This research was supported by National Institutes of Health, Institute of General Medical Sciences Grants GM-70004-01, GM-12868-08 and by the National Center for Health Statistics, HSMHA.

AN APPLICATION OF MULTIVARIATE ANALYSIS TO COMPLEX SAMPLE SURVEY DATA

by

Gary G. Koch and Stanley Lemeshow

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 802

February 1972

AN APPLICATION OF MULTIVARIATE ANALYSIS TO COMPLEX SAMPLE SURVEY DATA

by

Gary G. Koch
Department of Biostatistics
University of North Carolina
Chapel Hill, N.C. 27514

Stanley Lemeshow
National Center for Health Statistics
Public Health Service
Rockville, Maryland 20852

In this paper, we shall consider an example which resulted from the analysis of data from the second Health Examination Survey of the National Center for Health Statistics. This program is concerned with certain aspects of health of children between 6 and 11 years of age and the problem of interest here arose during an attempt to determine whether Negro children of a particular age had the same height and weight as White children of the same age.

The Health Examination Survey (HES) is based on a complex design which is highly stratified and involves a multistage probability method of selection. Hence, standard methods of multivariate analysis (e.g., see Anderson [1], Morrison [9], or Rao [11]) are not directly applicable since they, for the most part, pertain only to data obtained from simple random samples (with replacement from a population characterized by a multivariate normal probability distribution). On the other hand, since the sample sizes in such situation are often very large, it generally can be assumed (on the basis of Central Limit Theory) that the estimates of population means for a set of variables do have a multivariate normal distribution. Moreover, for purposes associated with comparisons of these population means, the covariance matrix of this multivariate normal distribution can be regarded as known in the sense of being approximately equal to any valid and consistent estimate. One approach to the formulation of valid and consistent estimates of variance for statistics calculated from sample survey data is the method of balanced repeated replication as discussed, for example, by Frankel [2], Kish and Frankel [4,5], and McCarthy [6,7,8].

In Table 1, estimates from this second HES are shown for the mean height (in centimeters) and mean weight (in kilograms) of both Negro and White 6 year old, male children (see Hamill, Johnston, and Grams [3]).

TABLE 1

ESTIMATES OF MEAN HEIGHT AND MEAN WEIGHT
TOGETHER WITH CORRESPONDING STANDARD ERRORS

	Negro Males, 6 Years Old		White Males, 6 Years Old	
	Estimated Mean	Standard Error	Estimated Mean	Standard Error
Height (cm)	119.12	0.72	118.54	0.30
Weight (kg)	21.76	0.37	22.04	0.18

These results were obtained as weighted averages of the type

$$\bar{\underset{\sim}{y}}_j = \{ \sum_{i=1}^{n} W_i U_{ij} \underset{\sim}{y}_i \} / \{ \sum_{i=1}^{n} W_i U_{ij} \} \qquad (1)$$

where

i. $n = 7119$ denotes the number of children given physical examinations

ii. W_i denotes an individual statistical weight corresponding to the i-th child in the survey so that $W = \sum_{i=1}^{n} W_i = 23{,}784{,}072$ where W represents the estimated number of children between ages 6 and 11 years living in the United States during the period 1963-1965.

iii. U_{ij} is an indicator random variable for the j-th sub-population category (e.g., Negro, 6 year old, male children) so that

$$U_{ij} = \begin{cases} 1 \text{ if i-th child in j-th category} \\ 0 \text{ otherwise} \end{cases}$$

iv. $$\underset{\sim}{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} \text{Height of i-th child} \\ \text{Weight of i-th child} \end{bmatrix}$$

v. $$\bar{\underset{\sim}{y}}_j = \begin{bmatrix} \text{mean height in j-th sub-population category} \\ \text{mean weight in j-th sub-population category} \end{bmatrix}$$

If j=1 corresponds to Negro, 6 year old, male children and j=2 to White, 6 year old, male children, then a consistent estimate of the covariance matrix of the composite mean vector $\bar{\underset{\sim}{y}}$ where $\bar{\underset{\sim}{y}}' = (\bar{\underset{\sim}{y}}_1', \bar{\underset{\sim}{y}}_2')$ can be obtained by using the method of balanced half samples. For Cycle II of HES, 20 balanced half samples were used, and the resulting estimates $\bar{\underset{\sim}{y}}^{(k)}$ of mean height and mean weight for Negro and White, 6 year old male children are shown in Table 2 for each half sample k where k = 1,2,...,20.

TABLE 2

HALF-SAMPLE ESTIMATES

Half Sample	Negro Males, 6 Years Old		White Males, 6 Years Old	
	Height(cm)	Weight (kg)	Height(cm)	Weight (kg)
1	119.38	21.76	118.30	21.93
2	119.25	21.51	118.27	22.03
3	119.33	21.85	119.25	22.44
4	118.33	21.32	118.73	22.23
5	118.37	21.50	119.13	22.23
6	119.38	21.73	118.53	21.98
7	119.11	21.62	118.45	22.00
8	118.92	21.72	118.83	22.16
9	119.16	21.87	118.86	22.42
10	119.40	21.91	118.44	22.03
11	119.53	22.04	118.58	22.07
12	118.85	21.57	118.40	22.05
13	119.82	22.23	118.36	22.13
14	118.73	21.82	118.65	22.03
15	118.10	21.19	118.47	22.14
16	119.81	21.67	118.67	21.86
17	116.72	20.71	119.22	22.33
18	118.79	21.57	118.34	21.78
19	119.86	22.40	118.66	22.00
20	118.86	21.50	118.61	22.10

More explicitly, $\bar{\underset{\sim}{y}}^{(k)}$ was determined by applying (1) only to the children in the k-th half sample; i.e.,

$$\underset{\sim}{\bar{y}}^{(k)} = \begin{bmatrix} \underset{\sim}{\bar{y}}_1^{(k)} \\ \underset{\sim}{\bar{y}}_2^{(k)} \end{bmatrix} \quad \text{where} \quad \underset{\sim}{\bar{y}}_j^{(k)} = \{ \sum_{i=1}^{n} W_i U_{ij} H_{ik} \underset{\sim}{y}_i \} / \{ \sum_{i=1}^{n} W_i U_{ij} H_{ik} \} \qquad (2)$$

with H_{ik} being a known indicator function such that

$$H_{ik} = \begin{cases} 1 \text{ if individual i is in k-th half sample} \\ 0 \text{ otherwise} \end{cases} \qquad (3)$$

From these results, an estimate $\underset{\sim}{V}$ for the covariance matrix corresponding to $\underset{\sim}{\bar{y}}$ is given by

$$\underset{\sim}{V} = \frac{1}{20} \sum_{k=1}^{20} [(\underset{\sim}{\bar{y}}^{(k)} - \underset{\sim}{\bar{y}})(\underset{\sim}{\bar{y}}^{(k)} - \underset{\sim}{\bar{y}})'] . \qquad (4)$$

Here, it should be noted that the appropriate choice of the matrix $H = (H_{ik})$ represents a very important part of this procedure for determining $\underset{\sim}{V}$. Some efficient strategies for this purpose are described in McCarthy [6 , 7]. In this regard, Simmons and Baird [12] have discussed certain aspects of the application of the balanced half sample approach to the Health Examination Survey.

For the data under consideration here, the covariance matrix $\underset{\sim}{V}$ obtained from (4) is given in (5).

$$\underset{\sim}{V} = \begin{bmatrix} 0.5201 & 0.2348 & -0.1030 & -0.0530 \\ 0.2348 & 0.1333 & -0.0354 & -0.0166 \\ -0.1030 & -0.0354 & 0.0907 & 0.0407 \\ -0.0530 & -0.0166 & 0.0407 & 0.0306 \end{bmatrix} \qquad (5)$$

The square roots of the diagonal elements of $\underset{\sim}{V}$ are the estimated standard errors for the corresponding estimated means; and these are displayed as such in Table 1.

Since the sample size n = 7119 is large, it can be argued that $\underset{\sim}{\bar{y}}$ has approximately a multivariate normal distribution with known covariance matrix $\underset{\sim}{V}$. This condition is equivalent to the statement that every linear function $\underset{\sim}{c}'\underset{\sim}{\bar{y}}$

(where $\underset{\sim}{c}$ is any known (4 x 1) vector of constants) has a univariate normal distribution with known variance $\underset{\sim}{c}'\underset{\sim\sim}{V}\underset{\sim}{c}$. Hence, for any hypothesis which implies that the expected value of $\underset{\sim}{c}'\bar{\underset{\sim}{y}}$ is zero, an appropriate test statistic is

$$X^2 = (\underset{\sim}{c}'\bar{\underset{\sim}{y}})^2 \; / \; \underset{\sim}{c}'\underset{\sim\sim}{V}\underset{\sim}{c} \qquad (6)$$

which has approximately a chi-square distribution with D.F. = 1 under this condition when n is large. To compare the mean height of Negro, 6 years old, male children with the mean height of White, 6 year old, male children we use

$$\underset{\sim}{c}' = [1 \quad 0 \quad -1 \quad 0]$$

and obtain

$$\underset{\sim}{c}'\bar{\underset{\sim}{y}} = 0.58 \qquad \underset{\sim}{c}'\underset{\sim\sim}{V}\underset{\sim}{c} = 0.8168$$

$$X^2 = 0.41 \quad ,$$

while to compare their mean weights, we use

$$\underset{\sim}{c}' = [0 \quad 1 \quad 0 \quad -1]$$

and obtain

$$\underset{\sim}{c}'\bar{\underset{\sim}{y}} = -0.28 \qquad \underset{\sim}{c}'\underset{\sim\sim}{V}\underset{\sim}{c} = 0.1971$$

$$X^2 = 0.40$$

For both of these comparisons, the values of X^2 were not significant ($\alpha = .25$), suggesting that the two sub-populations were similar with respect to average height and weight. However, before stating this result as a conclusion, it is worthwhile to consider a multivariate test in which these two variables are treated jointly and simultaneously. In this event, the appropriate test statistic is given in (7)

$$X^2 = \bar{\underset{\sim}{y}}'\underset{\sim}{C}' \; [\underset{\sim\sim\sim}{CVC'}]^{-1}\underset{\sim\sim}{C}\bar{\underset{\sim}{y}} \qquad (7)$$

where $\underset{\sim}{C}$ in (8) is a matrix which reflects both comparisons.

$$C = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \qquad (8)$$

Under the hypothesis that Negro and White, 6 year old, male children have the same average height and weight, X^2 has approximately a chi square distribution with D.F. = 2 when n is large. In this case $X^2 = 8.70$ which is significant at the $\alpha = 0.05$ level. Hence, we have a situation in which a multivariate test is more sensitive at detecting the difference between two populations than separate univariate tests. One reason for this result is that these data exhibit what can be called a "cross-over" effect. This condition arises when the directions of the differences between two groups with respect to two positively correlated variables are reversed (i.e., have opposite signs). Here, the estimated correlation of height and weight is 0.89 for Negro, 6 year old males and 0.77 for White, 6 year old males. However, the Negro children are 0.58 cm. taller but weigh 0.28 kg. less than the White children.

With these considerations in mind, it is worthwhile to reconsider the relative merits of the multivariate and univariate approaches. When this is done, four possible situations need to be recognized. These are designated in Table 3 with Case II being applicable to the data considered here.

TABLE 3

COMPARISON OF MULTIVARIATE AND UNIVARIATE ANALYSES

		Multivariate Test	
		Non-Significant	Significant
Separate Univariate Tests	None Significant	I. Groups are interpreted as similar	II. Groups are interpreted as different, but the difference cannot be explained in terms of any of the variables individually.
	Some Significant	III. Caution is exercised, but groups are interpreted as different with respect to the univariate variables that are significant. The problem here is that the greater scope of the multivariate test is often bought at the price of reduced power.	IV. Groups are interpreted as different, and the difference can usually be explained in terms of the univariate variables that are significant.

Before concluding this discussion, it is useful to recognize two questions which are consequences of this approach to multivariate analysis for complex sample survey data. The first pertains to recognizing that $\underset{\sim}{V}$ is essentially a sample covariance matrix function of the $\underset{\sim}{\bar{y}}^{(k)}$. Hence, alternative test statistics to those defined in (6) and (7) are F-transformed Hotelling-T^2-analogues determined from the formula in expression (9) where p is the rank of

$$Q = \frac{(20 - p)}{19p} \; X^2 \qquad (9)$$

the corresponding $\underset{\sim}{C}$ matrix. If the groups are similar, Q has approximately the F-distribution with D.F. = (p,20-p). For the comparison of height, $Q = 0.41$ with D.F. = (1,19); and for weight, $Q = 0.40$ with D.F. = (1,19). Finally, for the joint comparison, $Q = 4.12$ with D.F. = (2,18) which is significant at $\alpha = 0.05$ as was the corresponding X^2. Thus, these results are similar to what has been obtained previously. In certain situations where the estimate of $\underset{\sim}{V}$ is not considered stable, it may be more appropriate to use Q instead of X^2 as a manner of accounting for such variation. However, if the sample size is large and a good estimate of $\underset{\sim}{V}$ has been determined, then X^2 should be preferred because of its added power. For a further discussion of the Q and X^2 statistics, see McCarthy [6,7] and Nathan [10].

The other question results from the fact that if this type of analysis were applied to twenty or more variables, $\underset{\sim}{V}$ would usually be singular. Thus, statistics like Q and X^2 could not be computed for such higher dimensional comparisons (i.e., when rank of $\underset{\sim}{C}$ exceeds 20).

One approach to dealing with this problem, which is based on the results of Kish and Frankel [5], involves a two-stage method. First the weighted measurements $\underset{\sim i}{y^*} = (W_i \underset{\sim i}{y_i} / W)$ are determined for each individual, and from these values the corresponding sample covariance matrix $\underset{\sim}{V}^*$ under the assumption of simple random sampling is determined. In these calculations,

different groups of individuals are regarded as independent and are treated separately. In the second stage, the balanced half sample approach is used to determine more appropriate estimates of the variance for the respective variables in each group. If $\underset{\sim}{D}$ represents a diagonal matrix in which the diagonal elements are the square roots of the ratio of the variances determined from the balanced half sample approach divided by corresponding diagonal elements of $\underset{\sim}{V}^*$ respectively, then the estimate $\underset{\sim}{V}$ is $\underset{\sim}{V} = \underset{\sim\sim}{D}\underset{\sim}{V}^*\underset{\sim}{D}$. This estimate of $\underset{\sim}{V}$ has some intuitive appeal, but further research is needed to determine its statistical properties. Nevertheless, the other aspects of the multivariate analysis would be as previously described.

REFERENCES

1. Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley & Sons, 1958.

2. Frankel, M.R., Inference from Survey Samples: An Empirical Investigation, Institute for Social Research, The University of Michigan, 1971.

3. Hamill, P.V.V., Johnston, S.E., and Grams, W. Height and Weight of Children, United States, 1963-1965. Washington: National Center for Health Statistics, Series 11, No. 104, 1970.

4. Kish, L. and Frankel, M.R., "Balanced Repeated Replication for Analytical Statistics", Proceedings of the Social Statistics Section of ASA, 1968.

5. Kish, L. "Balanced Repeated Replications for Standard Errors", Journal of the American Statistical Association, 65 (1970), 1071-1094.

6. McCarthy, P.J., Replication: An Approach to the Analysis of Data from Complex Surveys. Washington: National Center for Health Statistics, Series 2, No. 14, 1966.

7. McCarthy, P.J., Pseudo-Replication: Further Evaluation and Application of the Balanced Half-Sample Technique. Washington: National Center for Health Statistics, Series 2, No. 31, 1966.

8. McCarthy, P.J., "Pseudo-Replication: Half Samples", Review of the International Statistical Institute, 37 (1969), 239-64.

9. Morrison, D.F., Multivariate Statistical Methods, McGraw-Hill Book Company, 1967.

10. Nathan, G., "Approximate Tests of Independence in Contingency Tables From Complex Stratified Cluster Samples", Unpublished report prepared for National Center for Health Statistics.

11. Rao, C.R., Linear Statistical Inference and Its Applications, John Wiley & Sons, 1965.

12. Simmons, W.R. and Baird, J., "Use of Pseudo-Replication in the NCHS Health Examination Survey", Proceedings of the Social Statistics Section of ASA, 1968.