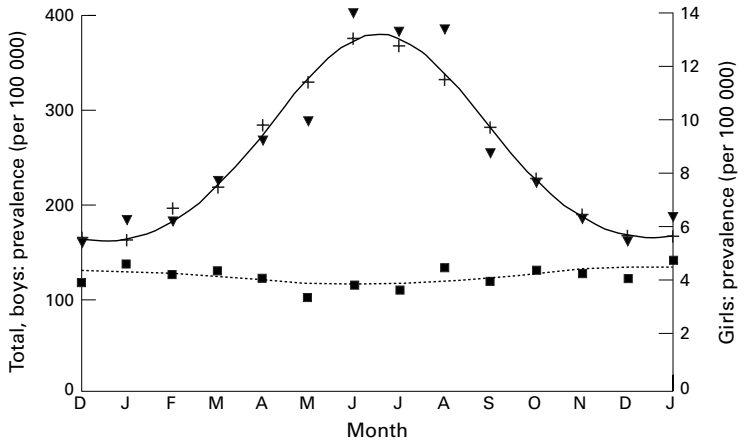

Figure 2 Probability of a child with anencephaly (P), expressed as the prevalence of anencephaly per 100 000 births, crude and after adjustment for sex (fictitious data). Dotted line, crude cosine function with period of 12 months; square symbols, prevalence in total population; solid line, sex adjusted cosine function with period of 12 months, inverted triangles, prevalence among boys; crosses, prevalence among girls.

cosine function with an amplitude of 0.070 and a shift of 0.14 (= $2\pi \times 0.27/T$) radials. Secondly, the test of seasonal pattern that could be described by a cosine function with a period of six months showed no better fit than the former model. Then we tested whether the model using the cosine function with a period of 12 months could be improved by adding a second cosine function to the model with a period of six months. From figure 2 and the likelihood ratio test result in table 2 it can be derived that this extra cosine function improved the model. This model with a mix of cosine functions indicated the highest peak about in February and the lowest through about in June. The goodness of fit statistic for this model is given by a deviance of 15.39, df=7, p=0.03, which shows a statistically lack of fit. This clear indication of overdispersion should be taken into account, for instance by a model based approach that considers a model with a random month effect throughout. We performed this analysis by means of a logistic normal module in EGRET.[14] The comparison of the model with two cosine functions and a random effect versus the "constant" model with random effect resulted in a χ^2 of 25.28–14.98=10.3, df=4, p=0.036. Thus, this appropriate analysis shows that there is seasonality that exceeds the apparent month to month variation that is statistically significant at α=0.05 and can be described by the mix of two cosine functions.

For means of illustration of the way to control for confounding, we ignore the fact that the mix of cosine functions fitted the data of Walter and Elwood better than a cosine function with a period of 12 months. We fictitiously divided the data into data from boys and from girls, so that confounding by sex was present (table 1). In figure 2 the unimodal crude and adjusted cosine functions with a period of 12 months are presented, showing considerable confounding by sex. As shown before, the crude cosine function had an amplitude of 0.070 and a shift of 0.14 radials, which gives the maximum prevalence of anencephaly in December and the minimum in June. The sex adjusted seasonal pattern in the prevalence of anencephaly could be described by:

$$\ln\left(\frac{P_t}{1-P_t}\right) = -9.358 - 0.1182 \times \sin\left(\frac{2\pi t}{12}\right) - 0.4086 \times \cos\left(\frac{2\pi t}{12}\right) + 3.358 \times sex$$

in which sex = 1 for boys and sex = 0 for girls. The maximum extreme was found in June (t_{max} = 0.54 + T/2 = 6.54). This results in a cosine function with an amplitude of 0.43 and the shift of $2\pi \times 6.54/T$ = 3.42 radials (fig 2). Thus after adjustment for the fictitiously introduced sex distribution, we found a totally opposite seasonal pattern with the maximum prevalence observed in June and the minimum in December.

We conclude that it is possible to test for seasonal patterns by means of applying sine and cosine functions into regression analysis. Not only a period of, for instance, 12 or 6 months can be described by such a cosine function, but also a mixture of cosine functions is possible. Moreover, this way of analysing seasonality allows for adjustment of confounding effects.

1 Matsuda S, Kahyo H. Geographical differences and time trends in the seasonality of birth in Japan. *Int J Epidemiol* 1994;**23**:107–18.
2 Weinberg CR, Moledor E, Baird DD, *et al*. Is there a seasonal pattern in risk of early pregnancy loss? *Epidemiology* 1994;**5**:484–9.
3 Bound JP, Harvey PW, Francis BJ. Seasonal prevalence of major congenital malformations in the Flyde of Lancashire 1957–1981. *J Epidemiol Community Health* 1989;**43**:330–42.
4 Edwards JH. The recognition and estimation of cyclic trends. *Ann Hum Genet Lond* 1961;**25**:83–7.
5 Cave DR, Freedman LS. Seasonal variations in the clinical presentation of Crohn's disease and ulcerative colitis. *Int J Epidemiol* 1975;**4**:317–20.
6 Walter SD, Elwood JM. A test for seasonality of events with a variable population at risk. *Br J Prev Soc Med* 1975;**29**:18–21.
7 Roger JH. A significance test for cyclic trends in incidence data. *Biometrika* 1977;**64**:152–5.
8 Jones RH, Ford PM, Hamman RF. Seasonality comparisons among groups using incidence data. *Biometrics* 1988;**44**: 1131–44.
9 Hewitt D, Milner J, Csima A, *et al*. On Edwards' criterion of seasonality and a non-parametric alternative. *Br J Prev Soc Med* 1971;**25**:174–6.
10 Rogerson PA. A generalization of Hewitt's test for seasonality. *Int J Epidemiol* 1996;**25**:644–8.
11 Freedman LS. The use of a Kolmogorov-Smirnov type statistic in testing hypotheses about seasonal variation. *J Epidemiol Community Health* 1979;**33**:223–8.
12 Woodhouse PR, Khaw KT, Plummer M, *et al*. Seasonal variations of plasma fibrinogen and factor VII activity in the elderly: winter infections and death from cardiovascular disease. *Lancet* 1994;**343**:435–9.
13 Fahrmeir L, Tutz G. *Multivariate statistical modelling based in generalized linear models*. Springer Series in Statistics. New York: Springer Verlag, 1994:187–218.
14 EGRET. *Reference Manual*. First draft, revision 3. Statistics and Epidemiology Research Corporation and Cytel Software Corporation, 1985–1992:278.