UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Environmental Criteria and Assessment Office (MD-52)
Research Triangle Park, North Carolina 27711

July 31, 1985

Dr. James Ware
Dept. of Biostatistics
Harvard School of Public Health
677 Huntington Ave.
Boston, MA 02115

Dear Dr. Ware:

I apologize for the long delay in getting back to you regarding the comments from Du Pont on the Corrigenda to the Lead Criteria Document and Joel Schwartz' response to them, but I have been away from the office for a week and a half on my honeymoon cruise to Mexico. Upon my return, I found the attached paper from Dick Landis that I had mentioned in my letter of July 1. If you would be so kind as to review this material, which was prepared specifically in response to the comments raised by Du Pont at the CASAC meeting on the Lead document, and integrate it into the opinion which you are preparing for us, we would be most grateful.

I will call you on Wednesday, August 7, to inquire when we might be able to complete this task. If you should have any questions in the meantime, please feel free to call me (919/541-4163), Joel Schwartz (202/382-2782), or Dick Landis (313/764-5450). Thank you for your help in this matter.

Sincerely,

David E. Weil, Ph.D.
Project Manager

ECAO-CD-81-2
IIA.D.8

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Environmental Criteria and Assessment Office (MD-52)
Research Triangle Park, North Carolina 27711

July 31, 1985

Dr. Richard Royall
Dept. of Biostatistics
Johns Hopkins University
615 N. Wolfe St.
Baltimore, MD 21205

Dear Dr. Royall:

I apologize for the long delay in getting back to you regarding the comments from Du Pont on the Corrigenda to the Lead Criteria Document and Joel Schwartz' response to them, but I have been away from the office for a week and a half on my honeymoon cruise to Mexico. Upon my return, I found the attached paper from Dick Landis that I had mentioned in my letter of July 1. If you would be so kind as to review this material, which was prepared specifically in response to the comments raised by Du Pont at the CASAC meeting on the Lead document, and integrate it into the opinion which you are preparing for us, we would be most grateful.

I am also enclosing a copy of a memo that Joel Schwartz sent to Jim Ware upon request; hopefully, it will be of use to you.

I will call you on Wednesday, August 7, to inquire when we might be able to complete this task. If you should have any questions in the meantime, please feel free to call me (919/541-4163), Joel Schwartz (202/382-2782), or Dick Landis (313/764-5450). Thank you for your help in this matter.

Sincerely,

David E. Weil, Ph.D.
Project Manager

NOTE: SEE IIA.D.1 FOR ATTACHMENTS

TEH 0412350

DUP050453402

N 30986.01

APPLICATIONS OF THE GENERALIZED MANTEL-HAENSZEL PROCEDURE TO THE ANALYSIS OF THE BLOOD PRESSURE/ BLOOD LEAD RELATIONSHIP IN THE NHANES II DATA

J. Richard Landis
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109 U.S.A.

Presented to the EPA Relative to Analyses Described
at the CASAC Meeting, May 13-14, 1985

Abstract

Many research investigations involve the analysis of categorical data in which primary attention is directed at the relationship between two of the variables, while controlling for the effects of a set of covariables. Frequently, the two primary variables correspond to treatment and response and one or both are measured and/or reported on ordinal scales. The Generalized Mantel-Haenszel procedure, based on a randomization model, can be used to address this relationship under appropriate scores for the row and column variable.

In contrast to this usual context, the row and column variables may be continuous, so that the scores are the actual measurements themselves. This use of the Generalized Mantel-Haenszel procedure is much less common, but does provide an alternative strategy for multiple regression models when the main focus is on the primary relationship averaged across the strata, rather than on a predictive model which includes effects for each of the strata.

This methodology is applied to the analysis of the blood pressure/blood lead relationship in the NHANES II data as an alternative to multiple regression modelling. Adjustments are made for age and body mass index, the two main covariables in the variation of blood pressure, and the 64 sampling sites.

1.1 Introduction

For many research situations the primary question of interest frequently involves the relationship between a set of s categories which correspond to distinct sub-populations (as defined in terms of pertinent independent variables) and a set of r categories which correspond to the response profiles associated with the specific dependent variables under study. Quite frequently, however, the distribution of the response profiles may also be influenced by the effects of a secondary set of t categories which correspond to distinct

levels of relevant covariables such as investigators, hospitals, clinics, or pretreatment status.

The resulting data obtained from such studies can be summarized in a set of t:(s x r) contingency tables which will be indexed by h = 1, 2, ..., t. In this formulation, the basic hypothesis then can be expressed in terms of ***no partial association*** between the sub-populations and the response profiles, after adjusting for the possible effects of the covariables.

1.2 Notation and Basic Model

Let h = 1, 2, ..., t index a set of (s x r) contingency tables which correspond to distinct levels of a covariable or combinations of several pertinent covariables. Let i = 1, 2, ..., s index a set of sub-populations such as treatment regimens which are to be compared with respect to a particular response variable for which the outcome categories are indexed by j = 1, 2, ..., r. Then let $\mathbf{n}_h' = (n_{h11}, ..., n_{h1r}, ..., n_{hs1}, ..., n_{hsr})$, where n_{hij} denotes the number of subjects in the sample who are jointly classified as belonging to the h*th* table, the i*th* sub-population and the j*th* response category. These frequency data, corresponding to the h*th* table, can be summarized as shown in Table 20.2, where $N_{hi\cdot}$ denotes the marginal total number of subjects classified as belonging to the i*th* sub-population, $N_{h\cdot j}$ denotes the marginal total number of subjects classified as belonging to the j*th* response category, and N_h denotes the overall marginal total sample size in the h*th* table.

The basic hypothesis under investigation involves the relationship between the response variable and the sub-populations adjusted for the levels of the covariable set. Under the assumption that the marginal totals $\{N_{hi\cdot}\}$ and $\{N_{h\cdot j}\}$ are fixed(either by design or conditional distribution arguments), the overall null hypothesis of ***no partial association*** can be stated as

H_0: For each of the separate levels of the covariable set h = 1, 2, ..., t, the response variable is distributed at random with respect to the sub-populations,

Table 1

Observed Contingency Table for Level h of the Covariables

Sub-pop'n	Response Variable Categories 1	2	...	r	Total
1	n_{h11}	n_{h12}	...	n_{h1r}	$N_{h1\cdot}$
2	n_{h12}	n_{h22}	...	n_{h2r}	$N_{h2\cdot}$
.	.	.	...	.	.
s	n_{hs1}	n_{hs2}	...	n_{hsr}	$N_{hs\cdot}$
Total	$N_{h\cdot 1}$	$N_{h\cdot 2}$	...	$N_{h\cdot r}$	N_h

i.e., the data in the respective rows of the h*th* table can be regarded as a successive set of simple random samples of sizes $\{N_{hi\cdot}\}$ from a fixed population corresponding to the marginal total distribution of the response variable $\{N_{h\cdot j}\}$. (1)

Under H_0 in (1), it follows that the vector of observed frequencies, $\mathbf{n}_h$, has the multiple hypergeometric distribution

$$\Pr(\mathbf{n}_h|H_0) = \frac{\prod_{i=1}^{s} N_{hi\cdot}! \prod_{j=1}^{r} N_{h\cdot j}!}{N_h! \prod_{i=1}^{s}\prod_{j=1}^{r} n_{hij}!}, \qquad (2)$$

as developed in Chapter 16 for a single s x r table.

For each table indexed by h = 1, 2, ..., t, let $P_{hi\cdot} = (N_{hi\cdot}/N_h)$ denote the marginal proportion of subjects classified as belonging to the i*th* sub-population(row) and let $P_{h\cdot j} = (N_{h\cdot j}/N_h)$ denote the marginal proportion of subjects classified as belonging to the j*th* response category (column). These proportions, which are assumed to be fixed quantities, can be summarized in vector notation for h = 1, 2, ..., t as $\mathbf{P}'_{h*\cdot} = (P_{h1\cdot}, \ldots,$

$P_{hs\cdot}$) and $P'_{h\cdot *} = (P_{h\cdot 1}, \ldots, P_{h\cdot r})$. Then by computing the first and second moments of the probability distribution in (2) as outlined in Chapter 16, it follows that the expected value of n_{hij} under H_0 in (1) is $m_{hij} = N_h P_{hi\cdot} P_{h\cdot j}$. These expected values of the elements of $\mathbf{n}_h$ can be expressed in vector notation as

$$\begin{aligned} \mathbf{m}_h &= E(\mathbf{n}_h|H_0) \\ &= N_h[\mathbf{P}_{h\cdot *} \otimes \mathbf{P}_{h*\cdot}], \end{aligned} \tag{3}$$

where $\otimes$ denotes left hand Kronecker product multiplication. Furthermore, the covariance between n_{hij} and $n_{hi'j'}$ under H_0 in (1) is

$$v_{h,ij,i'j'} = \frac{N_{hi\cdot} N_{h\cdot j} (\delta_{ii'} N_h - N_{hi'\cdot}) (\delta_{jj'} N_h - N_{h\cdot j'})}{N_h^2 (N_h - 1)},$$

where $\delta_{ii'} = 1$, if $i = i'$, $= 0$, otherwise and $\delta_{jj'} = 1$, if $j = j'$, $= 0$ otherwise. Thus, the covariance matrix of $\mathbf{n}_h$ under H_0 in (1) can be expressed in matrix notation as

$$\mathrm{Var}(\mathbf{n}_h|H_0) = \frac{N_h^2}{(N_h - 1)} [\mathbf{D}_{\mathbf{P}_{h\cdot *}} - \mathbf{P}_{h\cdot *} \mathbf{P}'_{h\cdot *}] \otimes [\mathbf{D}_{\mathbf{P}_{h*\cdot}} - \mathbf{P}_{h*\cdot} \mathbf{P}'_{h*\cdot}], \tag{4}$$

where $\mathbf{D}_{\mathbf{P}_{h*\cdot}}$ and $\mathbf{D}_{\mathbf{P}_{h\cdot *}}$ are diagonal matrices with elements of the vectors $\mathbf{P}_{h*\cdot}$ and $\mathbf{P}_{h\cdot *}$ on the main diagonal.

1.3 Correlation Test

Essentially, there are three different situations in which this randomization model structure can be used to test H_0. In order of increasing specificity, they can be summarized as

i) nominal row variable, nominal column variable
ii) nominal row variable, ordinal column variable
iii) ordinal row variable, ordinal column variable.

Each of these cases are summarized in considerable detail in Landis *et al.* (1978, 1979); however, only case iii) will be discussed further here.

For situations where both the response categories $j = 1, 2, \ldots, r$ and the sub-

population categories i = 1, 2, ..., s are ordinally scaled with progressively larger intensities, it is often useful to test H_0 in terms of composite mean scores which correspond to the respective products of an appropriate vector of response scores $a_h' = (a_{h1}, a_{h2}, ..., a_{hr})$ and an appropriate vector of sub-population scores $c_h' = (c_{h1}, c_{h2}, ..., c_{hs})$ for h = 1, 2, ..., q. Specifically, let $w_{hij} = c_{hi}a_{hj}$ be the score assigned to the joint outcome corresponding to the *ith* sub-population and *jth* response category in the h*th* table which can be summarized in vector notation by letting $w_h' = (w_{h11}, ..., w_{h1r}, ..., w_{hsr})$. Then let the composite mean score function for the h*th* table be defined by

$$F_h = w_h' n_h / N_h. \tag{5}$$

Moreover, let the mean score on the row marginal proportions be denoted by

$$\bar{c}_h = \sum_{i=1}^{s} c_{hi} P_{hi\cdot}. \tag{6}$$

and let the mean score on the column marginal proportions be as defined in (20.18). Also, let the total variance of the sub-population variable in the overall population for the h*th* table with respect to c_h be denoted by

$$S_{c,h}^2 = \sum_{i=1}^{s} (c_{hi} - \bar{c}_h)^2 P_{hi\cdot}, \tag{7}$$

and let the corresponding total variance of the response variable be given as in (20.20). In addition, let

$$S_{ac,h} = \sum_{i=1}^{s} \sum_{j=1}^{r} (a_{hj} - \bar{a}_h)(c_{hi} - \bar{c}_h)(n_{hij}/N_h) \tag{8}$$

be the covariance between the response scores and the sub-population scores in the h*th* table. Then from (3) and (4) it follows that

$$E(F_h|H_0) = \bar{a}_h \bar{c}_h \tag{9}$$

$$\mathrm{Var}(F_h|H_0) = \frac{S^2_{a,h}S^2_{c,h}}{(N_h - 1)}. \tag{10}$$

Furthermore, from (5) and (8) it can be shown that

$$F_h - E(F_h|H_0) = S_{ac,h}. \tag{11}$$

Given this framework in (5) – (11), H_0 in (1) can be tested in terms of a test statistic which is directed at alternatives pertaining to the association between the response variable the sub-population variable under the composite scores $\{w_{hij}\}$. In particular, let

$$\begin{aligned} Q_{MA,h} &= \frac{[\,F_h - E(F_h|H_0)\,]^2}{\mathrm{Var}(F_h|H_0)} \\ &= \frac{(N_h-1)S^2_{ac,h}}{S^2_{a,h}S^2_{c,h}} \\ &= (N_h - 1)R^2_{ac,h}, \end{aligned} \tag{12}$$

where $R^2_{ac,h}$ denotes the squared Pearson correlation coefficient corresponding to the observed bivariate distribution for the $\{a_{hj}\}$ and the $\{c_{hi}\}$ in the h*th* table. Under H_0 in (1), $Q_{MA,h}$ asymptotically follows the χ^2 distribution with d.f. = 1. As defined in (12), $Q_{MA,h}$ essentially is a standard correlation analysis test statistic for H_0. As a result, an overall test statistic for the simultaneous test of the measures of association $\{F_h\}$ in (5) across all t tables is

$$Q_{MA} = \sum_{h=1}^{q} Q_{MA,h}. \tag{13}$$

If each of the respective N_h for h = 1, 2, ..., t is sufficiently large that all of the $\{Q_{MA,h}\}$ have approximate chi-square distributions, then Q_{MA} has an approximate chi-square distribution with d.f. = t under H_0 in (1).

Alternatively, if the overall sample size N is large but the individual sample sizes $\{N_h\}$ for many tables are small, then the statistic Q_{MA} in (13) is no longer useful for testing H_0 in (1). Instead it is more appropriate to investigate H_0 in terms of a generalized Cochran/Mantel-Haenszel/Birch strategy with respect to the measures of association $\{F_h\}$ in (5) based on the composite mean scores $\{w_h\}$. For this purpose, let the weighted sum of the composite mean scores across the t tables be denoted be

$$F = \sum_{h=1}^{t} N_h F_h. \tag{14}$$

From (9) and (10) it follows that

$$E(F|H_0) = \sum_{h=1}^{t} N_h \bar{a}_h \bar{c}_h \tag{15}$$

$$Var(F|H_0) = \sum_{h=1}^{t} \frac{N_h^2 S_{a,h}^2 S_{c,h}^2}{N_h - 1}. \tag{16}$$

Consequently, the generalized Cochran/Mantel-Haenszel/Birch statistic for H_0 in terms of composite mean scores with respect to $\{w_h\}$ is obtained as

$$Q_{RMA} = \frac{\left[\sum_{h=1}^{t} N_h S_{ac,h}\right]^2}{\sum_{h=1}^{t} \frac{N_h^2 S_{a,h}^2 S_{c,h}^2}{N_h - 1}}. \tag{17}$$

Under H_0, Q_{RMA} asymptotically has the chi-square distribution with d.f. = 1. This statistic is directed at the extent to which there is a consistent positive (or negative) association between the response scores and the sub-population scores in the respective tables. Thus, it is directed at ***average partial association*** alternatives across the t tables between the response scores $\{a_h\}$ and the sub-population scores $\{c_h\}$. In particular,

for the case where all the scores are either 0 or 1, Q_{RMA} in (17) simplifies to the Cochran/ Mantel-Haenszel/Birch statistic in (20.16) which corresponds to the collapsed set of (2 x 2) tables which are obtained by pooling the categories which are assigned "0" and those which are assigned "1" separately for the response categories and the sub-populations.

Finally, if marginal rank or ridit-type scores are obtained from both the rows and columns of each table with midranks assigned for ties, the statistic Q_{RMA} in (17) is essentially equivalent to a partial Spearman Rank Correlation test, conditioning on the levels of the covariable set. Otherwise, the same types of comments could be made about power properties and central limit theory considerations for these statistics as were made for the multivariate test in Section 20.1 and the mean score test in Section 20.2, except that the concept of power is stronger relative to the association structure at which they are targeted and sample size requirements are minimal because only one degree of freedom is involved.

1.4 Continuous Variable Situation

The test statistic in (17) simplifies considerably in situations for which the row and column variable are analyzed as continuous variables. In particular, let

y_{hi} denote the observed response variable value for the i*th* individual in the h*th* stratum;

x_{hi} denote the observed predictor(row) variable value for the i*th* individual in the h*th* stratum.

Then the contribution to the numerator in the statistic (17) from the h*th* stratum simplifies to

$$F_h - E(F_h|H_0) = \sum x_{hi}y_{hi} - \frac{(\sum x_{hi})(\sum y_{hi})}{N_h},$$

and the variance contribution to the denominator from the h*th* stratum simplifies to

$$\mathrm{Var}(F_h|H_0) = \frac{(\sum y_{hi}^2 - N_h \bar{y}_h^2)(\sum x_{hi}^2 - N_h \bar{x}_h^2)}{N_h - 1}.$$

As a result, the overall statistic Q_{RMA} for H_0, using the actual values for x and y, is based on weighted sums of the usual components of the regression coefficient across the strata.

1.5 Application to the NHANES II Data

This generalized procedure can be used to test the significance of the linear relationship between blood lead(Pb) on the *ln* scale and diastolic blood pressure(DBP), adjusting for the key predictive variables, age and body mass index, and the 64 sampling sites. An attractive feature of this methodology for the NHANES II design is that it provides adjustments for the blood pressure/blood lead test statistic for the potential effects of the sampling sites, without requiring a predictive model with 63 site parameters. Furthermore, this statistic averaged over the strata formed by the cross-classification of age and body mass subgroups and sampling sites potentially is more robust than a multiple regression model with indicator variables for the sites.

In order to adjust for age and body mass index(quetlets), we need to create several ordinal variables that, in turn, are cross-classified to form strata. Although a large number of different possible partitions of these two risk factors could be considered, we will limit our attention to two alternatives for age and quetlets index. For notational convenience, let a denote age and q denote quetlets index as continuous variables. Then define the stratification subclasses to be

$$A_1 = \begin{cases} 1, & a \le 40 \\ 2, & a > 40. \end{cases}$$

$$A_2 = \begin{cases} 1, & a \le 20 \\ 2, & 20 < a \le 40 \\ 3, & a > 40. \end{cases}$$

$$Q_1 = \begin{cases} 1, & q \leq 22 \\ 2, & 22 < q \leq 26.5 \\ 3, & q > 26.5 . \end{cases}$$

$$Q_2 = \begin{cases} 1, & q \leq 23 \\ 2, & 23 < q \leq 27 \\ 3, & q > 27 . \end{cases}$$

These particular cutpoints for age and quetlets index were selected on the basis of meaningful substantive issues, as well as sample size considerations. To illustrate the effects of these two variables on diastolic blood pressure(DBP) and blood lead levels(h scale), we can summarize the mean DBP by the subgroups. For example, the sample sizes, and the corresponding unweighted mean DBP and Blood Lead Levels for A2 by Q2 are displayed in Table 2.

Table 2

Sample Size, Unweighted Mean Diastolic Blood Pressure(DBP) and Unweighted Blood Lead Level(h Scale) by Age and Quetlets Subgroups: Males, 12–74 years, NHANES II, United States,

Age (A_2)	Quetlets (Q_2)	Sample Size	Mean DBP	Mean Blood Lead
1	1	484	70.9	2.55
1	2	135	75.6	2.54
1	3	43	84.2	2.67
2	1	336	75.4	2.78
2	2	368	79.6	2.77
2	3	265	86.6	2.75
3	1	327	79.9	2.77
3	2	672	83.4	2.76
3	3	551	87.6	2.74

The generalized Mantel-Haenszel statistic, Q_{RMA}, was computed for the relationship between blood lead level(h scale) and diastolic blood pressure, adjusting for the possible effects due to sampling site, quetlets index and age, either singly or in

combination. The resulting statistics are summarized in Table 3.

Table 3

Correlation Test Statistic for Diastolic Blood Pressure on Blood Lead Level(*ln* scale) Based on Randomization Model Adjusted for Stratification Variables: Males, 12-74 years, NHANES II, United States, 1976-1980

Stratification	Number of Strata with $N_h \geq 2$	Sample Size	Correlation Test Statistic	d.f.	Signif. Level
None	1	3181	68.83	1	<0.01
Sites	64	3181	44.04	1	<0.01
Sites, Q_1	192	3181	23.13	1	<0.01
Sites, Q_2	192	3181	25.94	1	<0.01
Sites, Q_1, A_1	355	3181	13.47	1	<0.01
Sites, Q_2, A_1	368	3181	15.01	1	<0.01
Sites, Q_1, A_2	463	3181	4.08	1	<0.05
Sites, Q_2, A_2	478	3181	4.62	1	<0.05

The primary objective in the analysis of these data was to investigate the relationship between Pb level and DBP level, adjusting for other important sources of variation. As noted elsewhere, because of the strong multicollinearity between age, body mass and Pb levels on DBP, these adjustments are already adjusting for some of the lead effect on blood pressure. Moreover, because of the striking decline, both in Pb levels and blood pressure over time, adjustment for the sampling sites also reduces the potential for the Pb level variable to provide a significant incremental effect on this relationship. It is noteworthy that even in the multiple regression models with a large number of potential predictor variables, that quetlets index and age were the dominant sources of variation in DBP. Nevertheless, even with the severe adjustments by sampling sites and age and quetlets, the blood pressure/blood lead relationship is still statistically significant at the 5%

level. In fact, the chi-square statistic of approximately 4 in the analysis with the most stringent adjustments corresponds to the square of the t statistic of 2 in the most conservative multiple regression models.

References

Landis, J.R., Heyman, E.R., and Koch, G.G. (1978), Average Partial Association in Three-Way Contingency Tables: A Review and Discussion, *International Statistical Review 46*, 237–254.

Landis, J.R., Cooper, M.M., Kennedy, T. and Koch, G.G. (1979), A Computer Program for Testing Average Partial Association in Three-Way Contingency Tables(PARCAT), *Computer Programs in Biomedicine 9*, 223–246.