Journal of Epidemiology and Community Health, 1979, **33**, 223–228

The use of a Kolmogorov–Smirnov type statistic in testing hypotheses about seasonal variation

L. S. FREEDMAN

From the Medical Research Council Cancer Trials Office, Cambridge

SUMMARY This paper presents a non-parametric method for testing departures from a uniform seasonal variation in incidence of disease. The method may be used either with exact dates of incidence when they are available or with monthly totals. It is equally valid for sinusoidal and non-sinusoidal departures from uniformity. A simulation study shows it to be much more powerful than other non-parametric alternatives and nearly as powerful as Edwards's test in detecting sinusoidal departures.

There has been considerable interest recently in the study of seasonal patterns of incidence of, or mortality from, a wide variety of diseases. These include, to name a few, Burkitt's lymphoma (Williams *et al.*, 1974), peptic ulcer perforation (MacKay, 1966), ulcerative colitis (Evans and Acheson, 1965; Cave and Freedman, 1975), suicides (Barraclough and White, 1978), and congenital malformations (Elwood, 1975). Different statistical methods have been proposed and used to analyse such data. This paper briefly reviews the methods currently available and introduces a new method which may be preferable in certain circumstances.

SUMMARY AND DATA PRESENTATION

The data are usually presented and analysed in the form of a series of 12 monthly totals, these being the numbers of persons diagnosed with or dying from the disease of interest in a given month of the year. Often the data have been gathered over a number of years and the monthly totals are obtained by summing over the individual years.

It should be noted that this way of compressing the data can be used only in the absence of a long-term trend in the data. For example, if there is a long-term decline in the incidence of a disease and no seasonal variation, then combining monthly totals in the way described will result in an apparently steady decrease in the numbers from January to December and perhaps the erroneous conclusion that some type of seasonal variation is present. In fact, data with long-term trends must be quite common but the more sophisticated regression methods developed for application to economics (for example, Thomas and Wallis, 1971) which can be used to cope with this appear to have been rarely used in medicine. Pocock (1974) describes a method which can separate 'seasonal' from 'non-seasonal' variation and in this analysis long-term trends in the data would contribute only to the 'non-seasonal' variation. The methods discussed in the rest of this paper assume that there is no long-term trend in the data. The simple modification of applying these methods to the residuals after fitting a trend to the data cannot be used because the sampling distributions of the resultant statistics would in each case be changed and are as yet unknown. The one exception to this rule is the test of Hewitt *et al.* (1971) in which the sampling distributions of the statistics *would* remain unchanged.

There appear to be three reasons for the custom of grouping the data into months:

(i) Often the dates of the relevant events are not known more accurately, especially when data are drawn from official publications.

(ii) The monthly totals offer a convenient summary of the data, and are especially suitable for plotting in a histogram to demonstrate absence or presence of a seasonal trend.

(iii) For large amounts of data it is tiresome to have to work with exact dates of events, rather than months.

However, for small studies (less than 50 subjects, say) the exact dates of diagnosis or death are often available, and, as will be explained later, there are ways of using this more detailed information to advantage. In addition, the data can be depicted graphically by plotting not the histogram but the cumulative relative frequency. Such a graph is shown in Fig. 1 for some hypothetical data on the dates of diagnosis of lung cancer in a chest hospital, listed in Table 1. Large departures from the bold straight line (which represents the cumulative relative frequency

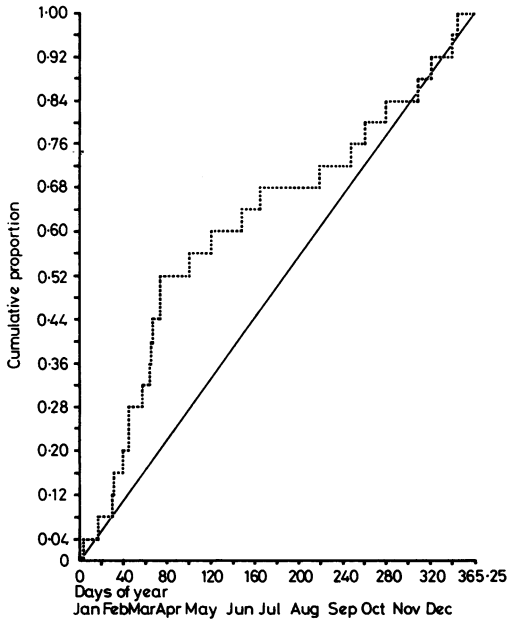

Fig. 1 *Cumulative distribution of sample values listed in Table 1 and that expected under hypothesis of no seasonal distribution.*

----- Sample cumulative distribution $F_N(t)$

——— Expected cumulative distribution $F(t)$

Table 1 *Hypothetical data: dates of diagnosis of 25 patients with lung cancer in a chest hospital*

Dates	*No. of days from beginning of year*
1. Jan 2	2
2. Jan 16	16
3. Jan 30	30
4. Jan 31	31
5. Feb 9	40
6. Feb 14	45
7. Feb 14	45
8. Feb 27	58
9. Mar 6	65·25
10. Mar 7	66·25
11. Mar 8	67·25
12. Mar 15	74·25
13. Mar 15	74·25
14. Apr 12	102·25
15. May 2	122·25
16. May 31	151·25
17. Jun 16	167·25
18. Aug 10	222·25
19. Sep 8	251·25
20. Sep 20	263·25
21. Oct 12	285·25
22. Nov 10	314·25
23. Nov 24	328·25
24. Dec 11	345·25
25. Dec 15	349·25

expected if there were no seasonal variation) indicate departures from a uniform seasonal distribution. An even clearer way of displaying the data is by plotting not the cumulative relative frequency itself but the difference between the cumulative relative frequency and the bold straight line. This is shown in Fig. 2. We can now look for departures from the horizontal axis to assess the extent of seasonal variation. In addition, upward slopes correspond to periods of peak incidence and downward slopes to periods of lowest incidence during the year.

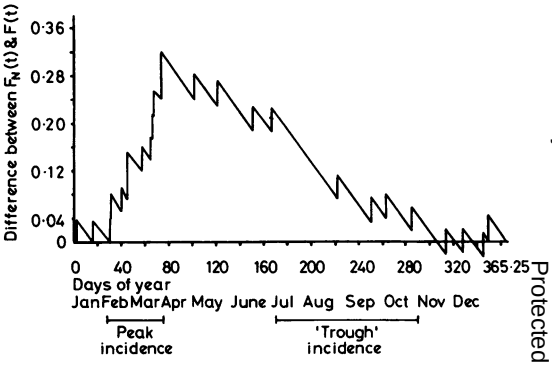

Fig. 2 *Difference between sample and expected cumulative distributions.*

EXISTING METHODS OF ANALYSIS

We assume that there are no 'long-term' trends in the data and that the monthly totals may be combined over the individual years. The statistical problem which has been the main subject of interest has been the development of a significance test to compare the hypothesis of (i) a seasonal fluctuation of frequency of incidence over the year with (ii) a uniform distribution throughout the year.

(i) The usual χ^2 test for heterogeneity is not particularly suitable because it is a general test of *any* departure from a uniform distribution and not a specific test of reasonably smooth upward and downward trends over a period of a year.

(ii) Edwards (1961) proposed a method which has been widely used. This tests the hypothesis (ii) against a particular version of hypothesis (i) namely that the frequencies follow a sinusoidal curve of period 12 months. Such a curve has just one peak and one trough during the year. Walter and Elwood (1975) noted that Edwards's test assumed that the length of each calendar month and the size of population-at-risk in each month were equal, and modified the method to correct for this.

(iii) Pocock (1974) has proposed a method which generalises the approach of Edwards in that it considers a combination of sinusoidal curves of different periodicities. Although the method is described in more general terms than the 12-month model, it can be used to analyse the type of seasonal data under discussion and has been employed in this way by Barraclough and White (1978a; 1978b). The method has the same drawback as Edwards's in that it does not allow for the different lengths of the months. Barraclough and White attempted to overcome this by working with frequencies which were themselves adjusted for the length of each month.
Cave and Freedman (1975) employed an idea similar to that of Pocock's in less general form by modifying Edwards's test to test the hypothesis that the frequencies follow a sinusoidal curve of period 6 months instead of 12 months.

(iv) St. Leger (1976) proposes the use of a refinement of Edwards's method using a maximum likelihood technique. This requires the use of a specially written computer programme.

(v) Hewitt *et al.* (1971) describe a non-parametric test using a procedure of ranking the months in the order of their corresponding frequencies. This avoids the problem of specifying a particular algebraic version of hypothesis (i). Unfortunately the test lacks the power of the parametric methods for moderate sample sizes (Walter and Elwood, 1975).

It would appear from the above description that if a non-parametric method of greater power could be devised it might in some cases be of use in the hypothesis-testing situation. The following sections describe such a method.

THE KOLMOGOROV-SMIRNOV TYPE STATISTIC V_N

Suppose we have N subjects in our sample and that the time of occurrence of the event in question (diagnosis or death) is, for the i^{th} patient, t_i ($i = 1, \ldots, N$) where t_i is the time from the beginning of the year. In practice t_i will very rarely be pinpointed to the exact hour, minute, and second, but in many circumstances it would be possible to measure t_i in days. We assume the study covers a large number of years and consider an 'average' year consisting of 365¼ days with the month of February having 28¼ days. (If an event occurs on 29 February (in a leap year), t will be equal to 59¼; if on 1 March of any year, t will be equal to 60¼, etc.). If the study covers no leap years, then the quarter day for 29 February may be dropped. If the study covers only one year which is a leap year, then 29 February is counted as one whole day. The results presented in this paper will be only marginally affected by such modifications.

Let F(t) be the cumulative distribution function on the assumption that there is a uniform seasonal distribution. Then

$$F(t) = P(t_i \leqslant t) = t/365\tfrac{1}{4}$$

Let $F_N(t)$ be the sample cumulative distribution function. Then

$$F_N(t) = j/N$$

where j is the number of patients with $t_i \leqslant t$.

The one-sided Kolmogorov-Smirnov statistic (Lindgren, 1962) D_N^+ is defined by

$$D_N^+ = \max_{0 \leqslant t \leqslant 365\frac{1}{4}} (F_N(t) - F(t))$$

Similarly D_N^- is defined as

$$D_N^- = | \min_{0 \leqslant t \leqslant 365\frac{1}{4}} (F_N(t) - F(t)) |$$

where $| \times |$ denotes the absolute value of x.

If we write $V_N = D_N^+ + D_N^-$ we have a statistic, first suggested by Kuiper (1962), which is suitable for seasonal data. The important property of V_N which is not shared by D_N^+ or D_N^- is that it is independent of the origin from which t is measured. In other words, the value of V_N would be the same whether we took 12.00 midnight 31 December as our starting point or 6.45 p.m. on 3 March or any other time for that matter. This is a property which is clearly needed for our problem.

The distributional form of V_N under the null hypothesis has been extensively investigated by Stephens (1970) and tables of percentiles are readily available (Stephens, 1970; Pearson and Hartley, 1972). However, to use these tables one needs to use ungrouped data. Once the values of t_i have been grouped into months, the method is no longer valid.

REDEFINITION OF THE STATISTIC V_N FOR GROUPED DATA AND ITS DISTRIBUTION

When the data are grouped into months, the cumulative distribution function F becomes a step function with 12 steps. The value of F at the end of January is 31/365¼: at the end of February, the value is 59¼/365¼, etc. Denote these values by F(1), F(2), etc. Similarly, the sample cumulative distribution function is also a step function with value n_1/N at the end of January, where n_1 is the number of events in January, $(n_1 + n_2)/N$ at the end of February, where n_2 is the number of events in February etc. Denote these values by $F_N(1)$, $F_N(2)$, etc.

Then we redefine V_N as follows:

$$V_N = \max_{1 \leqslant t \leqslant 12} (F_N(t) - F(t)) + | \min_{1 \leqslant t \leqslant 12} (F_N(t) - F(t)) |$$

V_N as defined could be thought of as a Kolmogorov-Smirnov type statistic for a discrete distribution. Unlike the continuous case, the

distribution of V_N under the null hypothesis is not independent of the discrete distribution specified by the null hypothesis (Conover, 1972). Moreover no generally applicable methods are available of deriving theoretically the distribution of V_N under the null hypothesis. It was therefore necessary to perform some Monte Carlo simulations to determine this distribution.

In order to determine the asymptotic distribution of V_N an approach similar to that of Wood and Altavela (1978) was used. Each simulation required 11 calls of a normal pseudo-random number generator using the Marsaglia Polar (1964) method which was available in the CAMLIB Subroutine Library (University of Cambridge Computing Service, 1976) on the Cambridge University IBM 370/165. Ten thousand simulations were performed. Estimated percentiles of the distribution of $V_N\sqrt{N}$ are shown in Table 2. Ninety per cent confidence limits for the percentiles are also given in Table 2.

Table 2 *Estimated percentiles of the asymptotic distribution of* $V_N\sqrt{N}$

Percentile	*Estimate*	*90% confidence limits*
10%	0·58	0·576—0·589
20%	0·67	0·670—0·680
30%	0·75	0·744—0·754
40%	0·82	0·811—0·822
50%	0·89	0·882—0·893
60%	0·96	0·950—0·961
70%	1·03	1·028—1·040
80%	1·14	1·128—1·145
85%	1·21	1·199—1·217
90%	1·29	1·280—1·297
95%	1·41	1·400—1·422
99%	1·66	1·641—1·683

Although the percentiles have not been estimated to a very high degree of precision, the estimates should be adequate for most purposes. Two further sets of 10 000 simulations were carried out to examine whether the asymptotic distribution of $V_N\sqrt{N}$ could be applied to reasonably small finite samples. The distribution of $V_N\sqrt{N}$ was investigated for sample sizes N = 50 and N = 25. The estimated percentiles of the distributions are shown in Table 3. It can be seen from graphs of the frequency distributions (not shown here) that the distribution becomes less smooth for small samples. Nevertheless the percentiles for N = 50 agree quite well with the asymptotic distribution. It would therefore seem quite safe to refer to the percentiles of the asymptotic distribution of $V_N\sqrt{N}$ for sample sizes greater than 50. For smaller samples it may often be possible to obtain more exact dates of the events and use the continuous method described above.

Table 3 *Estimated percentiles of the distribution of* $V_N\sqrt{N}$ *for sample sizes N = 50 and N = 25*

Percentile	*N = 50*	*N = 25*
10%	0·59	0·57
20%	0·67	0·68
30%	0·76	0·74
40%	0·81	0·82
50%	0·89	0·90
60%	0·95	0·94
70%	1·04	1·06
80%	1·14	1·13
85%	1·20	1·17
90%	1·29	1·31
95%	1·42	1·38
99%	1·67	1·70

EXAMPLES

(i) *Ungrouped data*

We illustrate the method with the data shown in Table 1. From the plot in Figure 2, it can be seen that the maximum displacement of $F_N(t) - F(t)$ above the line is 0·317, and below the line it is 0·025. The sum of these values, 0·342, is the magnitude of the statistic V_N. Following Stephens (1970), we calculate $V_N(\sqrt{N}+0·155+0·24/\sqrt{N})$ to be 1·78, and compare with the percentiles of this statistic shown in Table 5, reproduced from Stephens (1970). It is found that the probability level P lies between 5% and 2·5% ($0·025<P<0·05$).

Table 5 *Percentiles of the distribution of* $V_N(\sqrt{N} + 0·155 + 0·24/\sqrt{N})$

Per cent Percentile	85% 1·537	90% 1·620	95% 1·747	97·5% 1·862	99% 2·001

From Stephens (1970).

(ii) *Grouped data*

The method is illustrated using some data on the incidence of Burkitt's lymphoma gathered in the West Nile district of Uganda (Williams *et al.*, 1974). The years 1966–73 only are considered because during those years there appeared to be no long-term trend in incidence, either upwards or downwards. The frequencies and cumulative frequencies are shown in the second and third columns of Table 4. The fourth and fifth columns show the values of the functions F_N and F and the final column shows the difference between them. The calculation of $V_N\sqrt{N}$ is shown at the foot of the table. The value of 1·50 lies between the 95% and 99% points of the asymptotic distribution as shown in Table 1. Thus the deviation from a uniform seasonal incidence is significant at the 5% but not the 1% level ($0·01<P<0·05$).

Table 4 *Calculations required to calculate V_N for grouped data**

Month	*Frequency*	*Cumulative frequency*	F_N	F	F_N-F
Jan	11	11	0·0827	0·0849	−0·0022
Feb	6	17	0·1278	0·1622	−0·0344
Mar	9	26	0·1955	0·2471	−0·0516
Apr	8	34	0·2556	0·3292	−0·0736
May	8	42	0·3158	0·4141	−0·0983
Jun	7	49	0·3684	0·4962	−0·1278
Jul	11	60	0·4511	0·5811	−0·1300
Aug	19	79	0·5940	0·6660	−0·0720
Sep	12	91	0·6842	0·7481	−0·0639
Oct	16	107	0·8045	0·8330	−0·0285
Nov	6	113	0·8496	0·9151	−0·0655
Dec	20	133	1·0000	1·0000	0
TOTAL	133				

$\text{Max}(F_N-F) = 0|\text{min}(F_N-F)| = 0{\cdot}1300$

$V_N = 0 + 0{\cdot}1300 = 0{\cdot}1300$

$V_N\sqrt{N} = 0{\cdot}1300 \times \sqrt{133} = 1{\cdot}50$

$0{\cdot}01<P<0{\cdot}05$

* Data from Williams *et al.* (1974).

POWER OF THE TEST

In order to compare the power of the proposed test with that of some of the major alternatives, a small simulation study was carried out.

Data were generated to simulate a study of 500 observations where the underlying seasonal distribution conformed to a sinusoidal pattern with amplitude 0·25 (that is, the proportion of cases incident in month i was given by m_i $(1 + 0{\cdot}25 \sin(\pi i/6))/12$ where m_i is the correction factor for the number of days in month i). One thousand such simulated studies were generated. Four significance tests were compared and their power was estimated at the 10%, 5%, and 1% levels. In the case of Hewitt's test these levels could not be achieved exactly, and the corresponding levels used were approximately 12·6%, 4·6%, and 1·3%. The results are shown in Table 6.

Table 6 *Results of simulations to estimate the power of four significance tests against a sinusoidal alternative (1000 simulations)*

Test	*Significance level*		
	10%	*5%*	*1%*
χ^2	83·1%	75·7%	52·3%
Kuiper's V_N	96·3%	92·7%	79·8%
Edwards	97·8%	95·0%	86·6%
Hewitt*	87·2%	73·2%	54·6%

* The significance levels for this test are not exact: see text.

It can be seen from the Table that the Kuiper test is not much less powerful than Edwards's test against this alternative, while Hewitt's test and the χ^2 test are both much inferior.

Discussion

This paper presents a non-parametric test for departures of seasonal data from a uniform pattern of seasonal incidence. Another non-parametric test which has been proposed for this purpose (Hewitt *et al.*, 1971) was reported to have low power (Walter and Elwood, 1975) for moderate sample sizes and the χ^2 test suffers from a similar defect. The power simulations presented in this paper show that the test proposed is more powerful than Hewitt's test and the χ^2 test against a sinusoidal alternative where the number of observations in the sample is 500. In this situation the test is almost as efficient as Edwards's test. The superiority of a Kolmogorov-Smirnov test over the χ^2 test has been reported in a similar situation (Horn, 1977).

The proposed test should therefore be of use when there is no particular reason to expect a specific parametric alternative to the null hypothesis (such as a sinusoidal curve of period 12 months) to hold true.

The test can be used either for exact data, where the dates of incidence or death are known to the day, or for grouped data, where the dates are grouped into months of the year. In the former case, the test

statistic corresponds exactly to Kuiper's Statistic, V_N, one of the family of Kolmogorov-Smirnov statistics. In the latter case, the statistic V_N has had to be redefined for a discrete distribution and the asymptotic sampling distribution under the particular null hypothesis of a uniform seasonal pattern has been estimated by Monte Carlo methods.

The link between this test and the Kolmogorov-Smirnov statistics suggests that another non-parametric method based on the Watson (1961) statistic, U_N^2, may also be of use. Experience with these statistics in the 'continuous' case suggests that U_N^2 may be more efficient than V_N in detecting a slight but steady departure from the hypothesis (Pearson and Hartley, 1972). Thus further work on this topic may lead to a test of increased power. In the meantime, the test proposed in this paper should be of use when there is no particular reason to expect a specific parametric alternative (such as a sine curve of period 12 months) to the null hypothesis to hold true.

I thank Peter Fayers, Sue White, and Dr. A. S. St. Leger for their help, and also Bethan Chapman and Jane Donaldson.

Reprints from L. S. Freedman, MRC Cancer Trials Office, The Medical School, Hills Road, Cambridge CB2 2QH.

References

Barraclough, B. M., and White, S. J. (1978a). Monthly variation of suicide and undetermined death compared. *British Journal of Psychiatry,* **132,** 275–278.

Barraclough, B. M., and White, S. J. (1978b). Monthly variation of suicidal, accidental and undetermined poisoning deaths. *British Journal of Psychiatry,* **132,** 279–282.

Cave, D. R., and Freedman, L. S. (1975). Seasonal variation in the clinical presentation of Crohn's disease and ulcerative colitis. *International Journal of Epidemiology,* **4,** 317–320.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association,* **67,** 591–596.

Edwards, J. H. (1961). The recognition and estimation of cyclic trends. *Annals of Human Genetics,* **25,** 83–86.

Elwood, J. M. (1975). Seasonal variation in anencephalus in Canada. *British Journal of Preventive and Social Medicine,* **29,** 22–26.

Evans, J. G., and Acheson, E. D. (1965). An epidemiological study of ulcerative colitis and regional enteritis in the Oxford area. *Gut,* **6,** 311–324.

Hewitt, D., Milner, J., Csina, A., and Pateual, A. (1971). On Edwards's criterion of seasonality and a non-parametric alternative. *British Journal of Preventive and Social Medicine,* **25,** 174–176.

Horn, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics,* **33,** 237–248.

Kuiper, N. H. (1962). Tests concerning random points on a circle. *Proceedings. Koninklijke Nederlandse Akademie van Wetenschappen,* Series A, **63,** 38–47.

Lindgren, B. W. (1962). *Statistical Theory,* first edition, p. 303. Macmillan: New York.

Mackay, C. (1966). Perforated peptic ulcer in the West of Scotland: a survey of 5343 cases during 1954–63. *British Medical Journal,* **1,** 701–705.

Marsaglia, F., and Bray, T. A. (1964). A convenient method for generating normal variables. *SIAM Reviews,* **6,** 260–264.

Pearson, E. S., and Hartley, H. O. (1972). *Biometrika Tables for Statisticians,* volume 2, pp. 117–119 and Table 54. Biometrika Trust: London.

Pocock, S. J. (1974). Harmonic analysis applied to seasonal variations in sickness absence. *Applied Statistics,* **23,** 103–120.

St. Leger, A. S. (1976). Comparison of two tests for seasonality in epidemiological data. *Applied Statistics,* **25,** 280–286.

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramer-Von Mieses and related statistics without extensive tables. *Journal of the Royal Statistical Society,* Series B, **32,** 115–122.

Thomas, J. J., and Wallis, K. F. (1971). Seasonal variation in regression analysis. *Journal of the Royal Statistical Society,* Series A, **134,** 57–72.

University of Cambridge Computing Service (1976). *Camlib subroutine library specifications,* third edition, pp. 76 and 79. University of Cambridge Computer Laboratory: Cambridge.

Walter, S. D., and Elwood, J. M. (1975). A test for seasonality of events with a variable population at risk. *British Journal of Preventive and Social Medicine,* **29,** 18–21.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. I. *Biometrika,* **48,** 109–114.

Williams, E. H., Day, N. E., and Geser, A. G. (1974). Seasonal variation in the onset of Burkitt's Lymphoma in the West Nile district of Uganda. *Lancet,* **2,** 19–22.

Wood, C. L. and Altavela, M. M. (1978). Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions. *Biometrika,* **65,** 235-239.