



A Double-Scan Statistic for Clusters of Two Types of Events

Joseph I. Naus; Dan Wartenberg

Journal of the American Statistical Association, Vol. 92, No. 439 (Sep., 1997),
1105-1113.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28199709%2992%3A439%3C1105%3AADSFCO%3E2.0.CO%3B2-4>

Journal of the American Statistical Association is currently published by American Statistical Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

A Double-Scan Statistic for Clusters of Two Types of Events

Joseph I. NAUS and Dan WARTENBERG

We develop a scan-type statistic to measure the unusualness of the clustering of two types of events over time. The statistic allows for a lagged effect between the two types of events. We derive the expected number of nonoverlapping clumps of clusters under retrospective and prospective chance models of no association. Results are derived and approaches are given to handle both uniform and more general distributions of events over time. We investigate the power of the statistic against an alternative where the observed data is a mixture of linked and unassociated clusters. The statistic is applied to data on homicide/suicide clusters over a 7-year period for several counties and several sex/race combinations.

KEY WORDS: Homicide; Lagged effects; Suicide; Temporal association.

1. INTRODUCTION

Researchers studying series of multiple outcomes sometimes observe different types of events clustering together in time. The researchers ask whether the observed clustering is unlikely to have arisen due to chance. Cressie (1980), Glaz and Naus (1983, 1991), Krauth (1992), Loader (1991), Naus (1982, 1988), Wallenstein, Naus, and Glaz (1993), and Wallenstein, Weinberg, and Gould (1989) have described properties of a test that scans for unusually large clusters or multiple large clusters of events over time. But these tests do not include consideration of different types of events within the cluster. Page (1965), in a quality control setting, considered two types of events involving k out of n points falling outside warning lines or a point falling outside action lines. For very simple cases, he derived expected waiting times until at least one of the events occur. He described a general method that can be used to find expected waiting times for two types of events. This method involves generating all combinations of states of the related discrete Markov chain. Page noted that he did not use that method, and that for complex events such as k out of n points, the number of states makes the method complex. Huntington (1976) detailed this approach to finding the expected waiting time until occurrence of k events in n days, allowing constraints on the number of types of events. In this article we focus on the expected number of nonoverlapping clusters of two types of events, and derive this quantity directly. We derive the expectation and approximate distribution of the number of clusters for this test statistic for retrospective and prospective chance models, both for uniform and general distribution of the two types of events over time.

Our interest grew out of studying different causes of death, or similar cases among different races and genders. One example is homicide/suicide clusters. In Kentucky three cases of homicide each followed by a suicide occurred during a 3-month period in 1990 and received wide publicity (MMWR 1991, pp. 652–659). The resulting concern led to the formation of a state homicide/suicide task force. In

1990 in Quebec, a cluster of homicides followed by suicides were reported by the news media (Buteau, Lesage, and Kiely 1993). Another example is male suicide/female suicide clusters. Gould, Wallenstein, and Kleinman (1990) and Greenberg, Naus, Schneider, and Wartenberg (1991) and others have investigated the separate clustering of these types of suicide. We wish to study the joint or lagged clustering of these two types of suicide, and to investigate whether other similar such couplings occur in other data. Although many of the statistical scan clustering methods deal with the bunching of one type of event over time and space, one can readily apply occupancy theory to the clustering of two types of events on the same day. In other cases there may be a lagged relation; it is for these cases that we develop a two-event-type scan statistic.

Our data consist of number of deaths by cause of death (homicide, suicide) for each of several counties for each day over a 7-year period, broken down by gender/race combination. One approach in studying clusters is to look at the number of days in which two different types of death occurred (e.g., homicide, suicide) or in which the same type of death occurred between two types of individuals (e.g., suicide of males and of females). In looking at such links in time, we need not restrict ourselves to cases where the two events occur on the same day. If we anticipated a delayed or lagged effect, then we would want to look at cases where two types of events might occur within the same d -day interval. For certain choices of d , looking at overlapping d -day periods would average out certain periodic effects. For example, Lester (1979) noted that suicides are more likely to occur on certain days of the week. Taking $d = 7$ eliminates the effect of this periodicity on the scan-type statistic.

For certain types of events, past experience, the epidemiological model, data measurement, and the nature of the coincidences screened for suggest the choice of d . In screening our data for homicide/suicides within d days, we view the observed clusters as a mixture of unrelated clusters (close in time by chance), and related homicide/suicides. Without checking sources outside the data base, we do not know

Joseph I. Naus is Professor, Department of Statistics, Rutgers University, Hill Center, Busch Campus, Piscataway, NJ 08855. Dan Wartenberg is Associate Professor, Environmental and Occupational Health Sciences Institute, UMDNJ/Robert Wood Johnson Medical School, Busch Campus, Piscataway, NJ 08855.

Table 1. Cook County WFS/BFS Within 7 Days, By Day

BFS	39	107	<u>223</u>	<u>250</u>	374	<u>404</u>	<u>495</u>	554	<u>563</u>
WFS			<u>229</u>	<u>245</u>		<u>403</u>	<u>490</u>		<u>563</u> <u>568</u>
BFS	779	810	834	850	920	<u>992</u>	1,051	1,252	<u>1,271</u>
WFS						<u>988</u> <u>996</u>			<u>1,267</u> <u>1,272</u>
BFS	1,408	1,448	1,470	1,581	1,592	1,602	1,606	<u>1,871</u>	2,062 2,346
WFS								<u>1,866</u> <u>1,877</u>	

NOTE: Underlined items are clusters.

which clusters are related. The natural choice for d is based on the time interval in which the related homicide/suicides are anticipated to occur. Here we are primarily interested in screening for a type of homicide/suicide cluster in which a person murders someone and then kills himself or herself shortly thereafter. We anticipate that the recorded dates of death of the homicide and suicide are within ($d =$) 2 or 3 days.

In Section 2 we define the double scan statistic and in Section 3 apply it to a large dataset. In Section 4 we derive its expectation and its approximate or exact variance for simple retrospective and prospective (null) models. The simple models assume that each of the two types of events occurs independently of each other and completely at random (with even probabilities) over time. In certain applications, the two types of events may have varying probabilities over time. In Section 6 we discuss general formulas for this case and we derive these formulas in the Appendix. These formulas allow handling trends and other cyclic variations as special cases. We also detail alternative simpler procedures for handling trends in Section 6. These results are useful in further analyses on clusters found significant under the simple null models. In Section 7 we study the power of the double-scan screening procedure and note its usefulness against a type of mixture alternative.

The simple null model is useful for initial screening even in cases in which trends or cycles exist that cause the two types of events to vary together over time. For example, if homicides and suicides both vary proportionately with an increasing population size, then the expected number of homicide/suicide clusters will be greater than under the simple null model. Thus if the observed number of clusters is not significant relative to the simple model, then they would not be significant relative to common trends or cycles that result from variations in population size. An important use of the scan statistic's simple null distribution is to warn us when not to jump to conclusions of association.

Our sensitivity analysis in Section 6 shows that for modestly increasing trends in the number of two types of events over time, the expected total number of clusters does not increase very much over the simple null model. For cases in which the two types of events have sharp trends or more complex patterns of variability over time, the methods of Section 6 can be applied.

2. THE DOUBLE-SCAN STATISTIC

Given that one or more of both types of events occur within a d -day period, we say that a "two-type d -day cluster"

has occurred. Over a long period of D days, a scientist may observe several such clusters. The scientist seeks to determine whether the observed number of clusters is unusually greater than what would be expected under certain chance models.

For the case $d = 1$, the number of two-type 1-day clusters, N_1 , can be counted simply as the number of days in the D -day period that contain at least one or more of both of the two types of events. For the case $d > 1$, there are many alternative ways to count the number of two-type d -day clusters. We focus on a method that avoids multiple counting of the same or too-closely overlapping clusters. The advantage of this approach is that when the events are relatively rare, the number of "declustered" clusters counted in this way will be approximately Poisson distributed. (For a more detailed discussion of Poisson approximation and the declumping approach, see Aldous 1989, Arratia, Goldstein, and Gordon 1990, and Barbour, Holst, and Janson 1992.)

For $d > 1$, define the event E_i to have occurred if anywhere within the d consecutive days $i, i + 1, \dots, i + d - 1$ there are one or more of both types of events. The event E_i indicates the occurrence of a two-type d -day cluster. Our method counts the number of times that an E_i occurs with no previously overlapping E_j 's. Let $Z_i = 1$ if E_i occurs and none of $E_{i-1}, E_{i-2}, \dots, E_{i-d+1}$ occurs and $Z_i = 0$ otherwise. Let $N_d = \sum_{1 \leq i \leq D-d+1} Z_i$. We call N_d , the total number of nonoverlapping two-type d -day clusters, the *double-scan statistic*.

Section 4 gives the expectation of N_d for a retrospective and for a prospective model, along with simple approximations for the expectation and variance. Section 5 defines a directional scan statistic. The next section compares the expectation of the double scan statistic with its observed values for the 7-year period by county, sex, race, and homicide/suicide data.

3. COMPARING THE NULL MODEL TO THE DATA

Our data consist of several causes of death (suicide, homicide, accidents) among 15- to 24-year-old white and black Americans for each day for the 7-year period from January 1, 1979–December 31, 1985. The data are broken down by day, type of death, race, gender, and county. Greenberg et al. (1991) looked at clustering of homicide for this dataset for each of 22 selected counties. The counties selected were those with the largest number of black residents: Cook (Illinois), Los Angeles (California), Wayne (Michigan), Kings (New York), and Philadelphia (Pennsylvania).

We first illustrate the double-scan statistic for the two events suicide among white females (WFS) and suicide

Table 2. Female Homicide (FH) and Male Suicide (MS) Within d -Day Scans, for $d = 2, 3$ by Race, White (W) and Black (B), for 7-year Period, With $D = 2,557$ and Observed Versus Expectation of Number, N_d , of Clusters

County	A/B	$d = 3$		$d = 2$	
		Obs.	Exp.	Obs.	Exp.
WFH/WMS					
Cook	117, 407	59	60.4	46	45.0
Los Angeles	229, 688	140	138.8	121	124.1
Wayne	58, 212	18	19.4	17	12.9
Kings	56, 63	6	6.3	3	4.0
Philadelphia	19, 135	5	4.4	3	2.8
BFH/BMS					
Cook	191, 122	35	35.3	22	24.1
Los Angeles	182, 110	35	30.9	21	20.9
Wayne	162, 113	28	28.6	22	19.2
Kings	81, 38	4	5.5	1	3.4
Philadelphia	79, 61	13	8.4	11*	5.4

*Significant at .05 level.

among black females (BFS) for Cook County, Illinois. During the 7-year period (= 2,557 days) there were 99 days with WFS and 28 days with BFS. Table 1 presents the 28 days on which BFS occurred, and any day within 7 days of those BFS days in which a WFS occurred. (Day 1 is January 1, 1979.)

There is only one day (day 563) that contains at least one WFS and one BFS. The expected number of days that contain at least one WFS and one BFS is $(28)99/2,557 = 1.08$. For a scanning window $d = 2$, in three cases one each of the two types of events WFS and BFS fell within 2 days: (404, 403), (563, 563), and (1,271, 1,272). The expected number of clusters of the two types within 2 days is 3.1. For $d = 3$ and $d = 4$, these are also the only three cases. The expected number of clusters of the two types within 3 days is 4.9. For a scanning window with $d = 7$, $N_7 = 8$. This corresponds to the eight clusters underlined in the foregoing data and compares to an expected number of 10.5 clusters.

Table 2 gives the observed and expected number of two-type clusters in d -days for $d = 2, 3$ for the events female homicide (FH) and male suicide (MS). The column A/B tells us how many type I (FH) days there are (A), followed by the number of type II (MS) days there are (B). The table gives separate breakdowns for blacks and whites.

In all of the comparisons in Table 2, the observations are close to expectation, with only the 11 observed cases of black female homicide (BFH)/black male suicide (BMS) in Philadelphia, being somewhat high. The type I (BFH) and type II (BMS) cases for Philadelphia within 3 days are listed in Table 3.

Section 6 studies the effect of variation over time of homicides and suicides on the expectation of the double-

scan statistic, and illustrates its effect on our analysis of the Philadelphia data. Section 5 explores the Philadelphia BFH/BMS data further by means of a directional double-scan statistic.

Table 4 compares observed and expected values for suicide/suicide for all six combinations of gender-race combinations, for 2-day scans. The observed and expected values are remarkably close.

4. THE EXPECTATION AND VARIANCE OF THE NUMBER OF TWO-TYPE d -DAY CLUSTERS

4.1 The Retrospective Chance Model

Assume that there are exactly A of the D days in which a type I event occurs, and that there are exactly B of the D days in which a type II event occurs. Consider the chance model where all $\binom{D}{A} = D!/A!(D-A)!$ ways of picking the A type I days, and all $\binom{D}{B}$ ways of picking the B type II days are equally likely, and the occurrence of the two types of days are independent.

For this chance model and $d = 1$, $E(N_1) = AB/D$ and

$$P(N_1 = k) = \binom{D}{k} \binom{D-k}{A-k} \binom{D-A}{B-k} \div \binom{D}{A} \binom{D}{B}.$$

Theorem 1 derives an exact formula for $E(N_d)$, for $d > 1$.

Theorem 1. For $1 \leq A, B \leq D$, let

$$P_0(r, s) = \binom{D-r}{A-s} / \binom{D}{A},$$

$$P_1(r, s) = \binom{D-r}{B-s} / \binom{D}{B},$$

and

$$Q_i(r, s) = 1 - P_i(r, s), \quad \text{for } i = 0, 1.$$

Then

$$E(N_d) = \sum_{i=1}^{d-1} P(Z_i = 1) + (D - 2d + 2)P(Z_d = 1), \quad (1)$$

where

$$P(Z_1 = 1) = \prod_{i=0}^1 Q_i(d, 0), \quad (2)$$

Table 3. Philadelphia BFH/BMS Within 3 Days By Day

BFH	210	211	221	293	783	786	906	936	951
BMS	211		223	292	785	788	908	936	952
BFH	976		1,058	1,399	1,627	1,629	1,812	1,814	2,260
BMS	976		1,059	1,399	1,627		1,811	1,814	2,258

Table 4. Suicide by Suicide Within a 2-Day Period by Gender-Race Combinations for the Five Counties

	Cook		Los Angeles		Wayne		Kings		Philadelphia	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
WM/WF	46	38.4	111	105.6	13	12.7	1	1.4	3	4.9
WM/BM	46	46.8	65	63.4	26	24.6	5	2.7	7	8.9
WM/BF	10	11.2	27	26.8	4	6.1	1	.6	3	2.7
WF/BM	15	13.0	19	21.8	4	7.1	1	.9	2	2.3
WF/BF	3	3.1	6	9.2	3	1.7	0	.2	0	.7
BF/BM	4	3.8	7	5.5	4	3.4	0	.4	2	1.2

$$P(Z_2 = 1) = \left[\sum_{i=0}^1 P_i(d+1, 1)Q_{1-i}(d, 0) \right] - \prod_{i=0}^1 P_i(d+1, 1), \quad (3)$$

and, for $3 \leq k \leq d$,

$$P(Z_k = 1) = \left[\sum_{i=0}^1 P_i(d+k-1, 1)Q_{1-i}(d, 0) \right] - \prod_{i=0}^1 P_i(d+k-1, 1) + \sum_{j=0}^{k-3} \sum_{i=0}^1 P_i(d+k-1-j, 2) \times [P_{1-i}(d+j, 0) - P_{1-i}(d+k-2, 0)]. \quad (4)$$

Proof. See the Appendix.

4.2 The Prospective Chance Model

The retrospective chance model is conditioned on A and B , the total number of type I and type II events in the D day period. A prospective model might be more appropriate in the situation where one seeks to predict the number of clusters for the next D day period, assuming that the expected number of days with type I and type II events were A and B . The model would assume that on each day, the chance of a type I event is A/D and the chance of a type II event is B/D . The occurrence of a type I event on a given day is assumed (under the null chance model) to be independent of the occurrence of a type II event on that day, and the occurrences on different days are assumed to be independent. For this prospective chance model, Equations (1)–(4) hold with the modification of $P_i^*(r, s)$ replacing $P_i(r, s)$, where

$$P_0^*(r, s) = (D - A)^{r-s} A^s / D^r$$

and

$$P_1^*(r, s) = (D - B)^{r-s} B^s / D^r. \quad (5)$$

4.3 Simple Approximations for Expectation and Variance of N_d

Abbreviate $P(Z_i = 1)$ to P_i . The exact formula for $E(N_d)$ can be simplified by ignoring end effects and taking

all the P_i to be equal. Then Equation (1) simplifies to

$$E(N_d) \approx (D - d + 1)P_d, \quad (6)$$

where P_d is given by the right side of (4) (or modification (5)) taking $k = d$. In applications where $D \gg d$, this approximation loses little accuracy. This is illustrated at the end of this section.

To develop formula for variance of N_d , note that in terms of the 0–1 indicator variables Z_i ,

$$\text{var}(N_d) = \sum_{1 \leq i \leq D-d+1} \text{var}(Z_i) + 2 \sum_{i < j} \text{cov}(Z_i, Z_j),$$

where

$$\text{var}(Z_i) = E\{(Z_i)^2\} - (E\{Z_i\})^2 = P_i(1 - P_i)$$

and

$$\text{cov}(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j) = P_{ij} - P_i P_j,$$

where

$$P_{ij} = P(Z_i = 1 \cap Z_j = 1).$$

Note that from the definition of Z_i , $P_{ij} = 0$ if $|i - j| < d$. For the prospective model, $P_{ij} = P_i P_j$ for $|i - j| \geq d$, and $2 \sum_{i < j} \text{cov}(Z_i, Z_j)$ reduces to $-b_1$, where

$$b_1 = \sum_{\substack{|i-j| < d \\ i \neq j}} P_i P_j = 2 \sum_{j=1}^{d-1} \sum_{i=j+1}^{d-1-j+d-1} P_i P_j + (d-1)(2D - 5d + 4)P_d^2. \quad (7)$$

For the prospective model, we can compute exactly

$$\text{var}(N_d) = \sum_{i=1}^{d-1} P_i(1 - P_i) + (D - 2d + 2)P_d(1 - P_d) - b_1. \quad (8)$$

For either the retrospective or the prospective model, if we ignore end effects by approximating all P_i by P_d , we find

$$\text{var}(N_d) \approx (D - d + 1)P_d(1 - P_d) + 2 \binom{D - 2d + 2}{2} P_{d,2d+1} - 2 \binom{D - d + 1}{2} P_d^2. \quad (9)$$

For the prospective model, $P_{d,2d+1} = (P_d)^2$. For the retrospective model, $P_{d,2d+1} < P_d^2$, and approximating $P_{d,2d+1}$

Table 5. Female Homicide (FH) on the Same or on the Preceding Day as Male Suicide (MS), by Race, White (W) and Black (B), for a 7-Year Period, with $D = 2,557$, $d = 2$, Directional Double Scan

County	WFH/WMS		BFH/BMS		WFH/BMS		BFH/WMS	
	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
Cook	31	33.1	16	17.0	11	10.6	62	52.8
Los Angeles	91	97.8	16	14.7	18	18.3	91	79.1
Wayne	11	9.1	17	13.5	8	4.9	28	24.7
Kings	2	2.7	1	2.4	1	1.6	4	3.9
Philadelphia	0	1.9	9*	3.7	0	.9	6	8.0
TOTAL	135	144.6	59	57.5	38	36.3	191	168.5

NOTE: Observed versus expectation of number, N_d , of clusters.
*Significant at .05 level.

by P_d^2 for the retrospective model (conservatively) overestimates the variance. With this simplification, approximation (9) reduces to

$$\text{var}(N_d) \approx (D - d + 1)P_d - \{(D - d + 1)(2d - 1) - (d - 1)d\}P_d^2. \quad (10)$$

For the case where D is large, P_d is small, and $D \gg d$, $\text{var}(N_d) \approx E(N_d)$, and the relation of the first two moments agrees with the Poisson model. We follow the results of Arratia et al. (1990, pp. 405–406) on the Chen–Stein error bounds for the goodness of fit of the Poisson approximation. For the prospective model, we use these results to bound the variational difference between the distribution of N_d and the distribution of a Poisson variate \mathcal{V} , where $E(\mathcal{V}) = E(N_d) = \lambda$. For the prospective model, the b_i 's defined by Arratia et al. (1990) become $b_2 = b_3 = 0$, and b_1 is given by Equation (7); from theorem 1 of Arratia et al. (1990), for any set S of values,

$$|P(N_d \in S) - P(\mathcal{V} \in S)| \leq b_1(1 - e^{-\lambda})/\lambda. \quad (11)$$

For example, let $D = 2,557$, $d = 3$, $A = 56$, and $B = 63$. For the prospective model, using the exact formulas, $P_1 = .00463468$, $P_2 = .00247862$, $P_3 = .00245341$, $E(N_3) = 6.271 = \lambda$, and $\text{var}(N_3) = 6.194$. (Using approximations (6) and (10) gives $E(N_3) \approx 6.268$ and $\text{var}(N_3) \approx 6.192$.) Here $b_1 = .0615$, and the bounds (11) imply that for any k , $|P(N_d \geq k) - P(\mathcal{V} \geq k)| \leq .01$. Approximating the distribution of N_d by the Poisson gives reasonable accuracy here.

5. DIRECTIONAL DOUBLE-SCAN COMPARISONS

In Table 2, for BFH and BMS for Philadelphia, the observed double scans for $d = 2, 3$ were higher than expected. The comparisons made in Table 2 did not predict the order in which the two types of events would occur; thus the suicide could be on a preceding day, on the same day, or on a day after the homicide. Both the observed and expected values in Table 2 were calculated on this basis. In the case of a related homicide/suicide, there is a specific order of events in mind, with homicide followed by suicide. The *directional double-scan statistic* incorporates this order and is defined as follows. The event E_i^* is said to have occurred if anywhere within the d consecutive days $i, i + 1, \dots, i + d - 1$, there are at least one type I event and one type II event with

the type I event on the same or earlier day than the type II event. For $d > 1$, the directional double-scan statistic counts the number of nonoverlapping E_i^* 's as before.

We illustrate the application of the directional double-scan statistic for $d = 2$ and a retrospective chance model. For this case, the expected number of clumps of 2-day periods, where a type I event and a type II event occur, with the type I event either being on the previous or the same day as the type II event, is

$$P(Z_1^* = 1) + (D - 2)P(Z_2^* = 1).$$

Here D = the total number of days (in our example, 2,557), and

$$P(Z_1^* = 1) = 2P_0(1, 1)P_1(1, 1) - P_0(2, 2)P_1(2, 2) + P_0(2, 1)P_1(2, 1)$$

and

$$P(Z_2^* = 1) = 2\{P_0(2, 1)P_1(2, 1) - P_0(3, 2)P_1(3, 2)\} + P_0(2, 2)P_1(2, 1) - P_0(3, 3)P_1(3, 2) + P_0(2, 1)P_1(2, 2) - P_0(3, 2)P_1(2, 2).$$

For example, for the case of BFH/BMS in Philadelphia, Table 2 shows that, ignoring direction, there were 11 observed clumps of 2-day clusters, as compared to an expected 5.4 cases. The directional double scan has an expected number of 3.7 instances for the BFH on the previous or same day as the BMS. The sequence of data for BFH/BMS for Philadelphia, following Table 2, shows nine clumps of 2-day clusters where the BFH is on the previous or same day as the BMS. The comparison taking into account the predicted order of the events is more sensitive here. Table 5 compares the observed and expected values for the directional scan statistic for female homicide on preceding or same day ($d = 2$) as male suicide for several race combinations. In almost all cases the observed and expected values are very close. The only exception is the BFH/BMS for Philadelphia.

In the case of homicide/suicide clusters, it was noted at the beginning of this section that there was a particular order in mind. It was assumed that if a person murders another person and also kills himself or herself, that the homicide will be on a preceding or the same day as the suicide. We received a reality check in following up some of

Table 6. Philadelphia BFH BMS Cases and Expected Clusters by Year

Year	Observed BFH cases	Observed BMS cases	Expected BFH/BMS clusters	
			$d = 2$	$d = 3$
1979	13	12	1.20	1.87
1980	11	3	.26	.42
1981	14	15	1.59	2.47
1982	9	3	.21	.35
1983	13	11	1.10	1.73
1984	9	7	.49	.79
1985	10	10	.78	1.24
TOTAL	79	61		

the Philadelphia cases. The Philadelphia Inquirer (Thursday June 16, 1983, Section B, p. 2) under Metro Area News in Brief described how in a 3-day period there were two unrelated incidents where a person shot himself to death after shooting another person. In both cases, the person shot first survived longer than the attacker.

6. HANDLING TRENDS AND OTHER VARIATIONS OVER TIME

Section 4 derives the expectation of N_d , the number of two-type d -day clusters for retrospective and prospective simple chance models. Those chance models assume that each of the two types of events are distributed with constant probability of occurring over the time period. In certain applications, the numbers of one or both of the two types of events may increase or vary cyclically over time. This section gives and illustrates several approaches for handling such variation.

The first approach is to generalize Theorem 1 for the prospective model to the case in which the probabilities of each type of event can vary from day to day. These probabilities can be modeled to account for variations over time due to changes in population size, fitted to trends in data, or to take account of seasonality in the data. Theorem 2 in the Appendix summarizes these results.

The second approach to handling variations over time is to divide the time interval into smaller time periods and carry out the approach of Theorem 1 separately for each time period. We illustrate this approach for the significant Philadelphia BFH/BMS data, for $d = 2$ and 3 for the retrospective model. We divide the 7-year period into 7 calendar years. We add $d - 1$ days to each year (except the last), to allow for overlap possibilities between years. The observed number of BFH and BMS for each of the 7 years, together with the expected number of clusters, is presented in Table 6. For $d = 2$, the sum of the seven yearly expectations is 5.6. This is not materially different from the expectation of 5.4 cases given in Table 1. In either case, the 11 observed clusters is significant. For $d = 3$, the sum of the seven yearly expectations is 8.9, compared to the expectation of 8.4 given in Table 1. In either case, the observed 13 clusters are not significant.

This approach sometimes helps to further focus investigations. Breaking the Philadelphia data down into yearly components shows that 5 of the 11 observed ($d = 2$) clus-

ters occurred in year 3 (1981) and was significantly (at the .023 level) higher than the expectation of 1.59 clusters for that year.

The approach also gives a simple method for studying the effects of trends on the expectation. Consider the following hypothetical data for 7 years ($D = 2, 557$), showing a modest trend in both FH and MS:

FH	23	24	25	26	27	28	29	(total = 182)
MS	13	14	15	16	17	18	19	(total = 112)

For $d = 2$, the expected number of clusters based on the simple retrospective model using the total figures is 21.2; summing the individual years' expectation gives 21.4. For negatively correlated trends (switch the order of the MS to go from 19 to 13), the expectation is 21.1. For $d = 3$, the corresponding numbers are 31.4, 31.8, and 31.4. For steep trends, as in the following hypothetical example, the expectations differ more:

FH	2	5	8	11	14	17	20	(total = 77)
MS	2	5	8	11	14	17	20	(total = 77)

For $d = 2$, the expected number of clusters using the total figures is 6.5; summing the individual years' expectation gives 8.3. For $d = 3$, the respective numbers are 10.3 and 12.7.

7. THE POWER OF THE DOUBLE-SCAN STATISTIC

The simple null hypothesis is that the A type I events and the B type II events are independently distributed by chance over the D -day period. The double-scan statistic is designed for the alternative in which most of the A and B events are distributed as under the null hypothesis, but some (say C) of the type I events are causally paired with type II events, linked within a distance of d days. The test based on the double-scan statistic has the nice feature that an existing formula can be used to compute its power against the alternative for which it is designed.

For $D \gg A, B$, clusters will be relatively rare, and under the null hypothesis, N_d is approximately Poisson distributed with mean $\lambda_0 = f(A, B) = E(N_d|A, B)$ given by Theorem 1. Under the alternative hypothesis, we assume that the number of clusters = $C + X$, where X is approximately Poisson distributed with mean $\lambda_1 = f(A - C, B - C)$. From the nature of $f(\)$, $C + \lambda_1 > \lambda_0$. We now illustrate how to evaluate the power of the double-scan statistic as a function of preliminary hypotheses about C and knowledge of A and B .

In Philadelphia there were 79 BFH and 61 BMS over the 7-year period. Under the simple null model, with $d = 2$, the expected number of clusters is 5.4. Based on the Poisson distribution, the chance of getting 10 or more clusters is .049. The researcher chooses to reject the chance model if 10 or more clusters are observed. The researcher asks what is the power of the test to detect a situation where over the 7 years, a total of 7 suicides are causally ($d = 2$) linked with a homicide. Given 7 linked suicide/homicides, this leaves an unrelated $(79 - 7 = 72)$ homicides and 54 suicides, which under Theorem 1 would lead to an expected 4.34 (unrelated) clusters. The total number of clusters is 7

Table 7. Power of Double-Scan Test Against Alternative Where a Fraction f of the Events are d -Linked, $d = 2,3$, for $D = 2,557$

d	$A = B$	Null $E(\text{clusters})$	Critical value (significance level)	Power against fraction f linked		
				$f: .05$	$.10$	$.15$
2	60	4.029	8 (.053)	.30	.84	1.00
	80	7.048	12 (.056)	.31	.82	1.00
	100	10.833	17 (.050)	.28	.78	1.00
	140	20.537	29 (.045)	.25	.71	.98
	180	32.814	43 (.050)	.25	.68	.97
	200	39.817	51 (.049)	.24	.65	.96
3	60	6.409	11 (.062)	.23	.60	.95
	80	11.027	17 (.057)	.21	.55	.91
	100	16.666	24 (.053)	.19	.51	.87
	140	30.522	40 (.057)	.19	.47	.82
	180	47.060	59 (.052)	.17	.42	.75
	200	56.072	69 (.052)	.16	.41	.74

plus a Poisson variate with mean 4.34; the power of the test is the probability that a Poisson variate with mean 4.34 is 3 or more, and equals .81.

Table 7 displays the power of the double-scan statistic test against alternatives where a fraction f of the events are linked. We carry out the analysis for $f = .05, .10, .15$, $d = 2, 3$, for $A = B = 60, 80, 100, 140, 180, 200$, where A and B are the number of type I and type II events in $D = 2,557$ days. The critical values are chosen to make the level of significance as close to .05 as possible.

We also see from Table 7 some of the consequences of misspecifying the value for d . For illustration, take the case $A = B = 100$ and $f = .10$, and call the two types of events homicides and suicides. Suppose that there are 10 causally $d = 2$ linked homicide/suicide clusters mixed in with 90 other homicides and 90 other suicides. If we carry out the double-scan test with $d = 2$, then the test will have power of .78. If we carry out the double-scan test with $d = 3$, then the power will be .51. On the other hand, if all of the 10 causally linked homicide/suicide clusters are linked within $d = 3$ (but not $d = 2$) days, then the test using $d = 2$ has power of only about .05. With similar calculations, one finds here that using the $d = 2$ test with 7 (8) causally $d = 2$ linked clusters has power .47 (.57), about equivalent to using the $d = 3$ test with 10 causally $d = 3$ (or closer) linked clusters. The researcher can use the foregoing reasoning to help decide on a specific value for d for testing purposes.

The researcher can similarly use an existing formula to construct a confidence interval for C based on the observed total number of clusters. In Philadelphia we observed 11 BFH/BMS ($d = 2$) clusters, compared to the 5.4 expected under the null chance model. If there was exactly 1 linked cluster, then the remaining 78 BFH and 60 BMS would lead to an expectation of 5.2 unlinked clusters, and the probability of getting a total of at least 11 clusters as observed is .04. Similarly, for 2 and 3 linked clusters, the probabilities would be .07 and .12. Thus we could say with $(1 - .12 =)88\%$ confidence that there are at least 4 linked clusters in the Philadelphia data. This is a fairly strong state-

ment given that the observed number of clusters is only 5.6 more than expectation and part of this excess could be chance variation.

8. DISCUSSION AND CONCLUSIONS

The study of disease clusters has led to the identification of various factors that can lead to unusual aggregations of disease. In this article we have developed a scan statistic to explore and identify unusual clustering between two types of events. A directional version of the statistic was also given that incorporates information on the anticipated order of the events. Examples of possible applications include the occurrence of violent deaths among different ethnic and gender groups (as demonstrated herein), occurrence of media reports of events followed by recurrences of the reported event (so-called “copy cat” crimes), analysis of air pollution episodes and emergency room visits and unusually hot summer days and mortality.

We see the use of this approach as exploratory. As with most cluster studies, there are concerns about both false-positive and false-negative results, and these should be kept in mind in the interpretation of any data. To address this, the specifics of identified clusters should be researched to determine whether there is any connection between the observed cases.

APPENDIX: PROOF AND GENERALIZATION OF THEOREM 1

Proof of Theorem 1

Recall that $N_d = \sum_{1 \leq i \leq D-d+1} Z_i$ where Z_i is defined in Section 2. Because Z_i only takes the values 0 or 1,

$$E(N_d) = \sum_{i=1}^{D-d+1} E(Z_i) = \sum_{i=1}^{D-d+1} P(Z_i = 1). \quad (A.1)$$

By symmetry,

$$P(Z_i = 1) = P(Z_d = 1) \quad \text{for} \quad i = d, d + 1, \dots, D - d + 1. \quad (A.2)$$

Substitute the right side of (A.2) into (A.1) to find (1).

Let $X_i = 1$ if the i th day contains a type I event, and let $Y_i = 1$ if the i th day contains a type II event. Let A_i denote the event $X_i = 1$ and let B_i denote the event $Y_i = 1$.

To prove Equation (2) (for the case $Z_1 = 1$), note that A_i 's are independent of B_j 's, and that

$$P(Z_1 = 1) = \left[1 - P\left(\bigcap_{i=1}^d A_i^c\right) \right] \left[1 - P\left(\bigcap_{i=1}^d B_i^c\right) \right], \quad (A.3)$$

where (for the retrospective model)

$$P\left(\bigcap_{i=1}^d A_i^c\right) = \binom{D-d}{A} / \binom{D}{A} = P_0(d, 0) \quad (A.4)$$

and

$$P\left(\bigcap_{i=1}^d B_i^c\right) = P_1(d, 0). \quad (A.5)$$

Substitute (A.5) and (A.4) into (A.3) to find Equation (2).

The Case $Z_2 = 1$. Note that $Z_2 = 1$ if either of the events C_{AB} or C_{BA} occur, where

$$C_{AB} = \left(\bigcap_{i=1}^d A_i^c\right) \cap A_{d+1} \cap \left(\bigcup_{i=2}^{d+1} B_i\right) \quad (A.6)$$

and C_{BA} is just the expression C_{AB} , with A_i 's and B_i 's switched. Further,

$$C_{AB} \cap C_{BA} = \left(\bigcap_{i=1}^d A_i^c\right) \cap A_{d+1} \cap \left[\bigcap_{i=1}^d B_i^c\right] \cap B_{d+1}. \quad (A.7)$$

It follows that

$$P(C_{AB}) = P_0(d+1, 1)Q_1(d, 0), \quad (A.8)$$

$$P(C_{BA}) = P_1(d+1, 1)Q_0(d, 1), \quad (A.9)$$

and

$$P(C_{AB} \cap C_{BA}) = P_0(d+1, 1)P_1(d+1, 0). \quad (A.10)$$

Substitute (A.8), (A.9) and (A.10) into

$$P(Z_2 = 1) = P(C_{AB}) + P(C_{BA}) - P(C_{AB} \cap C_{BA}) \quad (A.11)$$

to find Equation (3).

The Case $Z_k = 1$, for $3 \leq k \leq d$. Let d_{AB} denote the event

$$d_{AB} = \left(\bigcap_{i=1}^{d+k-2} A_i^c\right) \cap A_{d+k-1} \cap \left(\bigcup_{i=k}^{d+k-1} B_i\right) \quad (A.12)$$

and let d_{BA} denote the similar event where A 's and B 's are switched in (A.12). The simplest way for the event $Z_k = 1$ to happen is if $d_{AB} \cup d_{BA}$ occurs,

$$P(d_{AB}) = P_0(d+k-1, 1)Q_1(d, 0) \quad (A.13)$$

and

$$P(d_{AB} \cup d_{BA}) = \sum_{i=0}^1 P_i(d+k-1, 1)Q_{1-i}(d, 0) - \prod_{i=0}^1 P_i(d+k-1, 1). \quad (A.14)$$

Equation (A.14) gives first two terms on the right side of Equation (4).

Another way for the event $Z_k = 1$ to happen is if for some r , $r = 1, 2, \dots, k-2$, the event $D_{AB,r} \cup D_{BA,r}$ occurs, where

$$D_{AB,r} = A_r \cap \left(\bigcap_{i=r+1}^{d+k-2} A_i^c\right) \cap A_{d+k-1} \cap \left(\bigcap_{i=1}^{d+r-1} B_i^c\right) \cap \left(\bigcup_{i=d+r}^{d+k-2} B_i\right) \quad (A.15)$$

and $D_{BA,r}$ is the similar event where B 's and A 's are switched in (A.15). Note that $D_{AB,r}$ and $D_{BA,s}$ are mutually exclusive for r and s both in set $\{1, 2, \dots, k-2\}$. Note also that upper limit of last union on the right side of (A.15) is $d+k-2$, rather than $d+k-1$, because including that term would double count what is already counted by d_{BA} (see (A.12) switching A and B):

$$P(D_{AB,r}) = P_0(d+k-r, 2)[P_1(d+r-1, 0) - P_1(d+k-2, 0)] \quad (A.16)$$

and

$$P\left\{\bigcup_{r=1}^{k-2} (D_{AB,r} \cup D_{BA,r})\right\} = \sum_{r=1}^{k-2} [P(D_{AB,r}) + P(D_{BA,r})] = \sum_{r=1}^{k-2} \sum_{i=0}^1 P_i(d+k-r, 2) \times [P_{1-i}(d+r-1, 0) - P_{1-i}(d+k-2, 0)]. \quad (A.17)$$

Letting $j = r-1$ in (A.17) yields the last term on the right side of Equation (4) and completes the proof of Theorem 1.

Generalization of Theorem 1 for Prospective Model to Varying Probabilities of Events

Define the events A_i and B_i as earlier. Let α_i denote $\Pr(A_i)$, the probability of occurrence of a type I event on day i , for $i = 1, 2, \dots, D$. Similarly define $\beta_i = \Pr(B_i)$ for type II events. Define

$$P_{A,J}(r, s) = \left\{ \prod_{i=J}^{r-s+J-1} (1 - \alpha_i) \right\} \left\{ \prod_{i=r-s+J}^{r-1+J} (\alpha_i) \right\}, \quad (A.18)$$

where for $s = 0$ second product is 1. Let $P_{B,J}(r, s)$ be similarly defined with β 's replacing α 's. Let $Q_{A,J}(r, s)$ denote $1 - P_{A,J}(r, s)$.

Theorem 2. Given the occurrence of events on the different days are independently distributed, with probabilities α_i and β_i for the two types of events on day i , then $E(N_d)$ is given by (A.1), where

$$P(Z_1 = 1) = Q_{A,1}(d, 0)Q_{B,1}(d, 0) \quad (A.19)$$

and

$$P(Z_2 = 1) = P_{A,1}(d+1, 1)Q_{B,2}(d, 0) + P_{B,1}(d+1, 1)Q_{A,2}(d, 0) - P_{A,1}(d+1, 1)P_{B,1}(d+1, 1). \quad (A.20)$$

For the case $3 \leq k \leq d$,

$$P(Z_k = 1) = P(d_{AB}) + P(d_{BA}) - P(d_{AB} \cap d_{BA}) + \sum_{r=1}^{k-2} \{P(D_{AB,r}) + P(D_{BA,r})\}, \quad (\text{A.21})$$

where

$$P(d_{AB}) = P_{A,1}(d+k-1, 1)Q_{B,k}(d, 0), \quad (\text{A.22})$$

$$P(d_{AB} \cap d_{BA}) = P_{A,1}(d+k-1, 1)P_{B,1}(d+k-1, 1), \quad (\text{A.23})$$

$$P(D_{AB,r}) = \alpha_r P_{A,r+1}(d+k-r-1, 1) \times P_{B,1}(d+r-1, 0)Q_{B,d+r}(k-r-1, 0), \quad (\text{A.24})$$

and $P(d_{BA})$ and $P(D_{BA,r})$ are just $P(d_{AB})$ and $P(D_{AB,r})$ with α and β switched.

For the case $d < k$, apply the formula (A.21) for $k = d$, letting α_i take the value α_{k-d+i} for $i = 1, \dots, 2d$, and similarly for β 's.

Proof. Equations (A.1) and (A.3) hold for the general case of varying probabilities α_i, β_i . For the case $Z_2 = 1$, the probability of the events defined by (A.6) and (A.7) for the general case becomes

$$P(C_{AB}) = P_{A,1}(d+1, 1)Q_{B,2}(d, 0). \quad (\text{A.25})$$

$P(C_{BA})$ is just $P(C_{AB})$ with α 's and β 's switched:

$$P(C_{AB} \cap C_{BA}) = P_{A,1}(d+1, 1)P_{B,1}(d+1, 1). \quad (\text{A.26})$$

Substitute Equations (A.25) and (A.26) into Equation (A.11) to find $P(Z_2 = 1)$.

For the case $3 \leq k \leq d$, use the previous definitions (A.12) and (A.15) of the events d_{AB} and $D_{AB,r}$ to find (A.22) to (A.24) and follow through as in the proof of Theorem 1.

The final line in Theorem 2 notes how, given that we have programmed (A.21) to compute $P(Z_d = 1)$, we can with appropriate α 's and β 's compute the remaining $P(Z_k = 1)$ for $k > d$.

REFERENCES

- Aldous, D. (1989), *Probability Approximations via the Poisson Clumping Heuristic*, New York: Springer-Verlag.
- Arratia, R., Goldstein, L., and Gordon, L. (1990), "Poisson Approximation and the Chen-Stein Method," *Statistical Science*, 5, 403-434.
- Barbour, A. D., Holst, L., and Janson, S. (1992), *Poisson Approximation*. Oxford Studies in Probability 2, Oxford, U.K.: Clarendon Press.
- Buteau, J., Lesage, A. D., and Kiely, M. C. (1993), "Homicide Followed by Suicide: A Quebec Case Series, 1988-1990," *Canadian Journal of Psychiatry*, 38, 552-556.
- Cressie, N. (1980), "The Asymptotic Distribution of the Scan Statistic Under Uniformity," *Annals of Probability*, 8, 828-840.
- Glaz, J., and Naus, J. (1983), "Multiple Clusters on the Line," *Communications in Statistics, Part A—Theory and Methods*, 12, 1961-1986.
- (1991), "Tight Bounds and Approximations for Scan Statistic Probabilities for Discrete Data," *Annals of Applied Probability*, 1, 306-318.
- Gould, M. S., Wallenstein, S., and Kleinman, M. (1990), "Time Space Clustering of Teenage Suicide," *American Journal of Epidemiology*, 131, 71-78.
- Greenberg, M., Naus, J., Schneider, D., and Wartenberg, D. (1991), "Temporal Clustering of Homicide and Suicide Among 15-24-Year-Old White and Black Americans," *Ethnicity and Disease*, 1, 342-350.
- Huntington, R. J. (1976), "Constrained Mean Recurrence Times for k Successes in m Trials," technical report, AT&T, Long Lines Business Research.
- Krauth, J. (1992), "Bounds for the Upper-Tail Probabilities of the Circular Ratchet Scan Statistic," *Biometrics*, 48, 1177-1185.
- Lester, D. (1979), "Temporal Variations in Suicide and Homicide," *American Journal of Epidemiology*, 109, 517-520.
- Loader, C. (1991), "Large Deviation Approximations to the Distribution of Scan Statistics," *Advances in Applied Probability*, 4, 751-771.
- MMWR Current Trends (1991), "Homicide Followed by Suicide—Kentucky, 1985-1990," *Morbidity and Mortality Weekly Report*, 40 (38), 652-653, 659.
- Naus, J. I. (1982), "Approximations for Distributions of Scan Statistics," *Journal of the American Statistical Association*, 77, 177-183.
- (1988), "Scan Statistics," in *Encyclopedia of Statistical Sciences*, 8, eds. N. L. Johnson and S. Kotz. New York: Wiley, pp. 281-284.
- Page, E. S. (1955), "Control Charts With Warning Lines," *Biometrika*, 42, 243-257.
- Wallenstein, S., Gould, M. S., and Kleinman, M. (1989), "Use of the Scan Statistic to Detect Time-Space Clustering," *American Journal of Epidemiology*, 130, 1057-1064.
- Wallenstein, S., Naus, J., and Glaz, J. (1993), "Power of the Scan Statistic for Detection of Clustering," *Statistics in Medicine*, 12, 1829-1844.
- Wallenstein, S., Weinberg, C. R., and Gould, M. (1989), "Testing for a Pulse in Seasonal Event Data," *Biometrics*, 45, 817-830.

[Received November 1995. Revised December 1996.]