Sankhyā : The Indian Journal of Statistics
1989, Volume 51, Series B, Pt. 2, pp. 263-269.

AN ESTIMATOR OF THE CUTOFF POINT MAXIMIZING SUM OF SENSITIVITY AND SPECIFICITY

By RADHESHYAM BAIRAGI
Indian Statistical Institute
and
C. M. SUCHINDRAN
University of North Carolina at Chapel Hill

SUMMARY. A new estimator of the cutoff point at which the sum of sensitivity and specificity of an indicator is maximum is derived and some statistical properties of this estimator are investigated. Using this method, an estimate and its variance of the cutoff point of the anthropometric index weight-for-age as an indicator of mortality is obtained from a Bangaladesh study. A situation is described where this seems to be the only estimator obtainable from data.

1. Introduction

In many cases, the prevalence of a disease or the proportion of persons at the risk of a disease is estimated by an indicator that is easier to measure than the disease itself. For example, protein-energy malnutrition (PEM) is often measured by an anthropometric indicator (Waterlow *et al.*, 1977), and in turn this indicator may be used as a predictor of mortality (Bairagi *et al.*, 1985) and morbidity (Black *et al.*, 1984). Two important characteristics of an indicator are its sensitivity (probability of diseased correctly identified) and specificity (probability of nondiseased correctly identified). If the true proportion of diseased persons in the population be P, the diagnosed proportion of diseased persons in the population, p, is (Habicht, 1980) :

$$p = P \times se + (1-P) \times (1-sp) \qquad \dots \quad (1.1)$$

Diagnosed proportion p is usually a biased estimator of true proportion P (Rogan and Gladen, 1976). However, to monitor the disease state in a population over time and to compare the proportions of diseased between two populations on the basis of the diagnosed proportions, one would like to see change in p is as large as possible for a change in P. Since this change is

$$\frac{\delta p}{\delta P} = (se + sp - 1), \qquad \dots \quad (1.2)$$

AMS (1980) *subject classification :* 62P30

Key words and phrases : Cutoff point ; Estimator ; Sensitivity ; Specificity.

one would like to set up the identification procedure in such a way that the sum of sensitivity and specificity (SSS) is maximum. The SSS of an indicator with interval scale depends on its cutoff point, as shown in Figure 1.

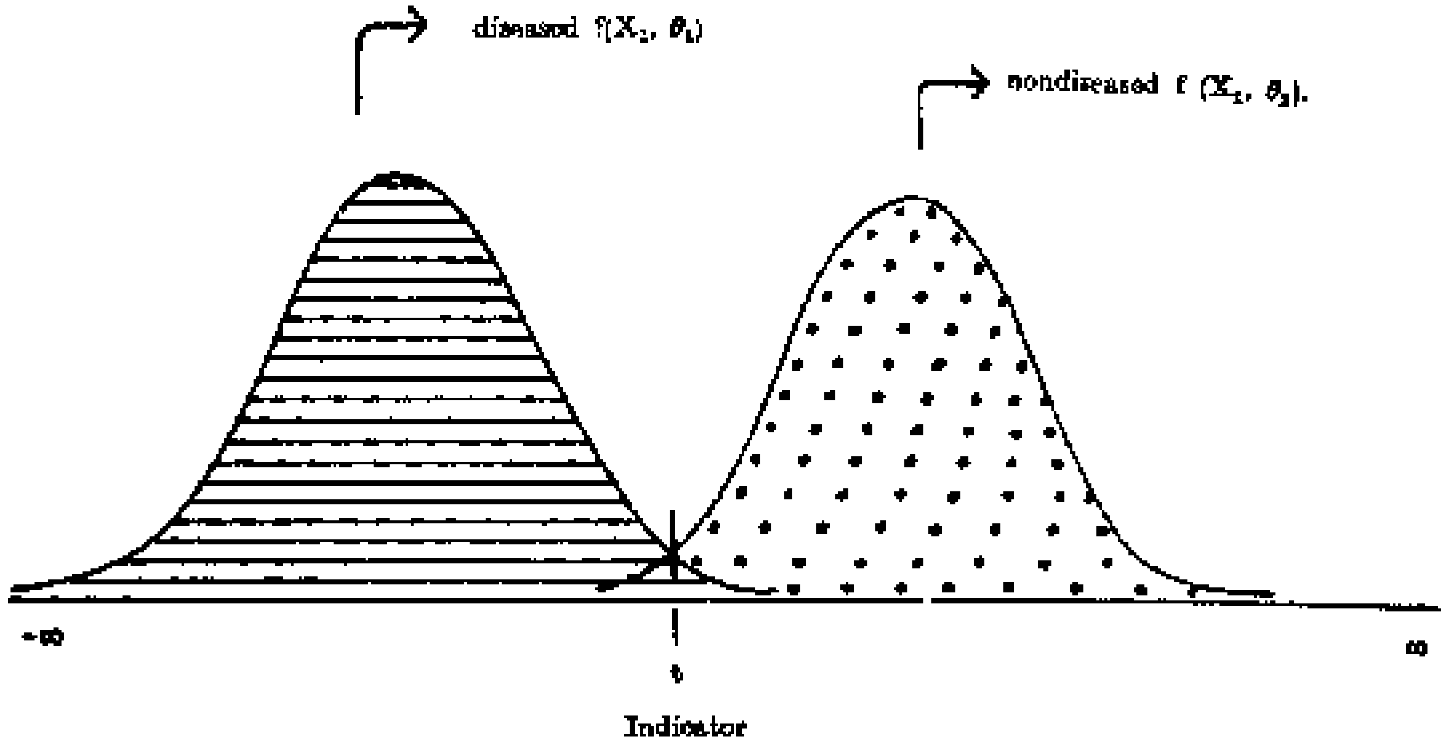

Fig. 1. A hypothetical distribution of the values of an indicator of diseased and nondiseased persons.

If t is the cutoff point, *se* is the lined area to the left of t and under $f(X, \theta_1)$; and *sp* is the dotted area to the right of t and under $f(X, \theta_2)$. If $f(X, \theta_1)$ and $f(X, \theta_2)$ both are Gaussian, SSS, Φ, will be maximum at (Bairagi, 1981 : 34-36)

$$t = \left[-(\mu_2\sigma_1^2 - \mu_1\sigma_2^2) + \left\{ (\mu_2\sigma_1^2 - \mu_1\sigma_2^2) - (\sigma_2^2 - \sigma_1^2)\ (\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2) - 2\sigma_1^2\sigma_2^2\, ln\, \frac{\sigma_2}{\sigma_1} \right\}^{1/2} \right] / (\sigma_2^2 - \sigma_1^2), \qquad \ldots \quad (1.3)$$

$$\theta_1 = (\mu_1, \sigma_1^2) \text{ and } \theta_2 = (\mu_2, \sigma_2^2).$$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

$$t = \frac{\mu_1 + \mu_2}{2}. \qquad \ldots \quad (1.4)$$

$\hat{t} = \dfrac{\bar{X}_1 + \bar{X}_2}{2}$ is the maximum likelihood estimator (MLE) of t in equation (1.4.)

Usually, the distribution functions of the indicators of diseased and nondiseased persons remain unknown, and a parametric MLE cannot be obtained. However, if the frequency distributions of diseased and nondiseased persons are known SSS can be obtained at different cutoff points, and the maximum sum of sensitivity and specificity (MSSS) and its corresponding cutoff point can be obtained numerically as shown in Table 1.

TABLE 1. DISTRIBUTION OF INDICATOR (WEIGHT-FOR-AGE) OF BANGLADESHI CHILDREN AGED 12–13 MONTHS BY THEIR MORTALITY STATUS IN FIRST 24 MONTHS FOLLOWING ANTHROPOMETRY AND THE CORRESPONDING SENSITIVITY, SPECIFICITY, AND PROPORTION DIED

indicator	dead	alive	se	sp	SSS	proportion died
(1)	(2)	(3)	(4)	(5)	(6)	(7)
< 49	13	69	12	96	108	0.159
50–54	17	89	27	91	118	0.160
55–59	15	190	41	81	122	0.073
60–64	24	296	63	65	128	0.075
65–69	11	399	73	44	117	0.027
70–74	12	324	84	27	111	0.036
75–79	10	315	93	10	103	0.031
80–84	6	136	98	2	100	0.042
85+	2	47	100	0	100	0.041
all	110	1,865				0.056
			$\bar{t}_s = 64.5$			

The weight-for-age (see Table 1) is defined as the ratio of the weight of the study child to the weight of the reference child multiplied by 100 (Bairagi, 1986). This indicator has been widely used throughout the World to investigate the nutritional status of children and it has been found to be a good perdictor of mortality in many studies (Bairagi *et al.*, 1985 ; Kielmann and McCord, 1978). Some characteristics of this indicator are available in Waterlow *et al.* 1977. In Table 1, the distribution of weight-for-age of children by their mortality status in the first 24 months following anthropometry, the corresponding sensitivity and specificity, and mortality rate are given from a Bangladesh study (Bairagi, 1981 : 55-70). Here sensitivity expressed in percent at a cutoff point t is the proportion of children having a weight-for-age value less than or equal to t among those children who died in the 24 months following anthropometry. This is the percent of children who died correctly classified at the cutoff point t at the time of anthropometry. Similarly, specificity expressed in percent at a cutooff point t is the proportion of children having a weight-for-age value greater or equal to t among those children alive after 24 months following anthropometry. This is the percent of children alive correctly classified at the cutoff point t at the time of anthropometry. Specificity at the cutoff point t in the table may also be defined as

100 minus the cumulative percent of children alive to the point t. MSSS in this table belongs to the weight-for-age value of 60–64. Therefore, the numeric estimate, t_n, equals 64.5, the mid-value between the end point of 60–64 and the beginning point of 65–69.

This cutoff point of MSSS can be obtained graphically using the information on proportion of diseased or dead (column 7 in Table 1) at different values of the indicator. That estimator is obtained below.

2. Graphic estimator $\hat{t}_g$

Let $f(x, \theta)$ be the density function of the indicator for the combined group (diseased and nondiseased), and let the proportion of diseased in the population at x be $R(x)$.

Therefore

$$N = \text{Total number of persons} = N \int_{-\infty}^{\infty} f(x, \theta)\, dx,$$

and

$$D = \text{Number of diseased persons} = N \int_{-\infty}^{\infty} f(x, \theta)\, R(x)\, dx.$$

Let t be the cutoff point. Therefore

$$se = N \int_{-\infty}^{t} f(x, \theta) \,.\, R(x)\, dx / D,$$

and

$$sp = \left[N - N \int_{-\infty}^{t} f(x, \theta)\, dx - N \int_{t}^{\infty} f(x, \theta) \,.\, R(x)\, dx \right] / [N-D].$$

The sum of se and sp is

$$\Phi = \frac{N \int_{-\infty}^{t} f(x, \theta) \cdot R(x)\, dx}{D} + \frac{N - N \int_{-\infty}^{t} f(x, \theta)\, dx - N \int_{t}^{\infty} f(x, \theta) \cdot R(x)\, dx}{N-D}$$

$$\frac{\partial \Phi}{\partial t} = \frac{N \cdot f(t, \theta) \,.\, R(t)}{D} + \frac{-N \cdot f(t, \theta) + N \cdot f(t, \theta) \cdot R(t)}{N-D}$$

Setting $\dfrac{\partial \Phi}{\partial t} = 0$,

$$R(t) = \frac{D}{N} \cdot \qquad \ldots \quad (2.1)$$

$$t = R^{-1} \left(\frac{D}{N} \right) \qquad \ldots \quad (2.2)$$

In the Gaussian cases described in (1.3)

$$R(t) = \frac{\left[\left(\frac{D}{N}\right)(\sigma_1^2 \cdot 2\pi)^{-1/2} \exp\left\{-(2^{-1}\sigma_1^{-2})(t-\mu_1)^2\right\}\right]}{\left[\left(\frac{D}{N}\right)(\sigma_1^2 \cdot 2\pi)^{-1/2} \exp\left\{-2^{-1}\sigma_1^{-2})(t-\mu_1)^2\right\} + \left(1-\frac{D}{N}\right)\cdot(\sigma_2^2 \cdot 2\pi)^{-1/2} \exp\left\{-(2^{-1}\sigma_2^{-2})(t-\mu_2)^2\right\}\right]^{-1}}$$

Putting this value of $R(t)$ in (2.1), t in (1.3) can be obtained easily.

From (2.2) it appers that at point t, where the proportion of diseased persons is $\frac{D}{N}$ on the $R(x)$ curve SSS will be maximum or minimum. Replacing population value D, N, and $R(x)$ by sample value d, n, and $r(x)$ where $r(x)$ estimates $R(x)$—the graphic estimator, $\tilde{t}_g$, can be obtained as shown in Figure 2 using data from Table 1.

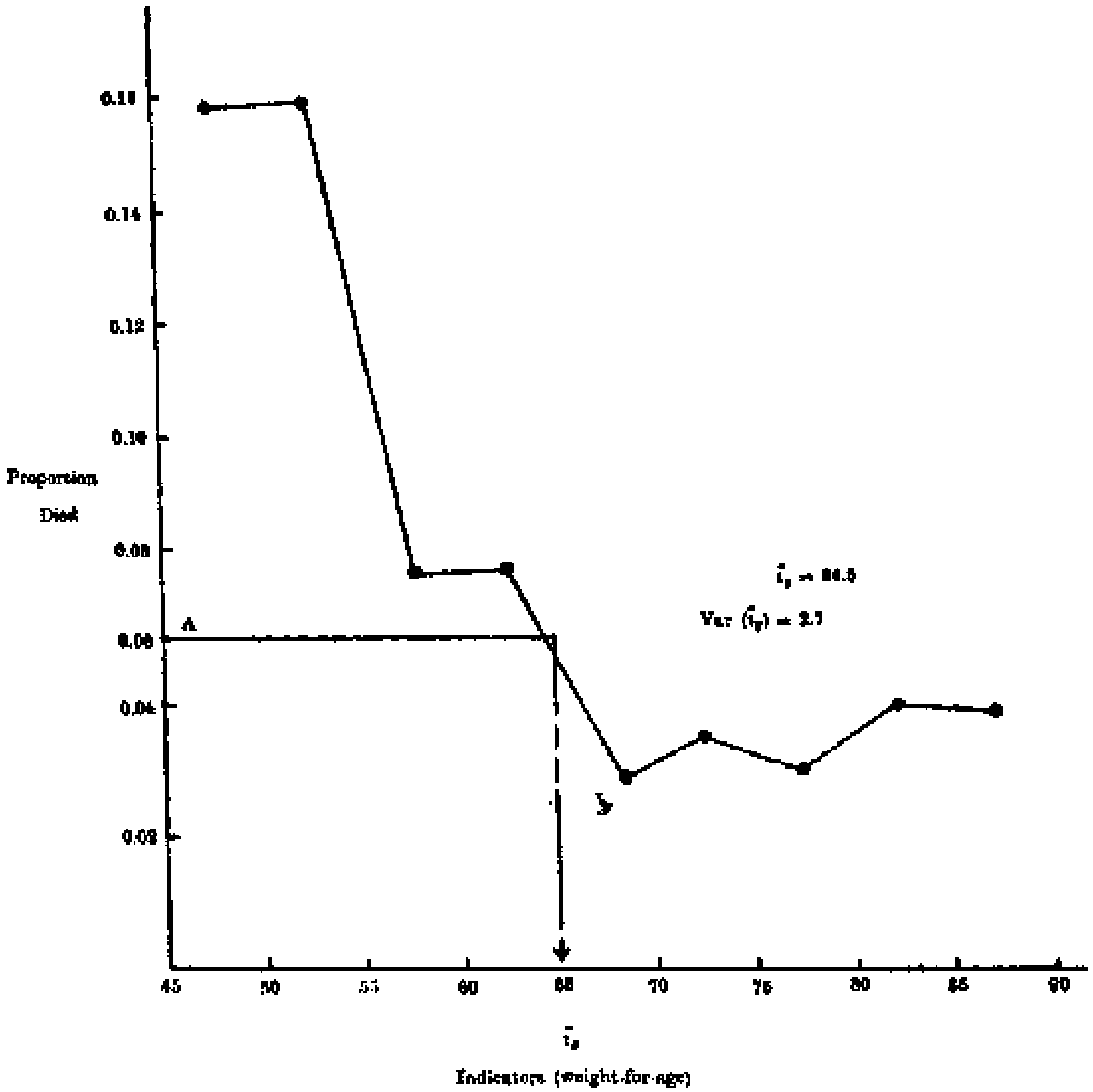

Fig. 2. Estimation of the graphic estimate.

A parallel line, AB, from point $\frac{d}{n} = 0.056$ on the y-axis to the x-axis cuts $r(x)$ at B. From B, the perpendicular on the x-axis will meet at the required point, $\widetilde{t}_g$. It should be noted that the numeric estimator obtained in Table 1 and the graphic estimator obtained in Figure 1 are the same. Some discrepancy between the graphic estimate and the numeric estimate in a sample may occur due to grouping of observations and rounding off values. And if the disease or death curve $R(x)$ is not a monotonically decreasing or increasing function of x, there may be more than one cutoff point at which SSS will be maximum or minimum depending on the sign of the slopes of $R(x)$.

3. Statistical properties of $\widetilde{t}_g$

The sample proportion of diseased person $\frac{d_i}{n_i}$ in the i-th class of the indicator is a maximum likelihood estimator of $\frac{D_i}{N_i}$. The disease curve $R(x)$ in the population is obtained by joining $\frac{D_i}{N_i}$ for different values of i. Similarly, $r(x)$ curve in the sample is obtained by joining $\frac{d_i}{n_i}$. On the other hand, $\frac{d}{n}$ is a maximum likelihood estimator of $\frac{D}{N}$. By the invariance property of the maximum likelihood estimator (Mood *et al.*, 1974 : 284-286), $\widetilde{t}_g = r^{-1}\left(\frac{d}{n}\right)$ is a maximum likelihood estimator of $t = R^{-1}\left(\frac{D}{N}\right)$. Therefore, $\widetilde{t}_g$ is consistent and asymptotically unbiased.

It is reasonable to assume that the $R(x)$ curve is a monotonically decreasing or increasing function of x at the neighbouring points of $t = R^{-1}\left(\frac{D}{N}\right)$. Under that assumption and according to Cramer's general convergence theorem (1946 : 253-255)

$$\text{var}\left(\frac{d}{n}\right) = \beta_1^2 \text{ var }(\widetilde{t}_g),$$

where β_1 is the slope of $R(x)$ at $\frac{D}{N}$ and can be estimated from a few neighboring points of $r^{-1}\left(\frac{d}{n}\right)$and their corresponding $r(x)$. Estimated variance of $\widetilde{t}_g$ is

$$\text{var }(\widetilde{t}_g) = \frac{1}{\widetilde{\beta}_1^2} \cdot \frac{1}{n} \cdot \frac{d}{n}\left(1 - \frac{d}{n}\right). \qquad \dots \quad (3.1)$$

In the example given in Figure 1, β_1 was estimated regressionally as 0.00318 from the following four points : (57.5, 0.073), (62.5, 0.075), (67.5, 0.027) and (72.5, 0.036). The estimated variance of $\tilde{t}_g$ was

$$\text{var}\ (\tilde{t}_g) = \frac{1}{0.00318^2} \times \frac{1}{1975} \times 0.056 \times 0.944 = 2.7.$$

4. Conclusions

In most situations, a parametric MLE or any other parametric estimators of t will not be obtainable, and one will have to depend on the numeric estimator, $\tilde{t}_n$ or graphic estimator, $\tilde{t}_g$. However, if the data are adequate to obtain $\tilde{t}_n$, those data necessarily permit one to obtain $\tilde{t}_g$. But the reverse is not true. A practical example can be given from Kielmann and McCord's study (1978), where the anthropometric indicator weight-for-age has been used as an index of risk of death in children. In that study 148 children died during 8,157 child-years' follow-up. Proportion of death at different values of the indicator, weight-for-age, is available in a figure from that study. Since no distribution of the indicator by mortality status is available, one can obtain $\tilde{t}_g = 73.2$ from the figure of that study, but no parametric estimator or $\tilde{t}_n$.

Acknowledgements. Comments of Professor P. K. Sen on an earlier draft were helpful.

References

Bairagi, R. (1981) : Anthropometric indicators of nutrition of young children. *Dissertation*, The Johns Hopkins University, Baltimore, Maryland 21205.

——— (1986) : Effects of bias and random error in anthropometry and in age on estimation of malnutrition. *Am. J. Epidemiol.*, **123** : 186-91.

Bairagi, R., Chowdhury, M. K., Kim, Y. J. and Curlin, G. T. (1985) : Alternative anthropometric indicators of mortality. *Am. J. Clin. Nutr.*, **42**, 296-306.

Black, R. E., Brown, K. H., and Becker, S. (1984) : Malnutrition is a determining factor in diarrheal duration, but not in incidence among young children in a longitudinal study in rural Bangaladesh. *Am. J. Clin. Nutr.*, **37**, 87-94.

Cramer, H. (1946) : *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

Habicht, J. P. (1980) : Some characteristics of indicators of nutritional status for use in screening and surveillance. *Am. J. Clin. Nutr.*, **33**, 531-5.

Kielmann, A. A. and McCord, C. W. (1978) : Weight-for-age as an index of risk of death in children. *Lancet*, **1**, 1247-50.

Mood, A. M., Graybill, F. A., and Bose, D. C. (1974) : *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, New York.

Rogan, W. J., and Gladen, B. (1978) : Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.*, **107**, 71-6.

Waterlow, J. C., Buzina, R., Keller, W., Lane, J. M., Nichaman, M. T., and Tanner, J. M. (1977) : The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years. *Bull. WHO* **55**, 489-98.

Paper received : *March*, 1988.
Revised : *June*, 1988.