A COMPUTER SIMULATION IN THE STUDY OF THE COMMUNICABILITY OF DISEASES

Herbert S. Wilf and Martin Golubitsky
Department of Mathematics, University of Pennsylvania

I. Introduction.

In the period 1957-1960 in Niles, Illinois, eleven cases of leukemia were observed. The population of Niles (20,393) was such that about two or three cases should have been expected in such a period. Evidently, in this situation, the question of paramount importance is to decide whether or not such an event can be explained on purely random grounds, or else whether some mechanism of communicability or common etiology is operating.

To examine the hypothesis that the event is explicable on probabilistic grounds alone one must ask exactly the right question. As an example of a wrong question we mention the following: "What is the probability that eleven cases should strike in Niles, Illinois during this time period?" A candidate for the right question is: "What is the probability that if statistics on leukemia are accumulated for, say, twenty-five years, that in some $2\frac{1}{2}$ year period in some city the size of Niles, eleven cases will be observed?" The distinction is clearly of considerable importance.

To answer such a question there are at least two possible

approaches: mathematical analysis [7, 13, 3, 6, 11, and 1] and computer simulation. In this paper we present the results of such a simulation, along with the belief that it may serve as a useful model for the analysis of similar questions in the future.

II. Statement of the Problem.

What makes the sequence of eleven cases of leukemia which occurred over a two and one half year period from the autumn of 1957 to the spring of 1960 in Niles, Illinois so extraordinary is that Niles having a population of approximately 20,000 people should expect only two or three cases over the relevant time period. The question then arises: "Is such an aggregation, which is obviously abnormal, unexpected?" In other words, since there are many cities the size of Niles, should we expect one of them to have eleven cases over some two and one half year period in a twenty-five year observation period. Both positions have been taken. Erderer et al[1] draw the analogy of finding four successive sevens in a ~~total~~ table of 1,000,000 random digits. Heath and Hosterlik[2], on the other hand, make the following statement: "The cluster of eight cases of leukemia among children in Niles cannot reasonably be attributed to the effects of random distribution. These cases constitute a clearly defined micro-epidemic which occurs within a particular community and which affects a particular segment of that community's population." Note that neither statement yields any proof. What we should like to do here is to present a

simplified model of the U.S., its population distribution, and the incidence of leukemia in order to find the probability of an event like that in Niles.

III. Description of the Simulation.

We divided the population of the United States into subdivisions, or cities, and the populations of these cities were stratified into population classes such that all the cities in a given population class had the same population. We then created 7,000 cases of leukemia and assigned them at random to the various cities, where the probability that a case was assigned to a certain city was proportional to its population. This defines one "year" of experience. Twenty-five such one year experiences were run and thereby comprised one simulated quarter century. We then recorded for every $2\frac{1}{2}$ year subperiod (consecutive $2\frac{1}{2}$ year periods began with consecutive cases, making 158,501 $2\frac{1}{2}$ year overlapping subperiods per quarter century) the number of cities in each population class which received exactly one case, exactly two cases, ... , exactly 49 cases, and 50 or more cases. For the entire simulated quarter-century we then asked: "Did eleven cases occur in some city in the population class of Niles, Illinois in some $2\frac{1}{2}$ year subperiod?" Thus the end result of all this computation (about 12 minutes on the IBM 7040 of the Computing Center of the University of Pennsylvania) was either a "Yes" or a "No". We ran a total of 25 simulated quarter-centuries, receiving 24 "Yes" answers and one "No". To the question stated in the Introduction above, we answer that the probability is about 96% that such a

phenomenon will be observed.

IV. Summary of the Results.

Letting:

N = the number of population classes

A_i = the median population of the i^{th} population class $1 \leq i \leq N$

B_i = the number of cities in the i^{th} population class $1 \leq i \leq N$

A = the total sample population = $\sum_{i=1}^{N} A_i B_i$

p_i = probability that the next case will land in some city in the population class = A_i/A, $1 \leq i \leq N$

t = the number of cases which constitute the critical sub-period (with respect to Niles t is the number of cases which correspond to a $2\frac{1}{2}$ year period)

c = the critical number of cases in the time period t (eleven in the case of Niles)

T = the number of cases which constitute the entire observation period

The United States was divided into approximately 3300 cities each having more than 5,000 people. The cities were divided into 32 ($N = 32$) population classes and the population assigned to each city in a particular population class was approximately equal to the median of the various cities in that population class. The figures for the population distribution were

taken from the 1960 U.S. Census. This process gave us a distribution for the urban population only; we ignored the rural population. We then obtained a figure for the yearly incidence of leukemia by adjusting the 1960 mortality figure of 12,725 for the exclusion of the rural population. The mortality figure was taken from the U.S. Census volume on Vital Statistics. The adjusted figure for the incidence was about 7,000. Therefore, we let t equal 17,500 and T equal 175,000. We ran the program for twenty-five twenty-five year periods to obtain the results contained in Table II. Note that approximately five hours of computer time was necessary.

We again state that these are the probabilities that exactly c cases will occur in some city of population approximately A_i in some consecutive 17,500 case intervals out of a total of 175,000 cases. Niles, Illinois, having a population of approximately 20,000 people lies between population classes thirty and thirty-one (see Table I). The probabilities that eleven cases will appear in some $2\frac{1}{2}$ year subperiod in a twenty-five year observation period are 1.00 and .96 respectively for the two population classes. Even though these probabilities are not exact due to the Monte Carlo techniques used they should be close enough to indicate that the Niles episode was not so extraordinary after all.

We mention, in conclusion, that this model can be refined in several ways to parallel more closely the true disease patterns and the true population changes and movements over the

relevant time period. Also, considerably more information can be extracted from the program than the mere "Yes" or "No" referred to above, and a variety of other related questions can easily be answered. We believe that this method offers considerable promise for the study of diseases other than leukemia in differentiating the "exceptional" from the "expected".

V. Tables.

A. Table I gives the ~~exact~~ population distribution of the U.S. which we used for our model.

B. Table II gives the probabilities for obtaining exactly c cases of leukemia in any one of the thirty-two population classes.

Table I: Division of U.S. into Cities

Total population (A) = 104,216,000
Number of population classes (N) = 32

Population class (i)	Population (A_i)	No. of cities (B_i)	Probability (p_i)
1	7,782,000	1	.075
2	3,550,000	1	.034
3	2,479,000	1	.024
4	2,002,000	1	.019
5	1,670,000	1	.016
6	950,000	2	.018
7	850,000	1	.008
8	750,000	4	.029
9	650,000	4	.025
10	550,000	5	.026
11	475,000	7	.032
12	425,000	2	.008
13	375,000	4	.014
14	325,000	8	.025
15	275,000	9	.024
16	212,000	10	.020
17	187,000	10	.018
18	162,000	10	.016
19	137,000	16	.021
20	112,000	33	.035
21	95,000	20	.018
22	85,000	29	.024
23	75,000	25	.018
24	65,000	50	.031
25	55,000	56	.030
26	45,000	90	.039
27	37,000	64	.023
28	32,000	101	.031
29	27,000	111	.029
30	22,000	200	.042
31	15,000	934	.134
32	7,000	1394	.094

Table II: Results

Population classes 1-9 had 50 or more cases in every t-case sub-period.

	Population Class					
c	10	11	12	13	14	15
50	1.00	1.00	1.00	1.00	1.00	1.00
49	.04	.20	.48	1.00	1.00	1.00
48	.04	.04	.40	1.00	1.00	1.00
47	.04		.32	.96	1.00	1.00
46			.32	.96	1.00	1.00
45			.32	.80	1.00	1.00
44			.24	.56	1.00	1.00
43			.20	.48	1.00	1.00
42			.16	.44	1.00	1.00
41			.12	.28	1.00	1.00
40			.04	.20	1.00	1.00
39			.04	.16	1.00	1.00
38			.04	.12	1.00	1.00
37			.04	.08	1.00	1.00
36				.04	.96	1.00
35				.04	.96	1.00
34				.04	.88	1.00
33				.04	.84	1.00
32				.04	.68	1.00
31					.48	1.00
30					.32	1.00
29					.16	.96
28					.12	.96
27					.08	.84
26					.08	.72
25					.04	.60
24					.04	.32
23						.16
22						.12
21						.04
20						.04

	Population Class							
c	16	17	18	19	20	21	22	23
50	1.00	1.00						
49	1.00	1.00	.08					
48	1.00	1.00	.12					
47	1.00	1.00	.16	.08				
46	1.00	1.00	.16	.20				
45	1.00	1.00	.32	.28				
44	1.00	1.00	.48	.40		.04		
43	1.00	1.00	.68	.52	.04	.04		
42	1.00	1.00	.80	.68	.04	.04		
41	1.00	1.00	.96	.80	.04	.04		
40	1.00	1.00	.96	.88	.16	.04		
39	1.00	1.00	1.00	1.00	.36	.04		
38	1.00	1.00	1.00	1.00	.48	.08	.04	
37	1.00	1.00	1.00	1.00	.72	.16	.04	
36	1.00	1.00	1.00	1.00	.88	.28	.08	
35	1.00	1.00	1.00	1.00	.96	.36	.20	
34	1.00	1.00	1.00	1.00	1.00	.60	.24	.04
33	1.00	1.00	1.00	1.00	1.00	.84	.28	.08
32	1.00	1.00	1.00	1.00	1.00	.88	.32	.16
31	1.00	1.00	1.00	1.00	1.00	1.00	.52	.44
30	1.00	1.00	1.00	1.00	1.00	1.00	.72	.64
29	1.00	1.00	1.00	1.00	1.00	1.00	.92	.84
28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.96
27	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
26	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16	.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	.72	.92	1.00	1.00	1.00	1.00	1.00	1.00
14	.52	.76	1.00	1.00	1.00	1.00	1.00	1.00
13	.20	.52	1.00	1.00	1.00	1.00	1.00	1.00
12	.12	.36	.92	1.00	1.00	1.00	1.00	1.00
11	.08	.16	.64	1.00	1.00	1.00	1.00	1.00

	Population Class							
c	16	17	18	19	20	21	22	23
10	.04	.12	.32	1.00	1.00	1.00	1.00	1.00
9		.04	.12	1.00	1.00	1.00	1.00	1.00
8			.08	1.00	1.00	1.00	1.00	1.00
7				.80	1.00	1.00	1.00	1.00
6				.44	1.00	1.00	1.00	1.00
5				.04	.72	1.00	1.00	1.00
4				.04	.36	1.00	1.00	1.00
3					.04	.36	.60	1.00
2						.16	.24	.80
1						.04	.04	.16
0								

Population Class

c	24	25	26	27	28	29	30	31	32
35									
34									
33									
32									
31									
30									
29	.04								
28	.04								
27	.12								
26	.20								
25	.36	.12							
24	.48	.44							
23	.84	.60	.04						
22	1.00	.76	.16	.04					
21	1.00	.92	.36	.08					
20	1.00	1.00	.60	.16	.04				
19	1.00	1.00	.88	.24	.08				
18	1.00	1.00	1.00	.60	.20				
17	1.00	1.00	1.00	.80	.48	.04			
16	1.00	1.00	1.00	.92	.84	.32	.08		
15	1.00	1.00	1.00	1.00	1.00	.76	.16		
14	1.00	1.00	1.00	1.00	1.00	.96	.52	.04	
13	1.00	1.00	1.00	1.00	1.00	1.00	.84	.16	
12	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.56	
11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.96	
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.36
8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.92
7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	.36	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0		.68	1.00	1.00	1.00	1.00	1.00	1.00	1.00

BIBLIOGRAPHY

1. Erderer, F., M. H. Myers, H. Eisenberg and P. C. Campbell, "Temporal-Spatial Distribution of Leukemia and Lymphoma in Connecticut", J. Nat. Cancer Inst., 35:625-629, 1965.

2. Heath, C. W., and R. J. Hosterlik, "Leukemia Among Children in a Suburban Community", Am. J. of Med., 34:796, 1963.

3. Naus, J.I., "The Distribution of the Maximum Cluster of Points on a Line", ASD Paper 8, Applied Science Division, Operations Evaluations Group, June 29, 1962.

4. Hayes, D. M., "The Seasonal Incidence of Acute Leukemia", Cancer, 14:1301-05, 1961.

5. Pinkel, D., and D. Nefzger, "Some Epidemiological Features of Childhood Leukemia in the Buffalo, N.Y. Area", Cancer, 12:351-58, 1959.

6. Barton, D. E. and F. N. David, "Randomization Bases for Multivariate Tests. I. The Bivariate Case, Randomness of Points in a Plane", Bulliten de l'Institute International de Statistique, 39(2):455-67, 1962.

7. Erderer, F., M. H. Myers and N. Mantel, "A Statistical Problem in Space and Time: Do Leukemia Cases Come in Clusters?" Biometrics, Vol. 20, No. 3, Sept. 1964.

8. Katz, L., and J. H. Powell, "Probability Distribution of Random Variables Associated with a Structure of the Sample Space of Sociometric Investigations", Ann. of Math. Statist., 28:422-48, 1959.

9. Katz, L., and C. H. Procter, "The Concept of Configuration of Interpersonal Relations in a Group as a Time-Dependent Stochastic Process", Phychometrika, 24:317-27, 1959.

10. Knox, G., "Epidemiology of Childhood Leukemia in Northcumberland and Durham", Brit. J. of Prev. and Soc. Med., 18:17-24, 1964.

11. Mantel, N., "Chi-square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure", J. Amer. Statist. Assoc., 58:690-700, 1963.

12. Pinkel, D., J. E. Dowd, and I. D. J. Bross, "Some Epidemiological Features of Malignant Solid Tumors of Children in the Buffalo, N. Y. Area", Cancer, 16:28-33, 1963.

13. Mantel, N., "The Detection of Disease Clustering and a Generalized Regression Approach", submitted to Biometrika, 1965.

14. Erderer, F., R. W. Miller and J. S. Scotto, "U. S. Childhood Cancer Mortality Pattern, 1950-1959. Etiological Implications", J. Amer. Med. Assoc., Vol. 192, 593-96, May 17, 1965.

15. Manning and Carroll, "Some Epidemiological Aspects of Leukemia in Children", J. Nat. Cancer Inst., 19:1087, 1957.