STATISTICAL ASPECTS OF A COMMUNITY HEALTH AND ENVIRONMENTAL SURVEILLANCE SYSTEM

WILLIAM C. NELSON, VICTOR HASSELBLAD,
and GENE R. LOWRIMORE
ENVIRONMENTAL PROTECTION AGENCY

1. Introduction

The Community Health and Environmental Surveillance System (CHESS), a program conducted by the Division of Health Effects Research, Office of Research and Monitoring, Environmental Protection Agency, has been described in some detail by Riggan and co-workers [4] and Shy and co-workers [5]. Briefly, CHESS is a continuing series of epidemiologic studies carried out in selected communities representing an exposure gradient for the most common air pollutants. The basic purpose is to relate community health to changing environmental quality. The program involves monitoring of various pollutants and simultaneous surveillance of health indicators known to be sensitive to variations in environmental quality. The CHESS program will be useful in quantitating pollutant burdens, evaluating environmental standards, and documenting the health benefits of pollution control.

Area sets, sensitive health indicators, and environmental monitoring are the three key elements of CHESS. An area set consists of a group of communities selected as representative of a pollution gradient and similar to each other with respect to climate and socioeconomic traits. Each community in an area set is a middle class residential neighborhood.

The health indicators used reflect a broad spectrum of human responses, including no demonstrable effect, increase in body burden, physiologic changes of uncertain significance, physiologic sentinels of disease, acute and chronic morbidity, and death. Indicators used currently include symptoms of chronic respiratory disease in adults, incidence of acute respiratory disease in families, pulmonary function testing and lower respiratory illness of elementary school children, daily symptom reporting of asthmatics and elderly patients with chronic heart or lung disease, and tissue concentrations of selected trace elements.

A pollution monitoring station is established within each study neighborhood. Such factors as topography, emission sources, and land use are considered in monitoring site selection to ensure that the measurements are representative of population exposure. Pollutants are usually monitored on a 24 hour basis, but some continuous instruments are now being used to provide data on short term peaks.

The CHESS program is interdisciplinary and involves clinicians, epidemiologists, statisticians, engineers, chemists, meteorologists, sociologists, and programmers. However, the administration of the studies can be separated into various phases such as field data collection, bioenvironmental laboratory analysis, data preparation, data processing, statistical analysis, and technical report preparation. We have defined the latter four stages as Information Synthesis. Certain phases of Information Synthesis will now be discussed in detail.

2. Data processing aspects

Most health indicator data are collected by questionnaire. Study questions are administered biannually (lower respiratory illness, chronic respiratory illness, pollutant burdens), bimonthly (pulmonary function testing), biweekly (acute respiratory disease), or daily (asthmatic and elderly panels).

The sample sizes for these studies necessarily are large because of the many covariables present, such as age, sex, race, cigarette smoking, and economic status. For the biannual studies, approximately 2,000 families are surveyed in each of the 30 sectors participating in CHESS. About 300 families per sector are enrolled in the biweekly surveys. Each panel study requires approximately 100 individuals per sector. The biannual survey is given first and provides family background information on covariables and chronic disease conditions; this information is utilized in the selection of candidates for the additional studies.

The computer is an essential part of the data processing operation. Not only does it perform the standard large storage and retrieval functions and statistical analyses, but it also aids extensively in the study logistics. The computer selects candidates in priority order for the repetitive studies (acute respiratory disease, asthma and elderly panels). It ensures that study groups are similar as to proximity to environmental monitoring site, socioeconomic status, age, family size, and length of residence.

Much clerical time is saved by using the computer to pre-print necessary identifying information on the questionnaire forms. To lessen an enormous amount of coding and keypunching, optical mark questionnaire forms are now used. The information on these forms can be read directly onto magnetic tape. Even so, a large amount of editing and correcting is still necessary. For each questionnaire, various edit programs are performed to identify errors. The computer also can be used to prepare mailing labels for the questionnaire.

As the program scope expands to include wider use of continuous pollutant monitors and more sensitive health indicators such as electrocardiograms, the importance of the computer system increases still further. Features such as online telemetry and analog-digital conversion will be necessary.

3. Statistical aspects

The appropriate statistical analysis of the CHESS data requires considerable effort. The large populations and the community setting impose severe restrictions on the experimental design. Standard analyses are usually not possible because of repeated measurements, serial correlation, missing observations, or discreteness of the variables.

A pulmonary function study was done on second grade children in eleven selected Cincinnati schools to estimate, separately, socioeconomic and air pollution effects. Cigarette smoking and age variation were designed out of the study. Pollution monitoring stations were set up at each school. Pulmonary function tests, as measured by three-quarter second forced expiratory volume ($FEV_{0.75}$), were performed weekly in November, February, and May. Table I shows the mean forced expiratory volume and variance for each month averaged over all schools. Many more children participated than the 198 used for our analyses. However, this group met several selective criteria, including at least three valid readings each month, absence of asthma or acute bronchitis, and socioeconomic status (SES) (determined from personal family interviews) consistent with neighborhood SES.

TABLE I

$FEV_{0.75}$ MEANS AND VARIANCES FOR 198 SECOND GRADE CHILDREN IN THE CINCINNATI SCHOOL STUDY

Month	Mean	Variance
Nov.	1.1968	0.04355
Feb.	1.1908	0.04705
May	1.2342	0.05176

The simple correlation matrix (Table II) illustrates the consistency of the FEV readings over the three months and the intercorrelation of the covariables

TABLE II

CORRELATION MATRIX OF PULMONARY FUNCTION READINGS FOR THE CINCINNATI SCHOOL STUDY

	Nov.	Feb.	May	SES	Sex	SO_x	Height
Nov. FEV	1.000	0.885	0.857	0.072	0.229	−0.222	0.685
Feb. FEV		1.000	0.900	0.109	0.327	−0.366	0.688
May FEV			1.000	0.192	0.285	−0.338	0.660
SES				1.000	−0.058	−0.168	0.177
Sex					1.000	−0.043	0.219
SO_x						1.000	−0.282
Height							1.000

and the particulate sulfate (SO_x) concentrations. Height was used to adjust the FEV for individual body build differences.

A multivariate analysis of variance was used to allow for the repeated measurements and adjust for the covariables. Table III shows the FEV covariances (upper triangular), variances (diagonal), and correlations (lower triangular) after adjusting for SES, sex, and height. The variances are approximately half of the unadjusted values. The correlations are reduced only slightly by the adjustment. The p-values associated with the linear hypothesis analysis (Table IV) show each factor to be statistically significant.

TABLE III

COVARIANCE AND CORRELATION MATRICES FOR THE CINCINNATI SCHOOL STUDY

Correlations given below the diagonal.

	Nov.	Feb.	May
Nov.	0.02223	0.01736	0.01818
Feb.	0.7912	0.02165	0.01932
May	0.7488	0.8064	0.02651

TABLE IV

RESULTS FROM LINEAR MODEL ANALYSIS OF THE CINCINNATI SCHOOL STUDY

Factor	D.F.	U Statistic	p-value
Econ	3,1,193	0.9433	0.0091
Sex	3,1,193	0.9305	0.0025
SO_x	3,1,193	0.8712	<0.0001
Ht	3,1,193	0.5648	<0.0001

The dependent variable in the Cincinnati school study was continuous. In many studies, however, the health indicator is categorical. A useful method for the linear model analysis of discrete data has been developed by Grizzle and co-workers [2]. This method is difficult to apply, however, when the independent variable is continuous, that is, in the mixed model case. One solution is to group the independent variable into categories, which unfortunately, rapidly increases the number of response vectors. For example, a study involving five cities, five cigarette smoking categories, and three age groups produces 75 cells. The analysis requires dealing with several 75 × 75 matrices, which exceeds the memory capacity of many computers. Furthermore, even with a large sample size, some vectors occur with zero variance.

A technique capable of handling these study designs was desired, even at the cost of some sophistication or power. Since the dependent variable is always ordered, the ridit transformation, which involves transforming to the cumulative frequency distribution, was used. While the original use of the ridit by Bross [1] involved a standard or control population, we use the total sample for all areas as our standard. The technique leads to ordinary ANOVA and generalized linear hypothesis analysis. The two-sample ridit tests leading to the standard t test can be shown to be equivalent to the rank sum test.

TABLE V

COMPARISON OF 1000 SIMULATED RIDIT t'S WITH STUDENT t'S FOR VARIOUS SAMPLE SIZES

Simulated from multinomial with $p_1 = 0.4$, $p_2 = 0.2$, $p_3 = 0.2$, $p_4 = 0.1$, $p_5 = 0.1$.

Sample sizes		Observed ridit t's (Expected = 50)		Chi-square for goodness of fit	
n_1	n_2	$<t_{n_1+n_2-2,\,.05}$	$>t_{n_1+n_2-2,\,.95}$	(19 degrees of freedom)	p-value
5	5	57	53	151.36	.0000
5	10	42	44	15.88	.6653
5	50	38	64	29.88	.0533
10	10	60	55	13.56	.8087
50	50	54	58	18.56	.4854
200	500	45	51	11.16	.9183

Table V shows the result of some empirical sampling. We wanted to see how well the α-levels are preserved if the ridit transformation is used. In particular, we were interested in knowing how the test performs for small sample sizes. A multinomial distribution was simulated with five categories: $p_1 = 0.4$, $p_2 = 0.2$, $p_3 = 0.2$, $p_4 = 0.1$, and $p_5 = 0.1$. Two samples, of size n_1 and n_2 were drawn and tested for differences using the ridit t test. Only the two tails are shown, although 20 intervals, each with expected value of 50, were used. The goodness of fit statistic (19 degrees of freedom) is shown. The results show excellent agreement with the Student t for n_1, $n_2 > 10$. Even for the extreme case of $n_1 = n_2 = 5$, there is no evidence of distortion of the tail probabilities, even though the overall fit is poor.

Further empirical sampling showed ridit procedures using ANOVA and F tests quite comparable to LINCAT programs [2] using chi square analyses. The reason for using ridits is not to replace the rank sum or LINCAT, but there may be times when capacity of programs may be exceeded by those techniques. The ridit is a simple transformation with broad applications. Robustness of the analysis of variance has been reaffirmed even with discrete or dichotomous dependent variables.

Table VI illustrates an ANOVA table developed after a ridit transformation on the dependent variable (chronic respiratory disease prevalence as measured by cough, phlegm, and shortness of breath symptom reporting). The data were collected on fathers of elementary school children in five cities in Montana and Idaho with varying levels of trace metals exposure from mining. The

TABLE VI

ANALYSIS OF SEVERITY OF RESPIRATORY SYMPTOMS IN THE MONTANA-IDAHO STUDY

Factor	D.F.	S.S.	%S.S.	F	p-value
Cities	4	0.3038	0.43	1.64	0.1604
Gradient	1	0.1500	0.21	3.24	0.0719
Age	1	0.1073	0.15	2.32	0.1277
Smoking	1	12.2154	17.15	264.16	$<10^{-6}$
Education	1	0.0060	0.01	0.13	0.7184
Error	1224	56.6004			

symptom responses were placed on a severity scale of one to seven with the most common response, "no symptoms," being one. The purpose of the study was to compare differences in severity of respiratory symptoms with the differences in trace element exposure. In this case, the city differences in severity of disease symptoms were in a direction consistent with the exposure gradient, but were not statistically significant. Cigarette smoking, as expected, was overwhelmingly significant.

In setting pollution control standards, the existence of a threshold concentration for effects is implicitly assumed. The ideal data then are dose response relationships which show no effect until some nonzero threshold concentration is achieved. Most statistical analyses assume a strictly monotonic function as the relationship between two variables, and, therefore, that a significant relationship between health and pollutant exposure occurs at all levels of the pollutant.

As a simple alternative, we hypothesized a segmented line with zero slope up to a point x_0, and with positive slope above x_0. The point x_0 is found by the least squares method. This function has been named a "hockey stick" function. The technique for obtaining the least squares solution to a more general problem is due to Quandt [3].

The hockey stick function was fitted to data on high school cross country runners at a Southern California high school. The per cent of team members with decreased performance (increased time) from their last race is plotted against oxidant value one hour before the race (Figure 1). Fortunately, there was a monitoring station near the school course. Only home meet times were used. For this particular problem, the point estimate of the threshold was 12.2 pphm, with a 95 per cent confidence interval from 6.8 to 16.3 pphm.

We have briefly mentioned the statistical topics of cumulative frequency distribution transformations, mixed model categorical data problems, and fitting the hockey stick function. Another topic that presents problems for us is truncated or censored data. For example, in measuring body burdens for comparison of group means, many persons may have levels below minimum detectable limits. Thus, we may be faced with situations where as many as 50 per cent of the observations are missing. Since we know the number of missing observations, we technically have a censoring problem. We have used the cumulative frequency

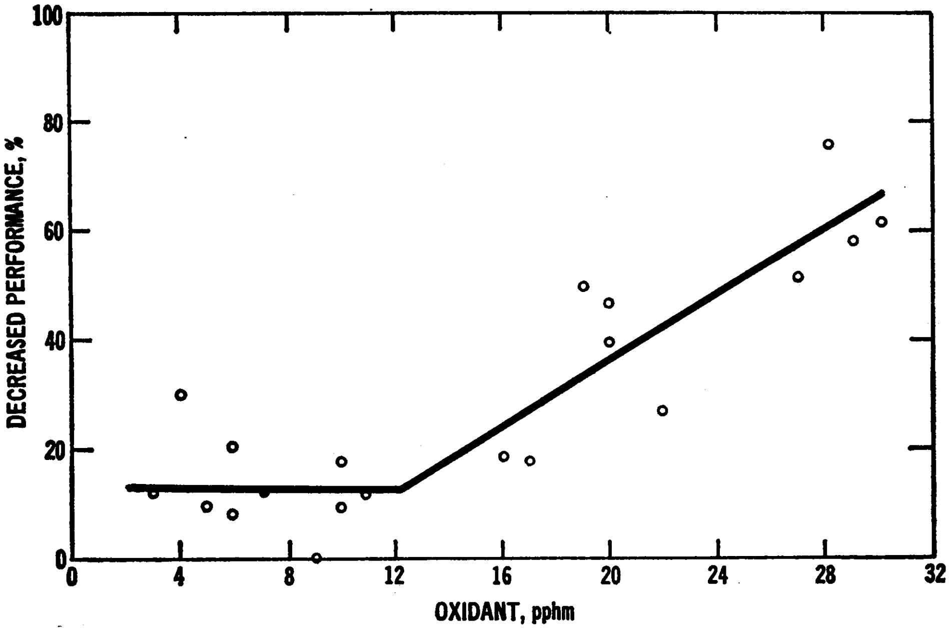

FIGURE 1

Per cent decreased performance *versus* oxidant level, with fitted "hockey stick" function.

transformation to handle this problem, but feel this area needs further development.

Another area requiring more statistical work is the problem of intercorrelated multivariate time series. Our panel studies illustrate the problem. These observations are made on groups of individuals through time. Because both our health indicators and pollutant exposures show seasonal variations, our results can be confounded. We make use of seasonal adjustments or season specific rates. Also we are experimenting with some types of probit analysis, but welcome additional results or suggestions.

Measuring the pollutant exposure of an individual is one of our most difficult problems. At best, we can measure pollution at an elementary school and assume that all children going to that school are exposed to that pollutant level. In some cases, we have to estimate a person's pollutant exposure from his address by using a few pollution values from several miles away. In other cases, we may not know how long he has lived at the address, where he works, whether he smokes, or a number of other important factors. Since humans do not live in controlled chambers, we are forced to estimate pollutant exposure from a few sites. For most pollutants, there are no emission inventories to give additional information on exposure. Statistical models are needed which will estimate exposure from this minimal information.

4. Summary

We have mentioned a few of the statistical topics associated with the community health effects studies. As statisticians, we feel there is no lack of challenging problems in this area.

Solutions to many of these problems must be found. The toxic effects at high concentrations of such atmospheric pollutants as carbon monoxide, sulfur dioxide, and nitrogen dioxide are well known. Several epidemiologic studies have suggested harmful effects of low level pollution. Further quantitation of these results is needed so that the dose effect relationships can be demonstrated and applied to the standard-setting process.

Environmental control is a tremendously expensive undertaking requiring difficult choices between many options. Better inputs are needed before the cost-benefit analysis approach for pollution control management can be effective. Statisticians must shoulder an increasing share of the effort in striving toward a cleaner environment.

REFERENCES

[1] I. D. J. Bross, "How to use the ridit analysis," *Biometrics*, Vol. 14 (1958), pp. 18–25.

[2] J. E. Grizzle, C. F. Starmer, and G. G. Koch, "Analysis of categorical data by linear models," *Biometrics*, Vol. 25 (1969), pp. 489–502.

[3] R. E. Quandt, "The estimation of the parameters of a linear regression system obeying two separate regimes," *J. Amer. Statist. Soc.*, Vol. 53 (1958), pp. 873–880.

[4] W. B. Riggan, D. I. Hammer, J. F. Finklea, V. Hasselblad, C. R. Sharp, R. M. Burton, and C. M. Shy, "CHESS: a community health and environmental surveillance system," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 6.

[5] C. M. Shy, J. F. Finklea, D. C. Calafiore, F. Benson, W. C. Nelson, and V. A. Newill, "A program of community health and environmental surveillance studies (CHESS)," *Determination of Air Quality*, New York, Plenum Press, 1972.

Discussion

Question: Burton E. Vaughan, Ecosystems Department, Battelle Memorial Institute, Richland, Wn.

Regarding the "hockey stick" effect shown in your slide. This is basically a correlation, in which other (unmeasured?) stressors presumably also affected running performance.

On the days when oxidant was measured, at say 8, 12, or 20 pphm, do you have evidence to show whether or not temperature, humidity, or some other stressor was also abnormally high? Could one of the latter factors have been the cause of 18% impairment in running performance, for the subthreshold region of your graph?

Reply: W. Nelson

Besides oxidant, hourly monitoring data were obtained for nitrogen oxides, carbon monoxide, particulates, temperature, relative humidity, and wind speed.

The correlations of these environmental factors with decreased running performance were examined. Correlation with oxidant was the most significant, though there was evidence of significant association with some of the other factors. Unfortunately, no extremes of temperature were observed for these races, so this effect could not be accurately assessed.

Question: Alexander Grendon, Donner Laboratory, University of California, Berkeley

The term used to describe the ordinate of your graph on the San Marino athletes was "per cent decrease in performance." It should evidently be, "percent of group who showed decrease in performance." Since the fraction showing a decrease from their last previous level of performance was related to the level of oxidants in air during the latest trial, I ask: was the last previous performance during a period of zero oxidants?

Reply: W. Nelson

There were no periods of zero oxidants observed on the race days. I would agree with you if you are suggesting by your question that, since the performance variable involves change from the last race, a better choice of pollution variable is "change in oxidant from last race." The choice of variables was made in an earlier published report from which we were only estimating a threshold level. However, we also tried using the "change in oxidants" variable and it made no difference for these data.