International Journal of Epidemiology
© International Epidemiological Association 1996

Vol. 25, No 2
Printed in Great Britain

Analysis of Seasonal Data Using the Lorenz Curve and the Associated Gini Index

WEN-CHUNG LEE

Lee W-C (Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Rd, 1st Sec, Taipei, Taiwan, Republic of China). Analysis of seasonal data using the Lorenz curve and the associated Gini index. *International Journal of Epidemiology* 1996; **25**: 426–434.

Background. Epidemiological inferences about the aetiology of a disease can often be made from its seasonal patterns. However, due to its multifactorial nature, the seasonality component can be obscured by other factors. It is therefore important to develop statistical techniques which are sensitive to minute temporal changes.

Methods. The Lorenz curve and the associated Gini index are applied for characterizing and testing seasonal variations. Computer simulations were conducted to compare the powers of the Gini test and other seasonality tests. We also show that the Gini index can itself be interpreted as a probability related to temporal clustering.

Results. The powers of the proposed tests are shown to be higher than or at least comparable to other tests under various conditions.

Conclusions. Though computer-demanding, the proposed method is well-suited for analysing seasonal data.

Keywords: Gini index, Lorenz curve, Monte Carlo simulation, seasonal variation, temporal clustering

The study of seasonal variation in disease frequency has been of great interest to epidemiologists and medical professionals. A seasonal pattern in the occurrence of a disease, say a temporal clustering, a cyclical variation, or a long-term trend, suggests the presence of environmental factors in its aetiology. Many infectious diseases exhibit well-known temporal patterns. However, the occurrences of some chronic diseases or conditions may also be influenced by seasonality. Examples are some types of cancers,[1,2] ischaemic heart disease,[3] congenital abnormalities,[4,5] peptic ulcer perforation,[6] and mental disorders,[7–9] etc. These diseases have multifactorial aetiology and thus the seasonality component is often obscured by other factors. It is therefore important to develop statistical techniques which are sensitive to minute temporal changes.

The data for this type of study are usually presented and analysed in the form of a series of 12 monthly totals. To test for temporal variation, the ordinary χ^2 goodness-of-fit test (with 11 degrees of freedom) is not particularly suitable due to its low power.[10] Edwards,[11] by way of mechanical analogy of weights, attaches square roots of monthly frequencies on the rim of a unit circle (divided into 12 equal sectors), and derives his test statistic as the distance from the centre of gravity of these weights to the centre of the unit circle. Such a test is indeed very sensitive to detecting a sinusoidal departure (one peak and one trough in a year) from the null hypothesis of uniform distribution of cases over 12 monthly intervals. However, when the disease occurrence follows a sinusoidal pattern with a period of 6 months rather than 12 months, an alternative test[12] is needed since the Edwards's test fails to detect this particular type of departure (the centre of gravity is right on the centre of the circle for this case). To be consistent for all types of departures from uniformity (able to detect any type of seasonal variations when the sample size is large) while retaining reasonably high power when the specific alternative is a sinusoidal one, Freedman suggests using the Kolmogorov-Smirnov type statistic, specifically Kuiper's statistic, for testing hypotheses of seasonality.[13] Several other tests have also been developed.[14–19] The performance of some of these seasonality tests has been investigated and compared through computer simulations.[13,17,20]

In this paper, the author applies the Lorenz curve and the associated Gini index for characterizing and testing seasonal variation of disease occurrence. The Lorenz curve and the Gini index, though perhaps new to epidemiologists, are widely used by economists to assess the distributional properties of family income and

Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Rd, 1st Sec, Taipei, Taiwan, Republic of China.

TABLE 1 *Births of anencephalics to primiparous women in Birmingham, England during 1940–1947*

Month	No . of cases	Day	Incidence (case/day)	Three-month moving average			
				with weights: 1/3, 1/3, 1/3		with weights: 1/4, 2/4, 1/4	
				Incidence	No. of cases	Incidence	No. of cases
Jan	10	248	0.0403	0.0706	17.52	0.0629	15.60
Feb	19	226	0.0841	0.0651	14.71	0.0696	15.74
Mar	18	248	0.0726	0.0728	18 06	0.0728	18.05
Apr	15	240	0.0625	0.0598	14.35	0.0605	14.51
May	11	248	0.0444	0.0536	13.28	0.0512	12.71
Jun	13	240	0.0542	0.0421	10.11	0.0451	10.82
Jul	7	248	0.0282	0.0408	10.11	0.0376	9.33
Aug	10	248	0.0403	0.0408	10.11	0.0407	10.08
Sep	13	240	0.0542	0.0625	15.00	0.0605	14.51
Oct	23	248	0.0927	0 0701	17.37	0.0758	18.81
Nov	15	240	0.0625	0.0815	19.56	0.0768	18.45
Dec	22	248	0.0887	0.0639	15.83	0.0701	17.39
Total	176	2922			176		176

wealth[21] and by demographers to quantify the degree of population concentration.[22] Their use will be illustrated with published monthly data of anencephalics. We will also show that the Gini index can be used not only as a statistic for testing seasonal variation but can itself be interpreted as a probability related to temporal clustering. The ability of the Gini index to detect some different seasonal patterns will be investigated using computer simulations.

THE PROPOSED METHODOLOGY

In the following, data on births of anencephalics to primiparous women in Birmingham, England during 1940–1947 (176 cases in total) are used to illustrate the methodology. The data are taken from Edwards's paper[11] and pooled into 12 monthly totals (Table 1).

The Construction of the Lorenz Curve

To construct the Lorenz curve, the incidence of anencephalics (defined as cases per day) for each month is first calculated. And then the months are ranked, from the lowest to the highest, according to the monthly incidences, that is, Jul, Jan, Aug, May, Jun, Sep, Apr, Nov, Mar, Feb, Dec, Oct (Table 1). The Lorenz curve is the plot of the cumulative percentage of 'cases' against the cumulative percentage of 'days' (Figure 1). Note that our use of the Lorenz curve is somewhat different from that of economists or demographers. Economists, in assessing the distributional properties of family income and wealth, often plot the cumulative percentage of income against the cumulative percentage of population. While for the purpose of quantifying the degree of population concentration, demographers plot the cumulative percentage of population number against the cumulative percentage of land area. In all these applications, the message conveyed in the curve is

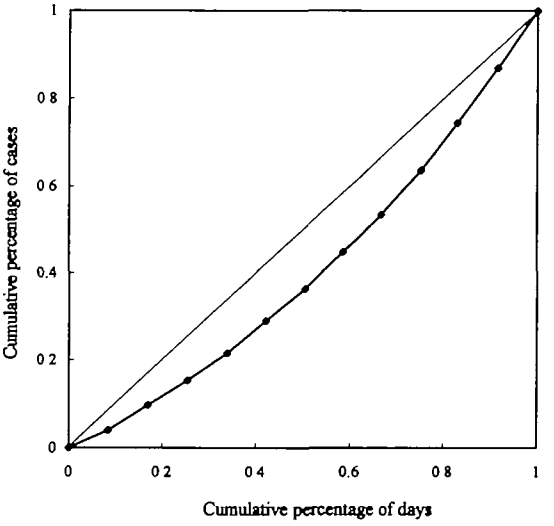

FIGURE 1 *The Lorenz curve of the example in Table 1. Twice the area between the curve and the diagonal line is defined as the Gini index*

essentially the same. That is, with the most even distribution obtainable (e.g. every family in a country earns the same income, or the population density is the same everywhere in a defined geographical area, and/or the monthly incidences of disease occurrence are equal), the curve would follow the diagonal line throughout (the diagonal line is the line connecting the points (0,0) and (1,1)). Whereas with a maximal concentration (or unevenness), the curve would, for the study of seasonal variation, coincide with the X-axis for the first 11 months and then jump up to the right uppermost point (1,1). This could happen if all cases occurred exclusively in one particular month. This observation suggests the use of the area between the Lorenz curve and the diagonal as the index of temporal clustering. A larger area, resulting from a more bowed Lorenz curve, means that the cases are more concentrated in particular months, while a smaller area or a flatter curve indicates a more uniform distribution.

The Meaning and Use of the Gini Index

The Gini index is defined as twice the area between the Lorenz curve and the diagonal line, or equivalently as the ratio of the aforementioned area to the area of the triangle below the diagonal line. Clearly this index is between zero and one, with larger values indicating greater concentration while a smaller one indicates greater uniformity.

Calculating the Gini index is straightforward and can be done using simple algebra. For the anencephalics data the index is calculated as 0.1881. To determine whether this figure is statistically larger than zero (no seasonal variation) is more difficult as the Gini index does not have a theoretical reference distribution as yet. We therefore perform 10 000 Monte Carlo simulations, each with sample size of 176, to derive its approximate distribution. The 99th percentile of the simulated sampling distribution is found to be 0.2072 and the 95th percentile, 0.1843. Therefore we see that the null hypothesis of no seasonality can be rejected at the $\alpha = 0.05$ level but not at the $\alpha = 0.01$ level.

The Gini index can be used not only as a statistic for testing seasonal variation but can itself be interpreted as a probability related to temporal clustering. To see this, consider the index, (1/2 + Gini/2). This refined index is between 0.5 and 1. The Appendix shows that it can be interpreted as the probability that a randomly selected case (among the pool of total cases) arrives from a month whose incidence of disease occurrence is higher than a randomly selected month (according to the probability proportional to the days in each month). Note that if the randomly selected case arrives from the same month as the randomly selected month, we toss a coin to decide whether the comparison is deemed higher. It is conceivable that when the cases are arriving predominantly from one or a few high risk months, the above-defined probability would be higher, whereas when the cases are distributed evenly, we would have no better chance of guessing to which months the randomly chosen case belongs other than tossing a coin (the probability index is 0.5, at its lowest value). Since the above-defined probability index is closely related to temporal clustering, it may be referred to as the 'clustering probability'. For the example of anencephalics, the clustering probability is 0.5941 (1/2 + 0.1881/2). Therefore one can infer that the cases of anencephalics are not very concentrated in particular months even though the seasonality test reveals a significant temporal variation (at $\alpha = 0.05$).

The Application of Smoothing Technique

From the above presentation, it can be seen that the first and crucial step in the plotting of the Lorenz curve and the calculation of the Gini index is the ranking of the months according to the monthly incidences. However, the incidences are estimated by dividing the monthly number of cases by the number of days in each month, and thus are subject to sampling variability. The problem is more serious when the monthly case numbers are small (or equivalently the incidences are low). This instability in the incidence estimates will affect the ranking of the months and hence the precision of the Gini index. To ease the problem, one may use the smoothing technique.[23] The rationale behind smoothing lies in the fact that there is strong *a priori* expectation for most diseases that the monthly incidences more often follow a fairly smooth curve rather than exhibit sudden and violent jumps. For the above example of anencephalics, we calculate the 3-month moving average incidence for each month, with two weighting schemes, (1/3, 1/3, 1/3) and (1/4, 2/4, 1/4). Taking the smoothed incidences, one then derives the expected cases for each month (Table 1). The smoothed Lorenz curves are similarly constructed by plotting the cumulative percentage of the expected cases against the cumulative percentage of days. The smoothed Gini indices for the two weighting schemes (referred to later as Gini-1 and Gini-2, respectively) can then be calculated, using simple algebra, as twice the area between the smoothed Lorenz curve and the diagonal line as before. For our example, the figures are 0.1198 and 0.1235, respectively. The Gini-1 and the Gini-2 indices can also be used as test statistics for testing seasonality provided that their sampling distributions under the null hypothesis are determined. To this end, we perform 10 000 Monte Carlo simulations again, each with

TABLE 2 *Models of seasonal variation investigated in the simulation study. The monthly incidences (in proportional terms) are shown*

Month	Models				
	1	2	3	4	5
Jan	1.065	1 125	0.875	1.125	1.000
Feb	1.177	1 250	0.875	1.125	1.000
Mar	1.241	1.125	0.875	1.125	1.000
Apr	1.241	0.875	0.875	1.125	1.000
May	1.177	0 750	0.875	1.125	1.000
Jun	1.065	0.875	1.100	0.900	1.200
Jul	0.935	1 125	1.500	0 500	1.400
Aug	0.823	1.250	1.300	0.700	1.200
Sep	0 759	1.125	1.100	0.900	0.800
Oct	0.759	0.875	0.875	1.125	0.600
Nov	0 823	0.750	0.875	1.125	0.800
Dec	0.935	0.875	0.875	1.125	1.000

sample size of 176. The 99th percentiles are found to be 0.1306 and 0.1377 for Gini-1 and Gini-2, and the 95th percentiles, 0.1090 and 0.1156, respectively. It turns out that for this particular example, the null hypothesis of no seasonality is rejected at the $\alpha = 0.05$ level but not at the $\alpha = 0.01$ level, using either the Gini, the Gini-1, or the Gini-2 index.

STATISTICAL POWERS OF THE VARIOUS SEASONALITY TESTS

The statistical powers of using the Gini indices (including Gini-1 and Gini-2) for testing seasonality as compared to other well-known seasonality tests are investigated through computer simulations. We consider five different models of seasonal variation as the alternative hypotheses (Table 2). Model 1 is the simple sinusoidal curve with one peak and one trough. To be specific, the incidence in month i is proportional to $1 + 0.25 \sin((2i - 1)\pi/12)$. Model 2 is also a sinusoidal curve but with two peaks and two troughs within a year ($1 + 0.25\sin((2i - 1)\pi/6)$). Model 3 is a one-peak model. This model reflects the situation when the particular disease occurs predominantly in some of the hottest months in the year. Model 4 is a one-trough model, which may happen when the infectivity of a transmittable disease is greatly reduced in some months due perhaps to specific climatic conditions in these months. The last one is a model with one peak followed by one trough. This model simulates a hypothetical epidemic process. The epidemic (the peak) wipes out the pool of susceptibles and/or builds up herd immunity in the population, hence a trough follows.

The tests being considered in this paper are Edwards's test,[11] Roger's test,[18] Kuiper's test,[13] the χ^2 test (11 degrees of freedom), and the proposed Gini tests. The powers of these tests are compared at sample sizes of 25, 50, 100, 200, and 500 with levels of significance (α level) fixed at 0.01 and 0.05. Although the asymptotic distributions of some of these tests are known, the approximation may not be satisfactory for small or even medium-sized samples (especially for Edwards's test).[18,20] To avoid making the type I error in these tests too disparate from the nominal α level and hence leading to a spurious power comparison, we determine the critical values (at α levels of 0.01 and 0.05) of all of these tests via 10 000 Monte Carlo simulations for each specified sample size. The powers of these tests at the various sample sizes are presented in Tables 3–7. These are estimated based on 1000 simulations for each alternative hypothesis per test. Most simulation studies for seasonality are also based on such numbers of simulations.[13,17,20] Certainly this magnitude of simulation will give rise to some sampling variation. To avoid over-interpreting small differences in power that may be non-significant, one can resort to, e.g. the McNemar test for hypothesis testing. For simplicity, we only perform some repeated calculations for selected instances as the likely magnitude of simulation error can also be (roughly) gauged this way. Also note that the number of days in each month is assumed equal in the power comparison, since a detailed consideration only creates an unnecessary complication for the present purpose.

Table 3 presents the case of a simple sinusoidal model (model 1). It can be seen that Edwards's and Roger's tests have highest power. The χ^2 test has low

TABLE 3 *Estimated power (%) of various tests for model 1 (sinusoidal curve with one peak and one trough)*

Sample size	Level of significance	Tests						
		Edwards	Roger	Kuiper	χ^2	Gini[a]	Gini-1[b]	Gini-2[c]
25	0.01	3.0	2.9	3.2	1.5	1.4	2.6	2.5
	0.05	10.3	12.9	12.1	6.0	6.8	10.6	10.0
50	0.01	4.0	4.9	3.9	3.0	2.4	4.8	4.3
	0.05	14.8	16.1	15.9	9.5	10.3	13.7	13.4
100	0.01	10.3	10.7	9.8	3.8	4.3	10.3	9.6
	0.05	30.5	30.4	26.6	13.9	15.0	28.3	25.9
200	0.01	39.0	40.0	29.6	12.9	13.7	34.5	30.3
	0.05	60.2	60.4	52.3	30.5	31.7	58.5	54.4
500	0.01	86.1	86.0	82.4	53.7	55.3	84.3	84.4
	0.05	95.7	95.6	93.4	75.6	76.5	94.8	94.6

[a] Test for temporal variation based on the Gini index.
[b] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/3, 1/3, 1/3.
[c] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/4, 2/4, 1/4.

TABLE 4 *Estimated power (%) of various tests for model 2 (sinusoidal curve with two peaks and two troughs)*

Sample size	Level of significance	Tests						
		Edwards	Roger	Kuiper	χ^2	Gini[a]	Gini-1[b]	Gini-2[c]
25	0.01	1.4	1.2	1.7	1.6	1.8	2.0	1.7
	0.05	4.2	5.4	7.2	5.9	7.2	6.2	6.1
50	0.01	1.0	1.4	2.1	3.2	2.8	3.0	3.9
	0.05	6 4	5.9	9.4	9.8	10.7	9 2	11.5
100	0.01	1.0	0.7	1.9	4.6	5.3	3.5	4.3
	0.05	4.7	4 1	10.1	13.7	14.0	15 7	17.4
200	0.01	1.5	1.3	6.5	15.7	15.1	13.2	17.2
	0.05	5.9	6 4	20.0	32.4	34.6	34.9	40.1
500	0.01	0.6	0.7	23.2	54.9	56.6	51.7	63.1
	0.05	5.1	5.1	49.1	77.5	78.5	80.5	85.5

[a] Test for temporal variation based on the Gini index.
[b] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/3, 1/3, 1/3.
[c] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/4, 2/4, 1/4.

power. Whereas the non-parametric Kuiper's test is nearly as powerful as Edwards's test. These findings are consistent with the study of Freedman.[13] However, we note that the Gini-1 and Gini-2 tests which are also non-parametric tests, achieve even higher power than Kuiper's test when sample size is larger than 100. Three repeated calculations (each with 1000 simulations) are consistent with this finding.

The powers when the alternative is the sinusoidal curve with a period of 6 months (model 2) are presented in Table 4. It can be seen as expected that the Edwards's and Roger's tests have hardly any power to detect this particular type of departure. Kuiper's test, however, retains some ability to this end, but has even lower power than the χ^2 test. It is noted that the Gini-2 test is the only one which reaches a reasonably high

TABLE 5 *Estimated power (%) of various tests for model 3 (one peak)*

Sample size	Level of significance	Tests						
		Edwards	Roger	Kuiper	χ^2	Gini[a]	Gini-1[b]	Gini-2[c]
25	0.01	2.4	3.0	2.8	2.3	2.9	3.3	2.8
	0.05	9.4	11.4	11.9	7.7	7.3	11.4	11.1
50	0.01	4.0	4.1	4.2	4.7	3.4	5.5	5.6
	0.05	13.7	16.8	16.7	12.7	12.8	16.4	16.2
100	0.01	9.1	10.9	10.6	9.9	7.7	13.8	13.4
	0.05	26.0	28.8	29.7	21.4	19.9	32.0	30.9
200	0 01	25.8	30.1	28.2	19.8	17.1	31.9	29.7
	0.05	45.0	48.7	48.1	38.1	36.8	53.8	52.0
500	0.01	68.7	71.9	79.6	69 7	64.5	80.0	81.5
	0.05	86.8	89.1	92.9	86.5	83.7	94.5	93.8

[a] Test for temporal variation based on the Gini index.
[b] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/3, 1/3, 1/3.
[c] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/4, 2/4, 1/4.

power for this alternative among the tests studied (three repeated calculations are consistent with this finding). There exist, however, parametric tests such as the test of Cave and Freedman[12] that are specially designed to test the two-peak-two-trough alternative. We did not compare the performance of the Gini tests with those tests in this paper since our intention here is to show that the Gini test performs well under various alternatives rather than to claim that it is the most powerful one.

The power comparison under a one-peak model (model 3) is shown in Table 5. It can be seen that the previous belief that the tests of Edwards and Roger have no ability whatsoever to detect seasonality other than in a sinusoidal model is clearly wrong. However, we note that it is the Gini-1 and Gini-2 tests that have the highest power in this situation again (three repeated calculations are consistent with this finding).

The case of the one-trough model (model 4) is presented in Table 6. Rather unexpectedly as the Edwards's test is not a specific test for this situation it achieves a power comparable to Kuiper's test. It can also be seen that Gini-1 and Gini-2 also perform well under this 'a-hole-in-time' model.

Table 7 presents the results of power comparison under model 5. This time, none of the traditional tests rival the proposed Gini-1 and Gini-2 tests in statistical power (three repeated calculations are consistent with this finding).

DISCUSSION

In this paper, we introduce the Lorenz curve and the associated Gini index for making inference about seasonal patterns. The methodology is easy to apply and, like many other seasonality tests, has graphical representation. Through our simulation study, the powers of the proposed tests are shown to be higher than, or at least comparable to, other tests under various conditions. Moreover, the proposed test statistic *per se*, i.e. the Gini index, can also be interpreted as a probabilistic measure of temporal clustering. These properties render the methodology well-suited for analysing seasonal data.

The proposed Gini tests have one additional virtue, that of being able to deal with 'calendar effects' explicitly. Calendar effects, such as the variation in month length, the irregular number of weekend days in each month, and the occurrence of holidays etc, can produce spuriously significant seasonal effects if they are not taken into account properly in the analysis.[24] For control of this problem in our framework, one needs only to plot the cumulative percentage of cases against the cumulative percentage of the relevant denominators. The others remain the same. For example, if the disease of interest is absence through sickness in elementary schools, then the relevant denominator may be taken to be the number of school days in each month. If the disease of concern is some type of congenital anomaly such as anencephaly as in our example, an alternative way of analysis is to take the cumulative

TABLE 6 *Estimated power (%) of various tests for model 4 (one trough)*

Sample size	Level of significance	Tests						
		Edwards	Roger	Kuiper	χ^2	Gini[a]	Gini-1[b]	Gini-2[c]
25	0.01	2.2	2.7	1.5	1.5	2.2	2.7	2.1
	0.05	11.7	9.4	8.9	5.1	6 2	9.3	9.3
50	0.01	4.7	4.2	4.0	1.9	1.7	5.1	5.6
	0.05	19.2	17.5	18.6	10.6	12.1	16.9	18.0
100	0.01	10.9	9.7	8.8	6.0	6.3	11.8	11.7
	0.05	30.3	25.8	27.7	18 6	19.4	30.1	29.9
200	0.01	34.3	29.5	25.9	18.4	17.6	33.2	32.5
	0.05	57 4	50.4	53.9	39.7	42.3	59.6	59 0
500	0.01	83.1	76.7	85.9	77.1	73.0	86.5	88.3
	0.05	93.6	91.3	94.5	92 3	89.7	96.1	96.2

[a] Test for temporal variation based on the Gini index.
[b] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/3, 1/3, 1/3.
[c] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/4, 2/4, 1/4.

TABLE 7 *Estimated power (%) of various tests for model 5 (one peak followed by one trough)*

Sample size	Level of significance	Tests						
		Edwards	Roger	Kuiper	χ^2	Gini[a]	Gini-1[b]	Gini-2[c]
25	0.01	1.3	1.4	1.7	2.1	2.3	2.3	2.1
	0.05	7.5	8.9	9 9	7.9	8.1	9 3	8.2
50	0.01	2.5	2.8	2.7	2.8	2.2	3.1	3 4
	0.05	11.2	11.7	14.1	11.9	12.8	12.8	14.8
100	0.01	6.6	5.5	8.2	6.5	6.9	9.1	8.8
	0.05	19.0	18.1	22.6	18.1	18.1	24 9	25.1
200	0.01	16.3	16.1	21.8	20.0	19.4	26.9	29.4
	0.05	35.6	34.3	41.9	41.6	42 4	51.1	50.9
500	0.01	48.7	46.5	72.8	70.9	70.3	76.0	81.5
	0.05	73.1	71.2	88 4	86.5	87.3	92.4	93.7

[a] Test for temporal variation based on the Gini index.
[b] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/3, 1/3, 1/3.
[c] Test for temporal variation based on the 3-month moving average Gini index with weights, 1/4, 2/4, 1/4.

percentage of live births in each month instead of number of days as the abscissa of the Lorenz curve. Edwards's and Roger's tests can also be modified to cope with the problem of calendar effects, however the modification is not as straightforward.[25]

In our opinion, the only drawback (or inconvenience) of using the proposed methodology is that the sampling distributions of the Gini indices under the null hypothesis have not yet been derived theoretically and thus Monte Carlo simulations are needed to perform the tests. However, we believe that nowadays this may not create a serious problem because of easy access to fast computations. A program written in SAS/IML language[26] is available from the author.

It is of interest to compare our approach to standard parametric tests such as Edwards's test. Parametric tests are generally more powerful when the alternative hypothesis conforms to the pre-specified form. Whereas our approach, a non-parametric one, does not call for any prior knowledge about the likely patterns of seasonality, yet retains reasonably high power under various conditions. This property is of special importance, particularly for the study of seasonality of diseases with obscure aetiology (such as cancers, congenital anomalies, etc.). On the other hand, Edwards's test leads to estimation of some interpretable parameters, i.e. the height of the seasonality peak and its angular location, whereas in our approach, the graph is drawn after ranking the data, which means that the graph contains no information whatsoever about patterns in the data, only information about variation. Nevertheless, as shown in this paper, the Gini index itself can be interpreted as a clustering probability. Such a clustering probability represents yet another useful parameter of the inherent characteristic of the seasonal variation. Inherent here means that it does not depend on the sample size. This, in contrast, is not true for other test statistics commonly used in seasonality tests. In this respect, the Gini index also abstracts important information about the magnitude and degree of temporal clustering from the data and therefore is useful not merely for the purpose of hypothesis testing.

The results indicate that the power of the unsmoothed Gini test is at best comparable to that of the ordinary χ^2 test, while the smoothed Gini tests generally achieve higher power than the non-parametric Kuiper's test. Such results are conceivable since the unsmoothed Gini test and the χ^2 test do not take into account the ordered structure of the seasonal variation while Kuiper's test, which makes use of the cumulative distributions, may mask and destroy some interesting local features in seasonal pattern. In between these two extremes thus lies our belief that an appropriate smoothing of the data which preserves some local orderings may be necessary for the construction of a useful seasonality test. In this paper, two different smoothing schemes are used, i.e., the 1/3, 1/3, 1/3 weighting scheme and the 1/4, 2/4, 1/4 scheme. The former treats the preceding and the following months equally to the current month and the degree of smoothing is greater, while the latter puts more weight on the current month itself and thus attains less smoothing. The difference in the degree of smoothing leads to different powers of the Gini-1 and the Gini-2 tests under different conditions. Take the two sinusoidal alternatives studied in this paper, for example. The sinusoidal curve of model 1 has less up-and-down turns within a year than that of model 2, meaning that the latter is less smooth than the former. Therefore the statistical power of Gini-1 appears higher (though only slightly) than Gini-2 for model 1 while the reverse is true for model 2. Further studies seem necessary to examine and compare the performance of other smoothing schemes such as using a 5-month moving average instead of the 3-month average and/or using kernel smoothing technique.[23]

Finally, beyond the characterization and testing of seasonal patterns, the Lorenz curve and the Gini index may be used to analyse other interesting characteristics of disease occurrence, such as geographical variations and/or spatio-temporal clustering. These new applications deserve deeper exploration.

REFERENCES

[1] Fraumeni J F. Seasonal variation in leukaemia incidence. *Br Med J* 1963; **ii**: 1408–09.

[2] Lee J A H. Summer and death from neuroblastoma. *Br Med J* 1967; **ii**: 404–07.

[3] Bowie C, Prothero D. Finding causes of seasonal diseases using time series analysis *Int J Epidemiol* 1981, **10**: 87–92.

[4] Wehrung D A, Hay S. A study of seasonal incidence of congenital malformations in the United States. *Br J Prev Soc Med* 1970; **24**: 24–32.

[5] Jongbloet P H, Mulder A M, Hamers A J Seasonality of preovulatory non-disjunction and the aetiology of Down syndrome. a European collaborative study. *Hum Genet* 1982; **62**: 134–38.

[6] Mackay C. Perforated peptic ulcer in the West of Scotland: a survey of 5343 cases during 1954–63. *Br Med J* 1966; **i**: 701–05.

[7] Symonds R L, Williams P. Seasonal variation in the incidence of mania. *Br J Psychiatry* 1976; **129**: 45–48.

[8] Hare E H Season of birth in schizophrenia and neurosis. *Am J Psychiatry* 1975; **132**: 1168–71.

[9] Hare E H, Walter S D. Seasonal variation in admissions of psychiatric patients and its relation to seasonal variation in their birth. *J Epidemiol Community Health* 1978; **32**: 47–52.

[10] Horn S D. Goodness-of-fit tests for discrete data: a review and an application to a health impairment scale. *Biometrics* 1977; **33**: 237–48.

[11] Edwards J J. The recognition and estimation of cyclic trends. *Ann Hum Genet* 1961; **25**: 83–86.

[12] Cave D R, Freedman L S. Seasonal variations in the clinical presentation of Crohn's disease and ulcerative colitis. *Int J Epidemiol* 1975; **4**: 317–20.

[13] Freedman L S. The use of a Kolmogorov-Smirnov type statistic in testing hypotheses about seasonal variation. *J Epidemiol Community Health* 1979; **33**: 223–28.

[14] David H A, Newell D J. The identification of annual peak periods for a disease. *Biometrics* 1965; **21**: 645–50.

[15] Hewitt D, Milner J, Csima A, Pakula A. On Edwards' criterion of seasonality and a non-parametric alternative. *Br J Prev Soc Med* 1971; **25:** 175–76.

[16] Pocock S J. Harmonic analysis applied to seasonal variations in sickness absence. *Appl Stat* 1974; **23:** 103–20.

[17] St Leger A S. Comparison of two tests for seasonality in epidemiological data. *Appl Stat* 1976; **25:** 280–86.

[18] Roger J H. A significance test for cyclic trends in incidence data. *Biometrika* 1977; **64:** 152–55.

[19] Best D J, Rayner J C W. Disease clustering in time. *Biometrics* 1991; **47:** 589–93.

[20] Marrero O. The performance of several statistical tests for seasonality in monthly data. *J Stat Comp Simul* 1983; **17:** 275–96.

[21] Ekelund R B, Tollison R D, *Economics*. Boston: Little, Brown and Co., 1986.

[22] Shryock H S, Siegel J S. *The Methods and Materials of Demography*. Washington: US Government Printing Office, 1975.

[23] Thisted R A. *Elements of Statistical Computing*. New York: Chapman and Hall, 1988.

[24] Walter S D. Calendar effects in the analysis of seasonal data. *Am J Epidemiol* 1994; **140:** 649–57.

[25] Walter S D, Elwood J M. A test for seasonality of events with a variable population at risk. *Br J Prev Soc Med* 1975; **29:** 18–21.

[26] SAS Institute Inc. *SAS/IML User's Guide, Release 6.03* Cary, NC, USA: SAS Institute Inc., 1988.

(Revised version received August 1995)

APPENDIX

Assume that the 12 months have been rearranged according to the monthly incidences (from the lowest to the highest) and are indexed by i (i = 1,2,..., 12). The number of days in each month is represented by y_i and the number of cases by n_i. Let y denote total days in a year ($y = \Sigma_i y_i$) and let n denote total cases ($n = \Sigma_i n_i$). The probability (clustering probability) that a randomly selected case arrives from a month whose incidence of disease occurrence is higher than a randomly selected month can easily be derived. That is,

$$\text{The clustering probability } = \sum_i \frac{y_i}{y} \cdot \left[\frac{\sum_{j>i} n_j}{n} + \frac{1}{2} \cdot \frac{n_i}{n} \right]$$

$$= \sum_i \frac{y_i}{y} \cdot \left[\frac{\sum_{j>i} n_j}{n} + \frac{\sum_{j \geq i} n_j}{n} \right] \cdot \frac{1}{2}$$

It can thus be recognized that this clustering probability is equal to the area above the Lorenz curve, which is, geometrically, just the sum of 0.5 and half of the Gini index (1/2 + Gini/2).