CATEGORICAL DATA ANALYSIS BY A RANDOMIZED RESPONSE TECHNIQUE FOR STATISTICAL DISCLOSURE CONTROL

Martin J. Rosenberg, Syntex Research

1. INTRODUCTION

In order to protect the confidentiality of respondents, collecting agencies are often restricted from releasing data bases to external investigators who desire to perform specialized analyses on the data. Merely stripping the data of identifying information such as names and addresses does not provide sufficient protection to respondents, since if a number of variables were known from public or other external sources, these variables could be used to match on and hence identify individual respondents. Similar problems can arise when it desired to merge two or more data sets collected by different agencies from essentially the same respondents, but it is necessary to protect the confidentiality of at least one data set.

One solution advanced in the literature is to inoculate the data with random errors in such a way that individual respondents and their responses cannot be identified but that statistical analyses can still be performed on the data set. Warner (1971) noted that his randomized response (RR) technique, originally introduced to collect data of a sensitive nature, could be used to contaminate existing data sets so as to allow their release. However, the chief drawback to this application of RR has been the lack of multivariable analytic techniques that measure relationships between variables.

This paper presents an Additive Randomized Response Contamination (ARRC) model suitable for contaminating categorical data (both ordinal and nominal) and shows how the Grizzle, Starmer, Koch (GSK 1969) categorical linear model can be applied to such contaminated data to yield a wide variety of possible analyses. The techniques are also applicable to data actually collected by RR models that meet certain basic criteria.

2. THE PRIVACY TRANSFORMATION

Let X, W and Z represent the true variable, a contaminating random variable used to introduce random erros, and the contaminated random variable to be released, respectively. Each variable has K categories indexed 1,2,...,K.

Although a contaminating variable need not be independent of its own true variable, two independence properties always hold. Let the data set to be contaminated contain v categorical variables $X_1,\ldots,X_v$, respectively. Then:

(1) W_i is independent of X_j, $i \neq j$

(2) $W_1,\ldots,W_v$ are independent.

Once the realizations of W have been selected,

$$Z = X + W \text{ (m mod K)} \qquad (2.1)$$

where (m mod K) indicates the addition has occurred by modified modulo arithmetic which is defined as

$$a + b \text{ (m mod K)} = [a + b - 1 \text{ (mod K)}] + 1. \qquad (2.2)$$

3. THE TRANSITION MATRIX

The concept of transition matrices will greatly simplify the analysis of contaminated categorical data. A transition matrix $\underset{\sim}{T}$ is a non-singular square matrix composed of elements t_{ij} where

$$t_{ij} = \Pr(Z=i|X=j) \qquad i,j=1,\ldots,K \qquad (3.1)$$

and t_{ij} is the element in the i^{th} row and j^{th} column of $\underset{\sim}{T}$. From the definition comes three properties of transition matrices:

(1) Each element, being a conditional probability, takes on values between 0 and 1, that is $0 \leq t_{ij} \leq 1$ for all $i,j=1,\ldots,K$.

(2) Each column sums to one, that is

$$\sum_{i=1}^{K} t_{ij} = 1 \text{ for all } j=1,\ldots,K. \qquad (3.2)$$

(3) $\underset{\sim}{T}^{-1}$ exists.

4. ESTIMATION

With X denoting a true variable having $K \geq 2$ categories indexed by 1,2,...,K, the parameters of interest are the probabilities $\underset{\sim}{\pi} = (\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2,\ldots,\underset{\sim}{\pi}_K)'$ defined as

$$\underset{\sim}{\pi}_j = \Pr(X=j), \quad j=1,\ldots,K \qquad (4.1)$$

where $0 < \pi_j < 1$ and $\sum_{j=1}^{K} \pi_j = 1$.

ARRC produces a hypothetical population Z whose probabilities are given by $\underset{\sim}{Q} = (Q_1,Q_2,\ldots,Q_k)'$ where

$$Q_i = \Pr(Z=i) = \sum_{j=1}^{K} \Pr(Z=i|X=j) \Pr(X=j) \qquad (4.2)$$

$$= \sum_{j=1}^{K} t_{ij}\pi_j, i=1,2,\ldots,K.$$

In matrix notation

$$\underset{\sim}{Q} = \underset{\sim}{T}\,\underset{\sim}{\pi} \qquad (4.3)$$

and since $\underset{\sim}{T}$ is defined as non-singular

$$\underset{\sim}{\pi} = \underset{\sim}{T}^{-1}\,\underset{\sim}{Q} \qquad (4.4)$$

The Q_i can be estimated using the observed proportions $q_i = n_i/N$ of the contaminated variable Z. Thus the probabilities π_i of X are estimated by

$$\underset{\sim}{p} = \underset{\sim}{T}^{-1}\,\underset{\sim}{q}. \qquad (4.5)$$

Using the linearized Taylor series expansion [which is exact for linear transformations such as (4.5)] the covariance matrix of $\underset{\sim}{p}$ is given by

$$V(\underset{\sim}{p}) = \underset{\sim}{T}^{-1}\, V(\underset{\sim}{q})\, \underset{\sim}{T}^{-1'} \qquad (4.6)$$

where $\underset{\sim}{V}(\underset{\sim}{q})$ is the covariance matrix of $\underset{\sim}{q}$ given by

$$\underset{\sim}{V}(\underset{\sim}{q}) = (1/N)\, [\underset{\sim}{D}_{\underset{\sim}{Q}} - \underset{\sim}{Q}\underset{\sim}{Q}'] \qquad (4.7)$$

and $\underset{\sim}{D}_Q$ is the diagonal matrix with the elements of $\underset{\sim}{Q}$ on the main diagonal.

5. THE GSK MULTIVARIATE CATEGORICAL LINEAR MODEL

In applying the GSK approach, it is useful to reformulate the multi-way conteingency table we desire to analyze, into canonical form in which the s rows are sub-populations and the r columns are response profiles. It is assumed that we sample from each sub-population independently, hence the row margins are fixed and the cell probabilities in each row sum to one.

Let the first subscript represent the row, and the second the column in the canonical table. Let $\underset{\sim}{\pi}$ denote the unknown sr x 1 parameter vector of population cell probabilities in row order, where

$$\underset{\sim}{\pi} = (\pi_{11},\pi_{12},\ldots,\pi_{1r},\pi_{21},\ldots,\pi_{2r},\ldots,\pi_{sr})' \qquad (5.1)$$

and let $\underset{\sim}{p}$ be an estimate of $\underset{\sim}{\pi}$ consisting of sample cell proportions based on sample cell frequencies where similarly

$$\underset{\sim}{p} = (p_{11},p_{12},\ldots,p_{1r},p_{21},\ldots,p_{2r},\ldots, p_{sr})'.$$

Then the GSK framework fits models of the form

$$E_A[\underset{\sim}{F}(\underset{\sim}{p})] = \underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}$$

where the symbol E_A denotes asymptotic expectation.

When, however, data contaminated by ARRC is used to construct a contingency table, the sample cell proportions are no longer unbiased estimates of $\underset{\sim}{\pi}$. Instead these contaminated sample cell probabilities $\underset{\sim}{q}$ are estimates of $\underset{\sim}{Q} = \underset{\sim}{g}(\underset{\sim}{\pi})$ where g is a vector valued function. The exact form of $\underset{\sim}{g}$ varies according to conditions described below, but in all cases the strategy is the same. Once $\underset{\sim}{g}$ is known, construct the inverse function $\underset{\sim}{g}^{-1}$ so that $\underset{\sim}{g}^{-1}(\underset{\sim}{Q}) = \pi$. Then $\underset{\sim}{g}^{-1}(\underset{\sim}{q}) = \underset{\sim}{p}$ is an estimate of $\underset{\sim}{\pi}$, and we can proceed to build the model

$$\underset{\sim}{F}[\underset{\sim}{g}^{-1}(\underset{\sim}{Q})] = E_A\{\underset{\sim}{F}[\underset{\sim}{g}^{-1}(\underset{\sim}{q})]\} = E_A\{\underset{\sim}{F}[\underset{\sim}{p}]\}$$
$$= \underset{\sim}{F}(\underset{\sim}{\pi}) = \underset{\sim}{X}\underset{\sim}{\beta}.$$

Often the variables which index the sub-populations and/or response profiles of a canonical table are not single variables but compound variables composed of two or more factors. Suppose one had to form a compound contaminated variable Z from v contaminated factor variables z_i, each indexed from $1,2,\ldots,k_i$, where $i=1,2,\ldots,v$. Then the profiles of the levels of Z are given by $(z_1,z_2,\ldots,z_v)$ where the rightmost variable increments fastest. Rosenberg (1979) has shown that if $\underset{\sim}{T}_i$ is the transition matrix of Z_i, $i=1,2,\ldots,v$ the transition matrix $\underset{\sim}{T}$ of the compound variable Z is given by

$$\underset{\sim}{T} = \underset{\sim}{T}_v \otimes \underset{\sim}{T}_{v-1} \otimes \ldots \otimes \underset{\sim}{T}_2 \otimes \underset{\sim}{T}_1 \qquad (5.2)$$

where the symbol $\otimes$ denotes the left direct (or Kronecker) product of two matrices.

Let the subscript A refer to the row variable and the subscript B to the column variable in the canonical table. Hence $\underset{\sim}{T}_A$ is the transition matrix of the row variable. These variables may be either simple or compound. Excluding the situation where both the true row and the true column variables are known there are three possible cases.

Case I: True Sub-populations,
Contaminated Response Profiles

Rosenberg (1979) has shown that in this case

$$\pi = \underset{\sim}{T}^{-1} \underset{\sim}{Q}$$

where $\underset{\sim}{T}^{-1} = (\underset{\sim}{T}_B \otimes \underset{\sim}{I}_s)^{-1}$ and $\underset{\sim}{I}_s$ is an identity matrix of dimension s. Thus, the model fit is

$$E_A\{\underset{\sim}{F}(\underset{\sim}{T}^{-1}\underset{\sim}{q})\} = \underset{\sim}{X}\underset{\sim}{\beta}.$$

Case II: True Response Profiles,
Contaminated Sub-populations

Rosenberg (1979) shows that

$$\underset{\sim}{\pi} = (\underset{\sim}{I}_r \otimes \underset{\sim}{A}^{-1})\underset{\sim}{Q}$$

where the matrix A is composed of elements a_{im} such that

$$a_{im} = \frac{T_{Aim}\ \pi_{Am}}{Q_{Am}}\ ,\ i,m=1,\ldots,s. \qquad (5.4)$$

The form of the functions $\underset{\sim}{g}^{-1}$ which can be computed in current implementations of the GSK methodology is shown in Case III.

Case III: Both Sub-populations and
Response Profiles Contaminated

Rosenberg (1979) shows that

$$\underset{\sim}{\pi} = (\underset{\sim}{T}_B^{-1} \otimes \underset{\sim}{A}^{-1})\ \underset{\sim}{Q}. \qquad (5.5)$$

where $\underset{\sim}{A}$ is defined as in formula (5.4).

The two most popular current computer programs that implement the GSK methodology are GENCAT (Landis et al 1976) and FUNCAT, a procedure in SAS (SAS Institute 1979). These programs can compute compound functions of the observed cell frequencies composed of four classes of functions: linear, logarithmic, exponential, or the addition of a vector of constants. Then for Cases II and III the s x r canonical table is entered into the program as a 1 x sr table and the estimate of $\underset{\sim}{p}$ of $\underset{\sim}{\pi}$ is computed as

$$\underset{\sim}{p} = \underset{\sim}{g}^{-1}(\underset{\sim}{q}) = \underset{\sim\sim\sim}{\exp}\ \underset{\sim}{A}_2\ \underset{\sim\sim}{\ln}\ \underset{\sim}{A}_1\ \underset{\sim}{q} \qquad (5.6)$$

where

$$\underset{s(r+1)\ \times\ sr}{\underset{\sim}{A}_1} = \left[\begin{array}{c} \underset{\sim}{T}_B^{-1} \otimes \underset{\sim}{T}_A^{-1} \\ sr \times sr \\ \hline \underset{\sim}{J}_r{}' \otimes \underset{\sim}{T}_A^{-1} \\ s \times sr \end{array}\right]$$

$$\underset{sr\ \times\ s\ (r+1)}{\underset{\sim}{A}_2} = [\underset{\sim}{I}_{sr} \mid \underset{sr\ \times\ s}{(-1)\underset{\sim}{J}_r \otimes I}\ ]$$

and

$\underset{\sim}{I}_{sr}$ = identity matrix of dimension sr x sr,

$\underset{\sim}{J}'_r$ = row vector of ones of dimension s x r,

$(-1)\underset{\sim}{J}_r$ = column vector of negative ones of dimension r x 1.

As stated previously, Case II is a specialization of Case III. In terms of the compound GENCAT function, the only difference occurs in matrix $\underset{\sim}{A}_1$ where the submatrix $\underset{\sim}{T}_B^{-1} \otimes \underset{\sim}{T}_A^{-1}$ now becomes $\underset{\sim}{I}_r \otimes \underset{\sim}{T}_A^{-1}$ since knowing the true X_B values is equivalent to $\underset{\sim}{T}_B$ being the identity matrix. Then for Case II, $\underset{\sim}{A}_1$ takes the form

$$\underset{\sim}{A}_1 = \left[\begin{array}{c} \underset{\sim}{I}_r \otimes \underset{\sim}{T}_A^{-1} \\ sr \times sr \\ \hline \underset{\sim}{J}'_r \otimes \underset{\sim}{T}_A^{-1} \end{array}\right]$$

6. AN EXAMPLE

Define a Form I transition matrix as having the form

$$\underset{\sim}{T}_1 = (d\text{-}b)\ \underset{\sim}{I} + b\ \underset{\sim\sim}{JJ'} = \begin{bmatrix} d & b & b & \ldots & b & b \\ b & d & b & \ldots & b & b \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ b & b & b & \ldots & b & d \end{bmatrix}$$

where the parameters d and b satisfy

$$d > b > 0 \text{ and } d + (K\text{-}1)\ b = 1.$$

and a Form II transition matrix as

$$\underset{\sim}{T}_2 = \begin{bmatrix} d & c & & & & \underset{\sim}{0} \\ 2c & d & c & & & \\ & c & d & \ddots & & \\ & & c & \ddots & c & \\ & & & \ddots & d & 2c \\ \underset{\sim}{0} & & & & c & d \end{bmatrix}$$

where c + 2d = 1.

A subset of data from cycle two of the Tecumseh Community Health Study (Epstein et al 1970) was contaminated by Form I and Form II transition matrices with d = 0.9. An analysis of hypertension by smoking category and relative weight was conducted. Hypertension and Relative Weight were contaminated by Form I transition matrices and Smoking Status by a Form II transition matrix.

Three tables were analyzed: all true data; true sub-populations, but contaminated response profiles (Case I); and all contaminated data (Case III). The three contingency tables are shown in Table 1. Using (5.3) for Case I and (5.5) for Case III, estimated cell probabilities corrected for the contamination may be computed and are shown in Table 1. Overall, the ARRC estimated tables show excellent agreement with the true table.

One application of the GSK approach is as a categorical analogue to ANOVA. This will be demonstrated in the case of ARRC data. The theoretical details of this application are not discussed fully here, but may be found in

TABLE 1

PREVALENCE OF HYPERTENSION BY RELATIVE WEIGHT AND SMOKING STATUS IN ADULTS AGES 18-95*

		True Data		Case I		Case III	
Sub-population		Hypertension		Hypertension		Hypertension	
Rel. Wt.	Smoking Status	Normal	Hyper	Normal	Hyper	Normal	Hyper
L	None	1001 .7020	425 .2980	925 .7095	474 .2905	781 .6936	421 .3064
L	Ex	199 .6461	109 .3539	185 .6258	123 .3742	273 .6568	162 .3432
L	Cigar & Pipe	106 .6023	70 .3977	105 .6207	71 .3793	201 .5338	128 .4662
L	Cigarette	1260 .7655	386 .2345	1160 .7559	486 .2441	958 .7602	403 .2398
U	None	280 .5214	257 .4786	280 .5268	257 .4732	319 .5469	247 .4531
U	Ex	75 .5435	63 .4565	68 .4909	70 .5091	106 .4974	91 .5026
U	Cigar & Pipe	30 .4225	41 .5775	31 .4208	40 .5792	75 .4147	68 .5853
U	Cigarette	248 .5822	178 .4178	244 .5910	182 .4090	312 .6246	183 .3754

*Cell proportions are shown below frequencies. Case I and Case III cell proportions are ARRC estimates.

Johnson and Koch (1971). As stated previously, the GSK approach fits models of the form $E_A\{\underset{\sim}{F}(\underset{\sim}{p})\} = \underset{\sim}{X}\underset{\sim}{\beta}$. The function $\underset{\sim}{F}(\underset{\sim}{p})$ used in this analysis is $\underset{\sim}{A}\underset{\sim}{p}$ where $\underset{\sim}{A} = (0 \quad 1) \otimes \underset{\sim}{I}_8$. This has the effect of selecting the proportion of hypertensives in each sub-population.

For each table, the first model fit was the saturated model containing terms for: overall mean (β_1); relative weight (β_2); smoking status (β_3,β_4,β_5); and smoking status x relative weight interaction (β_6,β_7,β_8). Since, in all three cases the interaction term failed to differ significantly from zero, an additive model which only includes terms for overall mean, relative weight, and smoking status, was fit for each case. Estimated model parameters, their standard deviations, and associated test statistics are shown in Table 2. Each additive model exhibits a satisfactory fit as evidenced by the lack of fit statistic which in each case satisfied Johnson and Koch's (1971) suggested criterion of $\chi^2 < 3.84$. The additive models show that both relative weight and smoking status have a significant association with hypertension.

In all three cases the conclusions drawn from the hypothesis tests are equivalent, although the anticipated loss of power in the ARRC estimated models is manifested by a substantial decrease in the value of the test statistics. The formula for computing Scheffe-type confidence intervals, discussed in GSK (1969), can be used to construct a 95% confidence interval for single comparisons given by $\underset{\sim}{\beta} \pm 1.96$ S.E. (β). All ARRC estimated coefficients lie within 95% confidence interval limits constructed about the coefficients estimated from the true data. This is indicative of excellent agreement in this example.

TABLE 2

ESTIMATED PARAMETERS AND CHI-SQUARE TESTS OF ADDITIVE MODEL

		True Data	Case I	Case III
Mean β_0		.4038	.4075	.4092
Rel. Wt. β_1		-.0867	-.0858	-.0711
Smoking β_2		-.0174	-.0284	-.0306
	β_3	.0161	.0415	.0093
	β_4	.0824	.0656	.1208
χ^2_{RW}		111.4	66.5	26.8
p-value		<.0001	<.0001	<.0001
χ^2_{smoking}		43.8	24.9	24.4
p-value		<.0001	<.0001	<.0001
$\chi^2_{\text{Lack of Fit}}$		2.21	0.61	0.09
p-value		.5302	.8946	.9931

7. DISCUSSION

ARRC can be an effective method of preventing statistical disclosure and is especially well suited for moderately large data sets such as the Tecumseh data set where the risk of disclosure without using a disclosure control technique is high and other techniques are not suitable. Rosenberg (1979) has developed other techniques for contaminating and analyzing continuous data.

The "attained transition matrix" of a variable Z can be defined as the transition matrix composed of elements t_{ij} equal to the number of observations with X = j. Then if the collector were to release the contaminated variable Z together with its attained transition matrix, it is possible that "exact" techniques, which yield results identical to analyses performed on the true data, could be developed. Work is proceeding in this area.

ACKNOWLEDGMENTS

The research for this paper was supported in part by a National Research Service Award training grant #S T32 HL07032-03. I'd like to thank the members of my doctoral committee for their help.

REFERENCES

1. Epstein, F.H., Napier, J.A., Block, W.D., Hayner, N.S., Higgins, M.P., Johnson, B.C., Keller, J.B., Metzner, H.L., Montoye, H.J., Ostrander, L.D. and Ullman, B.M. (1970). The Tecumseh study: design, progress, and perspectives. Archives of Environmental Health 21, 402-407.

2. Grizzle, J.E., Starmer, F., and Koch, G.G. (1969). Analysis of Categorical data by Linear models. Biometrics 25, 489-504.

3. Johnson, W.D., and Koch, G.G. (1971). A note on the weighted least squares analysis of the Ries-Smith contingency table data. Technometrics 13, 438-447.

4. Landis, J.R., Stanish, W.M., Freeman, J.L. and Koch, G.G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). Computer Programs in Biomedicine 6, 196-231.

5. Rosenberg, M.J., (1979). Multivariable Analysis by a Randomized Response Technique for Statistical Disclosure Control. Ph.D. thesis, University of Michigan, Ann Arbor, Michigan.

6. SAS Institute (1979). SAS User's Guide, 1979 Edition. J.T. Helwig and K.A. Council, editors. SAS Institute, Inc., Raleigh, N.C.

7. Warner, S.L. (1971). The linear randomized response model. Journal of the American Statistical Association 66, 884-888.