A MULTIVARIATE STATISTICAL APPROACH TO THE ASSESSMENT OF NUTRITION STATUS

by
S. A. FELLINGHAM

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. C. G. Troskie, for guidance and encouragement in preparation of the thesis.

Thanks are due to Prof. J.F. Potgieter, former head of the Division of Field Studies of the National Nutrition Research Institute, CSIR, and the other members of the survey team who carried out the nutrition status surveys. The present study was only possible because they completed the arduous task of collecting an extremely comprehensive data set with care and accuracy.

I am indebted to Prof. J.J. Theron and Dr. I.F.H. Purchase, successive Directors of the National Research Institute for Nutritional Diseases for facilities to carry out the research. The assistance of my colleagues in handling large volumes of data processing is gratefully acknowledged. Mr. T. McDonald, not only carried out the necessary programming and ran the programs on the computer, but also made a number of useful suggestions. All computations were carried out on the I.B.M. 360 Model 65 electronic digital computer of the National Research Institute for Mathematical Sciences, CSIR.

My thanks are due to Dr. M. Brown of the National Institute for Personnel Research for discussions on factor analysis which have helped to broaden my understanding of the technique, and to Drs. J.P. du Plessis and F.J. Burger of the National Research Institute for Nutritional Diseases for discussions which have helped to clarify my understanding of the biological aspects of the problem.

Thanks are due to the CSIR Technical Services Department for assistance in drawing the figures and in printing the thesis.

I would also like to thank Mrs. M. Henzen and S. Veldman for help in tabulating and checking results and Mrs. E.D.A. de Jager for expertly typing the manuscript.

PRETORIA.
October, 1972. The Author.

Supervisor: Prof. Dr. C. G. Troskie

S U M M A R Y

Attention is drawn to the confusion which surrounds the concept of nutrition status and the problem of selecting an optimum subset of variables by which nutrition status can best be assessed is defined.

Using a multidisciplinary data set of some 60 variables observed on 1898 school children from four racial groups, the study aims to identify statistically, both those variables which are unrelated to nutrition status and also those which, although related, are so highly correlated that the measurement of all would be an unnecessary extravagance.

It is found that, while the somatometric variables provide a reasonably good (but non-specific) estimate of nutrition status, the disciplines form meaningful groups and the variables of the various disciplines tend to supplement rather than replicate each other. Certain variables from most of the disciplines are, therefore, necessary for an optimum and specific estimate of nutrition status.

Both the potential and the shortcomings of a number of statistical techniques are demonstrated.

C O N T E N T S

4.1

---oooOooo---

T A B L E S

F I G U R E S

CHAPTER 1

INTRODUCTION

The term nutrition status is a familiar one to the biologist and, indeed, most laymen have an intuitive feeling for its meaning. There are, however, few terms in the biological literature which have so persistently evaded concise definition. The term means different things to different people. The clinician will gauge nutrition status by the presence or absence of a variety of clinical signs and symptoms; the anthropologist thinks of the extent to which body measurements descriptive of size and shape conform to an ideal; the biochemist envisages the concentrations of various biochemical entities measured in the blood and urine; the radiologist would think in terms of the degree of bone development as revealed by an X-ray photograph, while the dietician, associated with the most directly measurable aspect of nutrition status, thinks in terms of the adequacy of ingested nutrients. Will a concept as diffuse as this yield to statistical treatment? Is it possible, on the basis of the above-mentioned disciplines to produce a unique, coherent definition? This thesis represents an attempt to do so.

In broad terms the assessment of nutrition status implies the evaluation of a certain complex set of characteristics. A large number of variables, purporting to measure these characteristics and representing many different disciplines have been put forward. The problem is to know which of the variables are really important. Some may in fact merely represent "noise" and have little or nothing to do with nutrition status. Others may well be important, but two or more variables may be so highly correlated that the information contained in the one is merely duplicated by the other and the measurement of both is therefore, an unnecessary extravagance. The problem is to

2/.....

choose that particular subset by which nutrition status can best be defined and assessed.

An attempt to define the concept of nutrition status, and formulate an interdisciplinary procedure for its assessment, must be based on observational data. Before an optimum selection of variables can be made, all those variables which may be eligible for selection will have to be measured on an appropriate sample. In drawing the sample, the following aspects must be considered.

(i) Nutritional deficiency diseases are most prevalent in infants and young children. The sample should, therefore, represent these age groups.

(ii) The variables measured may be sex-dependent, it is, therefore, clear that both sexes must be represented.

(iii) For any results to be generally applicable, it must be proven that they hold for the different racial groups. It is, therefore, essential that each of the four racial groups in the Republic of South Africa be represented in the sample.

In a series of nutrition status surveys carried out during the years 1962-1965 by the National Nutrition Research Institute+ of the CSIR, observational data were obtained which are highly suited for the purpose of such a study.

+This Institute was subsequently disbanded and the team responsible for the survey work became part of the National Research Institute for Nutritional Diseases, S.A. Medical Research Council.

CHAPTER 2

THE PRETORIA SURVEYS

2.1 Background and purpose of the surveys

The nutrition status surveys were conducted on representative samples of Pretoria school children in the four racial groups. They were carried out by the Field Studies Division of the National Nutrition Research Institute, in collaboration with other Divisions. The author, then a member of the National Research Institute for Mathematical Sciences of the CSIR was responsible for the statistical planning of the surveys and the analysis of the results. The statistical planning of the surveys has been published in detail (see Fellingham, 1966). Certain aspects, necessary to an understanding of the present study will briefly be repeated here.

The surveys were done with the primary purpose of studying techniques and evaluating criteria for carrying out nutrition status surveys. Such techniques covered the whole range of problems that might be encountered in survey work, including the initial planning of the survey, the drawing of a representive sample of children, the obtaining of biochemical samples, the clinical examination of the children, the obtaining of representative food samples, the relevant chemical analyses and finally, the evaluation of the results.

It was accordingly decided to observe as many somatometric, clinical, biochemical, haematological, radiological, dietary, socio-economic and other variables as was practicable. From this mass of information it was hoped that it would be possible to sift out those variables that would furnish the most useful parameters for describing the nutrition status of the school child, and that on completion of the Pretoria schools survey the survey procedure could be streamlined and simplified so as to facilitate future surveys conducted on a nationwide scale.

2.2 The Population

2.2.1 White children

The White primary school children of Pretoria were the first surveyed. Although it was considered desirable to study as young an age group as possible, pre-school children had to be excluded from this pioneer survey since the practical difficulties involved in their inclusion would have been too great. The White primary school children on the other hand, appeared to constitute an ideal group for a nutrition status survey since school attendance is compulsory for Whites and the entire child population in the younger age group (barring abnormal children) is, therefore, listed on the primary school registers and easily accessible at the schools.

The survey was conducted on all White school children of both sexes in the age group 6-15 years who at the time of the survey were attending schools in the Pretoria area. This area was defined as the geographical region, excluding Voortrekkerhoogte, lying within the (pre-1964) municipal boundaries of Pretoria, together with any peri-urban area not separated from these boundaries by an uninhabited tract of land. This area was selected as defining a reasonable homogenous population group. White children in the age group 6-11 years were surveyed in 1962 and those in the age group 12-15 years in 1965. The fact that the 12-15 year old White children were surveyed 3 years after the 6-11 year old group presents certain statistical problems, since the 12-15 year old children of 1965 would constitute the same population (with the exception of children who had left Pretoria or newcomers who had arrived) as the 9-12 year old children of 1962. The reasons for this situation and the disadvantages inherent in it have, however, been discussed by Fellingham (1966). Since it is unlikely to have an adverse effect on the present study, they will not be repeated here.

5/.....

2.2.2 Non-White Children

In 1963 the survey was extended to Bantu, and in the following year, to Coloured and Indian children, sufficient experience having been gained from the survey of the White children. The population for each of these non-White race groups was defined as children attending schools in those areas whose working population was in the main, employed in the defined Pretoria area. Whilst school attendance is not compulsory for the non-White race groups as it is for the Whites, according to the school authorities an estimated 80-90% of all Bantu children in the Pretoria area would have commenced attending school by the time they reached the age of 8 or 9 years. In the case of the Indian and the younger Coloured children, the attendance was believed to be in the region of 100%. Since very few 6-year old Bantu children attended school, this age group was omitted in the survey of the Bantu. For this reason the present report deals only with children in the age range 7-15 years.

The population sizes of the four racial groups are presented in Table I. It can be seen that the Bantu population was approximately as great as the corresponding White population, whereas the Coloured and Indian populations were only a small fraction of that size.

2.3 Sampling Procedure

After consideration had been given to a number of sampling techniques it was decided to draw a simple random sample from each population as defined above. Since many of the variables observed would be influenced by the age and/or sex of the child, it was further decided to stratify the sample in respect of both age and sex. Further sampling details are described in terms of the 1962 survey of White primary school children. The procedures used in the other surveys were basically similar with the exception of certain differences which will be pointed out later.

6/.....

TABLE I : POPULATION SIZES ACCORDING TO AGE AND SEX OF PRETORIA SCHOOL CHILDREN OF THE 4 MAIN RACIAL GROUPS.

AGE	WHITE 1962 and 1965+			BANTU 1963			COLOURED 1964			ASIATIC. 1964		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
7	1 885	1 821	3 706	1 737	1 820	3 557	124	121	245	124	129	253
8	1 954	1 778	3 732	1 935	2 077	4 012	110	130	240	113	125	238
9	1 939	1 764	3 703	1 709	1 900	3 609	121	103	224	139	108	247
10	1 829	1 622	3 451	1 664	1 978	3 642	118	130	248	113	116	229
11	1 683	1 695	3 378	1 564	1 714	3 278	95	101	196	123	134	257
12	2 016	1 918	3 934	1 424	1 643	3 067	84	98	182	137	139	276
13	1 963	1 983	3 946	1 360	1 634	2 994	84	84	168	115	117	232
14	1 926	2 013	3 939	1 158	1 241	2 399	63	55	118	129	102	231
15	1 900	1 732	3 632	1 147	1 116	2 263	37	39	76	110	95	205
Total	17 095	16 326	33 421	13 698	15 123	28 821	836	861	1 697	1 103	1 065	2 168

+7-11 year olds were done in 1962 and 12-15 year olds in 1965.

7/.....

In order to obtain a simple random sample of the primary school children it was necessary that within an age-sex cell, every child should have the same probability of being drawn. Since the distribution of the children, over the age-sex cells was not too disparate (Table I), it was decided to draw a constant number of children from each cell. The miscellaneous character of the variables selected for measurement and the consequent variation that could be expected in the standard deviation, made it very difficult to come to any valid conclusion as to the required sample size. On the basis of previous experience in isolated surveys carried out by the NNRI, it was anticipated that 30 observations would probably be adequate per age-sex cell. This figure of 30 allowed for a possible further dichotomy of each cell on the basis of some observed variable, e.g. socio-economic status. In view of the somewhat rigorous demands that the survey would make on both parents and children, it was decided to allow for a nonresponse of 50%. Thus a maximum of 60 children per cell would be drawn.

For each survey, teachers of individual classes at each school were requested to enumerate the number of boys and girls in each age group. For each age-sex group a one-to-one correspondence was then set up between each enumerated member of that group, and a serial list of numbers. From this list, random samples of the required size in each age and sex group were drawn with the aid of random numbers. The sampling was sequential, so that, although a maximum sample size of 60 per cell was catered for, the survey could be stopped before the planned end and still provide a reasonably random and representative sample of the population. This objective was attained by drawing a sequential random sample of the schools, on the basis of which the schools were arranged in a random sequence. For practical purposes certain intervals were selected within the random sequence. Within any interval the schools could be visited in any convenient order, but all schools within a certain interval had to be completed before schools in the next interval were surveyed.

8/.....

After the schools within each interval were completed, an analysis was made to determine whether a sufficiently accurate estimation of the desired population parameters could be made for each of the variables observed. For this purpose it was assumed that each of the variables was normally distributed and the following formula for large populations (see Sukhatme, 1954) was used:

$$n \geq \frac{F_{1;n_1-1}(\alpha)\ S_1^2}{\varepsilon^2 \bar{x}_1^2} \qquad (1)$$

Where n_1 = existing sample size

$\bar{x}_1$ = mean

S_1 = standard deviation

$100\varepsilon\%$ is the maximum percentage deviation from the (true) population mean μ that can be tolerated. ε was chosen as 0.10

α = level of uncertainty, and was chosen as 0.05

F = the appropriate F-value with 1 and $n_1 - 1$ degrees of freedom

n = estimated new sample size for which it will be true with $100(1 - \alpha)\%$ certainty that the new sample mean $\bar{x}_n$ lies within $100\varepsilon\%$ of the population mean μ, i.e. that $P\{|\bar{x}_n - \mu| \geq \varepsilon\mu\} = \alpha$.

For the 1962 White survey it was ascertained by means of this procedure that when approximately 25 children per age-sex cell had been surveyed it could be stated with 95% certainty that in the case of most of the variables, the observed (sample) mean for a given variable lay within 10% of the true population mean. By this time, however, the practical machinery of the survey was in full swing. It was clear that little financial saving would result if it were stopped at

9/.....

that stage. It was, therefore, decided to continue until all schools had been surveyed, thus attaining the highest possible degree of precision. The number of children actually surveyed on whom the present analyses are based are presented in Table II.

In subsequent surveys carried out on the Bantu, Coloured and Asiatic primary school children (1963-1964) and on the older White children (1965) a similar sampling procedure was used in each case. In the surveys of the Bantu and older White children formula (1) above was again used to ensure that adequate sample sizes were drawn. Owing to the much smaller size of the Asiatic and Coloured populations, these could not be regarded as approximately infinitely large and population size was taken into account when determining sample size. For this purpose the following formula for finite populations (see Sukhatme, 1954) was used:

$$n \geq \frac{A}{1 + (1/N)A} \qquad (2)$$

where $$A = \frac{F_{1;n_1-1}(\alpha)\, S_1^2}{\varepsilon^2 \bar{x}_1^2}$$

N = population size

All symbols are as previously defined for equation (1) above.

In each case the full sequential sample of schools was completed and the minimum sample size for the great majority of variables was exceeded for each survey. In these surveys it was possible to make a more accurate assessment of the anticipated response. A summary of the estimated response, the sample sizes drawn, the numbers actually surveyed and the true percentage response is given in Table III.

2.4/.....

TABLE II : NUMBERS ACTUALLY SURVEYED ACCORDING TO SEX, AGE AND RACE

Race	WHITE			BANTU			COLOURED			ASIATIC		
Age	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
7	29	31	60	25	37	62	21	22	43	16	18	34
8	43	35	78	28	29	57	21	24	45	22	20	42
9	41	41	82	26	28	54	23	21	44	22	25	47
10	47	38	85	31	29	60	20	20	40	21	18	39
11	46	47	93	30	28	58	24	22	46	16	23	39
12	30	27	57	29	30	59	22	24	46	21	19	40
13	32	25	57	35	31	66	22	24	46	21	17	38
14	28	30	58	26	26	52	19	18	37	19	20	39
15	32	32	64	31	28	59	22	17	39	16	17	33
Total	328	306	634	261	266	527	194	192	386	174	177	351

11/.....

TABLE III : RELATIONSHIP BETWEEN POPULATION SIZE AND SAMPLE SIZE

	WHITE 1962			BANTU 1963			COLOURED 1964			ASIATIC 1964			WHITE 1965		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
Population (N)	9 290	8 680	17 970	13 698	15 123	28 821	836	861	1 697	1 103	1 065	2 168	7 805	7 646	15 451
Estimated response %	50	50	50	75	75	75	80	80	80	80	80	80	85	85	85
Sample size drawn: Per cell	60	60	120	40	40	80	25	25	50	25	25	50	35	35	70
Total(n)*	300	300	600	360	360	720	225	225	450	225	225	450	140	140	280
% Population drawn (n/N)%	3,2	3,5	3,3	2,6	2,4	2,5	26,9	26,1	26,5	20,4	21,1	20,8	1,8	1,8	1,8
Sample actually surveyed (n')	206	192	398	261	266	527	194	192	386	174	177	351	122	114	236
% response (n'/n)	69	64	66	73	74	73	86	85	86	77	79	78	87	81	84
% population actually surveyed (n'/N)	2,2	2,2	2,2	1,9	1,8	1,8	23,2	22,3	22,7	15,8	16,6	16,6	1,6	1,5	1,5

* Excluding 6 year olds as stated in text.

12/.....

2.4 Representational value of sample

The extent to which the sample will (within the limits of chance variation) be representative of the population, will depend on three factors viz. the sample size, the extent to which the sample is unbiased, and the response to the sampling. These are discussed below:

2.4.1 Sample size

According to the tests carried out by means of formula (1) above during the 1962 survey of the younger White children, it was found that (assuming the variables to be normally distributed) one could state with 95% certainty that after about 25 children per age-sex cell had been surveyed, sample means for all except 8 of the variables deviated by less than 10% from the corresponding true population means. The variation in those variables which did not comply was, however, so great that an increase of 20 children in each age-sex group would have probably been necessary before the real population averages could be estimated within an accuracy of only 20%. In general, the scatter in values for each variable observed in the non-White surveys, and in the White survey on older children, proved to be of a similar order to that found in the first survey. It can, therefore, be stated that, for the majority of variables, the numbers surveyed were more than adequate in all five surveys for accurately estimating the population averages on each age-sex group.

2.4.2 Bias in sampling procedure

Since the sampling procedure was such that each child in a given age-sex group of the population had an equal probability of being drawn, the sample should be free of bias. The percentage of the population actually surveyed is shown in Table IV. It can be seen that the probability of the child being drawn varied somewhat from one age group to another,

since /.....

TABLE IV : PERCENTAGE OF POPULATION ACTUALLY SURVEYED

AGE	WHITE			BANTU			COLOURED			ASIATIC		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
7	1,5	1,7	1,6	1,4	2,0	1,7	16,9	18,2	17,6	12,9	14,0	13,4
8	2,2	2,0	2,1	1,4	1,4	1,4	19,1	18,5	18,8	19,5	16,0	17,6
9	2,1	2,3	2,2	1,5	1,5	1,5	19,0	20,4	19,6	15,8	23,1	19,0
10	2,6	2,3	2,5	1,9	1,5	1,6	16,9	15,4	16,1	18,6	15,5	17,0
11	2,7	2,8	2,8	1,9	1,6	1,8	25,3	21,8	23,5	13,0	17,2	15,2
12	1,5	1,4	1,4	2,0	1,8	1,9	26,2	24,5	25,3	15,3	13,7	14,5
13	1,6	1,3	1,4	2,6	1,9	2,2	26,2	28,6	27,4	18,3	14,5	16,4
14	1,5	1,5	1,5	2,2	2,1	2,2	30,2	32,7	31,4	14,7	19,6	16,9
15	1,7	1,8	1,8	2,7	2,5	2,6	59,5	43,6	51,3	14,5	17,9	16,1
Grand mean	1,9	1,9	1,9	1,9	1,8	1,8	23,2	22,3	22,7	15,8	16,6	16,2

14/.....

since the number of children in each group was not constant. This was more noticeable in the case of the Bantu and Coloured groups where there was a marked falling off in the number of children in the older age-groups (Table I). The fluctuation in the percentage sampled, however, remained within reasonable limits (Table III).

In general, the sampling procedure for the White and Asiatic surveys and for the Bantu and Coloured children of 7-11 years should yield a reasonably unbiased sample. The representational value for the child population as a whole, of the samples of the 11-15 year old Bantu and Coloured groups, is affected to some extent by the fact that not all children in these age groups were reached in the sampling. Since it is likely that the socio-economic factor determines at what age or standard the child leaves school, it is probable that the sampling of these age-groups will have been biased to some degree in favour of the children of the more affluent families, and that the nutritional picture which is obtained may be a somewhat optimistic one. The only way that this bias in a sample could have been corrected would have been to sample both the non-school-going and the school-going children in the older age groups. This was not practicable at the time and although the bias in favour of the more well to do families will clearly affect the *representational value* of the data, it is not likely to have any marked effect on the interrelationship between the variables with which the present study is concerned.

2.4.3 Response to sampling

It can be seen from Table III that the response for all the racial groups was exceptionally good. It varied for the White children from 66% in 1962 to 84% in 1965 (both sexes). The response for the Bantu and Asiatic surveys lay between these two values, whilst the response to the Coloured survey (86%) exceeded that of the 1965 White survey. In view of the

considerable demands made on both parent and child, these figures are surprisingly good and the representational value of the surveys should be little affected by the non-response. It should be noted that the figures for non-response vary somewhat from those published by Fellingham (1966). In the present study only those children have been included for whom a complete set of data in respect of *all* variables to be considered, was available. In the 1966 report a few children were included who did not respond in respect of the dietary and socio-economic surveys, the results of which were not available at that time. Also, the 6-year old children have been omitted from the present study.

The above comments on the efficiency of the sampling procedure are applicable to an independent consideration of each survey. When, however, we consider the data for all racial groups in Pretoria in the age groups 7-15 years, the time lapse between the surveys should be taken into account. Also, when comparing the two White surveys, it should be remembered that the 12-15 year olds that were surveyed, were actually the same population as that surveyed three years previously. Whilst these aspects may well affect a comparison in respect of a certain parameter or set of parameters between the various racial groups or even between the younger and older White children, it is, however, unlikely that they will have any serious effect on the interrelationship between the variables with which the present study is concerned.

2.5 Nature of recorded variables

The variables available for the present investigation were recorded on 1898 children in the four racial groups. In all,154 variables were recorded. A complete listing of these variables showing the coding schedules and key to the coding, comprises 34 pages and is available on request. Of the variables, 23 were not recorded for at least one of the four racial groups and could, therefore, not be used in the present analysis. The remaining 131 variables were observed for each

16/.....

of 1898 children. Of these 8 give general information such as race, age, sex and school. There were 15 dietary variables, 30 socio-economic variables, 11 somatometric variables, 22 biochemical variables, 9 haematological variables, 28 clinical characteristics and 8 clinical syndromes which constitute a summary of the clinical characteristics.

The bulk of the socio-economic variables were classificatory in nature and not relevant to the present investigation. From those socio-economic variables relating to family size, income, and expenditure an index of socio-economic status (rate/head/day) was calculated and has been used in the present analyses.

The clinical variables were mainly classificatory in nature. The treatment of such variables presents special problems outside the scope of the present report. They have, therefore, not been dealt with. All the relevant variables in the remaining disciplines have been studied. A complete list of all the variables upon which the report is based showing the units in which they were measured, is given in Table V.

At the present time the basic statistical analyses applied to the data have been completed and the results have been individually reported for each discipline in the following publications: Van der Merwe *et al.*(1965); Smit (1965); Du Plessis *et al.*(1965a & b); Lubbe, and Pretorius (1965); Oudkerk (1965); Potgieter (1965); Du Plessis *et al.* (1966a, b & c); Smit *et al.* 1967a & b); Potgieter and Fellingham (1967); Du Plessis (1967); Du Plessis *et al.* (1967a & b); Neser (1968a & b); Lubbe (1968); Stead (1968); Smit (1968) and Louw *et al.* (1969).

At the interdisciplinary level, a stepwise regression analysis has been done on the relationship between the intake of the various nutrients and certain biochemical variables (see Fellingham, (1969); Louw *et al.*(1969)). Apart from these analyses, nothing in the nature of a multidisciplinary study has,

17 /.....

TABLE V : LIST OF VARIABLES DEALT WITH IN THE ANALYSIS, ACCORDING TO DISCIPLINE

		IDENTIFICATION
1.	(1)	Race
2.	(2)	Age (years)
3.	(3)	Sex
		DIETARY
4.	(1)	Animal protein (g)
5.	(2)	Vegetable protein (g)
6.	(3)	Mixed protein (g)
7.	(4)	Calories (kilocal)
8.	(5)	Protein (g)
9.	(6)	Fat (g)
10.	(7)	Carbohydrate (g)
11.	(8)	Calcium (mg)
12.	(9)	Phosphorus (mg)
13.	(10)	Iron (mg)
14.	(11)	Vitamin A (I.U.)
15.	(12)	Thiamine (mg)
16.	(13)	Riboflavin (mg)
17.	(14)	Nicotinic acid (mg)
18.	(15)	Vitamin C (mg)
		SOCIO-ECONOMIC
19.	(1)	Rate/head/day. (R/D/H)
		SOMATOMETRIC
20.	(1)	Weight (kg)
21.	(2)	Height (cm)
22.	(3)	Cristal height (cm)
23.	(4)	Intercristal width (cm)
24.	(5)	Biacromial width (cm)
25.	(6)	Ulnar length (cm)
26.	(7)	Upper arm circumference, arm bent (cm)
27.	(8)	Calf circumference (cm)
28.	(9)	Triceps skinfold thickness (mm)
29.	(10)	Subscapular skinfold thickness (mm)
30.	(11)	Para-umbilical skinfold thickness (mm)

		BIOCHEMICAL
31.	(1)	Cholesterol (mg per 100ml serum)
32.	(2)	Phospholipids (mg per 100ml serum)
33.	(3)	Cholesterol/Phospholipids (ratio)
34.	(4)	Alkaline phosphatase (King-Armstrong units/100ml serum)
35.	(5)	Inorganic phosphorus.
36.	(6)	Amylase (SOMOGYI units per 100ml serum)
37.	(7)	Urinary amylase/creatinine
38.	(8)	Vitamin A (micrograms per 100ml serum)
39.	(9)	Carotene (micrograms per 100ml serum)
40.	(10)	Thiamine (urinary micrograms per gram creatinine)
41.	(11)	2-Pyridone (urinary N'-methyl-2-pyridone-5-carboxylamide, mg per gram creatinine).
42.	(12)	N'-Me (urinary N'-methyl nicotinamide, mg per gram creatinine)
43.	(13)	Urinary pyridone/N'-Me
44.	(14)	Total protein (grams per 100ml serum)
45.	(15)	Albumin (grams per 100ml serum)
46.	(16)	Globulin (grams per 100ml serum)
47.	(17)	α-globulin (grams per 100ml serum)
48.	(18)	β-globulin (grams per 100ml serum)
49.	(19)	γ-globulin (grams per 100ml serum)
50.	(20)	Albumin/Total Protein (%)
51.	(21)	γ-globulin/Total protein (%)
52.	(22)	Riboflavin (urinary, micrograms per gram creatinine)
		HAEMATOLOGICAL
53.	(1)	Haemoglobin (g%)
54.	(2)	Sedimentation rate (mm/hr)
55.	(3)	Haematocrit (%)
56.	(4)	M.C.H.C. (%)
57.	(5)	White cells x 10^{-1}
58.	(6)	Diff. count: Neutrophils (%)
59.	(7)	Monocytes (%)
60.	(8)	Lymphocytes (%)
61.	(9)	Eosinophils (%)

18/.....

as yet, been carried out. Little, if any progress has been made in selecting, on a multidisciplinery basis, a subset of those variables by which nutrition status can best be defined and observed.

2.6 The analytical problem

We have then, a certain complex characteristic called nutrition status, measured in terms of 61 measurements, representing no less than six different disciplines. We wish, in a multidisciplinary sense to select a small subset of the variables by which nutrition status can best be defined and assessed. There are two bases upon which variables can be eliminated:

(i) Certain variables bear little or no relationship to nutrition status and may be regarded as "noise" variables.

(ii) Certain variables are very much affected by nutrition status but are so highly related one to the other that the measurement of all is an unnecessary extravagance.

The problem presents both logical and statistical difficulties. Firstly, we have no outside criteria apart from our measurements to tell us which child has a satisfactory nutrition status and which does not. Had we such an external criteria the problem would have been reduced to a straightforward statistical exercise. We are, in fact, requiring from a set of observed variables, that the variables themselves tell us which are important and which are not. Secondly, we must note that any selection of variables based purely on an analysis of the variables of each discipline individually, is likely to differ from a selection based on an analysis carried out simultaneously over all disciplines. This is important, for it would appear that the bulk of criteria for the assessment of nutrition status put forward to date, have been chosen on a unidisciplinary rather than a multidisciplinary basis.

19 /.....

Although no outside criteria of nutrition status is available, there are certain characteristics, germane to the populations from which the survey samples were drawn, or to the variables themselves, which may be useful. For example, four different population groups have been measured which are known to differ widely in terms of socio-economic status and hence probably also in terms of nutrition status. A dichotomy of the variables can be made into those related to cause, ("causal" variables) and those which described a consequence ("consequential" variables) as follows:-

CAUSAL VARIABLES	CONSEQUENTIAL VARIABLES
Socio-economic	Somatometric
Dietary	Biochemical
	Haematological
	Clinical

The two sets of variables on the left are the only ones which can bring about or maintain any particular nutrition state. Those on the right serve only to describe some or other effect. It should,furthermore, be born in mind that the variables have been measured on children whose ages range from 7-15 years. This was fundamental to the purpose of the survey as many of the parameters measured change with age. It does, however, also present a potential pitfall, since two variables may be highly correlated, not because of any relationship between them, but simply because they are both dependent upon the age of the child. For a truly optimum selection of the "best" variables it is clear that the cost of measuring the variables should also be considered. It would clearly be more advantageous to measure six variables if these could be observed cheaply, than to measure two variables if the measurement proved to be extremely expensive, even if the two provided the same information as the six. It will be clear that the cost is closely related to the discipline. The variables associated with some disciplines, e.g. the somatometric variables,can be readily observed at insignificant

20/.....

cost. The assessment of nutrient intake, on the other hand, is extremely laborious and costly.

In the first instance an analysis of the data will be directed solely at the information content of the variables, irrespective of the cost of making the observation. After a choice of variables has been made on this basis, the cost aspect will be discussed.

It can readily be seen that of the existing multivariate statistical techniques no single one is ideally suited to the problem. For this reason a brief review will be given of available multivariate statistical techniques, with the emphasis on the logical concepts underlying their application.

Chapter 3/.....

CHAPTER 3

REVIEW OF APPLICABLE MULTIVARIATE STATISTICAL TECHNIQUES

3.1 Techniques for studying the interrelationship between two sets of variables

Suppose we have observed the variables y_1, y_2, ... y_p, x_1, ... x_n where y_1 to y_p can be identified as response variables and x_1 to x_n as predictor variables. If the set of response variables contains only one member y (the so-called dependent variable), we have the well known problem of *multiple regression analysis*. If that subset of the predictor variables is required, which best predicts the response variable, the problem is one of *stepwise regression analysis*. (Efroymsen, 1960). A measure of association between the response variable and the predictor variables is given by the *multiple correlation coefficient*.

In the general case, where both sets of variables have more than one member, several correlations have been defined as measures of the relationship between the two sets. The best known of these is the *canonical correlation coefficient*, but a number of others have also been proposed. The generalised multiple correlation matrix (Khatri 1964) has the remarkable property that nearly all other correlation coefficients follow as special cases (Troskie 1969).

3.2 Techniques for discriminating between two or more groups

In this situation the variables are observed on two or more groups of e.g. people. Suppose we have k groups of individuals n_1, n_2, ... n_k and on each individual we measure p variables x_1, x_2, ... x_p. Linear discriminanant functions of the variables can be derived which maximise the variance between the groups as compared to variance within the groups. These functions provide a rule for allocating a new individual to one of the k groups (Kendall and Stuart, 1966). If it is known that certain of the variables measured are likely to be

redundant and it is required to select the best subset for the purpose of allocation, this can be done by means of a stepwise discriminant analysis (Efroymsen, 1960). In such an analysis, the variables are added, one by one to the discriminant functions in order of their contribution to the discrimination between the groups.

3.3 Techniques for studying the interrelationships amongst a set of variables

The techniques of principal component analysis and factor analysis belong to this group. The former technique was put forward by Pearson (1901) and Hotelling (1933), and has subsequently been dealt with in depth by a number of writers, see e.g. Anderson (1958); Seal (1964) and Kendall (1966). The latter technique stems from the work of Spearman (1904, 1926), and a variety of different approaches have been developed, see Harman (1967). The two techniques are closely related but have somewhat different aims. A principal component analysis is conceptually the more simple, since it is merely a 'breaking down' of a co-variance or correlation matrix into a set of orthogonal components or axis, equal in number to the number of variables concerned. These correspond to the latent roots and accompanying latent vectors of the matrix. Suppose we have p variables x_1, x_2, ... x_p measured on n individuals. A principal component analysis attempts to combine these p variables into a smaller number of new variables which will provide almost all the information about the way in which one individual differs from another. A drawback of the technique, is that all measurements should be measured in the same units. If this is not the case, the original measurements are sometimes expressed in standard units.

In general terms then, the problem is to define

$$
\begin{aligned}
y_1 &= a_{11}\, x_1 + a_{12}\, x_2 + \ldots a_{1p}\, x_p \\
y_2 &= a_{21}\, x_1 + a_{22}\, x_2 + \ldots a_{2p}\, x_p \\
&\vdots \\
y_p &= a_{p1}\, x_1 + a_{p2}\, x_2 + \ldots a_{pp}\, x_p
\end{aligned}
\qquad (1)
$$

y_1 is chosen in such a way that it has the biggest possible variance, thereby representing, better than any other linear combination of the x's, the general difference between the individuals upon whom the observations have been made.

y_2 is chosen so as to be uncorrelated with y_1 and have the next largest variance, and so on.

In matrix notation the equation (1) may be written:

$$Y = A'X$$

where:

$$X = (x_1, x_2, \ldots x_p)$$
$$Y = (y_1, y_2, \ldots y_p)$$
$$A = a_{ir} \quad i = 1, \ldots p;\ r = 1 \ldots p$$

In order to illustrate the relationship with factor analysis, this could be re-written as:

$$X = BY$$

where:

$$B = (A')^{-1}$$

In factor analysis a definite model is assumed for the way in which the variables x_i (i = 1, ... p) are influenced by certain underlying factors. It is assumed that each individual has a certain value for these factors f_1, f_2, f_3,... f_k (k < p) which are correlated but cannot be measured directly. The variables x_i which can be measured, are assumed to represent a linear function of the factors.

24/.....

$$
\begin{aligned}
x_1 &= b_{11} f_1 + b_{12} f_2 + b_{13} f_3 + \ldots b_{1k} f_k + \varepsilon_1 \\
x_2 &= b_{21} f_1 + b_{22} f_2 + b_{23} f_3 + \ldots b_{2k} f_k + \varepsilon_2 \\
&\vdots \\
x_p &= b_{p1} f_1 + b_{p2} f_2 + b_{p3} f_3 + \ldots b_{pk} f_k + \varepsilon_p
\end{aligned} \qquad (2)
$$

The quantities b_{11} are constants known as the factor loadings which indicate to what extent the observed variables x_i are affected by the various factors. The ε_i ($i = 1, 2, \ldots p$) are residual sources of variation which only affect the observed variables x_i. They are supposed to be independent of one another, and also of the f_r ($r = 1,2, \ldots k$). There are various methods for factor analysis which investigate how many factors should be assumed, and estimate the factor loadings b_{ij}. (Lawley & Maxwell, 1963).

As a technique, factor analysis is conceptually far more involved than principal component analysis. Although it has been practised in one form or another for about 50 years, it still tends to generate mixed feelings in professional statisticians. On the one hand, protagonists of the technique have made far reaching claims as to its power. According to Warburton (1962): "A factor analysis takes into account all the relationships between all the variables .... its basic purpose is to increase logical clarity and cogency by using a few independent terms". The antagonists, on the other hand, have doomed factor analysis almost indiscriminantly. According to Ehrenberg (1962): "Factor analysis is technically underdeveloped and at times appears almost cretinous. Its practitioners seem to be largely unaware of the technical and methodological problems which they have let themselves in for. The techniques are confused, interpretive guidance is lacking and there is little one can do with the numerical results".

The reader will have an opportunity to judge whether the present study has been able to contribute any light on the

25/.....

value of factor analysis.

3.4 Techniques for dealing with data when it is not known whether they belong to category 2 or 3 above

The technique of cluster analysis has been derived for dealing with situations when categories are not clearly defined but are only suspected to exist. It represents an attempt e.g. to cluster subjects into groups, so that in terms of the observed variables, the differences in subjects between groups will be much larger than the differences within groups. Distinct clusters may then correspond with distinct biological categories. Suppose we have p variables x_1, ... x_p measured on each of n subjects. The problem may be represented geometrically as n points plotted in a p dimensional (variable) space. A cluster would then be a swarm of points.

In order to perform such an analysis, a measure of the similarity between any two subjects in respect of the set of variables is first required. A number of such measures have been suggested and the best one to choose will frequently depend on the circumstances.

A second form of cluster analysis, is one which aims at clustering not subjects, but variables. That is, bringing together variables which measure the same sort of characteristic. Geometrically this could be represented as p points in an n dimensional (subject) space,(the variables having been standardised). A distance measure between variables is now required and measures based on the correlation coefficient are popular. This form of cluster analysis is somewhat similar to a principal component or factor analysis, but is free of the stringent underlying assumptions of factor analysis. Ball (1965), gives a detailed review and an extensive bibliography, but fails to indicate clearly for which of the two types of clustering a particular technique is suited.

26/.....

Computer programs have been written which automatically form clusters. In general, however, the technique is still in a rather experimental stage and tends to be impracticable where a large number of variables have been observed on a large number of persons.

3.5 Linear scoring systems

These represent an attempt to combine the p-variables $x_1 \ldots x_p$ into a single score and thereby reduce a multivariate problem to a univariate one. Thus p measurements for each subject could be replaced by the single measurement g, where:

$$g = l_1 x_1 + \ldots + l_p x_p \qquad (3)$$

If the subjects belong to different groups which have to be compared, the coefficients $l_1 \ldots l_p$ are usually chosen so as to maximise the variance between the groups as compared to that within the groups. In the case of two groups the linear score found in this way is the linear discriminant function mentioned in Chapter 3.2.

A similar application of linear scores can be made when information is available on an ordinal scale. Suppose that subjects can be ranked in respect of some attribute on a p-point scale from 1 to p, and that k different treatment groups have been assessed in this way. It is often convenient to replace the p-point scale by a continuous variable taking values $l_1, l_2 \ldots l_p$. First consider a set of dummy variables $x_1 \ldots x_p$, chosen so that if a subject is graded 3, he will be given the dummy measurements:

$$x_1 = 0;\ x_2 = 0;\ x_3 = 1;\ x_4 = 0;\ \ldots\ x_p = 0$$

In (3) above $l_1 \ldots l_p$ are chosen to maximise the sum of squares between treatments, compared to the sum of squares

27/.....

within treatments. The variable g takes a value of l_1 for grade 1 on the p-point scale, l_2 for grade 2 and so on. Thus, the p-points of the scale are replaced by p scores l_1, l_2 ... l_p which maximise the treatments comparison.

Such linear scoring systems are examples of the general technique called canonical analysis (see Seal, 1964, Chapter 7).

3.6 Limitations of existing techniques

The above forms a cursory survey of available multivariate statistical techniques. In view of the stated purpose of the present report, viz. to select the most meaningful subset of variables for defining and describing nutrition status, it will readily be seen that no clear cut application exists for any of the above techniques. We have no criteria outside the data set itself of what nutrition status really is. Thus, the techniques of discrimination or multiple regression analysis are not immediately applicable.

No doubt the dimension of the problem could be reduced by the application of a principal component analysis, and a new and smaller set of hypothetical variables formulated which best described the difference between the children on whom the observations have been made. The number of original measurements which must be observed could, however, only be effectively reduced if:

(i) the first few principal components explained most of the variance (as reflected by the eigen *value*).

(ii) high loadings (as reflected by the eigen vectors) were found on only one or two variables for each of the first few components and low loadings on all the others - a situation unlike to arise for biological variables which are highly interrelated.

28/.....

The technique of factor analysis, although conceptually different from principal component analysis, is mathematically similar. A similar computational procedure can be followed except that the 1's in the diagonal of the correlation matrix are replaced by so-called communalities (the squared multiple correlation coefficients of each variable in respect of the rest). A 'rotation' of the factor matrix is then carried out in an attempt to achieve one or two high loadings for each factor with the remaining loadings of that factor as near to zero as possible. This technique may hold some promise as it should enable one to group the variables into categories and select the most important variable or variables in each category for measuring a specific attribute. In the context of assessing nutrition status, however, it may be extremely difficult to postulate in advance certain underlying factors, the existence of which are basic to the analysis.

In general, a cluster analysis would seem promising, as this should be able to form clusters of subjects (in the variable space) or variables (in the subject space) under conditions where these are only suspected to exist. Thus, for example, those variables could be grouped together which measure a specific aspect of nutrition status. In practice, however, as has been stated, such techniques are at present rather experimental. The application of a cluster analysis technique on some 60 variables measured on 1898 persons, would prove a major undertaking even for the high speed digital computers available today.

Whilst linear scoring systems might assist one in representing a large number of variables by a single variable, the original variables, as in the case of a principal component analysis, would still have to be measured before this new variable could be calculated.

29/.....

3.7 Outline of the analytic approach

In the light of the above, it must be clear that no single statistical analysis is likely to yield a clear cut and final answer to our problem. Since the variables we are studying belong to many and varied disciplines, it is reasonable that the first step in the analysis should be to take a close look at the variables of each discipline separately. Armed with a better understanding of these variables, and having obtained some indication as to which ones are important within the context of each respective discipline, we will then proceed to study the interrelationships of the variables across the various disciplines. Finally, we will proceed with an analysis simultaneously involving all the relevant variables.

Lacking any external information, it is clear that the variables within a particular discipline cannot, in themselves, tell us anything about nutrition status. It seems reasonable, therefore, to determine those variables which best describe the way in which one individual differs from another. For this purpose a principal component analysis is indicated.

It could be argued that certain, at least, of the disciplines, have already been individually reported on in some detail. This is particularly true of the somatometric variables and the biochemical variables which have already been the subject of a number of publications (see Chapter 2.6).

These reports have, however, mainly been concerned with describing the results obtained on each age-, sex-, and racial group, for each variable, and in interpreting these results against accepted biological criteria. They reflect a biological rather than a statistical viewpoint and did not employ any coherent and uniform statistical procedure.

30 /.....

It is intended in the present report to see what information can be derived from the data by objective statistical treatment. Having derived certain results, note will than be taken of the extent to which these results compare with current biological knowledge of the subject.

Chapter 4/.....

CHAPTER 4

THE INFORMATION CHARACTERISTICS AND PRINCIPAL COMPONENTS OF THE VARIABLES IN EACH DISCIPLINE

The technique of principal component analysis which we have selected, has a basic limitation: The results are only meaningful if the different variables to which it is applied are all measured on the same scale. In an attempt to circumvent this problem we have applied the analyses to standardised data (each observation being reduced by its mean and divided by its standard deviation). This is a meaningful way of dealing with the problem but not altogether satisfactory. See Seal (1964).

A principal component analysis was applied separately to each of the four racial groups in respect of the following disciplines: Somatometry, biochemistry, haematology and dietary. The rate/head/day, representing the socio-economic variables will be dealt with later.

The analyses were applied separately to each racial group in order to see whether any particular pattern emerged which was common to the various racial groups.

4.1 Principal component analysis of the somatometric measurements for each of the four racial groups

It will be noted that all the somatometric measurements with the exception of weight have a linear dimension. Since the weight of a person is approximately proportionate to his volume, it is reasonable to suppose that the relationship between a linear body dimension, such as height, and weight would be approximately cubic. It has, in fact, been shown that the relationship between weight (W) and height (H) is of the form $W = CH^{\alpha}$ where C is a constant. α is less than 3 and tends to vary somewhat with age. Defrise-Gussenhoven (1954, p.49), in a study of Belgium school children, found values for α varying from 2,00 for boys aged 6-8 years to 2,84 for boys

aged 9,75 to 15 years. We chose a value for α of 2,67 as being appropriate to the age range with which we were concerned in our analysis. We did, however, experiment with other values of α in the range $1 \leq \alpha \leq 2,67$ and found that the results of the principal component analysis were relatively little affected by the value of α.

The results of the principal component analysis on the somatometric variables is shown in Table VI. In order to keep the table down to a reasonable size, only those variables and associated eigen vectors have been tabulated where the eigen vector was reasonably large for at least one of the racial groups. The eigen vector can be interpreted as the coefficient of correlation of the original variable with the relevant principal component (Rees, 1969). The rank of the eigen vector within the component is also shown.

The separation into the three components is remarkably similar for the four racial groups. In each case weight was the most important variable in the first component as indicated by the modulus of the eigen vector, followed by calf circumference. Upper arm circumference, intercristal width and height, formed the next triplet for Whites and Bantu and fell amongst the first 7 variables for Coloureds and the first 6 for Asiatics.

We now consider whether the first principal component has biological significance. Weight is clearly an indication of general body size as also are both height and intercristal width, the one being a vertical and the other a horizontal skeletal measurement (See Fig. 1). Calf circumference and upper arm circumference appear to be out of place in the first component, since these two measurements have frequently been put forward as indicative of musculature or soft body tissue. Smit (1968) has, however, pointed out that these measurements are also increased by a large cross-sectional bone area or a

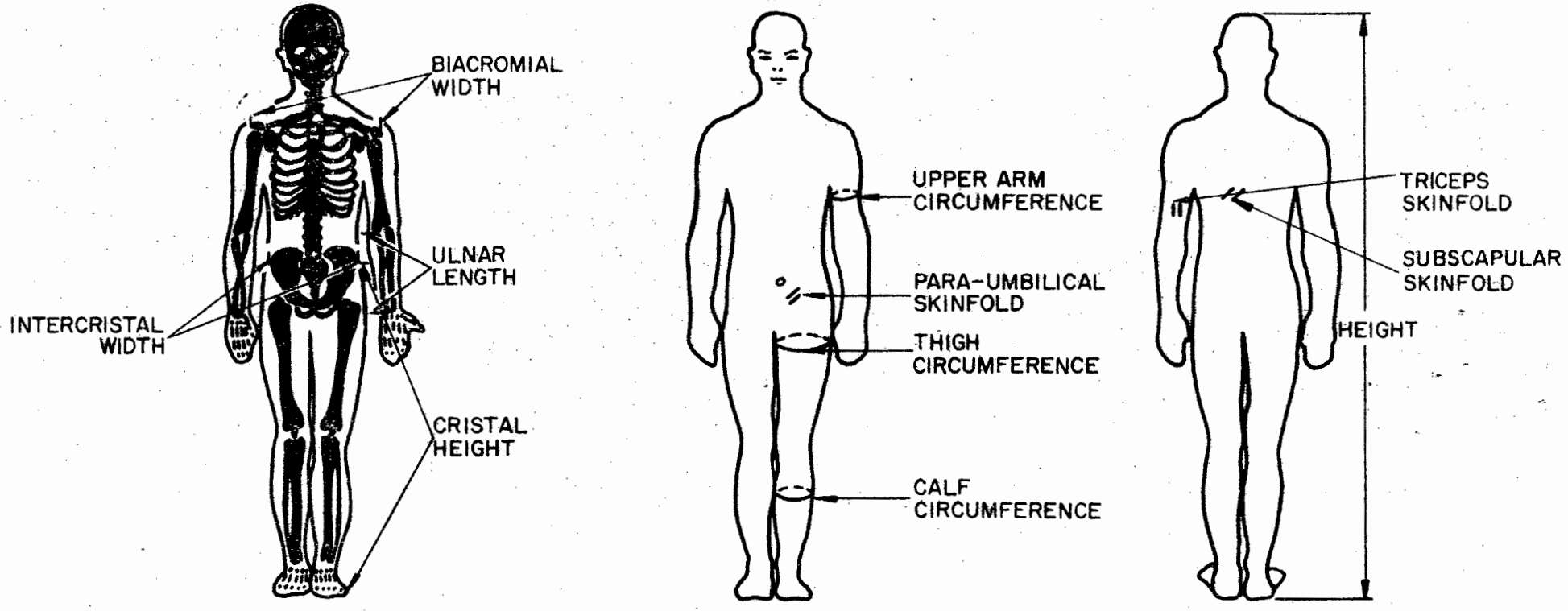
33/.....

TABLE VI : PRINCIPAL COMPONENT* ANALYSIS OF THE SOMATOMETRIC VARIABLES

Component			White		Bantu		Coloured		Asiatic	
No.	Character-istic.		Eigen Vector	Rank	Eigen vector	Rank	Eigen Vector	Rank	Eigen Vector	Rank
1.	General body size	Weight	-0,3454	1	0,3456	1	-0,3439	1	0,3447	1
		Calf circumference	-0,3298	2	0,3298	2	-0,3276	2	0,3295	2
		Upper arm circumference	-0,3257	3	0,3258	3	-0,3247	4	0,3260	3
		Intercristal width	-0,3250	4	0,3237	4	-0,3230	7	0,3198	5
		Height	-0,3205	5	0,3218	5	-0,3267	3	0,3139	6
			(0,74)		(0,73)		(0,75)		(0,75)	
2.	Tissue body fat	Triceps skinfold	0,5337	1	0,5546	1	0,5604	1	-0,4988	1
		Para-umbilical skinfold	0,4785	2	0,4764	2	0,4870	2	-0,4566	3
		Subscapular skinfold	0,4636	3	0,4490	3	0,4822	3	-0,4731	2
			(0,16) [0,90]		(0,17) [0,90]		(0,16) [0,91]		(0,17) [0,92]	
3.	Soft body tissue	Upper arm circum.	-0,5917	1	-0,5643	1	-0,5111	1	0,4798	1
		Calf circumference	-0,4247	2	-0,4810	2	-0,4313	2	0,4431	2
		Para-umbilical skinfold	0,3781	3	0,4070	3	0,4157	3	-0,3664	4
			(0,03) [0,93]		(0,03) [0,93]		(0,03) [0,94]		(0,02) [0,94]	

*The value in round brackets () is the proportion and that in square brackets [ ] the cumulative proportion of the total variance explained by the component

34/.....



35/.....

FIGURE I

Site of somatometric measurements

large amount of subcutaneous body fat. Thus it is clear that a strong relationship would also exist between general body size and both upper arm circumference and calf circumference, quite apart from the extent to which these measurements vary with a large or small amount of soft body tissue. It seems reasonable then to consider the first component as representing *general body size*.

It is interesting to note that for the first component the eigen vectors all have positive signs for the Bantu and Asiatic groups, but negative signs for the White and Coloured groups. In the case of the Bantu and Asiatics, the first principal component would increase as the variables indicative of general body size, increased. In the case of the Whites and Coloureds, however, it would decrease. The first component is responsible for approximately 74 percent of the variation in the variables.

On looking at the second principal component (Table VI) we see that, for the White, Bantu and Coloured racial groups, the skinfold thickness of the triceps has the highest eigen vector followed by the para-umbilical and subscapular skinfold thicknesses, in that order. In the case of the Asiatics, the subscapular skinfold thickness has a slightly higher eigen vector than that of the para-umbilicus. The eigen vectors are positive for the White, Bantu and Coloured groups, but negative for the Asiatics. They do not differ greatly for each variable (0,4785 - 0,5337 for Whites). Since the three skinfold thicknesses are widely acknowledged as being indicative of body fat (see e.g. Chinn and Allen, 1960; Durnin and Ramahan, 1967), we conclude that the second component represents *body fat* and is responsible for some 16 percent of the variation in the variables.

In the case of the third principal component, we note that for each racial group the upper arm circumference (bent) and calf circumference have been selected. Both these variables, as we have seen, relate to *soft body tissue* and it seems reasonable to presume that this is the characteristic of the third component. The presence of the para-umbilical skinfold thickness with a fairly high eigen vector, particularly in the case of the Coloureds, may suggest that this variable tends to increase with soft body tissue, and is, therefore, a less precise measure of body fat than the other two skinfold thicknesses. The third principal component is responsible for some 3 percent of the total variation, in the variables.

In view of the fact that the variables observed in the White racial group were obtained in two separate studies, the principal component analysis on the White group was repeated separately for the 7-11 year age group. The results were, however, consistent with those presented above and will not be repeated here. Thus the fact that the White children were surveyed at two separate times does not appear to have affected the present analysis.

It would appear that the observation of certain of the measurements listed in each of the first three components (above), should give a comprehensive description of the somatotype of an individual. While the"best" variables to measure to estimate a component are those with the largest loading for that component, the size of the loadings on each variable for a particular component do not vary greatly. The choice of which variables should be measured within a component could, therefore, to some extent be guided by convenience.

It should be noted at this stage, that the selection of "best" variables, based on the above, is not necessarily optimum for the assessment of nutrition status, but in fact,

37 /.....

represents that subset of variables which can give most information *about the way in which one individual differs from another.* It is remarkable that the components are so similar for the four racial groups, particularly as the intradisciplinary analyses have shown considerable variation in respect of nutrition status (Du Plessis, 1967 p.126). It would appear that the most useful purpose served by the principal component analysis, is to identify those variables which contain the best information regarding a certain characteristic. We have, furthermore, not yet looked at the somatometric variables within the context of the other disciplines. This will be done in Chapters 5-7.

It is instructive to note in passing, that the somatometric variables yield readily to intuitive interpretation. The results which we have obtained by an independent statistical analysis, and which are consistent for four widely different racial groups, are compatible with current biological knowledge.

4.2 Principal component analysis of the biochemical variables for each of the four racial groups

In contrast to the somatometric variables, the biochemical variables do not readily lend themselves to an intuitive interpretation. The results of the principal component analysis on these variables are shown in Table VII. Four components have been listed, representing in total some 50 percent of the variance. After the fourth component, the other components tended to consist of a repetition of those variables already listed.

A mere 22 percent (for Whites) of the total variation is explained by the first component, as compared to 74 percent in the case of the somatometric variables. The results for the four racial groups are similar although not identical for the first component. The degree of similarity seems

38/.....

TABLE VII : PRINCIPAL COMPONENT* ANALYSIS OF THE BIOCHEMICAL VARIABLES

COMPONENT			White		Bantu		Coloured		Asiatic	
Protein status	Characteristic	Name	Eigen Vector	Rank	Eigen Vector	Rank	Eigen Vector	Rank	Eigen Vector	Rank
1.	Blood serum Protein fractions reflecting protein status and exposure to antibody producing infections	Total globulin	0,4085	2	0,4812	1	0,4580	1	0,4648	1
		Alb./Total Prot.	-0,4143	1	-0,4541	2	-0,4365	2	-0,4288	3
		γ-Globulin	0,3112	3	0,4132	3	0,4222	3	0,4384	2
		γ-Globulin/Total protein	0,2938	4	0,3890	4	0,4073	4	0,4140	4
			(0,22)		(0,18)		(0,20)		(0,20)	
2.	Circulating serum body fats and fat soluble vitamins	Cholesterol	-0,5226	1	-0,5022	1	-0,5273	1	-0,4862	1
		Phospholipids	-0,4869	2	-0,3956	4	-0,4149	2	-0,4036	2
		Carotene	-0,3595	3	-0,4203	2	-0,3269	4	-0,2888	5
		Vitamin A	-0,1461	7	-0,3980	3	-0,2732	6	-0,1641	10
		Choles/Phospholipids	-0,3153	5	-0,2547	5	-0,3472	3	-0,2257	8
			(0,12)		(0,11)		(0,13)		(0,11)	
			[0,34]		[0,29]		[0,33]		[0,31]	
3.	Nicotinic acid metabolites and protein fractions	Urinary amylase	-0,3775	1	0,1686	9	0,0795	16	-0,1670	11
		γ-globulin	-0,3334	2	0,1523	15	-0,0540	18	0,0344	15
		2-Pyridone	-0,2564	7	-0,4257	1	-0,5812	1	-0,4961	1
		Pyridone/N'-Me	-0,0640	20	-0,3988	4	-0,3189	3	-0,3084	4
		Total Protein	-0,2565	6	0,4112	2	-0,0700	17	0,2065	10
		Albumin	-0,1261	15	0,4068	3	-0,1022	14	0,2928	5
		N'-Me	-0,2052	11	0,0563	19	-0,3987	2	-0,3659	2
			(0,10)		(0,10)		(0,09)		(0,09)	
			[0,44]		[0,39]		[0,42]		[0,40]	
4.	Water soluble vitamins.	2-Pyridone	0,5153	1	-0,2589	5	-0,1152	14	0,0275	20
		N'-Me/Creat.	0,3494	3	-0,5399	1	0,3135	4	-0,1048	16
		Urinary amylase	-0,1815	10	-0,3774	2	0,4866	1	-0,3688	2
		Urinary riboflavin	-0,2841	5	-0,3062	3	0,3268	3	0,0147	22
		Serum amylase	-0,2567	6	-0,3021	4	0,2159	8	-0,1757	11
		Pyridone/N'-Me	0,3063	4	0,1257	10	-0,4412	2	0,1404	13
		Carotene	0,0990	13	-0,0743	18	-0,2425	6	0,3938	1
		β-Globulin	-0,0179	19	-0,0987	15	0,1193	12	-0,3562	3
		Albumin	-0,3775	2	0,0957	16	0,0953	13	-0,2002	10
			(0,09)		(0,08)		(0,07)		(0,08)	
			[0,53]		[0,47]		[0,49]		[0,48]	

*The value in round brackets (.) is the proportion and that in square brackets [ ] the cumulative proportion of the total variance explained by the component.

39/.....

to decrease progressively from the first to the fourth component. The results, do nevertheless, show a pattern which can be explained biologically:

The serum protein can be divided into albumin and total globulin. The globulin can further be subdivided in α, β and γ-globulin fractions. The first component appears to represent the concentration of *blood serum protein fractions*. The three more important variables are total globulin, the albumin/total protein ratio and γ-globulin. Albumin is generally accepted as the best biochemical indicator of protein status. The γ-globulin level is related to the effect of continued antigen stimulation on antibody production, and hence would reflect the degree of exposure to infection (Lubran, 1966).

The second component, explains some 12 percent of the total variation. It contains cholesterol and phospholipids, which are fats. The absorption of vitamin A is dependent on the presence of fat, and carotene is the precursor of vitamin A. It would thus appear that this component represents the *concentration of fat and fat soluble vitamins* in the blood.

The third component explains some 9 percent of the total variation. The more important variables viz. 2-Pyridone, N^1-Me, Pyridone/N^1-Me ratio are all nicotinic acid metabolites and represent different ways in which *nicotinic acid status* can be measured. Since 2-pyridone has the largest eigen vector for each of the three non-White groups, it would appear to be the best indicator of nicotinic acid status in these groups. The other variables in the third component mostly represent the protein fractions which characterised the first component.

The fourth component explains some 8 percent of the total variation. It appears mostly to represent the *water soluble vitamin complexes*.

40/.....

4.3 Principal component analysis of the haematological measurements for each of the four racial groups.

The results have been summarised in Table VIII.

Once again in the case of the haematological variables, the proportion of total variation explained by each component is low, but reasonably consistent across the four racial groups. We can see that the first component explains some 25 to 27 percent of the total variation, the second some 17 to 23 percent and the third some 12 to 16 percent.

There is an interesting interchange between the first and second components in the four racial groups. For the Whites and Asiatics, haemoglobin and haematocrit have the highest eigen vectors in the first principal component and both eigen vectors have the same sign for a particular racial group. Neutrophils and lymphocytes, have the highest eigen vectors in the second principal component but have opposite signs. In the case of the Bantu and the Coloured racial groups, however, the position is reversed: Neutrophils and lymphocytes have the highest eigen vectors in the first principal component with opposite signs whilst haemoglobin and haematocrit have the highest eigen vectors in the second principal component, but with similar signs. They are accompanied, in the case of the Bantu group, by sedimentation rate.

The grouping of the haemoglobin and haematocrit together in the first component (White and Asiatics) appears to have biological significance. The haemoglobin, (the colouring matter in the red blood corpuscle), is an iron-binding substance, involved in the *absorption of oxygen* into the blood stream. The haematocrit represents the concentration of red blood corpuscles expressed as a percentage of whole blood. The first component would, therefore, represent the capacity of the blood to absorb oxygen or *oxygen uptake*. In the second component (Whites and Asiatics) lymphocytes and neutrophils which are both fractions of the total white cell count, have high eigen vectors (but of opposite sign).

41/.....

TABLE VIII : PRINCIPAL COMPONENT* ANALYSIS OF THE HAEMATOLOGICAL VARIABLES

Component			White		Bantu		Coloured		Asiatic	
Comp. No.	Characteristic	Variable	Eigen Vector	Rank	Eigen Vector	Rank	Eigen Vector	Rank	Eigen Vector	Rank
1	Oxygen uptake (Whites and Asiatics). Combatting antibody-producing infection (Bantu and Coloureds)	Haemoglobin	-0,6152	1	-0,2571	3	-0,3845	4	0,6200	1
		Haematocrit	-0,5065	2	-0,2473	4	-0,4125	3	0,5337	2
		Neutrophils	-0,2191	5	-0,6550	1	-0,5186	1	-0,1581	5
		Lymphocytes	0,1883	6	0,6225	2	0,4876	2	0,1362	6
			(0,25)		(0,27)		(0,25)		(0,26)	
2	Combatting antibody producing infection (Whites & Asiatics). Oxygen uptake. (Bantu and Coloureds)	Neutrophils	0,6569	1	0,2319	6	0,4020	3	0,6660	1
		Lymphocytes	-0,6331	2	-0,1634	8	-0,3664	4	-0,6339	2
		Haemoglobin	-0,1563	6	-0,5772	1	-0,4846	1	0,1662	5
		Haematocrit	0,1239	8	-0,5126	2	-0,4471	2	0,1456	6
		Sedimentation rate	0,2289	3	0,4068	3	0,3641	5	0,1047	8
			(0,22)		(0,17)		(0,23)		(0,23)	
			[0,47]		[0,44]		[0,48]		[0,49]	
3	Combatting parasitic infections and allergies (Whites, Coloureds and Asiatics) Oxygen uptake (Bantu).	White cells	0,1049	7	-0,0751	8	0,3848	4	-0,5740	2
		Monocytes	0,8202	1	0,2361	5	-0,5386	2	0,5719	3
		M.C.H.C.	-0,2472	3	0,7060	1	0,5650	1	-0,0028	9
		Eosinophils	0,3648	2	0,1352	6	0,3872	3	-0,5815	1
			(0,12)		(0,16)		(0,13)		(0,13)	
			[0,59]		[0,60]		[0,61]		[0,62]	

*The value in round brackets () is the proportion and that in square brackets [ ] the cumulative proportion of the total variance explained by the component.

42/.....

The role of lymphocytes is to combat antibody-producing infections, while neutrophils are involved in combatting any foreign bodies in the blood. This component would, therefore, represent the body's ability to fight infection, particularly *antibody-producing infection.* The finding that for the Bantu and Coloureds the variables related to combatting infection, appear to take precedence over those related to oxygen uptake, may indicate that in these populations, a greater heterogeneity exists than in the case of the Whites and Asiatics in respect of the extent to which infection is present. Thus, the Bantu and Coloured population groups may well be subjected to infection, particularly antibody-producing infection, to a greater extent than the Whites and Asiatics.

The third principal component is responsible for some 13 percent of the variation but there is little consistency across the four racial groups, both monocytes which combat any foreign bodies in the blood and eosinophils, which combat parasitic infections and allergies, have high eigen vectors for 3 of the 4 groups (the Bantu being excepted). The total white cell count is high for the Asiatics. The MCHC (mean corpuscular haemoglobin content) has high eigen vectors for all racial groups except the Asiatics. This variable is a function of the haematocrit and the haemoglobin and would, therefore, also reflect the uptake of oxygen. The third component, therefore, seems to combine the aspects of fighting infections, *particularly parasitic infection and allergies with that of oxygen uptake.*

The fact that the eosinophil count does not have a high loading for the Bantu implies that this race group does not have a high degree of heterogeneity in respect of this variable. This suggests a uniform degree of parasitic infection at any level. In the Bantu it is likely to be high.

It would appear that the primary haematological characteristics of the blood, viz. to *absorb oxygen* and to *fight infection*, in which the red blood corpuscles and the white cells, respectively figure prominently, have been described by the first two components. These together represents some 47 percent of the variation.

4.4 Principal component analysis of the dietary measurements for each of the four racial groups

It is clear that no valid basis for comparing dietary intake can be made in terms of actual *food intake*. It is first necessary to know the nutrient content of all foods eaten and then to calculate an estimate of the total daily intake of the various nutrients such as fats, proteins, carbohydrates etc. Estimates of daily nutrient intake for each of 15 nutrients were made for each child surveyed. The nutrients have been listed in Table V (for the sake of convenience, calories will be termed a nutrient). In the case of proteins, some problem existed, particularly in respect of certain composite dishes, in distinguishing between animal and vegetable protein. Since this was not always possible it was sometimes necessary to class the protein as "mixed" protein, hence the separate category under this heading.

The results of the principal component analysis have been summarised in Table IX.

In the case of the dietary measurements 46-52 percent of the total variation is explained by the first principal component; 9-14 percent by the second component and 8-10 percent by the third. In the first principal component, protein ranks either first or second for each of the four racial groups. Phosphorus, likewise, ranks first to third and calories second to fourth. Thiamine has a ranking of third for Whites and fourth for Bantu and Coloureds but only sixth for Asiatics.

44/.....

TABLE IX : PRINCIPAL COMPONENT* ANALYSIS OF THE DIETARY VARIABLES

Component			White		Bantu		Coloured		Asiatic	
Comp. No.	Characteristic	Name	Eigen Vector.	Rank	Eigen Vector.	Rank	Eigen Vector.	Rank	Eigen Vector.	Rank
1.	General intake of protein and energy producing foods	Protein	0,3301	1	0,3347	2	0,3353	2	-0,3310	1
		Phosphorus	0,3266	2	0,3355	1	0,3343	3	-0,3228	3
		Thiamine	0,3143	3	0,3314	4	0,3122	4	-0,2831	6
		Calories	0,3140	4	0,3327	3	0,3507	1	-0,3297	2
			(0,52)		(0,51)		(0,46)		(0,52)	
2.	Relative intake of animal protein and vegetable protein foods	Animal protein	0,3213	4	0,3317	5	-0,5498	1	0,5850	1
		Vegetable prot.	-0,5115	1	-0,0863	12	0,2201	8	-0,3986	2
		Carbohydrates	-0,4128	2	0,1038	11	0,2275	6	-0,2587	5
		Vitamin A	0,2411	7	-0,4469	1	-0,3105	4	-0,2233	6
		Iron	-0,0285	15	-0,3797	2	0,2185	9	0,0310	15
		Calcium	0,3022	5	-0,3642	3	0,1665	10	0,3414	3
		Vitamin C	0,1939	8	-0,3524	4	-0,2244	7	-0,1221	12
		Mixed protein	0,1596	9	0,1310	10	0,3763	2	0,0687	13
		Nicotinic acid	-0,0363	14	0,0145	15	-0,3424	3	-0,1534	10
			(0,12)		(0,14)		(0,13)		(0,09)	
			[0,64]		[0,65]		[0,59]		[0,61]	
3.	Relative intake of mixed protein and animal or vegetable protein foods.	Mixed protein	0,7913	1	0,3453	4	0,5671	1	-0,7428	1
		Animal protein	-0,4516	2	0,3924	1	-0,0900	12	0,3052	4
		Vitamin A	-0,1036	7	0,3904	2	0,2876	5	-0,0248	13
		Vegetable prot.	-0,1532	5	-0,3755	3	-0,4072	2	0,2946	5
		Calcium	-0,0081	15	-0,0062	15	0,3943	3	0;1201	8
			(0,08)		(0,10)		(0,10)		(0,09)	
			[0,72]		[0,75]		[0,69]		[0,70]	
	-	Vitamin A	0,6576	1	-0,1091	8	-0,5315	1	-0,5202	2
		Animal protein	-0,1832	7	-0,3912	2	0,4160	2	-0,0723	10
		Vitamin C	0,5051	2	-0,1056	9	-0,3883	3	-0,6180	1
			(0,06)		(0,08)		(0,08)		(0,07)	
			[0,78]		[0,83]		[0,77]		[0,77]	

*The value in round brackets () is the proportion and that in square brackets [ ] the cumulative proportion of the total variance explained by the compoment.

45/.....

The significance of the first principal component requires careful consideration. One might, on first thoughts have expected those variables to cluster together in a particular component which had the *same biological characteristics*, as in the case of the biochemical variables. After careful consideration, however, we must realise that this is not likely; the intake of specific nutrients is clearly determined by the intake of the *foodstuffs* in which they are borne. Since, in general, the type of food eaten by the four racial groups differed considerably, one could expect that this would have a marked effect on the pattern which emerged in the nutrient intake.

In order to investigate this aspect, each food item eaten by a racial group was classified into one of 6 food groups viz. *cereal group; meat and milk group; vegetable group; fruit group; fats and oils group and miscellaneous items.* We then assessed for each racial group, which particular food group was mainly responsible for supplying protein and phosphorus, the two nutrients with the highest eigen vectors in the first component. We found that for the Whites and Asiatics both protein and phosphorus were derived primarily from the *meat and milk group*. For Bantu and Coloureds, however, both these nutrients were derived primarily from the *cereal group*. It would thus appear that protein and phosphorus are grouped together in the first component, not because they are related in terms of their biological characteristics, but because, for a particular population, they are derived mainly from the same food group.

It would then seem that the first component represents the general *intake of protein containing and energy producing foodstuffs.*

Our findings bear a remarkable resemblance to those of Drion (1961, p.326) who carried out a study of the nutrients consumed by a group of families in the Netherlands. Drion

46/.....

applied a principal component analysis, separately to various family sizes (varying from 2 to 7 and over). In his case the analysis was applied, not to the nutrient intake, but to the nutrient intake expressed as a proportion of the recommended daily allowance - a quantity which Drion called the "reduced consumption". This aspect is discussed later (Chapter 6.1). In his analysis Drion found that the first principal component had high eigen vectors (0,85 and higher) for total protein, calories and thiamine, in that order. Ignoring phosphorus, which was not assessed by Drion, we can note that we obtained essentially the same result. This is indeed a remarkable similarity. Drion (p.329) interpreted his results as follows:

> "The factor loading of total protein is in general higher than that of calories; this indicates a rather well-balanced diet so far as the major nutrients are concerned. The factor loading of thiamine is high; it is well known that this vitamin is needed for internal combustion, so a diet with a high caloric value needs a high thiamine content for complete health. The high factor loading shows that in general, the Dutch diet contains enough thiamine in relation to its caloric value".

If we accept his interpretation we can only add that his remarks were also found to be true for the White population group. The fact that for the Asiatics protein also has the highest eigen vector, whilst in the case of the Bantu and Coloureds, protein had the second highest eigen vector, might be an indication that the Whites and Asiatics satisfy their protein requirements from the *meat and milk group*, whilst the Bantu and Coloureds, who are less well off socio-economically and cannot afford high protein foodstuffs, satisfy their appetites from the *cereal food group*.

The second principal component explains some 9-14 percent of the total variation. There is less similarity across the four racial groups. Animal protein ranks first, for both Coloureds and Asiatics, whilst vegetable protein ranks first

for Whites and second for Asiatics. Calcium ranks third for the Bantu and third for Asiatics. It is interesting to note that the signs of the eigen vectors for animal and vegetable protein differ for each racial group. This may be explained as follows: Assume a person eats protein-containing foodsstuffs of animal or vegetable origin, until he is satisfied, if he eats more animal protein he would tend to eat less vegetable protein, and vice versa. The second component would thus appear to represent a contrast between the *meat and milk food group* (containing animal protein) and the *cereal and vegetable food groups* (containing vegetable protein). For Whites, Bantu and Asiatics, the second component would increase as the proportion of animal to vegetable protein in the diet increased, and decrease as the proportion decreased.

The third principal component explains some 8-10 percent of the total variation. Here mixed protein(occurring in composite dishes where animal and vegetable proteins were mixed in unknown quantities) ranks first for Whites, Coloureds and Asiatics, but fourth for Bantu. Animal protein ranks second for Coloureds and third for Bantu. It is interesting to note that for the Whites, Coloureds and Asiatics the eigen vector for mixed proteins has the opposite sign to that for animal and vegetable proteins. In the case of the Bantu it has the same sign as animal protein but the opposite to that of vegetable protein. It would then appear that the third component could represent the proportion of protein derived from *composite protein dishes* as compared to that derived from either *animal* or *vegetable protein*.

For the fourth principal component, vitamin A ranks first for Whites and Coloureds and second for Asiatics. Animal protein ranks second for Bantu and Coloureds, and vitamin C first for Asiatics and second for Whites. The fourth component explains 6-8 percent of the variation (Table IX) but does not seem to have any particular significance.

48 /.....

It would then appear that the principal component analysis has provided information as to the type of foodstuffs ingested and the relative intake of animal, vegetable or mixed protein foodstuffs. The grouping of the nutrient intake variables into components, makes sense in terms of the foodstuffs from which they originate rather than in terms of their biological function in the human body. The biological interrelationship of the nutrients can only be ascertained once they have been absorbed into the blood stream, as has been seen in the case of the biochemical variables.

4.5 Critical appraisal of the principal component analyses

In view of the considerable time that the technique of principal component analysis has been around (Pearson, 1901), there are remarkably few documented examples of the application of this technique in the biological sciences. The greater majority of these applications have been made in respect of the measurement of the morphological characteristics of a wide range of living organisms.

Applications to taxonomic problems have been reviewed by Jeffers (1963), who lists a number of papers in which the authors aimed at identifying plant species by their physical characteristics. Cassie (1963) applied the technique to numerical plankton data. Ouellette and Qadrie (1966) used it to describe a pattern of growth in the dog-fish *(Cristivomer namaycush)*.

Jolicoeur and Mosimann (1960) used a principal component analysis to study size and shape variation in the painted turtle, whilst Rees (1969) used the technique to study morphological variation of the mandible of the white-tailed deer *(Odocoileus virginianus)*. Drion, (1961) as already mentioned, has used the technique to study the dietary components of nutrients ingested by families in the Netherlands.

49/.....

It is interesting to note that all except the last two applications are related to the morphological characteristics of organisms. Jollicoeur and Mosimann were able to identify two separate components, the one relating to the *size* of the painted turtle and the other to the *shape*. Rees was able to show that the two major components of *mandible shape variation* among White-tailed deer (within breeding groups) involved a contrast between the mandible and the dentition and between the pre-molars and the molars. *Size* variation was found to account for 34 percent of the total variation and the two major shape variations to account for 23 and 8 percent respectively. Ouelette and Qadrie, however, could only distinguish the general size factor as described by the first component in both male and female dog-fish.

The components which we have been able to identify in the somatometric variables, appear to describe the physical characteristics of the human body in a manner, similar to that in which the physical characteristics of other living organisms, be they plant, plankton, fish, turtles or deer, have been described by principal component analyses. Our results, also in fact, bear a strong resemblance to those obtained by Burt and Banks (1947), who carried out a *factor analysis* on the body measurements of British adult males. They identified a first factor of *general body size* which contributed over 50% of the total variation. A second factor, was noted to be bipolar contributing about 13% of the total variation. This classified traits into: *longitudinal and transverse or circumferential*. It would thus represent a lean versus a thick-set body build. Since Burt and Banks did not measure the skin-fold thicknesses, the factor relating to body fat was not identified by them.

As can be seen above, our application of the principal component analysis to the biochemical variables has yielded less explicit, though we judge, still meaningful results. No previous application of the technique to such biochemical observations could be traced.

50/.....

In the case of the haematological variables a similar situation exists. The analysis appears to yield meaningful though not clear cut results. The interchange between the first and second components for the Bantu and Coloured groups as compared to the White and Asiatic groups is most interesting. As far as we know, this is the first time a principal component analysis has been applied to haematological data.

The analysis carried out on the dietary variables appeared to contribute some information about the main food group from which certain nutrients were derived. This sort of information could, however, be derived more directly and more readily by other means.

In general, it would appear from the literature, that the more frequent and more successful biological applications have been in respect of morphological characteristics descriptive of size and shape. This is in agreement with our findings that the principal component analysis on the somatometric measurements yielded more clear cut, and more readily interpretable results than did those on the other disciplines which we have studied.

The relevance of these results to the problem we are studying, will become clear in the subsequent chapters.

CHAPTER 5

ANALYSIS OF CERTAIN DISCIPLINES USING INFORMATION GERMANE TO THE POPULATIONS SURVEYED

Having come to some understanding of the information content of the variables associated with each discipline, we can now proceed with the second phase of our analysis, viz. the analysis of certain subsets of the data, using information germane to the populations on which the variables were observed. Since the measurements were made on different racial groups, these are likely to reflect certain inherent differences. These will first be investigated for the somatometric variables in respect of Whites and Bantu.

5.1 Investigation of racial differences in the somatometric variables for White and Bantu school children by means of a discriminant analysis

The somatometric variables can be measured more readily and cheaply than any other in the available data set. We have seen from Chapter 4.1 that subsets of these variables can be selected, descriptive of *general body size, soft body tissue, and body fat.* It would clearly be most useful if an index for assessing nutrition status could be based on the somatometric variables, but we must first establish how the characteristics which we have identified as components in the principal component analysis, are related to nutrition status.

A stepwise discriminant analysis was, therefore, applied to the 11 somatometric variables, to establish firstly, to what extent these could be used to distinguish between children from the White and Bantu population groups. The rationale behind this analysis was as follows: There is no doubt that the majority of the White children in the sample had a much higher socio-economic status than the majority of the Bantu children. It can be shown that at the low socio-economic level of many Bantu, it is virtually impossible to maintain an adequate level of nutrition (Watts, 1967). It is thus to be expected that the Whites would, in the main, also

have a better nutrition status. The Coloured and Asiatic groups have a socio-economic level between that of the Whites and the Bantu. They were, therefore, omitted from the present analysis in order to increase the chance of obtaining discrete homogenous groups. One cannot, however, forthwith assume that any differences which may be demonstrated between White and Bantu, are definitely due to a difference in nutrition status. They could, for example, be caused by genetic differences.

In the application of the stepwise discriminant analysis, the age variable was deliberately omitted, although many of the somatometric observations are clearly age dependent. The reason for this will become clear later.

The following discriminant functions for Whites (X_W) and Bantu (X_B) were found:

$$
\begin{aligned}
X_W &= -414{,}219 - 12{,}108x_1 + 4{,}962x_2 - 2{,}375x_3 \\
&\quad + 5{,}637x_4 + 3{,}677x_5 - 0{,}560x_6 + 12{,}588x_7 \\
&\quad + 8{,}118x_8 + 0{,}065x_9 + 4{,}531x_{10} - 0{,}274x_{11} \\
X_B &= -399{,}342 - 12{,}076x_1 + 4{,}805x_2 - 2{,}261x_3 \\
&\quad + 4{,}469x_4 + 3{,}391x_5 + 0{,}626x_6 + 12{,}866x_7 \\
&\quad + 8{,}108x_8 - 0{,}389x_9 + 4{,}863x_{10} - 0{,}248x_{11}
\end{aligned}
\qquad (1)
$$

Where x_i is the i-th somatometric variable (i = 1, 2, ... 11) (The variables are identified by the *"within disciplines"* variable number (in brackets) in Table V, Chapter 3).

In order to assess in which group a particular subject would fall, the subject's somatometric measurements (x_i) are substituted in each of the two equations (1) above. The subject is then allocated to that group which yields the highest value of X.

53/.....

If in equation (1) we take differences, term by term, this would give:

$$\begin{aligned} X_{W-B} = {} & 14,877 - 0,032x_1 + 0,157x_2 - 0,114x_3 \\ & + 1,168x_4 + 0,286x_5 - 1,186x_6 - 0,278x_7 \\ & + 0,010x_8 + 0,454x_9 - 0,332x_{10} - 0,026x_{11} \end{aligned} \quad (2)$$

X_{W-B} will be positive for those cases which should be allocated to the White group and negative for those cases which should be allocated to the Bantu group.

In a stepwise discriminant analysis, the variables are included in the order of their contribution to the differentiation between the groups. Thus, since x_4 gives the greatest ratio of the variance between groups to that within groups and is, therefore, best able to discriminate between the two groups, it is selected first. The next variable to be selected (x_6) is that variable which, when used together with x_4, results in the best improvement in discrimination. This does not mean that x_6 is the second best variable for discriminating between the two groups. There may well be another variable, say x_p which is highly correlated with x_4 and could, if used alone, discriminate almost as well as x_4 and much better than x_6. Since, however, x_4 has already been entered, most of the information contained in x_p has already been utilized, and it has little "new" information to offer. It may, therefore, be relegated to a place low-down in the list of selected variables. Thus, in general, that variable is selected which, when partialled on the previously entered variables, has the highest multiple correlation between the groups (Dixon, 1968 p.214a). This point is important since, if given another sample of the data, the position of x_4 and x_p could possibly be reversed.

The order in which the somatometric variables were included in the discriminant analysis is given in Table X. The F-value (with degrees of freedom) in the table reflects

54 /.....

TABLE X : ORDER IN WHICH THE SOMATOMETRIC VARIABLES WERE SELECTED FOR DISCRIMINATING BETWEEN WHITES AND BANTU

Step No.	Variable Selected		Sign of coefft.	Characteristic	F-value	Degrees of freedom	P%
	No	Name					
1	4	Intercristal width	+	Skeletal	339,7	1 1159	<0,001
2	6	Ulnar length	-	proportion.	400,2	2 1158	<0,001
3	5	Biacromial width	+		25,2	3 1157	<0,001
4	7	Upper arm circumference	-	Proportion of soft body tissue to body fat.	20,5	4 1156	<0,001
5	9	Triceps skinfold thickness	+		41,7	5 1155	<0,001
6	10	Subscapular skinfold thickness	-		48,6	6 1154	<0,001
7	2	Height	+	Proportion of leg length to height.	13,0	7 1153	<0,001
8	3	Cristal height	-		10,4	8 1152	<0,001
9	1	Weight	-	No contribution to discrimination	0,8	9 1151	62
10	11	Para-umbilical skinfold	-		0,4	10 1150	95
11	8	Calf circumference	+		0,0	11 1149	100

55/.....

the importance of the contribution of each of the 11 variables to the discrimination.

The first three variables to be included, viz. intercristal width, (hip width) ulnar length and biacromial width (shoulder width) are all skeletal measurements. The first and third are horizontal measurements (See Fig. 1). They have probably both been given an important place, since the first tends to be relatively wide in females and the last in males. The first was one of the variables which we found to be descriptive of general body size (See Chapter 4.1 and Table VI). The second variable viz. ulnar length (or length of the forearm excluding the hand, Fig. 1) is an extremity which is not featured in any of the characteristics dealt with in Chapter 4.

It is instructive to note the signs of the coefficients of the variables in (2) above. These have been repeated in column 4 of Table X for convenience. The coefficients of both intercristal width (x_4) and biacromial width (x_5) are positive. Thus, if a person's intercristal width and/or his biacromial width *increased* relative to the values assumed by the other variables, then his value for X_{W-B} would move towards the (positive) White group. Likewise, his value for X_{W-B} would move towards the (negative) Bantu group as his values for these two variables decreased. The coefficient of ulnar length (x_6) is, however, negative. Thus, if a person's ulnar length decreased, relative to his other measurements, his value for X_{W-B} would move towards the White group, and vice versa. Therefore, an increase in the *ratio* of biacromial or intercristal width relative to ulnar length would result in a move of X_{W-B} toward the White group and vice versa. Since these are the three most important variables for discriminating between the two racial groups, this suggests *that the discrimination is based primarily on differences in skeletal proportion.*

This result could also be argued as follows: Children

56/.....

in the age range 7-15 years have been included in the analysis and the affect of age has not been partialled out. (See above). The variation in the body measurement from a young child to an older one, is clearly far greater than it could be from a child of one race group to that of another. It would follow, therefore, that no discrimination could be based solely on an absolute measure of general body size and must, therefore, depend on changes in body proportion.

The next variable to be included is upper arm circumference (negative coefficient) which we found (Chapter 4.1) to be an important indicator of both *general body size* and also of *soft body tissue*. This is followed by the triceps and subscapular skinfold thicknesses which we found to be indicative of *body fat*. Although the coefficient of the latter has a negative sign, the sum of the two coefficients is positive. By a similar reasoning to that presented above, it would follow that there are proportional differences between White and Bantu in respect of *soft body tissue* and *of body fat*.

It must be born in mind when considering the value and sign of the coefficients of each variable in (2) above that these are determined on the basis of all 11 variables. The order in which the variables are included is, as discussed previously, dependent on their contribution to the discrimination which is assessed, variable by variable. It might, therefore, have been better to base the discriminant functions only on those variables which make a significant contribution to the discrimination.

The next two variables, height and cristal height, are the last to make a significant contribution to the discrimination. They are both skeletal measures of vertical body size. The first has a positive sign and the second a negative. It would, therefore, seem that they contribute some further information in respect of proportional differences in skeletal structure between Whites and Bantu.

57/.....

In summary then, we have been able to demonstrate proportional differences between Whites and Bantu in respect both of skeletal proportion, and proportion of soft body tissue, to body fat. Since, on the surface, it seems unlikely that nutritional deficiency would result in disproportional skeletal growth, one might postulate, as is often done, that the skeletal differences are of genetic origin, while those in respect of soft body tissue and body fat are nutritional. We shall, however, investigate this point further below and return to it in the following chapter.

In order to assess the ability of discriminant functions to distinguish between the racial groups, each case was allocated to one of the two groups using the discriminant function (2) above. This procedure will, in general, give a slightly better classification than would be found with a new set of observations. In spite of this bias it does, however, provide a fair indication of the reliability of the discriminant technique.

A summary of the allocation is given in Table XI.

TABLE XI : CLASSIFICATION INTO RACIAL GROUPS BASED ON THE SOMATOMETRIC VARIABLES

		Group in which case was classified		Total no. of cases
		White	Bantu	
Group in which case belonged	White	529	105	634
	Bantu	56	471	527

Of the total of 634 White cases, 529 (83%) were correctly classified. Of the 527 Bantu cases 471 (89%) were

correctly classified. It is thus possible to distinguish between the two racial groups with a considerable degree of accuracy.

The first and second canonical variates based on the somatometric variables were calculated for the White and Bantu racial groups. In Fig. 2 the second canonical variate (vertical axis) has been plotted on the first (horizontal axis). It can be seen that the plotted points associated with the two racial groups form separate clusters, though with some degree of overlap. The mean of each group (indicated by an O) lies roughly on the perimeter of the cluster of the other.

5.2 Investigation of the relationship between socio-economic status and the somatometric variables for Whites and Bantu by means of a discriminant analysis.

In an attempt to establish to what extent the differences observed above are due to genetic or nutritional factors, a discriminant analysis was performed separately on the White and Bantu groups after, in each case, a trichotomy of the sample had been made on the basis of the rate per head per day (R/H/D), see Table V.

The R/H/D reflects the amount of money available per person per day after the obligatory expenses relating to rent, income tax, transport, rates and taxes, and water and lights, have been deducted. Thus the R/D/H shows the amount of money that is available on a daily basis to provide for all needs in respect of food, clothing and fuel. It is, therefore, reasonable to suppose that there would be a strong positive association between R/H/D and the nutrition status of the child in the case of those groups where the money available was only marginally sufficient to supply basic needs. The distribution curve of the R/H/D for the Whites is shown in Fig. 3, and that for Bantu in Fig. 4. The considerable disparity in the R/H/D for the two population groups is immediately evident.

59 /.....

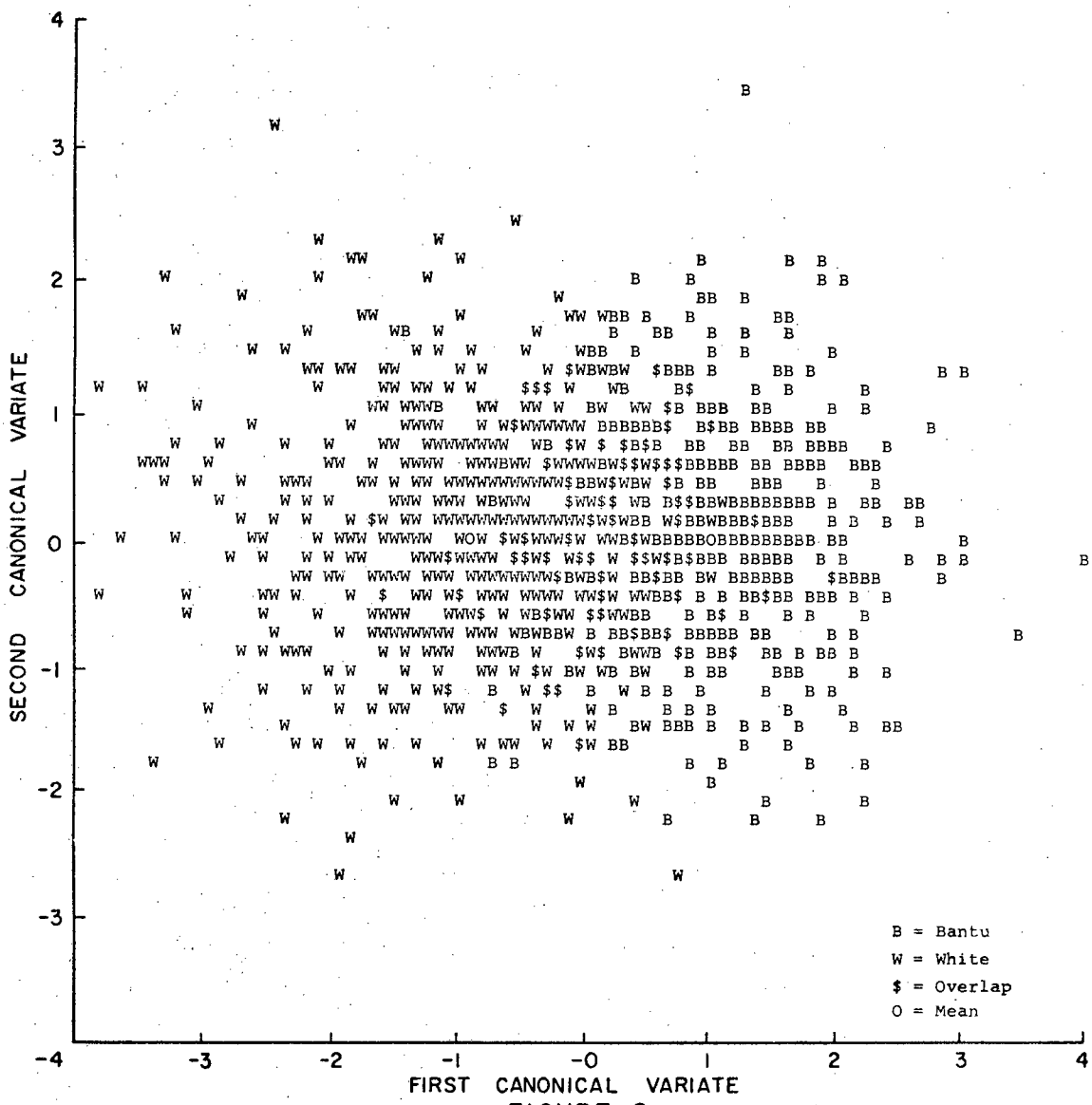


FIGURE 2

Scatter diagram of first and second canonical variates based on the somatometric variables for White and Bantu

60/.....

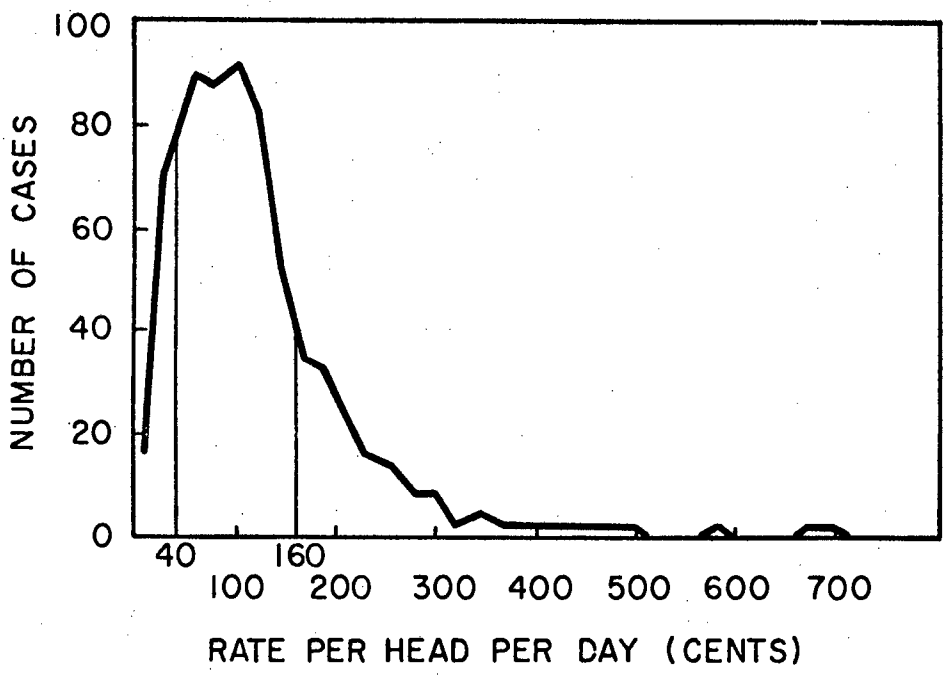


FIGURE 3

Distribution of R/H/D for White Pretoria schoolchildren

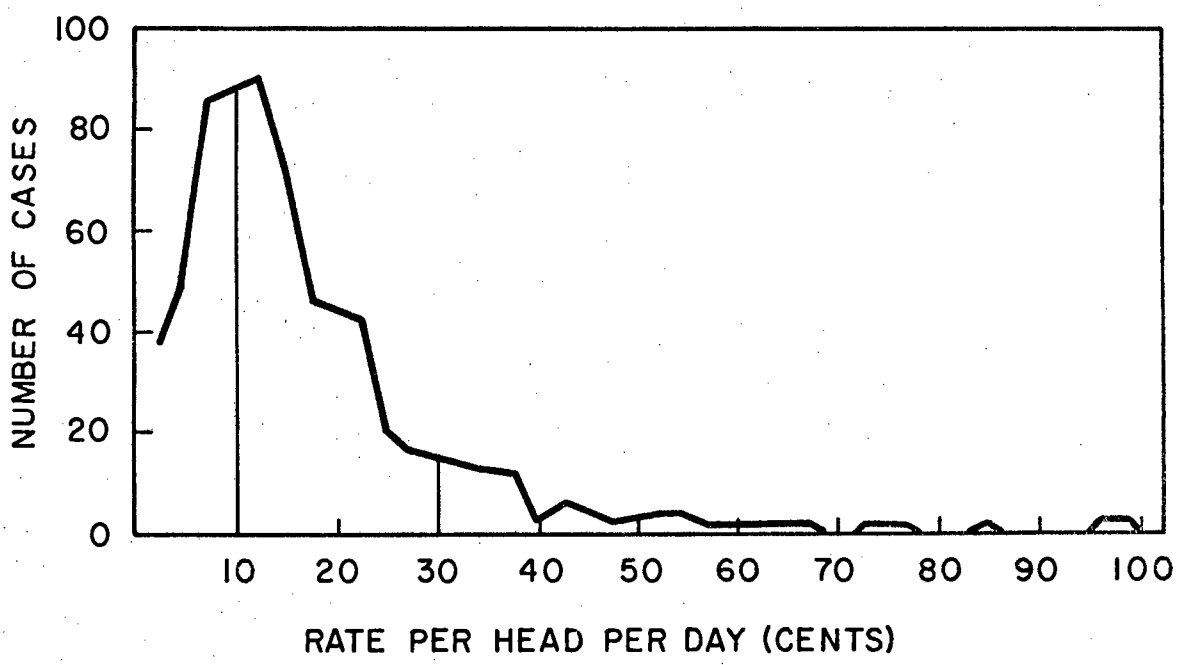


FIGURE 4

Distribution of R/H/D for Bantu Pretoria schoolchildren

61/.....

The peak for the White group lies in the region of 100 cents/day, while that of the Bantu is about 12 cents/day.

The relationship between the R/H/D and the somatometric variables could clearly be investigated by means of a stepwise regression analysis, since R/H/D is measured on a continuous scale. This would, however, only detect *linear* relationships and will be done in the next section. We felt, however, that it was possible that the relationship might be *non-linear*. Such a situation could, for example, exist if both a low and a high R/H/D lead to malnutrition: The low R/H/D due perhaps to inadequate food supplies, and the high R/H/D due to an unbalanced diet, such as tends to occur in the case of children of rich parents.

The children of each population group were, therefore, divided into three socio-economic groups as shown in Table XII.

TABLE XII : CLASSIFICATION INTO SOCIO-ECONOMIC GROUPS

		Socio-economic group (cents/day)		
		Low	Medium	High
White	Range	[0-40]	(40-160]	(160-]
	Mean	26,8	94,0	246,2
	No. of cases	69	425	140
Bantu	Range	[0-10]	(10-30]	(30 -]
	Mean	6,0	16,8	47,2
	No. of cases	229	243	55

Stepwise discriminant analyses were separately applied to the somatometric variables for each racial group to assess the value of these variables in distinguishing between the three socio-economic groups. It should be born in mind that we have here made an artificial trichotomy of a continuous vari-

able (R/H/D), and thus cannot expect to find clear cut discrimination.

In the case of the White population, the discriminant analysis on all 11 variables showed a highly significant difference between group means on the basis of an approximate F-test (p = 0,003%). The results were, however, not significant in the case of the Bantu (p = 37%), and no importance can, therefore, be attached to the order in which the variables have been included. This may be due partly to the way in which we have made the trichotomy for the Bantu (Fig. 4). The first boundary line chosen, bisects the distribution of R/H/D approximately at its mode and is, therefore, highly unlikely to form discrete homogenous groups in terms of any observed variables associated with R/H/D. This boundary line was chosen since it was felt that a sum of money of less than 10 cents per head per day would be totally inadequate to provide even basic nutritional needs, and that malnutrition would occur in this group. Since the possibility of non-linearity due to overfeeding is highly unlikely in the case of the Bantu, the stepwise regression analysis, which will be presented in the next section, should be adequate for investigating the relationship for this race group.

The order in which the variables have been included for the White group is summarised in Table XIII.

It can immediately be seen that only in the case of the first variable entered (biacromial width), was the F-value significant. The fact that the next two variables to be included, viz. upper arm circumference and calf circumference both represent *soft body tissue* (See Chapter 4.1) might suggest that the differences in the three White socio-economic groups are due to a varying proportion of soft body tissue in relation to skeletal size. In view, however, of the lack of significant F-values, further speculation in this respect is unwarranted.

63/.....

TABLE XIII : ORDER IN WHICH THE SOMATOMETRIC VARIABLES WERE SELECTED TO DISCRIMINATE BETWEEN THE THREE WHITE SOCIO-ECONOMIC GROUPS

Step No.	Var. No.	White	F-value	Degrees of Freedom		P%
1	5	Biacromial width	21,4	1	631	<0,001
2	7	Upper arm circumference	1,3	2	1260	27
3	8	Calf circumference	1,4	3	1258	24
4	4	Intercristal width	1,0	4	1256	41
5	11	Para-umbilical skinfold	0,7	5	1254	62
6	6	Ulnar length	1,1	6	1252	36
7	3	Cristal height	2,0	7	1250	5
8	2	Height	1,0	8	1248	43
9	9	Triceps skinfold	0,3	9	1246	97
10	10	Subscapular skinfold	0,2	10	1244	100
11	1	Weight	0,4	11	1242	96

64/.....

5.3 Investigation of the relationship between socio-economic status and the somatometric variables for Whites and Bantu by means of a stepwise regression analysis

In a further attempt to elucidate the nature of the relationship between socio-economic status as reflected by R/H/D, and the somatometric variables, the technique of stepwise regression analysis was used. For both Whites and Bantu, a stepwise regression analysis was carried out taking R/H/D as the dependent variable, and the somatometric observations as the independent variables. As mentioned previously (Chapter 3.1), a stepwise regression analysis investigates the linear relationship between the dependent variables (y) and the set of independent variables (x_i). That value of x (say x_1), is first selected which best explains the variation in y, or is most highly correlated with y, and the regression of y on x_1 is computed. The remaining x variables, are then tested one by one and that one is included in the regression equation which contributes the best improvement to the regression relationship. Once again, as in the case of the discriminant analysis, if x_1 is entered first and then x_2, the effect of x_1 is partialled out of the variance covariance matrix, before x_2 is included. Thus x_2 may not be that variable which has the second highest correlation with y but is the one which contribute most "new" information.

It is, generally speaking, difficult to know just how far a stepwise regression procedure should be taken. In both the present analyses the program was allowed to run until the F-value to enter the outstanding variables fell below the value of 0,01. The degrees of freedom for this F-value are 1 and 515 for the Bantu, and 1 and 622 for the Whites. Thus in the case of both races the stepwise inclusion of the variables proceeded well beyond the stage where the variables contributed a significant improvement in the regression relationship. For both races all but one of the 11 somatometric variables have been included.

65/.....

The regression equation for the Whites was:

$$Y_W = -0,1711 + 0,0019x_1 + 0,0107x_2 - 0,0220x_3 - 0,0472x_4 + 0,0277x_5 + 0,0702x_6 + 0,0543x_7 - 0,0303x_8 + 0,0058x_9 - 0,0046x_{10} \quad (3)$$

For the Bantu it was:

$$Y_B = +0,2507 + 0,0055x_1 + 0,0003x_2 - 0,0078x_4 - 0,0015x_5 + 0,0040x_6 - 0,0012x_7 - 0,0067x_8 + 0,0171x_9 - 0,0166x_{10} - 0,0049x_{11} \quad (4)$$

Where in each case (x_i) is the i-th somatometric variable (No. in brackets, Table V).

The order in which the variables have been included for both Whites and Bantu is shown in Table XIV. For convenience the sign of the regression coefficient has been tabulated together with the multiple correlation coefficient. It can be seen that the result of the stepwise regression analysis for Whites conforms reasonably closely with that obtained from the discriminant analysis on three socio-economic groups. In both cases the first 3 variables are the same. The order of biacromial width and upper arm circumference have been reversed. The five variables which appear 6th to 10th in the regression analysis appear 7th to 11th in the stepwise discriminant analysis in the same order.

In view of the very small increase in the multiple correlation coefficient, it is, however, doubtful whether much importance can be attached to the order in which the successive variables have been included. It does seem likely that,what relationship exists between R/H/D and somatometric measurements relates to a *change in soft body tissue for a specific skeletal size*. For the Bantu the increase in the multiple correlation coefficient is more gradual, but it never reaches the

66/.....

TABLE XIV : ORDER IN WHICH THE SOMATOMETRIC VARIABLES WERE SELECTED ON THE BASIS OF THEIR RELATIONSHIP TO R/H/D BY MEANS OF A STEPWISE REGRESSION ANALYSIS

WHITE					BANTU				
Step No.	Variable No.	Variable	Sign of coefft.	Multiple correlation	Step No.	Variable No.	Variable	Sign of coefft.	Multiple correlation
1	7	Upper arm circumference	+	0,21	1	9	Triceps skinfold	+	0,09
2	5	Biacromial width	+	0,21	2	10	Subscapular skinfold	-	0,12
3	8	Calf circumference	-	0,22	3	1	Weight	+	0,16
4	6	Ulnar length	+	0,22	4	8	Calf circumference	-	0,17
5	4	Intercristal width	-	0,23	5	11	Para-umbilical skinfold	-	0,17
6	3	Cristal height	-	0,23	6	4	Intercristal width	-	0,18
7	2	Height	+	0,24	7	6	Ulnar length	+	0,18
8	9	Triceps skinfold	+	0,24	8	5	Biacromial width	-	0,18
9	10	Subscapular skinfold	-	0,24	9	2	Height	+	0,18
10	1	Weight	+	0,24	10	7	Upper arm circumference	-	0,18

67/.....

simple correlation between R/H/D and upper arm circumference in the case of the Whites. For the Bantu the more important variables seem to be those characteristic of *body fat* seen in relation to *total body weight*.

The results presented in Chapter 5.1 have served to indicate that marked differences exist between the White and the Bantu racial groups in respect of somatometric measurements. These differences appear firstly, to be in relation to *skeletal proportion* and secondly, in relation to *soft body tissue* and *body fat*. It seems possible that the difference in skeletal proportion could have a genetic cause whilst those in soft body tissue and body fat might be due to nutrition status.

Since in the case of both Whites and Bantu, even a stepwise regression analysis (which is a more sensitive test for linear relationship than a stepwise discriminant analysis), has failed to reveal any but the most tenuous relationship between R/H/D and the somatometric variables, it would appear, either that R/H/D has little relationship to nutrition status or that the differences found between Whites and Bantu are largely due to non-nutritional causes. More light will be thrown on this problem in the following chapter.

Having established that differences exist between the White and Bantu racial groups in respect of their somatometric variables, and having suggested a possible explanation of these differences, we will now extend our investigation to the biochemical variables, which form the second major group of consequential variables which have been observed.

5.4 Investigation of racial differences in the biochemical variables for *all races* by means of a discriminant analysis

In order to study racial differences in the biochemical variables (Table V) a stepwise discriminant analysis was

68/.....

carried out to see whether these variables could be used to discriminate between the four racial groups. The stepwise procedure was limited to the inclusion of 15 variables in order to conserve computation time. The order in which the variables were selected is summarised in Table XV. It can be seen from the F-value that the last variable to be selected still made a significant contribution to the discrimination. The approximate F-test for overall discrimination between the four racial groups on the basis of the first 15 variables, yielded a F-value of 88 with associated degrees of freedom, 45 and 5586 ($P < 0,00001\%$). Thus a clear cut differentiation, particularly between the White and Bantu racial groups can be made on the basis of the first 15 biochemical variables.

The first variable to be included in the analysis (Table XV) was α-globulin. We have seen in Chapter 4.2 that the blood serum protein fractions reflecting protein status, constituted the first principal component of the biochemical variables. The most important variable for the three non-White racial groups was globulin, which consists of the sum of the α-, β-, and γ-globulin fractions. Thus α-globulin is one of the protein fractions which characterise the first component. α-Globulin has never, to our knowledge, been suggested as an index of protein status. It is, however, known to increase with the presence of *inflammatory conditions*, such as body sores or abrasions. Raised values are also found in various kinds of rheumatic diseases (Lubran, 1966). Since α-globulin is expressed per unit volume of blood serum, it is not likely to be directly related to the age of the child. There is, furthermore, no physiological reason why it should vary with age (Oberman *et al.* 1956). It may well reflect, therefore, an absolute difference between the racial groups in respect of the extent to which *inflammatory conditions* are present.

The next variable viz. phospholipids, is one of the *body fats*, which is completely synthesised in the body, and is highly correlated with cholesterol. Racial differences in

69/.....

TABLE XV : ORDER IN WHICH THE BIOCHEMICAL VARIABLES WERE SELECTED TO DISCRIMINATE BETWEEN THE FOUR RACIAL GROUPS

Step No.	No.	Variable selected Name	F-value	Degrees of Freedom		P%
1	17	α-globulin	373,8	1	1893	<0,001
2	2	Phospholipids	167,8	2	1892	<0,001
3	7	Urinary amylase/creatinine	144,5	3	1891	<0,001
4	8	Vitamin A (serum)	102,5	4	1890	<0,001
5	6	Amylase (serum)	63,9	5	1889	<0,001
6	21	γ-globulin/Total protein	55,0	6	1888	<0,001
7	5	Inorganic phosphorus	49,5	7	1887	<0,001
8	12	N'-Me (urinary)	58,0	8	1886	<0,001
9	22	Riboflavin (urinary)	29,2	9	1885	<0,001
10	10	Thiamine (urinary)	30,3	10	1884	<0,001
11	4	Alkaline phosphatase	20,3	11	1883	<0,001
12	14	Total protein	14,9	12	1882	<0,001
13	1	Cholesterol	12,9	13	1881	<0,001
14	13	Pyridone/N'-Me	11,4	14	1880	<0,001
15	9	Carotene (serum)	9,9	15	1879	<0,001

70/.....

serum cholesterol have frequently been pointed out. This variable, however, only appears in the 13th step, probably because the bulk of the information it contains has already been represented by phospholipids.

The third and fifth variables measure the activity of the same enzyme, amylase in urine and serum. This enzyme is responsible for breaking down dietary starches into simple sugars, and probably relates to the amount of *dietary starch* habitually consumed. The fourth variable is a fat soluble vitamin.

The racial groups appear then, to differ most markedly in respect of the proportion of protein fractions related to *acute (inflammatory) infections, chronic (antibody-producing) infections,* and *fat, starch and vitamin intake.*

Since total protein only appears in step no. 12 and albumin, which is generally regarded as the best indicator of an inadequate protein intake, has not been included in the first 15 variables, the results on the biochemical variables would seem to suggest that the *adequacy of protein intake is not an important factor in differentiating between the four racial groups.* This confirms Du Plessis' observation (Du Plessis, 1967, Chapter 3) that, according to the biochemical variables, no sign of acute protein deficiency could be found in any of the four racial groups.

The first two canonical variates have been plotted in Fig. 5. There is a surprisingly good separation between the White and Bantu racial groups in respect of the first canonical variate (x axis), but no difference in respect of the second canonical variate. The Coloureds and Asiatics are superimposed on one another. They lie in an intermediate position between the Bantu and the Whites, in respect of the first canonical variate and are slightly higher in respect of the second.

71/.....

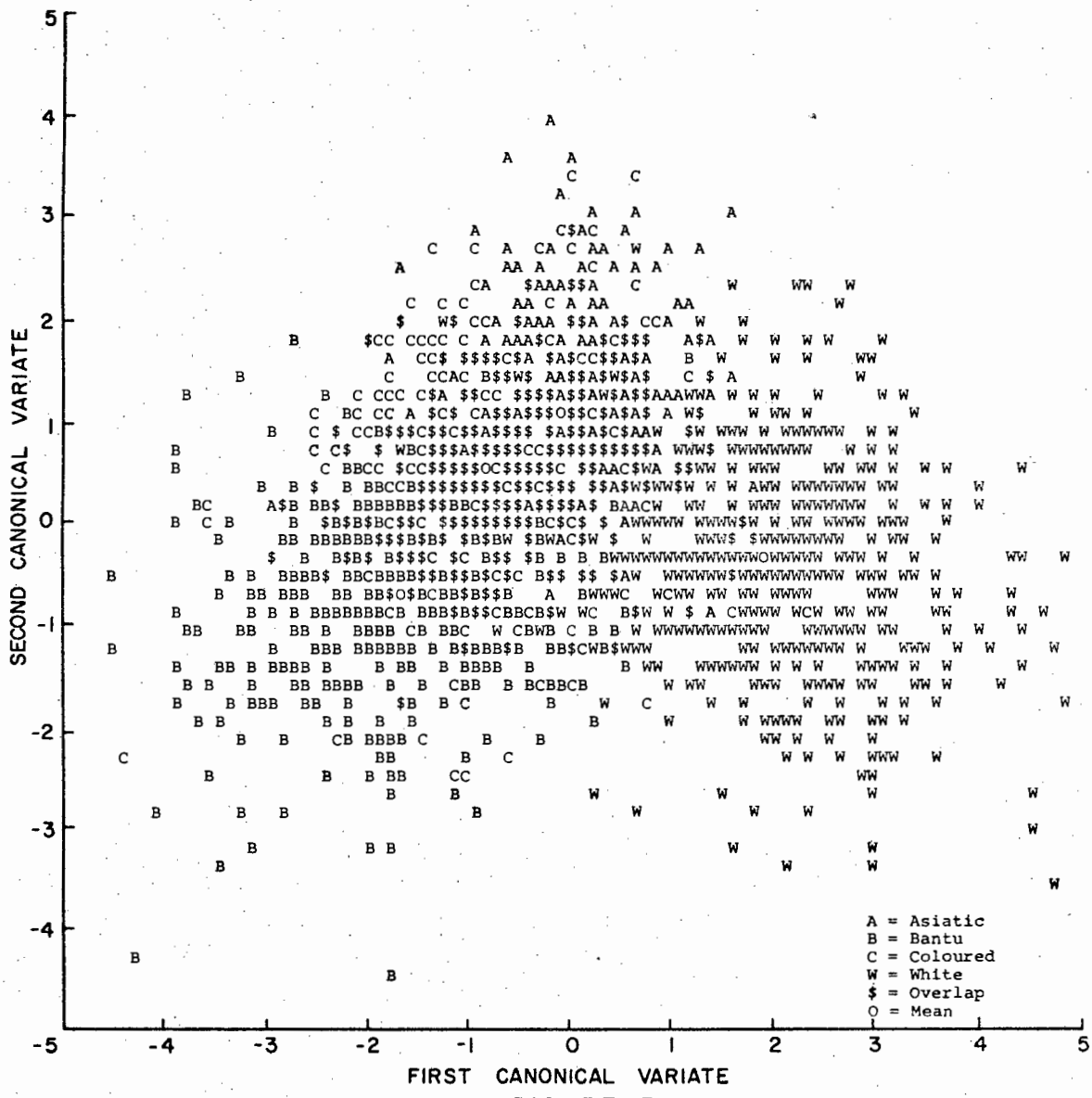


FIGURE 5

Scatter diagram of first and second canonical variates based on the biochemical variables for the four racial groups

72/.....

As in the case of the discriminant analysis on the Whites and Bantu, the subjects in the four racial groups have been reclassified using the discriminant functions obtained in the first 15 biochemical variables. The results are summarised in Table XVI. It can be seen that out of 634 Whites 84 percent were correctly classified. For the Bantu, Coloureds and Asiatics, the figures were: 78 percent, 57 percent and 70 percent respectively. This procedure, as stated previously, is subject to some bias but it does appear that a surprisingly good classification is possible on the basis of the biochemical variables. The relevance of these results will become clearer in later chapters.

Having established that differences exist between the White and Bantu racial groups in respect of their somatometric variables, and between all racial groups in respect of their biochemical variables, we will now investigate what degree of discrimination can be obtained between the four racial groups using all observed variables (Chapter 2.5, Table V), with the exception of age, race and sex.

5.5 Investigation of racial difference in all suitable variables for the four racial groups by means of a discriminant analysis

A stepwise discriminant analysis was carried out on all the relevant variables (no's 4-61, Table V) to see to what extent they could be used to discriminate between the four racial groups. In order to conserve computing time, the stepwise procedure was stopped after the inclusion of the first 25 variables.

The order in which the variables were selected is summarised in Table XVII. It can be seen that the 25th variable still made a significant contribution to the discrimination. The approximate F-test for the overall discrimination between the four racial groups on the basis of the first 25

73/.....

TABLE XVI : CLASSIFICATION ACCORDING TO RACE BASED ON THE BIOCHEMICAL VARIABLES

		Group in which case was classified				Total no. of cases
		White	Bantu	Coloured	Asiatic	
Group in which case belonged	White	531	7	27	69	634
	Bantu	10	409	80	28	527
	Coloured	11	63	219	93	386
	Asiatic	11	11	83	246	351

74/.....

TABLE XVII : ORDER IN WHICH VARIABLES WERE SELECTED FOR THE FOUR RACIAL GROUPS

Step No.	No.	Variable selected Name	Discipline	F-value	Degrees of Freedom	P%
1	47	α-globulin (serum)	Biochemical	373,8	1 1894	<0,001
2	19	R/H/D	Socio-economic	230,8	2 3786	<0,001
3	10	Carbohydrate (ingested)	Dietary	127,8	3 4605	<0,001
4	9	Fat (ingested)		157,6	4 5003	<0,001
5	14	Vitamin A (serum)	Biochemical	116,8	5 5218	<0,001
6	37	Amylase (urinary)		98,0	6 5343	<0,001
7	4	Animal Protein (ingested)	Dietary	78,9	7 5422	<0,001
8	32	Phospholipids (serum)	Biochemical	68,6	8 5474	<0,001
9	54	Sedimentation rate (blood)	Haematological	64,8	9 5509	<0,001
10	15	Thiamine (ingested)	Dietary	40,9	10 5534	<0,001
11	11	Calcium (ingested)		48,9	11 5551	<0,001
12	36	Amylase (serum)	Biochemical	37,3	12 5564	<0,001
13	42	N'-Me (urinary)		34,9	13 5574	<0,001
14	35	Inorganic phosphorus (serum)		41,0	14 5581	<0,001
15	53	Haemoglobin (blood)	Haematological	27,7	15 5586	<0,001
16	23	Intercristal width	Somatometric	28,7	16 5589	<0,001
17	25	Ulnar length		37,6	17 5592	<0,001
18	27	Calf circumference		26,5	18 5594	<0,001
19	30	Para-umbilical skinfold		28,3	19 5594	<0,001
20	51	γ-globulin/Total protein (serum)	Biochemical	22,9	20 5595	<0,001
21	14	Vitamin A (ingested)	Dietary	19,6	21 5595	<0,001
22	13	Iron (ingested)		22,8	22 5594	<0,001
23	24	Biacromial width	Somatometric	14,2	23 5593	<0,001
24	43	Pyridone/N'-Me (urinary)	Biochemical	12,9	24 5592	<0,001
25	44	Total protein (serum)		12,4	25 5591	<0,001

75/.....

variables, yielded an F-value of 99 with associated degrees of freedom 75 and 5591 ($P < 0,0001\%$). There is thus no doubt that a very clear cut differentiation can be made between the racial groups on the basis of the first 25 variables. Once again, as in Table XV, the first variable to be included was α-globulin, which as we have seen, is one of the protein fractions known to be affected by the presence of *inflammatory conditions* in the body. The second variable to be selected was R/H/D. This variable, like α-globulin, is invariant on the age of the child. In view of the marked differences in R/H/D, particularly between Whites and Bantu (See Fig. 3 & 4), the selection of α-globulin, rather than R/H/D, as the most important variable for discriminating between four groups is striking.

The third and fourth variables, namely, *ingested carbohydrates* and *fats*, represent the energy producing nutrients. Carbohydrates played a reasonably important part in the second principal component of the nutrient intake variables, (see Chapter 4.4) but fat did not. Since nutrient intake will tend to increase with age, we must conclude that it is the proportional nutrient intake rather than the absolute amount which differs for the races.

The fifth and sixth variables to be included, viz. *vitamin A* (serum) and urinary amylase, are both biochemical variables. It is interesting to note that these occurred fourth and third respectively in the discriminant analysis on the biochemical variables.

The seventh variable is the intake of animal protein. This would indicate differences between the racial groups in respect of the proportion of animal protein intake (though not necessarily of total protein intake which, as we have seen in the preceding section, appears to have been adequate). The eighth variable was phospholipids, a biochemical variable which appeared in the second place in the discriminant

76/.....

analysis on the biochemical variables. This is followed by sedimentation rate, a haematological variable which tends to increase when *acute infection* is present in the body and has a role somewhat similar to that of α-globulin.

It is interesting to note that in both the discriminant analysis on the biochemical variables only (Chapter 5.4) and that on all the variables, the characteristics of *degree of inflammation, fat intake, carbohydrate intake* (including starch) and *vitamin intake* appear to be the best discriminators, in that order.

It can be seen that in the first nine variables to be included, four of the five disciplines are represented, the somatometric variables being excepted. These first appear in the 16th-19th steps. This is rather surprising, since, as we have seen above, they can be used to discriminate clearly between the Whites and the Bantu. This suggests that the information in the somatometric variables which enables us to discriminate between the Whites and Bantu, is "weaker" than that in the socio-economic, biochemical and dietary variables which enables us to discriminate between the four racial groups. It is interesting to note, however, that the first and second somatometric variables which appear here, are the same as those which appeared first and second in the analysis to discriminate between Whites and the Bantu on the basis of the somatometric variables.

Out of the 25 variables shown in Table XVII, it can be seen that 10 out of a possible 22 are biochemical variables measured in blood or urine. Seven out of a possible 15 are dietary variables. Five out of a possible 11 are somatometric variables. Two out of a possible 9 are haematological variables, and the one socio-economic variable has been included. Thus in the case of the first three disciplines mentioned, each is represented by approximately half of the variables in the discipline.

77/.....

The differences between the racial groups is clearly demonstrated in Fig. 6, in which the first two canonical variates have been plotted. The separation between the Bantu and Whites is clear cut with almost no overlap. This difference is, furthermore, almost entirely in relation to the first canonical variate (horizontal axis) and is remarkably similar to that based on the biochemical variables.

The Asiatics now differ clearly from both the Bantu and Whites but not from the Coloureds. The difference is reflected both in the first and second canonical variates. The Coloureds are well separated from the Whites but overlap both the Asiatics and Bantu to a considerable degree since they lie in an intermediate position between the two racial groups.

As previously, the subjects in the 4 racial groups have been reclassified using the discriminant functions obtained on the first 25 variables. The results are summarised in Table XVIII. It can be seen that out of 634 Whites, 91 percent were correctly classified. Only 1 White was misclassified as a Bantu, whilst 29 were misclassified as Coloured and 29 as Asiatic. Of the 527 Bantu, 88 percent were correctly classified and all but two of those misclassified were placed in the Coloured group. Of the 386 Coloureds, 72 percent were correctly classified. The great proportion of those misclassified were placed in the Bantu and Asiatic groups. Of the 300 Asiatics, 85 percent were correctly classified. The bulk of the misclassification being in the Coloured group.

In summary, it may be noted that the most important differences between the four racial groups appear to relate to variables known or thought to be indicative of *degree of inflammation; socio-economic status; carbohydrate intake; fat intake; vitamin intake* and *animal protein intake*. The relevance of these findings will become clearer in succeeding chapters.

78 /.....

TABLE XVIII : CLASSIFICATION ACCORDING TO RACE BASED ON 58 VARIABLES

		GROUP IN WHICH CASE WAS CLASSIFIED				TOTAL NO. OF CASES
		WHITE	BANTU	COLOURED	ASIATIC	
Group in which case belonged	White	575	1	29	29	634
	Bantu	0	465	60	2	527
	Coloured	7	44	278	57	386
	Asiatic	6	1	44	300	351

80/.....

CHAPTER 6

A STUDY OF THE RELATIONSHIP BETWEEN "CAUSAL" AND "CONSEQUENTIAL" SETS OF VARIABLES, BASED ON THE CANONICAL CORRELATION COEFFICIENT

6.1 Problems relating to the assessment of nutrient intake

Clearly, if the intake of each *nutrient* could be precisely assessed in relation to the needs of an individual, this would yield an optimum method for the assessment of nutrition status. The accurate assessment of nutrient intake is, however, extremely difficult, due firstly to the problems surrounding the measurement of *food* intake, and secondly, to the problems involved in the estimation of the nutrient content of foods. It is also a very expensive undertaking. The difficulties, however, do not end there. It is even more difficult to judge whether any specified level of nutrient intake is adequate. All vegetable proteins, for example, have been classified under a single heading, yet it is well known that the quality, and therefore, the required amount of any such protein, can differ considerably. One cannot simply give a fixed quality rating to any given protein. Proteins consists of amino acids, some of which are regarded as essential and others as non-essential. A protein would be a "good" protein, if it contained the essential amino acids in an ideal or near ideal proportion. Few plant proteins exist for which the proportion of amino acids is anything approximating to the ideal, yet the quality can still vary remarkably from one plant protein to another. The quality of the protein, however, depends on the amino acids being in the right proportion in the stomach rather than in the plant. It is, therefore, at least theoretically possible that an optimum combination of two or more poor quality proteins, might produce a mixture of high quality. Thus, in the final count, *the quality of any specific protein depends on the context in which it has been eaten;* that is on the other protein foods eaten at or around the same time.

A similar problem may well exist in the case of many other nutrients but this has not yet been clearly investigated by biologists.

In spite of these often serious difficulties, it is customary when assessing the adequacy of the diet of any given population group, to ignore differences in the quality of nutrients and to compare the actual nutrient intake of a person of a given age and sex with a *recommended daily allowance* for that age-sex group, see e.g. National Research Council (1968). The recommended daily allowances are, however, by no means a clear indication of the ideal nutrient intake of an individual, they are rather designed to afford a margin, *sufficiently above average physiological requirements, to cover variations amongst practically all individuals in the general population*, of the age group specified. Something of the uncertainty surrounding these allowances can be gauged by the extent to which they have been changed from time to time.

These problems are important when considering what form of nutrient intake should be used when correlating nutrient intake with the observations of the other disciplines. Clearly, if recommended daily allowances could be relied upon to provide an accurate estimate of the amount of a given nutrient *required* by the individual, and not merely a safe upper limit of this amount, it would be best to express the actual nutrient intake of a child as a proportion of the recommended daily allowance for that child, resulting in a "*reduced*" *nutrient intake*, free of variation due to age and sex. This procedure was, in fact, followed by Drion (1961) when studying the intercorrelation between the nutrients consumed by a group of families in the Netherlands. As we have seen above, our results relating to the first principal component of the nutrient intake variables, showed a remarkable similarity with those obtained by Drion, although we did not use "reduced" variables.

6.2 Problems arising out of the effect of age on the relationship between two sets of variables

Before we can present the results of the canonical correlation between causal and consequential sets of variables, some attention must be given to a further problem to which superficial reference has been made above. It is clear that a young child will tend to have small somatometric measurements relating to skeletal and general body size, and to eat little. An older child will have larger somatometric measurements and will tend to eat more. A danger clearly exists, therefore, that any relationship which may emerge between the nutrient intake (or dietary) variables and the somatometric variables, could be due to both these sets being related to age. A spurious correlation might thus be found to exist between two sets of variables which are in fact completely unrelated.

For the biochemical variables the situation is different. These variables are normally expressed as a concentration of a particular constituent per unit volume of blood serum or urine. Thus, by and large, biochemical variables should not be directly correlated with age, except in the case of those variables which relate to some physiological property of the child which changes with age. Such variables are not likely to have a simple linear relationship, but rather to show a marked change in value at, for example, the age of puberty. In the absence of any clear knowledge of such a relationship, the biochemical data were not modified in any way. A similar situation will exist for the haematological variables.

Having reviewed these problems in some detail, we can now proceed to a study of the relationships between causal and consequential sets of variables.

83/.....

6.3 The canonical and partial canonical correlations between nutrient intake and the somatometric variables

The concept of a canonical correlation was stated in Chapter 3.1. The *first* or largest canonical correlation coefficient is the *maximum linear correlation* between two linear functions, $g_1(y)$ and $f_1(x)$. In the present case $g_1(y)$ is the first canonical function of the nutrient intake variables (y), and $f_1(x)$ the first canonical function of the somatometric variables (x). Suppose we have p variables in the one set (or discipline) and q variables in the other with $p \leq q$. In general, p pairs of functions can be found:

$$\{g_1(y),\ f_1(x)\},\ \{g_2(y), f_2(x)\},\ldots\{g_p(y), f_p(x)\} \qquad (1)$$

Each pair is independent of the proceding one, in much the same way as each principal component in a principal component analysis, is independent of the previous component.

These functions give rise to a decreasing series of canonical correlation coefficients, r_i $(i = 1, 2,\ldots p)$. If the first canonical correlation is large in relation to the rest, this means that a large proportion of the relationship between the variables y_i and x_i has been explained by the correlation between $g_1(y)$ and $f_1(x)$ (Anderson, 1958, p.305).

An overall measure of the relationship between the two sets of variables is given by the so-called *trace correlation coefficient*, $\bar{r}$. (Troskie, 1969 and 1971).

This is defined as follows:

$$\bar{r} = \sqrt{\frac{1}{p} \Sigma_{i=1}^{p} r_i^2} \qquad (2)$$

84/.....

Various criteria for testing the canonical correlations have been proposed, based either on the first canonical correlation, or on functions of the canonical correlations. The best one to choose tends to depend on the extent to which the first canonical correlation coefficient explains (or more correctly, is expected to explain) the relationship between the two sets of variables.

Since in the present case the proportion of the relationship explained by the first canonical correlation tended to vary, we have based significance tests of the relationship between the two sets of variables on the trace correlation coefficient (see Pillai, 1960, and Troskie, 1971, p.47). The following test statistic was used:

$$F = \frac{(2n + s + 1)}{(2m + s + 1)} \cdot \frac{\bar{r}^2}{1-\bar{r}^2} \qquad (3)$$

F has an F-distribution (approximately) with $f_1 = s(2m + s + 1)$ and $f_2 = s(2n + s + 1)$ degrees of freedom,

Where: $n = \frac{1}{2}(N - p - q - 2)$

$m = \frac{1}{2}(q - p - 1)$

N = number of subjects studied

s = the number of non zero canonical correlations (almost everywhere = p)

The canonical and trace correlation coefficients were first calculated separately for each racial group between the dietary (nutrient intake) and the somatometric variables. The data for all four racial groups was then pooled, and the calculations repeated. The first five canonical correlations for each race are given in the first five rows of Table XIX, followed by the trace correlation, the relevant F-value, degrees of freedom and associate probability, P. (In the case of the Bantu, one of the canonical correlations was zero, hence the smaller value for the first degree of freedom).

85/.....

TABLE XIX : CANONICAL AND TRACE CORRELATIONS BETWEEN THE DIETARY AND SOMATOMETRIC VARIABLES SHOWING EFFECT OF AGE

		Whites	Bantu	Asiatics	Coloureds	All races
Raw data	Largest canonical correlations	0,60	0,45	0,37	0,35	0,51
		0,28	0,33	0,30	0,32	0,45
		0,24	0,23	0,28	0,28	0,20
		0,21	0,21	0,26	0,22	0,15
		0,17	0,19	0,24	0,20	0,12
	Trace correlation	0,24	0,22	0,22	0,21	0,23
	F	2,41	2,11	1,11	1,10	6,74
	Degrees of freedom	165 6798	140 5100	165 3685	165 4070	165 20702
	P%	<0,001	<0,001	15,7	19,25	<0,001
Age partialled out	Largest canonical correlations	0,43	0,35	0,35	0,34	0,58
		0,30	0,27	0,29	0,32	0,41
		0,22	0,22	0,28	0,26	0,16
		0,21	0,20	0,27	0,20	0,13
		0,17	0,18	0,23	0,19	0,11
	Trace correlation	0,20	0,19	0,21	0,20	0,23
	F	1,73	1,25	1,12	1,02	6,89
	Degrees of freedom	165 6787	165 5089	165 3674	165 4059	165 20691
	P%	<0,001	1,77	12,17	42,6	<0,001

86/.....

It can be seen that the first canonical correlation ranged from 0,35 for Coloureds to 0,60 for Whites, with a value for all race groups of 0,51. The extent to which the first canonical correlation explained the relationship between the two sets of variables, varied from one race group to another. It was best for Whites and for all races combined.

The trace correlation coefficient varied from 0,21 to 0,24. According to the approximate F-test, it differs highly significantly from zero ($P < 0,001$) for the Whites, Bantu and all races. The fact that no significant difference could be demonstrated in the case of the Asiatics and Coloureds appears to be due to the smaller sample sizes for these groups. In general, a significant relationship would seem to exist between the dietary and the somatometric variables. We shall focus our attention on the first or largest canonical correlation but it must be borne in mind that the significance test relates to all the canonical correlations.

Since, as mentioned above under 6.2, the canonical correlations could perhaps be spurious due to both sets of variables being related to age, the *partial canonical correlations* were then computed, with effect of age partialled out. This was done by replacing each element r_{ij} of the correlation matrix by the partial correlation $r_{ij.k}$, in the programme which was used to calculate the canonical correlations.

$$r_{ij.k} = \frac{r_{ij} - r_{ik}\, r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}$$

Where k denotes the age variable.

Similar tests as those described above were carried out on the *partial trace correlation*. It can be shown that in the case of the partial trace correlation the test statistic F, as defined in (3) above, also has an approximate F distribution (Troskie, 1972). As above, p represents the number of

87/.....

variables in the one set and q, the number in the other with $p \leq q$. If the effect of r "new" variables (not included in p or q) has been partialled out, then N is replaced by N-r. Since we have partialled out the effect of one variable, age, which was not included in the (p + q) variables on which the previous analysis was carried out, the values for p and q in each case will remain unchanged. N is reduced to N-1.

The results are shown in the lower half of Table XIX. For each race group, the partial canonical correlation was somewhat smaller than the canonical correlation. The difference was larger for some racial groups than for others. When, however, the partial canonical correlation was computed for the data on all four racial groups, we found a larger canonical correlation coefficient than that observed for any single race group. This somewhat surprising result, which was reflected to a lesser degree in the trace correlation, may be due to the fact that by pooling the various groups, we have considerably increased the range over which the nutrient intake variables, in particular, have been observed. The trace correlation was still highly significant for the Whites and for all races (P <0,001%). It was significant (P < 1,77%) for the Bantu, but no significant difference could be shown in the case of the Asiatics and Coloureds. As mentioned, the lack of significance in the case of the latter two racial groups may be due to the smaller sample sizes.

In general, therefore, there appears to be a definite relationship between the nutrient intakes or dietary variables and the somatometric variables *which does not depend on age*.

As stated above, the adequacy of dietary intake is normally assessed in terms of a number of different nutrients (15 in the present case). It would clearly be of great advantage if a *single index of nutrient intake* could be determined.

88/.....

One possible such index would be the first canonical function with the effect of *age* partialled out, based on the pooled data for all racial groups. This is given by $g_1(y)$ for the nutrient intake variables below:

$$\begin{aligned} g_1(y) = \; & 0{,}0608y_1 - 0{,}1081y_2 - 0{,}0156y_3 \\ & - 0{,}6788y_4 - 0{,}0315y_5 + 0{,}8628y_6 \\ & + 0{,}3244y_7 + 0{,}3515y_8 - 0{,}0582y_9 \\ & - 0{,}1542y_{10} + 0{,}2126y_{11} - 0{,}1955y_{12} \\ & + 0{,}0458y_{13} + 0{,}0483y_{14} + 0{,}2434y_{15} \end{aligned} \qquad (5)$$

where y_i is the i-th nutrient intake variable (see Table V of Chapter 2, the variable number is given in brackets).

The function for the somatometric variables is given by $f_1(x)$ below:

$$\begin{aligned} f_1(x) = \; & - 0{,}1020x_1 + 0{,}4980x_2 - 0{,}1036x_3 \\ & + 0{,}6555x_4 + 0{,}3220x_5 - 0{,}3335x_6 \\ & - 0{,}1456x_7 + 0{,}0075x_8 + 0{,}4974x_9 \\ & - 0{,}2735x_{10} - 0{,}0640x_{11} \end{aligned} \qquad (6)$$

where x_i is the i-th somatometric variable (see Table V of Chapter 2, variables No. in brackets).

The function $g_1(y)$ is an optimum function of the nutrient intake variables, whilst $f_1(x)$ is an optimum function of the somatometric variables in terms of the relationship between the two data sets. It is, therefore, a reasonable postulate that each of these functions can, in a sense, serve as an index for the discipline it represents.

89/.....

The main purpose of our study is to find a simple technique for assessing nutrition status, and we have seen some of the problems related to the assessment of nutrient intake. Since the somatometric variables, as we have shown, are related to the nutrient intake variables, it is of interest to determine which ones are important in the relationship with nutrient intake.

Since, in the functions (5) and (6) above, all variables have been standardised in respect of their respective means and standard deviations, the size of the coefficient is directly proportional to the importance of the variable. The rank order of the variables according to the modulus of the coefficients, together with the sign of the coefficient is shown in columns 3 and 4 of Table XX. It will be remembered that in Chapter 5.1, we found that it was possible to discriminate clearly between the White and Bantu racial groups on the basis of the somatometric variables. It is interesting to compare those variables which best discriminate between the two racial groups with those which are most highly related to the canonical function of nutrient intake. To facilitate comparison, the rank order of the variables in the discriminant analysis, together with the sign of the coefficient, are shown in columns 5 and 6 of Table XX. It can be seen that intercristal width occurs first in both cases and the coefficients have the same sign. This variable, which occurred in our first principal component representing general body size (see Chapter 4.1), was also the most important somatometric variable in the stepwise discrimination between all four racial groups.

The 6th and 8th - 11th variables also occur in the same order and the coefficients have the same sign.

Before commenting on these results we would like to note

TABLE XX : COMPARISON OF THE ORDER OF THE SOMATOMETRIC VARIABLES, RANKED ACCORDING TO IMPORTANCE BY EACH OF THE THREE PROCEDURES

Variable		Modulus of coefficient		Discriminant analysis		Stepwise Regression		
No.	Name	Rank	Sign of coefft.	Rank	Sign of coefft.	Rank	Sign of coefft.	Multiple R
	Age					0	-	0,06
4	Intercristal width	1	+	1	+	1	+	0,53
5	Biacromial width	5	+	3	+	2	+	0,54
6	Ulnar length	4	-	2	-	3	-	0,55
2	Height	2	+	7	+	4	+	0,56
9	Triceps skinfold	3	+	5	+	5	+	0,56
10	Subscapular skinfold	6	-	6	-	6	-	0,58
7	Upper arm circ.	7	-	4	-	7	-	0,58
3	Cristal height	8	-	8	-	8	-	0,58
1	Weight	9	-	9	-	9	-	0,58
11	Para-umbilical skinfold	10	-	10	-	10	-	0,58
8	Calf circumference	11	+	11	+	11	+	0,58

91/.....

an important difference concerning the factors determining the *order* in which the variables, based on the modulus of the canonical coefficient, have been tabulated and the order arising from a stepwise discriminant or regression analysis. As stated previously in the stepwise analysis, each new variable is added because it can contribute more "new" information than any of the remaining variables. In the case of the canonical function, however, the modulus of the coefficient of the variables and hence the order given in the relevant column of Table XX, is applicable only to their absolute importance in the relationship. Thus in this case, two variables could both be given high priority in spite of the fact that they contained virtually the same information. If, however, they contained exactly the same information, i.e. if one was a linear combination of the other, the covariance matrix would be singular and the analysis would not be possible.

In order to obtain a sequence of the somatometric variables arranged according to the "new" information contributed by each variable in the canonical function, a further computation is necessary. This is described below.

6.4 Stepwise regression of the "partial" canonical function of nutrient intake on the somatometric variables

The function $g_1(y)$ (No.(5) in 6.3 above), which is based on the partial canonical correlation, was evaluated for the nutrient intake of each subject in all four racial groups. This function was then taken as the dependent variable, and a stepwise regression analysis carried out taking the somatometric variables as independent variables. Since the effect of age had been partialled out of the canonical functions (5) and (6) in 6.3 above, it was necessary to remove this effect from the somatometric variables. This was done by including age as a variable in the stepwise regression, *but forcing it in as the first selected variable.* Thus, the effect of age

92/.....

was then partialled out of the covariance matrix before the somatometric variables were considered for inclusion.

The variables are tabulated in Table XX in the order in which they were included. The rank of the variable, the sign of the regression coefficients, and the multiple correlation coefficients, are tabulated in the last three columns of Table XX. It is interesting to note the low correlation (0,06) of age with the canonical function of nutrient intake, which clearly demonstrates that the effect of age has been partialled out of the relationship between the nutrient intake, and somatometric variables. It can immediately be seen that a revised sequence of somatometric variables emerges which shows a remarkable similarity to the order obtained for the somatometric variables which best discriminate between Whites and Bantu (columns 4 and 5 of Table XX). If biacromial width is exchanged with ulnar length, and height with upper arm circumference, the order would be identical. Furthermore, in every case the sign of the coefficient for each variable is the same.

This result deserves careful consideration: It appears to suggest that the factors which permit discrimination between White and Bantu are almost the same as those which relate to nutrient intake.

In Chapter 5.1 we argued that the most important differences between Whites and Bantu were differences in skeletal proportion, possibly due to genetic causes. Since the coefficients of both intercristal width and biacromial width are positive, while that of ulnar length is negative, it would now appear that the differences in skeletal proportion are also primarily responsible for the relationship with the canonical function of nutrient intake variables. This result may well suggest that *the differences in skeletal proportion have a nutritional, rather than genetic cause.* Since the

effect of age has been partialled out, it does not appear that the above findings could be attributed to an age effect. It seems highly improbable that they could be attributable to some race effect impinging on the nutrient intake variables but this possibility cannot be completely excluded.

It is accepted by anthropologists that different racial groups may differ in respect of certain body proportions (see e.g. Metheny, 1939). Certain research workers in the field of nutrition, have postulated that these proportional differences might be the long-term effect of a particular level of nutrition. It has e.g. been noted that although the Japanese are traditionally a race of little stature, those who emigrate to America have children whose height and weight approximate to American standards. Our results would seem to contribute somewhat to the pool of evidence suggesting that nutrition is responsible, at least in part, for size and shape differences between the various races.

We will now turn to a second subset of the data which also permits a study of the relationship between causal and consequential variables.

6.5 The canonical correlations between nutrient intake or "reduced" nutrient intake, and the biochemical variables

In the same way as that described above, the canonical correlation coefficients were calculated between the nutrient intake variables and the biochemical variables. The calculations were carried out separately for each racial group and then on the combined data set for all races. The canonical correlations were calculated firstly on the raw data. The first five canonical correlations followed by the trace correlation coefficient and relevant test statistics, are shown in the upper half of Table XXI. It can be seen that the first canonical correlation ranged from 0,38 for Bantu to 0,55 for Whites, with a value for all races of 0,72. The increase

TABLE XXI : CANONICAL AND TRACE CORRELATIONS BETWEEN THE DIETARY AND BIOCHEMICAL VARIABLES SHOWING EFFECT OF USING " REDUCED " NUTRIENT INTAKE

		Whites	Bantu	Asiatics	Coloureds	All races
Raw data	Largest canonical correlations	0,55	0,38	0,63	0,52	0,72
		0,50	0,34	0,49	0,40	0,38
		0,41	0,32	0,40	0,37	0,28
		0,32	0,28	0,36	0,32	0,24
		0,25	0,26	0,31	0,31	0,24
	Trace correlation	0,28	0,23	0,30	0,27	0,25
	F	2,27	1,22	1,62	1,34	5,87
	Deg. of freedom	330 9165	330 7560	294 4578	330 5445	330 28125
	P%	<0,001	0,423	<0,001	0,007	<0,001
Standardised nutrient intake	Largest canonical correlations	0,66	0,51	0,60	0,51	0,71
		0,47	0,35	0,49	0,45	0,49
		0,43	0,31	0,44	0,39	0,30
		0,31	0,27	0,37	0,34	0,26
		0,26	0,24	0,34	0,29	0,21
	Trace correlation	0,32	0,25	0,33	0,30	0,29
	F	3,06	1,59	1,78	1,62	8,04
	Deg. of freedom	264 7332	264 6048	264 3936	264 4356	264 22524
	P%	<0,001	<0,001	<0,001	<0,001	<0,001

95/.....

in the canonical correlation for all races is probably due to the greater scatter in the nutrient intake variables when pooled for all races. Once again there is some variation between the racial groups in the extent to which the first canonical correlation explains the relationship between the two sets of variables. The trace correlation varied from 0,23 to 0,30, with a value of 0,25 for all races. It can be seen from Table XXI that in every case the trace correlation differed highly significantly from zero, *providing substantial proof that a relationship exists between the dietary and the biochemical variables.*

In the second place, the analyses were carried out after an attempt to eliminate the effect of *age*. We did not, however, partial out the age effect as in the case of the previous analysis. As we have mentioned earlier, the biochemical variables will largely be independent of age. Those variables which are dependent on age due to physiological reasons, are not likely to have a simple linear relationship, but rather to show a marked change in value at, for example, the age of puberty. In the absence of any clear knowledge of such a relationship, the biochemical data were not modified in any way. For the nutrient intake data, the observed value for each subject was expressed as a proportion of that subject's recommended daily allowance, giving the so-called "reduced" nutrient intake variables. The recommended daily allowances of the National Research Council (1968) were used. Since these were not available for animal, vegetable and mixed protein intakes, these variables were excluded from the analysis.

The results obtained are shown in the lower part of Table XXI. The effect of using "reduced" variables is rather surprising. For Whites and Bantu this resulted in an increase in the first canonical correlation coefficient, markedly so in the case of the Bantu. For Asiatics and Coloureds, there was a slight decrease. For all races, the first canonical correlation changed from 0,72 to 0,71 as a result of using

reduced variables. The trace correlation increased slightly and showed a highly significant difference from zero in every case. In general, using the standardised nutrient intake seems to yield a slightly better relationship between the two sets of variables. The improvement does not, however, appear to be concentrated in the first canonical correlation, *i.e.* is not in terms *of a single linear relationship* between the two sets of variables.

In view of the doubts expressed in 6.1 concerning the applicability of recommended daily allowances, and the fact that it has little overall effect on the *first* canonical correlation, attention will be mainly devoted to the results obtained on the "unreduced" or raw nutrient intake variables in the paragraphs that follow.

We may note that the first canonical correlations describing the relationship between nutrient intake and the biochemical variables are of the same order for each racial group, as those describing the relationship with the somatometric variables. They are, however, considerably higher for all races combined. This is probably due to the fact that the variation in both sets of variables is now increased by pooling the data for the different race groups.

The first canonical correlation coefficient is not, strictly speaking, comparable to a simple correlation coefficient. It is, nevertheless, interesting to note that the first canonical correlations are considerably higher than any simple correlation between nutrient intake and a biochemical variable. The highest simple correlation (0,38) was found between serum carotene and the intake of both animal protein and riboflavin. This result would tend to emphasise the fact that there is no unique relationship between the intake of a specific nutrient and a biochemical variable. There is rather, a complex interrelationship between the intake of a number of nutrients and the concentration of a number of biochemical

variables. The relationship which we have been able to establish, is well described by the two first canonical functions calculated on all racial groups. These were found to be:

$$\begin{aligned}\beta_1(y) = {} & - 0{,}0481y_1 - 0{,}3550y_2 - 0{,}0966y_3 \\ & - 0{,}3957y_4 + 0{,}2411y_5 + 0{,}5532y_6 \\ & + 0{,}1758y_7 + 0{,}4214y_8 - 0{,}0388y_9 \\ & - 0{,}1302y_{10} + 0{,}1694y_{11} - 0{,}1789y_{12} \\ & + 0{,}0967y_{13} - 0{,}0372y_{14} + 0{,}2025y_{15}\end{aligned} \qquad (7)$$

where y_i is the i-th nutrient intake variable (see Table V of Chapter 2, the variable No. is given in brackets).

$$\begin{aligned}\gamma_1(z) = {} & - 0{,}0280z_1 + 0{,}1147z_2 + 0{,}0363z_3 \\ & - 0{,}0912z_4 - 0{,}0721z_5 - 0{,}1809z_6 \\ & - 0{,}1263z_7 + 0{,}0024z_8 + 0{,}3226z_9 \\ & - 0{,}1292z_{10} + 0{,}0307z_{11} - 0{,}0186z_{12} \\ & + 0{,}0427z_{13} + 0{,}0618z_{14} - 0{,}0499z_{15} \\ & + 0{,}0318z_{16} - 0{,}1291z_{17} + 0{,}1335z_{18} \\ & - 0{,}2335z_{19} + 0{,}2877z_{20} + 0{,}1510z_{21} \\ & + 0{,}2607z_{22}\end{aligned} \qquad (8)$$

where z_i is the i-th biochemical variable (see Table V of Chapter 2, the variable No. is given in brackets).

Since our variables have all been standardised relative to their respective means and standard deviations, once again the modulus of the regression coefficient for each biochemical variable is a measure of its importance in establishing the relationship with the nutrient intake variables. The rank order of the modulus of the coefficients together with the sign have been tabulated in the 3rd and 4th columns of

98/.....

Table XXII. As described above in the case of the somatometric variables, it is not possible, from an inspection of these coefficients, to establish an optimum subset of biochemical variables. In order to do this, we have as previously, carried out a stepwise regression analysis.

6.6 Stepwise regression of the canonical function of nutrient intake on the biochemical variables

The function $\beta_1(y)$ (function no. (7) in 6.5 above), was evaluated for the nutrient intake of each subject in all four racial groups, (in the same way as $g_1(y)$ in Chapter 6.4). This function was then taken as the dependent variable, and a stepwise regression analysis was carried out taking the biochemical variables as independent variables. The variables are tabulated in Table XXII in the order in which they were included. The ranks of the variables, the signs of the regression coefficients and the multiple correlation coefficients are tabulated in the last three columns of the table.

In Chapter 6.4 we found that, when a stepwise regression of the canonical function of the nutrient intake variables, $g_1(y)$ on the somatometric variables was carried out, the ordering of the variables showed a remarkable similarity, to that obtained for the somatometric variables which best discriminated between the Whites and the Bantu.

In order to see whether a similar result would obtain for the biochemical variables, the order in which the biochemical variables were included in the discriminant analysis between the four racial groups (Chapter 5.5, Table XVII) has been repeated in the 5th column of Table XXII. Since only 15 variables were included in the discriminant analysis, the rank numbers only go up to 15. The discriminant analysis was done on the four racial groups, and a discriminant function of the same form as X_{W-B} (equation (2), Chapter 5.1), could be obtained for each *pair* of groups. Since the sign of the coefficient of a particular variable could vary from one function to another, it is not possible to tabulate it.

99/.....

TABLE XXII : COMPARISON OF THE ORDER OF THE BIOCHEMICAL VARIABLES, RANKED ACCORDING TO IMPORTANCE BY EACH OF THE THREE PROCEDURES

Variable		Modulus of coefficient.		Discriminant analysis	Stepwise regression		
No.	Name	Rank	Sign of coefft.	Rank	Rank	Sign of coefft.	Multiple R
9	Carotene (serum)	1	+	15	1	+	0,53
17	α-globulin	8	-	1	2	-	0,61
6	Amylase (serum)	5	-	5	3	-	0,65
22	Riboflavin (urinary)	3	+	9	4	+	0,68
10	Thiamine	10	-	10	5	-	0,69
21	γ-globulin/Total protein %	6	+	6	6	+	0,70
7	Urinary amylase/Creatinine	9	-	3	7	-	0,71
4	Alkaline phosphatase	12	-	11	8	-	0,71
2	Phospholipids (serum)	11	+	2	9	+	0,71
5	Inorganic phosphorus	13	-	7	10	-	0,72
13	Urinary Pyridone/N^1-Me	16	+	14	11	+	0,72
14	Total protein	14	+	12	12	+	0,72
19	γ-globulin	4	-		13	-	0,72
3	Cholesterol/Phospholipids	17	+		14	+	0,72
20	Albumin/Total protein	2	+		15	+	0,72
18	β-globulin	7	+		16	+	0,72
15	Albumin	15	-		17	-	0,72
11	2-Pyridone	19	+		18	+	0,72
12	N^1-Me	21	-	8	19	-	0,72
16	Globulin	18	+		20	+	0,72
1	Cholesterol	20	-	13	21	-	0,72
8	Vitamin A (serum)	22	+	4	-		

100/.....

At first little, if any, relationship appears to exist between the sequence obtained when discriminating between the four racial groups and the sequence obtained in the stepwise regression analysis which related the biochemical variables to nutrient intake. It can, however, be seen that in the first six variables, three of those which occurred in the discriminant analysis (variables Nos. 6, 17 and 21 in Table XXII), also occurred in the stepwise regression analysis. The maximum multiple correlation in the stepwise regression, is reached at the inclusion of the 10th variable. Eight of these variables were included in the first 10 in the discriminant analysis between the racial groups. Thus, in spite of the lack of correlation between the rank numbers, *virtually the same subset of biochemical variables appears to be important in the two procedures.*

The first variable to be included was serum carotene. It is interesting to note that the simple correlation between serum carotene and the function of nutrient intake $\beta_1(y)$, is the same value (0,53) as that observed (Table XX) between intercristal width and the canonical function of nutrient intake $g_1(y)$. The fact that serum carotene is selected first, makes good sense. Carotene is one of the few *nutrients* which occur in the *blood serum* and cannot be synthetized by the body. It seems, therefore, reasonable that it should be closely related to nutrient intake.

α-Globulin, which was the first variable to be included in the discriminant analysis, is the second variable to be included in the stepwise regression analysis. (It is, however, only ranked 8th in absolute importance, when judged by the modulus of the canonical coefficient, (see column 3 of Table XXII). The placing of this variable is at first somewhat puzzling. The result would appear to indicate that α-globulin is related to nutrient intake, whereas we have regarded it (Chapter 5.4) as a measure of infection caused by an inflammatory condition. The apparent relationship between α-globulin and nutrient intake has an interesting explanation.

101/.....

It has been shown by various research workers that malnutrition and infection go hand in hand. For example, Wittmann *et al.* (1967), found in a study of Coloured families at Bonteheuvel, Cape Province, a negative correlation between nutrition status and infectious diseases. Various research workers have shown that gastroenteritis and other infections occur frequently and have a high morbidity and mortality, particularly in malnourished children. It would, therefore, appear that the relationship between the canonical function of nutrition status and α-globulin can be explained as follows: A low level of nutrient intake for certain children has resulted in a lowering of resistance and hence an increased level of infections and inflammatory conditions. This in turn has resulted in raised α-globulin levels.

If our interpretation of the above relationship is correct, it forms an interesting example of how a correlation can be found between two variables, not because they are directly related to one another, but because both are related to a third variable.

CHAPTER 7

RELATIONSHIP BETWEEN SETS OF "CONSEQUENTIAL" VARIABLES

7.1 Relationship between the somatometric and the biochemical variables

We have already dealt extensively with the somatometric and the biochemical variables and have shown how both these disciplines are related to nutrient intake. If this be the case, it is only reasonable to suppose that these two disciplines would be related to each other. In order to investigate this aspect, the canonical correlation coefficients were calculated between the somatometric variables and the biochemical variables. Once again this was done firstly on the raw data dealing with each racial group separately, and then with the data on all races combined. The results are shown in the upper half of Table XXIII. In general, the first canonical correlation is somewhat higher for each race group than that found between the dietary and the somatometric variables or the dietary and the biochemical variables. It was highest for Whites (0,74) and lowest for Bantu (0,58). As previously, the remaining canonical correlations explain quite a high proportion of the relationship, particularly in the case of the Asiatics. The suggestion that the somatometric and biochemical sets of variables are more closely related than either the nutrient intake and the somatometric, or the nutrient intake and the biochemical variables, is also reflected in the trace correlation which ranged from 0,28 to 0,37. In every case the trace correlation differed highly significantly from zero (Table XXIII).

The partial canonical correlations were then calculated as in Chapter 6.3 with *the effect of age partialled out.* These results are shown in the lower half of Table XXIII. It can be seen that, in general, the partial correlation coefficient with the effect of age removed, is somewhat smaller for each racial group than the ordinary canonical correlation. When, however, particularly in the case of the Whites, the data for all racial groups is combined, the two coefficients

TABLE XXIII : CANONICAL AND TRACE CORRELATIONS BETWEEN THE SOMATOMETRIC AND BIOCHEMICAL VARIABLES SHOWING THE EFFECT OF AGE

		Whites	Bantu	Asiatics	Coloureds	All races
Raw data	Largest canonical correlations	0,74	0,58	0,68	0,62	0,63
		0,44	0,34	0,59	0,49	0,50
		0,29	0,30	0,42	0,40	0,33
		0,25	0,28	0,37	0,38	0,21
		0,20	0,25	0,32	0,28	0,19
	Trace correlation	0,31	0,28	0,37	0,33	0,29
	F	2,94	2,00	2,37	2,05	7,67
	Degrees of freedom	242 6721	242 5544	242 3608	242 3993	242 20625
	P%	<0,001	<0,001	<0,001	<0,001	<0,001
Age partialled out	Largest canonical correlations	0,48	0,35	0,60	0,51	0,63
		0,35	0,34	0,48	0,43	0,41
		0,27	0,29	0,42	0,39	0,24
		0,26	0,28	0,35	0,36	0,21
		0,24	0,25	0,33	0,28	0,17
	Trace correlation	0,25	0,24	0,34	0,30	0,26
	F	1,84	1,44	1,96	1,64	6,36
	Degrees of freedom	242 6710	242 5533	242 3597	242 3982	242 20614
	P%	<0,001	<0,001	<0,001	<0,001	<0,001

104/.....

remain the same. As pointed out in Chapter 6, this can probably be explained by the fact that the pooled data in respect of the biochemical variables have a considerably greater scatter than that for any given racial group. The trace correlation ranged from 0,24 to 0,34 and is, in every case, slightly smaller after the effect of age has been partialled out. It is, however, in each case somewhat larger than the trace correlation between the nutrient intake and the somatometric variables after the effect of age was partialled out. In every case it still differs highly significantly from zero.

Since we have already indicated that both the somatometric and biochemical variables form reasonably good estimators of nutrient intake, this relationship between these two sets is only to be expected. It is, however, encouraging since it would imply that in situations where it is desired to assess nutrition status by the cheapest and simplest procedure, the measurement of the somatometric variables may well provide adequate information.

Chapter 8/.....

CHAPTER 8

A FACTOR ANALYTIC APPROACH

8.1 The relevance of a factor analysis model

In Chapter 3, a review was given of applicable multivariate statistical techniques, and in the preceeding chapters we have delved heavily into the available statistical methodology. We have used techniques for studying the interrelationships between two sets of variables in Chapters 6 & 7. We have used techniques for discriminating between two or more groups in Chapter 5. One of the techniques for studying interrelationships amongst a set of variables, viz. a principal component analysis was used in Chapter 4. The remaining technique for dealing with this type of problem is, as was mentioned, a factor analysis. Thus far, we have hesitated in applying this procedure. We pointed out previously (Chapter 3.3), that it was based on rather stringent assumptions. One of the most important was that it should be possible to *postulate in advance that certain factors exist, which cannot be observed explicitly, but which are measured by the variables.* At the commencement of our study, we had no indication that such factors existed. The work of the preceeding chapters would, however, seem to justify the assumption that physiological factors do in fact exist.

In Chapter 4, we were able to demonstrate that in the somatometric variables there were three principal components which could be identified with the physiological aspects of *general body size, body fat, and soft body tissue.*

We could detect in the biochemical variables a total of four components relevant to *protein status and the degree of exposure to antibody producing infections; circulating serum body fats and fat soluble vitamins; nicotinic acid metabolites and water soluble vitamins.* In the haematological variables two components relevant to the *capacity to absorb oxygen* and the *capacity to combat infections* could be identified. In

the dietary variables three components emerged which were indicative of the kind of diet which was eaten. It seems, therefore, reasonable to accept that physiological factors exist such as could be described by a factor analysis model.

8.2 Problems arising out of the presence of various age groups

The type of problem which could arise due to the variables having been measured on children of various age groups, has already been detailed on various occasions. Whilst many of the variables studied are clearly related to age, it would appear, on the basis of the analyses presented above, either that age does not adversely affect the statistical procedure (as in the case of the discriminant analyses reported in Chapter 5), or that the effect of age can be partialled out, or removed in some other way, (as in the case of the canonical correlations in Chapter 6). The effect of age on a factor analysis are, however, not yet clear. For this reason we have followed three analytical procedures. In the first, the effect of age was ignored, in the second, age was included as a variable, and in the third, the effect of age was partialled out of the correlation matrix.

The results of the first two procedures were so similar that they will be discussed together below:

8.3 Factor analysis on all variables for all racial groups showing the effect of including age as a variable

The computer program used, both in this analysis and in those which follow, performed a principal factor solution followed by an orthogonal rotation of the factor matrix (Dixon, 1968, p.169; Harman, 1967, Chapter 8). The factor analysis was applied to all the relevant variables (Nos. 2 and 4-61 of Table V, Chapter 2.5) for all four racial groups. In the first analysis, the age variable (No. 2) was omitted.

107/.....

From the principal component analysis presented in Chapter 4, a total of 12 factors could be expected in the variables of the four disciplines. In order to get an independent confirmation of the numbers of factors which could be expected, the eigen values were calculated. It was found that there were 11 eigen values greater than 1. A factor analysis was, therefore, carried out and 11 factors were extracted. The squared multiple correlations were used to provide initial communality estimates, and iterative rotations of the factor matrix were carried out until a criterion for convergence was met. The criterion required that in four successive rotations the change in the sum of the factor loadings be less than 10^{-7} (Dixon 1968, p.177). Six rotations were necessary.

The results for the first 10 factors are shown in Table XXIV. The eleventh factor had only one high loading (0,70) on mixed protein. It would thus appear as if too many factors were extracted. The last factor was, therefore, omitted in the table and only 10 factors were extracted in the subsequent analyses. The variables and relevant factor loadings have been listed in the body of the table in the order of the modulus of the factor loading. In each case, one or two variables with low loadings and unimportant to the particular factor, have been tabulated to show the relative size of the factor loadings; these are separated from the larger loadings by a dotted line. This line has been drawn where, if evident, a sudden fall could be discerned in the modulus of the factor loadings, or where the factor loadings fell below 0,5. At the head of each factor the applicable physiological interpretation as we see it, has been given.

In the second place the analysis was repeated on the same data set but including *age* as a variable. This time, 10 factors were extracted and 7 rotations of the factor matrix were necessary to meet the convergence test. The results are tabulated in Table XXV. It can immediately be seen that there is a remarkable similarity in the two sets of results.

108 /.....

TABLE XXIV : FACTOR ANALYSIS OF ALL RELEVANT VARIABLES FOR ALL RACES (EFFECT OF AGE IGNORED)

VARIABLES WITH HIGH FACTOR LOADINGS (Loadings in brackets).									
Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
General body size	Circulating serum body fats.	Food intake (protein)	Serum protein fractions (anti-body-producing infections).	Combatting infection.	Oxygen uptake	Nicotinic acid metabolites	-	Tissue body fat	Protein status
Height (0,95)	Cholesterol (0,92)	Protein (0,92)	γ-globulin (0,92)	Neutrophils (0,99)	Haemoglobin (0,88)	2-Pyridone (0,88)	β-globulin (-0,85)	Triceps (0,80)	Albumin (0,92)
Cristal height (0,93)	Cholest/Phospholipids (0,75)	Calories (0,91)	γ-globulin/Total protein (0,81)	Lymphocytes (-0,93)	Haematocrit (0,74)	Pyridone/N'-Me (0,60)	Animal protein (-0,33)	Para-umbilical skinfold (0,79)	Albumin/Total protein (0,72)
Ulnar length (0,92)	Phospholipids (0,65)	Phosphorus (0,88)	Total protein (serum) (0,81)	Eosinophils (-0,34)	M.C.H.C. (0,45)	N'-Me (0,46)	Globulin (-0,29)	Subscapular skinfold (0,78)	Phospholipids (0,46)
Weight (0,91)	Carotene (0,43)	Thiamine (0,86)	Globulin (0,81)	White cells (0,16)	α-globulin (0,27)	Urinary riboflavin (0,37)	Total protein (serum) (-0,27)	Upper arm circ. bent (0,42)	Vitamin A(serum) (0,45)
Intercristal width (0,89)	Thiamine (-0,14)	Nicotinic acid (0,80)	Albumin/Total protein (-0,60)		Sedimen. rate (-0,20)			Weight (0,30)	Globulin (-0,45)
Biacromial width (0,89)	γ-globulin/Total protein (-0,14)	Carbohydrate (0,73)	Sedimen. rate (0,37)						
Calf circum. (0,85)	Veg. protein (-0,14)	Riboflavin (0,69)							
Upper arm circ. bent (0,81)		Fat (0,69)							
Para-umbilical skinfold (0,46)		Veg. protein (0,67)							
Subscapular skinfold (0,45)		Calcium (0,66)							
Urinary amylase (0,43)		Animal protein (0,60)							
		Iron (0,55)							
		Vitamin A (0,37)							
		Vitamin C (0,35)							

109/.....

TABLE XXV : FACTOR ANALYSIS OF ALL RELEVANT VARIABLES FOR ALL RACES (AGE INCLUDED AS A VARIABLE)

VARIABLES WITH HIGH LOADINGS (Loadings in brackets).									
Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
General body size	Circulating serum body fats	Food intake (vegetable protein).	Serum protein fractions (anti-body producing infections).	Combatting infection.	Tissue body fat	Nicotinic acid metabolites.	Oxygen uptake	Food intake (animal protein)	Protein status
Height (-0,96)	Cholesterol (0,96)	Vegetable prot. (-0,88)	Globulin (0,91)	Neutrophils (-1,00)	Triceps skinfold (-0,81)	2-Pyridone (-0,90)	Haemoglobin (0,91)	Riboflavin (0,78)	Albumin (0.85)
Cristal height (-0,94)	Phospholipids (0,62)	Thiamine (-0,82)	Y-globulin (0,83)	Lymphocytes (0,91)	Para-umbilical skinfold (-0,81)	Pyridone/N'-Me. (-0,46)	Haematocrit (0,69)	Calcium (0,77)	Albumin/Total protein (0,61)
Ulnar length (-0,93)	Cholest/Phospho-lipids (0,58)	Carbohydrate (-0,80)	Total protein (0,82)	White cells (-0,16)	Subscapular skin fold (-0,80)	N'-Me (-0,42)	M.C.H.C. (0,38)	Phosphorus (0,72)	α-globulin (-0,43)
Weight (-0,90)	Carotene (0,43)	Calories (-0,77)	Y-globulin/Tot. protein (0,70)	Y-globulin/Tot. protein (0,08)	Upper arm circ. bent (-0,43)	Urinary riboflavin (-0,33)		Fat (0,71)	Total protein (serum) (0,41)
Biacromial width (-0,89)	Rate/Head/Day (0,20)	Protein (-0,69)	Albumin/Total protein (-0,70)		Weight (-0,32)			Animal protein (0,66)	
Intercristal width (-0,88)	Y-globulin/Total protein (-0,16)	Nicotinic acid (-0,60)	β-globulin (0,37)		Calf circumference. (-0,31)			Protein (0,62)	
Age (-0,85)	Albumin (0.16)	Iron (-0,59)	Sedimentation rate (0,37)					Rate/Head/Day (0,53)	
Calf circum. (-0,84)		Phosphorus (-0,57)						Vitamin A (0,52)	
Upper arm circ. bent (-0,79)		Fat (-0,34)						Vitamin C (0,51)	
Para-umbilical skinfold (-0,43)		Riboflavin (-0,32)							
Subscapular skin-fold (-0,43)									
Triceps skinfold (-0,34)									

110/.....

In both cases the first factor is one of *general body size*. The same variables appear with similar factor loadings, and in almost the same order. The loadings which may be interpreted as *the correlation between the factor and the variable* are, however, positive in the analysis where age was ignored, and negative in the analysis where the age was included as a variable. There is also a faster fall off in the modulus of the loadings in the former analysis than in the latter. Although the factor relates to *general body size*, it can be seen that the variables indicative of *tissue body fat* have also been listed, but with considerably smaller loadings.

In both cases the second factor relates to *circulating serum body fats*. Once again, the same variables are involved in almost the same order, and the factor loadings are very similar.

The third factor in each case, represents *food intake*, particularly that of the protein, and energy-producing foodstuffs. Once again, the loadings in the second analysis are all negative for this factor.

The fourth factor represents the *serum protein fractions*. In the analysis where age was ignored, the globulin fractions relating to the degree of exposure to *antibody-producing infections* are given slightly more prominence than in the analysis where age is included as a variable. The two factors, however, remain extremely similar. They contain the same variables, but in a somewhat different order.

The fifth factor in both cases relates to the *combatting of infection* by means of certain cells which are part of the differential white cell count. In the one case, the lymphocytes(which combat antibody-producing infections) have a positive loading, and in the other case the neutrophils (which combat any foreign bodies in the blood stream).

111/.....

In the sixth factor the first difference occurs in the results of the two analyses. In the first analysis (effect of age ignored), the factor represents the ability of the blood to absorb oxygen or *oxygen uptake*. This factor only comes in as the 8th in the second analysis (age included as a variable). In the second analysis, the 6th factor represents the *tissue body fat* with a very clear demarcation in the size of the loadings between the three skinfold measurements, and the other somatometric measurements also listed under the factor but indicative of *general body size*.

For both analyses, the 7th factor represents the *nicotinic acid metabolites*. Only in the first analysis, however, do the 2-pyridone and pyridone/N^1-Me variables both have reasonably high loadings. In the second analysis only 2-pyridone has a high loading.

In the first analysis, the 8th factor is difficult to label, a result which may be due to the fact that 11 factors rather than 10 were extracted. As already mentioned in the second analysis, the 8th factor clearly represents *oxygen uptake*.

In the case of the first analysis, the 9th factor represents *tissue body fat* and is equivalent to the 6th factor in the second analysis. The 9th factor in the second analysis represents *food intake*. The emphasis appears to be on *animal protein* foodstuffs in this factor, whereas it appears to be on *vegetable protein* in the 3rd factor of the same analysis.

For both analyses, the 10th factor clearly represents *protein status* as indicated by the serum albumin concentration (expressed either as an absolute value, or as a proportion of the total serum protein). It would appear, since albumin in both analyses has a considerably higher loading, that it is probably the better of the two indices for estimating protein status.

112/.....

The measure of agreement between the two analyses, the one ignoring age, and the other including it as a variable, is remarkable. The inclusion of age as a variable in the second analysis and its subsequent inclusion in the first factor, implies that the effect of age has been largely removed from the other factors in Table XXV. This results suggested the possibility of eliminating from the correlation matrix, the effect of age before the analysis was performed. The way in which this was carried out and the result obtained, is described below.

8.4 Factor analysis on all variables for all racial groups with the effect of age partialled out

The partial correlation matrix was calculated with the effect of age partialled out by replacing each element r_{ij} of the correlation matrix by the partial correlation $r_{ij.k}$ (as defined in (4) in Chapter 6.3). The factor analysis, as described previously in Chapter 8.3, was then carried out on the *partial correlation matrix*. The results have been tabulated in Table XXVI in the same way as previously. All the factors which occurred in the first and second analyses can be discerned in the analysis with the effect of age partialled out, though the factors are in a somewhat different order (as based on the size of the eigen values).

In order to facilitate the comparison of the factors, the physiological characteristics which they represent have been tabulated in Table XXVII. It is interesting to note that the first and last factors for all the analyses are identical. The first representing *general body size* and the last *protein status*. In the third analysis with the effect of age partialled out *food intake*, however, appears in the 2nd factor rather than the 3rd, and the *serum protein fractions* occurs in the 3rd factor rather than the 4th. The *circulating (serum) body fats* which were in the second factor in the first two analyses occur in the 4th factor in the third analysis.

113/.....

TABLE XXVI : FACTOR ANALYSIS OF ALL RELEVANT VARIABLES FOR ALL RACES (EFFECT OF AGE PARTIALLED OUT).

VARIABLES WITH HIGH FACTOR LOADINGS (Loadings in brackets).									
Factor 1	Factor 2,	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
General body size	Food intake	Serum protein fractions (anti-body producing infections).	Circulating body fats	Oxygen uptake	Combatting infection.	Tissue body fat	Nicotinic acid metabolites.	Vegetable protein intake	Protein status
Height (-0,91)	Calories (0,91)	Globulin (-0,92)	Cholesterol (0,94)	Haemoglobin (-0,95)	Neutrophils (1,00)	Subscapular skin-fold (-0,87)	2-Pyridone/ creatinine (0,92)	Veg. protein (-0,52)	Albumin (-0,82)
Cristal height (-0,88)	Protein (0,91)	γ-globulin (-0,83)	Phospholipids (0,67)	Haematocrit (-0,75)	Lymphocytes (-0,91)	Para-umbilical skinfold (-0,85)	N'-Me creatinine (0,55)	Riboflavin (0,39)	Albumin/Total protein (-0,61)
Ulnar length (-0,84)	Phosphorus (0,87)	Total protein (serum) (-0,80)	Choles/Phospho-lipids (-0,56)	M.C.H.C. (-0,39)	White cells (0,18)	Triceps skinfold (-0,85)	Pyridone/N'-Me (0,41)	Calcium (0,39)	α-globulin (0,46)
Weight (-0,72)	Thiamine (0,87)	γ-globulin/Tot. protein (-0,71)	Carotene (serum) (0,52)	α-globulin (-0,18)		Upper arm circ. bent (-0,67)	Urinary riboflavin (0,38)	Alkaline phosphatase (-0,36)	Total protein (serum) (-0,41)
Biacromial width (-0,69)	Nicotinic acid (0,79)	Albumin/Total protein (0,71)	Vitamin A (serum) (0,31)			Weight (-0,58)			Vitamin A (serum) (-0,40)
Intercristal width (-0,67)	Carbohydrate (0,71)	β-globulin (-0,40)				Calf circum. (-0,54)			
Calf circum. (-0,62)	Riboflavin (0,68)	α-globulin (-0,33)							
Upper arm circ. bent (-0,52)	Fat (0,67)								
Urinary amylase (0,27)	Veg. protein (0,65)								
Serum amylase (0,26)	Calcium (0,65)								
Thiamine (-0,26)	Animal protein (0,58)								
	Iron (0,53)								
	Vitamin A (0,37)								
	Vitamin C (0,35)								

114/.....

TABLE XXVII : COMPARISON OF THE FACTORS FOR EACH OF THE THREE FACTOR ANALYSES

Factors	1	2	3	4	5	6	7	8	9	10
Age ignored	General body size	Circulating serum body fats.	Food intake (protein)	Serum protein fractions (antibody-producing infections)	Combatting infection	Oxygen uptake	Nicotinic acid metabolites	-	Tissue body fat	Protein status
Age as a variable	General body size	Circulating serum body fats.	Food intake (Veg. prot.)	Serum protein fractions (antibody producing infections).	Combatting infection	Tissue body fat	Nicotinic acid metabolites	Oxygen uptake	Food intake (animal protein)	Protein status
Age partialled out	General body size	Food intake	Serum protein fractions (antibody-producing infections)	Circulating serum body fats.	Oxygen uptake	Combatting infection	Tissue body fat.	Nicotinic acid metabolites	Food intake (Veg. protein)	Protein status

115/.....

The 5th factor represents *oxygen uptake* which occurred 6th in the first analysis and 8th in the second analysis. The 6th factor represents the ability of blood to *combat infection*. This occurred 5th in both the first and second analyses. The 7th factor represents *tissue body fat* with the upper arm circumference variable, which we have seen to be indicative of soft body tissue, also having a reasonably high loading.

The 8th factor represents the *nicotinic acid metabolites* which were included in the 7th factor for the first two analyses.

The 9th factor again represents *food intake*, this time perhaps with the emphasis on *vegetable protein*. In the 9th factor of the second analysis, a similar characteristic emerged but with the emphasis on *animal protein*. The 10th factor, as already mentioned, is identical for all three analyses and represents *protein status*.

In considering the results of these three analyses, it is difficult to select any one of them as having given the best results. Theoretically, since many of the variables are dependent on age, the last analysis (Table XXVI) should be the most reliable. It is, however, comforting to note that whether age is ignored or included as a variable, it does not have any serious effect. Since, in research in the biological field, the research worker may often be reluctant to omit the measurement of variables whose importance has not yet been established, it is indeed encouraging to find that a variable such as age, which is so evidently related to many of the variables (e.g. the somatometric variables descriptive of general body size) can be omitted from the analysis without serious loss to the results.

It is interesting to note that the somatometric variables descriptive of *tissue* body fat and the biochemical variables

116/.....

descriptive of circulating *serum* body fats, do not occur in the same factor. The reason for this is probably two-fold. First, it must be borne in mind that although all the variables with the exception of the dietary variables, were measured simultaneously, there is a physiological time-lag between the somatometric variables and the biochemical variables. Secondly, it has never been clearly shown to what extent the circulating serum fats are indicative of total body fat. The present results may well suggest that they are less closely related to total body fat than has sometimes been suggested.

8.5 The identification of redundant variables

It will be remembered that, when stating our purpose in Chapter 2, namely, to select an optimum subset of variables by which nutrition status can best be defined and assessed, we posed two criteria for the rejection of variables. The first being to reject "noise" variables, i.e. those variables that are little related to nutrition status. The second being to reject those variables which, although related to nutrition status, are so highly correlated to one another that the measurement of both is unnecessary.

It would appear that the variables which have been grouped together with high factor loadings for a particular factor measure *the same aspect* to a large extent. This can clearly be seen in respect of those variables indicative of general body size. These were height, cristal height, ulnar length, weight, biacromial width, intercristal width and calf circumference, all with a factor loading whose modulus was in excess of 0,60. Clearly, if we wished to select an optimum subset of these variables for the purpose of measuring general body size, the first two or three would be sufficient. In a similar way an optimum subset of variables could be selected for each of the other factors. The actual *selection* of variables of any given survey *would depend on which factors*

were of interest and how accurately each of them needed to be measured. The replicating of information by different disciplines would also have to be considered. We will pursue this line of thought further when summing up the results of all the statistical analyses in Chapter 10.

The factor analysis should, however, also be able to make a contribution in respect of identifying those variables which are *unrelated to the general information contained in the data set.* These are the variables with low communalities after the final rotation of the factor matrix. They have been ranked in Table XXVIII in the order of *increasing* communalities for the first 14 variables. The 3rd column of the table identifies the relevant discipline, whilst the last three columns show the rank number of the particular variable for the three analyses in which the effect of age was ignored; included as a variable; or partialled out.

It can be seen that a high measure of agreement exists between the last two techniques. The order obtained for the first analysis in which age was ignored disagrees markedly after the first three variables. The reasons for this may be twofold. It is probably due to the "age effect" on the variables, which has not been taken into account. It could, however, be partly related to the fact that in the first analysis we extracted 11 factors, whereas only 10 were extracted in the other two analyses.

In the unlikely event of the different order being due to some extent to the difference in the number of factors extracted, the results would be highly variable and would, therefore, have limited value. Although this is probably not the case, we will confine our attention to the first few variables.

All the variables were initially selected with a view to assessing nutrition status. The variables with low communalities will, therefore, be unrelated to nutrition status, but

TABLE XXVIII : VARIABLES WITH LOW COMMUNALITIES RANKED IN ORDER OF INCREASING SIZE FOR EACH OF THE THREE METHODS

Variable		Discipline	Rank		
No.	Name		Age ignored	Age included as variable.	Age partialled out
59	Monocytes	Haematology	2	1	1
57	White cells		1	3	2
61	Eosinophils		3	2	3
6	Mixed protein	Dietary	20	4	4
35	Inorganic phosphorus (serum)	Biochemical	4	6	5
40	Thiamine (serum)		6	8	6
56	M.C.H.C.	Haematology	8	7	7
34	Alkaline phosphatase		5	5	8
48	β-globulin	Biochemical	30	14	9
43	Pyridone/N^1-Me		18	11	10
30	Para-umbilical skinfold	Somatometric	38	36	11
37	Urinary amylase/Creatinine	Biochemical	12	16	12
52	Urinary riboflavin/Creatinine		11	15	13
14	Vitamin A intake	Dietary	10	13	14

119/.....

may have value in assessing the clinical condition or general health of the child. They should thus be carefully considered before being rejected in any future survey.

The first variable was the monocyte count. Monocytes also combat foreign bodies in the blood stream but have not been included in any of the factors for any of the analyses. The second variable is the total white cell count. White cells, have likewise, never had a high loading. The third is eosinophils, which are involved in combatting parasitic infections or allergies. Both monocytes and eosinophils were included in the 3rd principal component in the principal component analysis on the haematological variables, but appear to be less important in the overall context than lymphocytes and neutrophils.

The fourth variable is mixed protein. Whilst this is a nutrient intake variable, it is clearly a very poorly defined one since, when a food is classed as a mixed protein, this is in effect an admission that the dietician was unable to determine the proportion of animal to vegetable proteins in the food. It is, therefore, not surprising that this variable would be unimportant in estimating nutrition status.

The further sequence of the variables can be seen from Table XXVIII. It would appear reasonable, if variables are to be rejected, to consider those which appear early in the table as the first candidates.

8.6 Biological applications of factor analysis

As stated in Chapter 3, many variations of the factor analysis technique have been developed, and we have used only one of them above. Whilst the applications of factor analysis in psychology where this technique had its origin, are legion, relatively few applications in the biological sphere can be traced in the literature. Such applications have, however, covered a wide field as will be illustrated by the examples given below:

120/.....

Sokal and Daly (1961) e.g. have used a factor analysis to study insect behaviour. A total of 19 biological variables were measured on six species of insects. These were of two kinds; those involving the pulsation of the heart or gut, and those involving fleeing reactions of the insect when exposed on a glass surface. Six physical variables were measured concurrently with the biological variables. The authors note six factors that appeared to "cause" most of the correlations observed in the study. Two of these related to physical aspects and four related to the type of insect or its locomotion. They concluded that: ".... at least some of the factors could be reified to meaningful physical and biological variables while the unidentified factors served to indicate the need for further enquiry".

Wallace and Bader (1967), have used a factor analysis to investigate 27 dental and cranial measurements in the house mouse. They were able to identify five common factors with respect to the 27 variables which they concluded were related to *width, anterior length, posterior length, skull and "M3"*. They concluded that their analysis yielded "a more general view of the forces, or morphogenetic fields, affecting tooth size and interrelationships than can be obtained from the individual correlation coefficients".

Rohlf and Sokal (1962) have used the factor analysis technique to assist in the description of taxonomic relationships. They illustrated the technique on data arising out of a study of the *hoplitis* complex of bees, and conclude that the factors resulting from the analysis correspond in most cases to the previously established taxonomic groups.

Pearce and Holland (1960), discussed the application of a factor analysis to problems of fruit tree growth and cropping, and concluded that multivariate methods such as factor analysis, could lead to a better understanding of the tree *as a whole*. Gould (1965), has used a factor analysis technique to study evolutionary patterns in pelycosaurian reptiles.

121/.....

Kraus and Choi (1958) studied the prenatal growth of the human skeleton by means of a so-called factorial analysis, using the simplified method for the calculation of principal components put forward by Hotelling (1936). They concluded that the long bones of the human foetal skeleton are under a "major regulatory control" which they called the principal component but that there were, in addition, "regional characteristics of growth" which were under the influence of "secondary factors or components".

Doll and Bukatsch (1952), used a factor analysis to study age standardised mortality from various causes for a number of towns in England and Wales. We have already referred (Chapter 4.5) to the work of Burt and Banks (1947) who carried out a factor analysis of body measurements for British adult males. As we pointed out, there is a similarity between their results and the results of our principal component analysis on the somatometric measurements.

The above give some idea of the heterogeneous nature of factor analysis applications in the biological field. The technique has, however, never to our knowledge, been used before to aid in the selection of an optimum subset of variables by which nutrition status can best be assessed.

CHAPTER 9

A CLUSTER ANALYSIS APPROACH

The technique of cluster analysis was described in Chapter 3.4. It was pointed out that there are basically two types of cluster analyses. The first represents an attempt to cluster *subjects* into groups so that in terms of the observed variables the differences in subjects between groups, will be much larger than the differences within groups. The second form of cluster analysis is one which aims at grouping together *variables* which measure the same or a similar characteristic. The application of the first form of cluster analysis viz. the clustering of subjects could be applicable to our present data set. It would represent an attempt to cluster the school children into different physiological types in respect of the observed variables but this is outside the scope of the present report.

It is, however, of interest to assess to what extent the variables measured can be grouped together on the basis of the information they contain. This approach will be discussed below.

9.1 The rationale for clustering variables

The concept of clustering variables is not new to our study. This, in a sense, is what we have done in Chapter 4 by means of the principal component analyses. We have identified those variables which can be associated with a particular principal component. These are the variables which collectively measure a particular characteristic, and we have seen that in many cases this characteristic has biological identity. The concept of clustering variables was further developed in the factor analyses dealt with in the previous chapter. Here, on the basis of previous results, we presupposed the existence of factors and then demonstrated how these could be identified in terms of the observed variables. The factor loading, which is the correlation between the factor and the observed variable, provided an indication of the *importance of the variable*

in describing the factor.

The technique we will now follow, however, has certain important differences. Firstly, it requires none of the stringent assumptions which are inherent in the factor analysis approach. Secondly, the variables will be grouped in clusters in such a way that the position of each variable is unique. That is to say, *a variable will occur in one and only one cluster* in contrast to both the principal component analysis and the factor analysis which permit a variable to have a high loading on more than one factor. In a sense then, clustering of variables implies the identification of a "simple" attribute which can be measured explicitly in terms of one or more variables. This is of course contrary to one of the basic assumptions in factor analysis.

A third difference between the clustering of variables and the techniques we have used previously, is that, when using a cluster analysis approach, it is possible to observe *nested groups of clusters*. Thus, we might have two small clusters each measuring a specific attribute, which combine together to form a larger cluster with a composite attribute.

The cluster analysis procedure does, however, also have distinctive disadvantages over those which we have used previously. Computationally, the cluster procedures have not been well worked out and tend to proceed in a somewhat clumsy and time-consuming fashion. As such they form a distinct contrast with the elegant mathematical procedures of principal component or factor analysis. Because of this, cluster procedures tend to use a great deal of computer time and, are therefore, costly to use. The analysis which we are about to describe ran for *45 minutes*, fast core, on an I.B.M. System 360 Model 65 computer, and cost some 15 times more than a comparable factor analysis. Because of this, the analysis was only run once although a variety of approaches are possible.

The program used, (BMDP1M), was one of a new BMD series (Dixon, 1968), written in the Health Sciences Computing Facility, UCLA but not yet published. It was designed to produce a cluster analysis of variables by associating those variables which are highly similar in "clusters". Each cluster in turn, is treated as a single variable and associated with another variable to produce a new cluster. Similarity between the variables is judged on the basis of a distance matrix

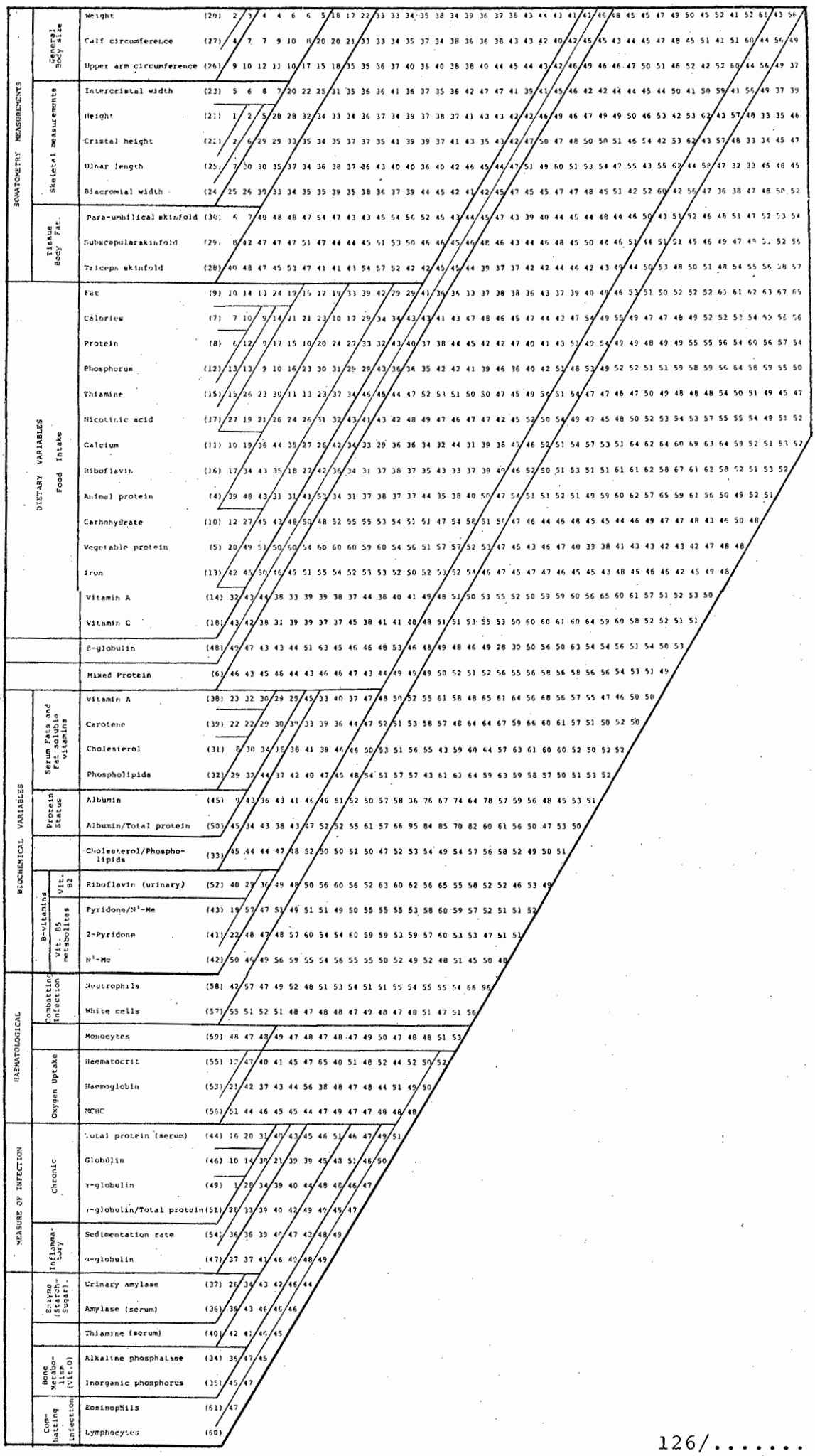
$$d_{ij} = 50(1 - r_{ij}) \qquad (1)$$

where r_{ij} is the correlation between the i-th and j-th variables. The algorithm in the clustering procedure, defines the two most highly correlated variables (i.e. those with the smallest distance between them) to be a cluster. It then regards this cluster as a simple variable and again finds the two variables which are closest together in terms of the distance function. This procedure is continued iteratively until all the variables have been clustered.

9.2 Results of cluster analysis applied to all variables for all racial groups

The cluster analysis procedure was applied to all variables for all racial groups. The results have been summarised in Table XXIX. The use of the table requires some explanation. The variables with appropriate identification numbers are tabulated in a specific order, with those variables which form part of the same cluster close together. To the right hand of the variable numbers, the matrix of distances has been printed. If, for example, we wish to observe the distance between *biacromial width* (No. 24) and *weight*, (No. 20), we would start from the point where the diagonal line meets the horizontal line associated with biacromial width and proceed up the line (in a north-easterly direction, assuming north to be at the top of the table), until we intercept the horizontal line associated with weight. At this point, a distance of 5 can be

TABLE XXIX : CLUSTER ANALYSIS OF ALL VARIABLES FOR THE FOUR RACIAL GROUPS USING THE DISTANCE MEASURE 50 (1-CORRELATION). (VARIABLE NO'S IN BRACKETS).



126/.....

read off from the table. The correlation between the two variables can also be calculated from the table since it follows from (1)

$$r_{ij} = 1 - d_{ij}/50$$

In this case $r_{24,20} = 0,90$

It can be observed that variables which are close together, or highly correlated, have been tabulated near to each other. The diagonal and horizontal lines drawn in on the graph indicate clusters identified by the program. Thus, at the top of the table, the smallest cluster is that of weight and calf circumference with a distance of 2, (r = 0,96). This cluster, however, is nested within the cluster formed by weight, calf circumference and upper arm circumference with maximum distance of 4,(r = 0,92). It can be immediately identified as a parameter of *general body size*. Just below this, a cluster representative of the skeletal measurements can be observed stretching as we move down the list of variables, from intercristal width to biacromial width. The maximum distance in this cluster viz. the distance between biacromial width and intercristal width, is 7. It appears to represent *skeletal size*.

In a similar fashion, a cluster below this, can be observed which is representative of *tissue body fat*. Again in this cluster the maximum distance is 8,(r = 0,84). It can, however, be seen that the cluster representative of body size, skeletal measurement and tissue body fat forms a composite cluster of somatometric measurements, with maximum distance 35,(r = 0,30), representative of the *somatometric measurements*.

Below this we have a large cluster representative of 14 of the nutrient intake variables. This would clearly represent *food intake*. Only mixed protein is excluded from the

127/.....

group and this variable, together with β-globulin is equally closely related to the somatometric variables as to the dietary variables. The problems relating to mixed protein have been discussed in Chapter 8.4, and the unsatisfactory nature of this variable has been indicated.

Below this, we have a large cluster representing the biochemical variables. Within this are nested clusters relevant to *serum fats and fat-soluble vitamins, protein status and B vitamins*. The latter cluster can be sub-divided into *vitamin B5* (or nicotinic acid metabolites) and *vitamin B2*. It can be seen that all the distance measures between the biochemical variables in the large clusters are less than 50 indicating that they are positively correlated with each other.

Below this, we have the neutrophils and white cells which form part of a collection of clusters, consisting of all the variables already listed. Below this we have monocytes, which are not related to any specific variable, but can be grouped with all the remaining variables. Below these, we have a cluster of two variables, haematocrit and haemoglobin which we have seen (Chapters 4.3 and 8.3) to be indicative of *oxygen uptake*. The M.C.H.C., which is in fact a function of haematocrit and haemoglobin, is also related to this cluster but not as closely.

Below this, we have a cluster consisting of some more biochemical variables. The larger cluster reaching from total protein down to α-globulin, represents the total protein and protein fractions in the blood serum. Within this, are nested two clusters, one relating to the γ-globulin fraction which, as we have seen, is associated with combatting *chronic (antibody-producing) infections*, and one relating to α-globulin and sedimentation rate. We have seen that α-globulin is thought to be associated with *inflammatory conditions*. The inclusion of sedimentation rate in this cluster is, therefore, most interesting, since it is a well-known fact that when infection is present in the body, the sedimentation rate is

128/.....

increased.

Below this we have the urinary and serum amylase activity, forming a small cluster. As we have seen in Chapter 5.4, this enzyme is responsible for breaking down the dietary starch into simple sugars. Below this we have a cluster consisting of alkaline phosphatase and inorganic phosphorus, which are both active in bone metabolism and also form parameters of estimating vitamin D status. Below this, we have the remaining variables in the differential white cell count, eosinophils, and lymphocytes. The former is related to *combatting parasitic infections*, and the latter to *combatting antibody-producing infection*.

It is evident that a simple clustering procedure such as we have used can produce a remarkably coherent and clear cut indication of the interrelationships amongst the variables. The similarity between these results and those obtained both in the principal component analysis for each racial group (Chapter 4) and the factor analyses carried out on all racial groups (Chapter 8) is striking and will be discussed below.

9.3 Comparison with previous results

Since the principal component analyses were carried out on each discipline individually, they are more likely to bring out detailed characteristics than the factor or cluster analyses which were done on all disciplines combined. We will, therefore, mostly confine ourselves to a comparison of the cluster analysis results with those obtained for the factor analyses.

If we compare the summary in Table XXVII with Table XXIX we can see that each factor mentioned in the former can be found in the latter. The cluster analysis, however, permits observing the factors in greater detail. In the factor analyses, for example, the factor of serum protein fractions

emerged in all three analyses. It was the fourth factor in the first two analyses and the third factor in the last factor analysis. In the cluster analysis, however, we note that a further dichotomy of the variables is possible into those relating to *chronic (antibody-producing) infections* and those relating to *inflammatory conditions*. Furthermore, sedimentation rate, which is known to relate to inflammatory conditions, has been included in the cluster analysis. The cluster analysis was unable to separate out the concepts of general body size and soft body tissue or musculature, but then, neither was the factor analysis. These were only discernable in the principal component analysis carried out solely on the somatometric variables.

The comparison of the factor analysis and cluster analysis techniques is interesting. As we have observed, the factor analysis permits (at least in theory) the measurement of a complex factor which cannot be explicitly measured. In this process a variable may, in fact, be likely to contribute information on more than one factor. In the cluster analysis, however, the concept which is measured, is one which can be uniquely expressed by a group of variables, and these variables are not related to any other concept or factor.

Whilst the factor analysis model is probably nearer the truth in a biological problem of this kind, it would seem, since we have been able to name every cluster produced in the cluster analysis, that biologists, in general, are used to thinking in terms of simple rather than complex factors. Both the ability to postulate complex factors, and then to unravel these by means of a factor analysis, must of course be credited to the psycologist. He can postulate concepts such as *arithmetic ability, intelligence, memory* and then devise a series of tests, none of which can specifically measure any of the concepts he has postulated, but each of which measure a proportionate part of a number of concepts.

130/.....

It would seem possible that such a situation could well exist also in the biological field in terms of the variables we have been studying but, by and large, the results of the factor analyses seem to suggest that the biological situation can be adequately described by *simple* rather than *complex* factors.

The factor analysis provides some indication of the redundant variables based on a study of the communalities. It is interesting to note that the variables with low communalities (Tables XXVIII) are usually those with a distance measure in the region of 50 from all other variables in the cluster analysis (Table XXIX).

In general then, it would appear that the two techniques of cluster analysis and factor analysis are, to a considerable degree, *complementary* the one to the other. If, however, only *one* of the two techniques is to be used, then factor analysis has little to offer which is superior to cluster analysis, except computational facility. The stringent underlying assumptions associated with the factor analysis model must, furthermore, be held against it. The cluster analysis, in comparison, is conceptually simple. It involves a straightforward (if somewhat lengthy) procedure and gives a clear cut and unambiguous end result. According to Engelman (1972), it should be possible to improve the speed of the cluster analysis procedure 20 fold by a refinement of the computing procedure. This would make it an extremely useful and usable technique.

The present study is, to our knowledge, the first in which the techniques of factor analysis and cluster analysis have been compared on a large number of biological variables associated with a variety of disciplines. The similarity in the results would seem to suggest that both techniques are reliable in the basic information they reveal.

CHAPTER 10

GENERAL SUMMARY AND CONCLUSIONS

Whilst the primary purpose of this study has been the solution of a biological problem relating to the choice of variables for the assessment of nutrition status, the work has yielded an interesting illustration of the usefulness and reliability of a number of multivariate statistical techniques. This aspect will first be dealt with before the biological importance of the results is discussed.

10.1 The reliability and comparability of the multivariate statistical procedures

The data set on which the present study is based has provided a useful means of demonstrating both the potential and the limitations of a number of statistical techniques. It was, as we have pointed out in Chapter 2, based on representative samples of Pretoria school children in the age range 7-15 years from the four racial groups. Care was taken to obtain a random sample stratified according to age, sex and race, and no effort was spared to ensure that the response to the sample was as complete as possible. The recording of the variables was carried out with care and the techniques of assay or measurement used, were the most reliable available at the time. We had thus, a coherent and meaningful data set in which a volume of biological information was hidden. The extent to which this information has been deciphered is, at least in measure, a patent commentary on the value of multivariate statistical methodology.

Viewed in this light, the value of the *stepwise discriminant analysis* has been clearly demonstrated. The fact that differences could be established between, for example, the White and Bantu racial groups on the basis of their somatometric variables, notwithstanding the dependence of many of the variables on age, is a tribute to the inherent robustness of the technique. Its power to detect proportional differences

in the variables, has, as far as we know, not been noted in the literature. The ability of the discriminant analysis to identify the White and Bantu racial groups on the basis of such differences has, however, been clearly demonstrated.

The results have, furthermore, served to show the inherent similarity between the stepwise regression analysis and stepwise discriminant analysis procedures. They have also served to high-light an important characteristic of both these procedures, which has, on occasion, led the biological research worker to misinterpret his results: The sequence in which the variables are included depends on the amount of "new" information which each can contribute, and is not related to the absolute importance of the variables in the relationship which is being studied. The study has shown that a discriminant analysis is an intelligible, stable and highly useful statistical procedure, which has wide application in the biological field.

The study of the relationship between causal and consequential sets of variables has provided an interesting demonstration of the usefulness of the *canonical and trace correlation coefficients*. Since it is incumbent upon the biological research worker to tread warily, for fear of spurious relationships, the facility provided by the partial canonical and trace correlation coefficients for removing the effect of a variable such as age (which might have been responsible for a spurious correlation), is indeed a valuable one. This statistical procedure is relatively new and unknown, but should have wide range of application in biological research. The use of the *stepwise regression analysis* in conjunction with the canonical function, which as far as we know has never been done before, has provided an interesting example of how two statistical techniques can be combined to throw light on an important aspect. Were it not for the facility provided by the combination of these two tests, we would not have been able to derive the result (to be discussed more fully in the

following section), which suggest that differences in skeletal proportion may have a nutritional origin.

The best insight into both the pitfalls and potential of multivariate statistical techniques has, however, been obtained from the principal component, factor and cluster analyses.

The consistency of the *factor analysis* procedure in the three analyses which were carried out, in which age was ignored, included as a variable, or partialled out of the correlation matrix, is highly satisfying. Clearly, this result does not permit any form of generalisation, but it may be a comforting thought for the biological research worker that even if there is some unknown variable which he cannot measure, which is related to the variable he is studying, it may after all not have too disturbing an effect on his results.

We have described in Chapters 3 & 4 how a *principal component analysis* is essentially a mathematical procedure for breaking down a covariance or correlation matrix into a set of orthogonal components or axes. It is thus basically a technique which assumes that the variation observed, is mainly due to the component having a characteristic effect on each of the variables. In contrast, the factor analysis attempts to take into account the possibility of *extraneous variation*. This is indeed the purpose behind replacing the "ones" in the diagonal of the correlation matrix with the so-called *communalities*. These communalities would represent the proportion of the variation of each variable which is due to factors operating on some or all of the other variables. The rest of the variation is assumed to be due to chance (Pearce and Holland, 1960).

The methods used for determining the communalities have, on occasion, been severely criticised. For example, Ehrenberg (1962) states "... the *'best' communality* cannot be determined easily, if at all. A good many rules for choosing some number

or other have been invented, but these rules all tend to give different answers".

We have no intension of going into the many varieties of factor analysis procedures, much less of presenting a case in their defence. The results of our investigation can speak for themselves. In our analyses, the principal component analyses and the factor analyses were not carried out on the same data set, and are therefore, not directly comparable. The similarity of the results of the two procedures is, nevertheless, quite remarkable. What is even more remarkable, however, is the extent to which the results of the *cluster analysis*, which was done on the same multidisciplinary data set as the factor analyses, compare with the results of the factor analyses.

The principal component analysis is an elegant mathematical procedure. The factor analysis is even more sophisticated and dependent in addition, on stringent assumptions. In contrast, the cluster analysis is, at best, an empirical and rather pedestrian procedure for "making sense" out of the correlation matrix. And yet, the strength of the procedure lies in its simplicity. It is conceptually so straightforward that it can be clearly described and understood, even by the non-mathematical research worker. Since it makes no assumptions, there are none which require defending. Yet, as we have seen, it has given basically the same information as the principal component or factor analyses. Our results then, are encouraging in that the three procedures *present a coherent and interpretable result or series of results.*

The factor analysis procedure may have a usefulness over the other techniques in respect of the problem of rejecting redundant variables. This aspect will be dealt with more fully in the following section.

We have pointed out that a factor analysis tends to measure complex factors (where one variable may be related to more than one factor) while a cluster analysis displays

135/.....

simple factors.

To a certain extent, therefore, the technique of cluster analysis appears to be supplementary to that of either a principal component or factor analysis. If possible, it would be useful to carry out a cluster analysis in conjunction with either a principal component or factor analysis. There is, however, a considerable measure of duplication in the results they provide. Hence the calculation of both procedures may not always prove economically justifiable. As far as the choice between a principal component or factor analysis is concerned, we cannot, as we have stated, make a direct comparison between the results of these two procedures. Our results do however, tend to raise a question whether the greater sophistication of the factor analysis procedure is really justified. In the context of the problem we are studying, the answer (except for the possibility of eliminating redundant variables on the basis of low communalities), might well be in the negative. Perhaps some of the critisms levelled at factor analysis would never have been made, had its exponents showed more restraint or been more selective in its application.

Having expressed some doubt as to the necessity of carrying out the factor analysis, we must hasten to point out that the combined contribution of the three procedures has provided results which are of real value to the biologist. This aspect will now be discussed.

10.2 The biological importance of the results

Before going on to discuss the biological importance of the results, it is well to note in passing that there are certain factors outside the control of this study which may affect its validity. Firstly, the result must in the final count, be a function of the *variables observed in the survey*. No statistical procedure can present information relevant to

a variable which has been omitted from the data set (though it is possible that a procedure might suggest that not all the relevant variables had been included). To what extent then could this aspect have influenced our results? As far as the somatometric variables are concerned, the variables measured were those which could most readily be taken and which are best known. The skinfold measurements were, however, taken on the left side of the body (Smit, 1968,p.108), whereas it is customary to take them on the right side. It is not, however, likely that this could affect the predictive capability of the variables. As far as nutrient intake is concerned, the intake of most, if not all, known nutrients was measured. The situation is less clear cut for the biochemical variables but, it would appear that most of the well known biochemical parameters have been measured. The same would appear to be true for the haematological variables. It would therefore appear that on the whole, we are safe in assuming that most, if not all, of the *potentially useful variables for estimating nutrition status*, have been included in our data set.

There is, however, a further possibility for error, viz. if too many redundant variables which happen to relate to the same or similar aspect, have been included, these may bias our results. This is difficult to assess. If we knew which variables were redundant, we would not be faced with the problem of selecting an optimum subset. We can only say that the variables selected, represent an honest attempt at choosing all known variables thought to be relevant to the problem of assessing nutrition status.

There is a further aspect which can effect the validity of our results. If, for example, all the children in the samples taken from each of the racial groups had an optimum nutrition status, then the variables which measure nutrition status would all be relatively constant. It is necessary, therefore, that a range of nutrition levels should occur in the population we are studying, before we can detect a rela-

137 /.....

ship between the level of nutrition and a particular variable. Since we have covered four racial groups which vary considerably in respect of socio-economic status and dietary habits, it seems highly likely that a sufficiently wide range of nutrition status levels will have been observed. We have remarked earlier that the racial groups did not differ markedly in respect of protein status. This would explain why in the factor analyses, the best biochemical estimator of protein status viz. serum albumin, only came out as the 10th factor. The quality of our investigation could have been improved had the sample also included *overtly malnourished* children, but the range of nutritional levels observed is probably wide enough for reasonably reliable deductions to be made.

Having put forward the above reservations, we can now proceed with a discussion of the biological implications of the results.

In general, it would appear that principal component analyses have played a useful part in *identifying* the biological characteristics *measured by each discipline*. The procedure was highly successful in the case of the somatometric variables, where some 93 percent of the variation was explained by three clearly definable and intelligible components. Thus, if we measure the most important variables of these components, we can rest assured that we are obtaining almost all the information concerning the way in which one individual differs from another. The consistency of the result across the four racial groups would imply that the results are generally applicable.

The principal component analyses further provided insight which assisted in interpreting the result of the discriminant analysis between the White and Bantu racial groups. *The finding that these groups differed primarily in respect of skeletal proportion should be of considerable value to the biologist.* This result has never, to our knowledge been demonstrated in this way. The further application of a canonical ana-

138/.....

lysis, followed by a stepwise regression analysis (Chapter 6.4) which suggested that the same difference in skeletal proportion could be related to nutrient intake, has served to provide further evidence on a highly controversial issue, viz. *whether differences between the various racial groups are primarily nutritional or genetic in origin.*

The finding when, considering the biochemical variables, that the four racial groups differed primarily in respect of the parameter known to relate to inflammatory conditions, is interesting. This result had hitherto, not been mentioned in reports on the Pretoria surveys.

The clear differentiation between the racial groups which was possible on the basis of the variables from all the different disciplines, is remarkable and should stimulate further research into the cause of the differences.

The canonical and trace correlations between the nutrient intake variables and the somatometric variables have provided an impartial and objective indication *that a relationship exists.* This is most useful in that it indicates that, in some measure at least, nutrition status can be estimated by the somatometric variables. In a similar way, it has been shown that nutrition status can be estimated somewhat more accurately by the biochemical variables. These are, however, somewhat more expensive and difficult to obtain.

The three factor analyses have indicated very strongly that the various disciplines in themselves tend to form conceptual entities, This conclusion is supported by the results of the cluster analysis. The disciplines then, tend to supplement each other rather than to replicate the same information. (The extent to which each discipline measures the same information has, of course, been indicated by the canonical correlation coeffient). This result should prove encouraging to the biologist since it confirms that the categories (disci-

plines) which have evolved over many years are, after all, reasonable and coherent. This is, for example, particularly true of the biochemical and haematological variables which are both measured in the blood, but which in the factor analysis have emerged as different, and well-defined factors.

We now come to the main subject of the present study: To define *that* subset of variables which should be selected for the assessment of nutrition status. If the reader is hoping for a single, final and clear cut answer to this question, we fear he will be disappointed. Our results, do however, provide certain guide lines as to what choice should be made under a given set of circumstances.

Since the disciplines tend to supplement one another, it is clear that if *it is desired to assess the nutrition status to the highest possible degree of accuracy, most if not all of the disciplines would have to be involved.* Because, however, of the interrelationships which exist as detailed in the preceding chapters, a reasonable index of nutrition status can be obtained with fewer disciplines. The simplest and most direct approach would be to measure certain of the somatometric variables which, as we have seen, reflect the properties of *general body size, body fat and soft body tissues.* It would appear desirable that each of the three components should be measured.

The optimum subset of the somatometric variables which best relate to nutrient intake has been indicated in Table XX. These variables were virtually the same as those which best separated the White and Bantu racial groups. We would then suggest that a selection of the best somatometric variables be made in such a way as to ensure that all three components are represented. For example, *intercristal width, biacromial width, ulnar length and height* could be taken as representing both *general body size* and *skeletal proportion*, which we have seen, may be affected by nutrition. The *triceps* and *subscapu-*

140/.....

lar skinfold measurements could be taken for estimating *body fat*. The *upper arm circumference* could be taken for estimating *soft body tissue*.

This represent more than half the somatometric variables, and it could well be argued that the research worker might as well measure them all and should at least include weight. One need not, however, necessarily follow the above selection. In the principal component analysis on the somatometric variables we found that the loadings within the components did not vary greatly in size. Therefore, the choice of a particular variable can, to some extent, be based on convenience. If we therefore, suggest that *weight* and *height* be measured as an index of *general body size*, *triceps skinfold* as an index of *body fat* and *upper arm circumference* as an index of *soft body tissue*, we will have a selection of variables which is almost as good as the ones proposed above, but which can be taken far more readily, since these measurements can all be observed without the subject having to undress. (An allowance could be made for the contribution of the subject's clothes to the weight).

The *somatometric variables* can, however, only give a general indication of nutrition status and, if a poor nutrition status is observed, will not indicate the *nature* of the dietary deficiency. To do this, it would be necessary also to measure certain biochemical variables. It can be seen from Table XXVI (Chapter 8.5), that the factors 1 & 7, relating to *general body size* and *tissue body fat*, can be estimated by the somatometric variables. The inclusion of certain biochemical variables would further permit the assessment of factors nos. 3, 4, 6, 8 and 10, relating to *serum protein fractions* and *antibody-producing infections*, *circulating body fats*, *nicotinic acid metabolites* and *protein status*.

141/.....

The inclusion of the *biochemical variables* thus represents a considerable increase in the scope of information presented but at the same time implies a considerable increase in the cost. The research worker is now faced with a choice: Does he wish to measure a few factors with a high degree of accuracy or a number of factors with a lower degree of accuracy? In the former case, he would measure all variables with high factor loadings (Table XXVI) for those factors in which he is interested. In the latter case, he would measure only one or two variables with high loadings in each of a number of factors. This decision can obviously best be made by the biological research worker in the context of a specific problem. It may well be that in the population he wishes to study, he has reason to suspect the existence of a vitamin deficiency. In this case, the inclusion of the appropriate factor is obviously required. A reasonable general choice of the biochemical variables would be: *serum cholesterol*, to provide an indication of *serum fats*, the γ- and α-globulins to provide a *measure of inflammatory and chronic infection*, *serum albumin* to provide an indication of *protein status* and *2-pyridone* with possibly *urinary riboflavin* to provide an indication of *vitamin status*. This represents only 6 of the 22 biochemical variables measured in the present study.

A similar argument would hold in respect of the *haematological variables*. These do not appear to play a direct role in the assessment of nutrition status. They do, however, reflect the general health of the subject, since ill health might lead to poor eating and this, over a long period would result in malnutrition. The inclusion of certain haematological variables could thus be motivated, especially since a blood sample would, in any case, have to be taken if certain biochemical variables were to be measured. If included, the best variables would probably be *haemoglobin* and *haematocrit* for estimating *oxygen uptake* and possibly *neutrophils* and *lymphocytes*, for measuring the extent to which the body is *combatting infection*. This result is confirmed by both the

142/.....

factor analyses and the relevant principal component analysis.

The *dietary variables* clearly have considerable importance in the assessment of nutrition status. This is confirmed by the fact that they have come out either as the second or third factors in the factor analysis. They are, however, both costly and difficult to measure and their inclusion could probably not be motivated on the basis of a careful study of the cost in relation to information gained.

Clearly the above results depend to some extent, on the population groups studied. They should, however, be reasonably applicable to any population group in South Africa.

In Chapter 8.5 we pointed out that communalities could be used to identify redundant variables i.e., variables unrelated to the general information content of the data set. These were listed in Table XXVIII. We prefer, on the whole, however, to approach the problem from the view point of *only including those variables whose inclusion can be motivated* because they are needed to measure a specific factor. We would, therefore, rather base our choice on the loadings in the factor analysis, or if greater detail is required, on the eigen vectors in the principal component analysis.

It should be no cause for disappointment that the present study has not provided a final answer to the problem of assessing nutrition status, as formulated in the opening chapters. We have, in effect, required that variables themselves tell us something about the aspects which they are measuring, - an almost impossible task. There is no doubt that the study has contributed considerably to our understanding, both of the pitfalls and potential of the various statistical techniques and

143/.....

of the underlying characteristics of the biological variables. *It is surely worthy of note that information which has been painstakingly gathered by biologists over many centuries, can be confirmed by a multivariate statistical analysis of a well-planned multidisciplinary survey.*

R E F E R E N C E S.

1. Anderson, T.W. (1958). *An introduction to multivariate statistical analysis.* New York : Wiley.

2. Ball, G.H. (1965). *Data analysis in the social sciences :* American Federation of Information Processing Societies Conference Proceedings, Fall Joint Computer Conference, Washington : Spartan Books, 533-559.

3. Burt, Cyril and Banks, Charlotte. (1947). A factor analysis of body measurements for British adult males. *Ann. Eugen., London,* 13, 238-256.

4. Cassie, R.M. (1963). Multivariate analysis in the interpretation of numerical plankton data. *New Zealand Journal of Science,* 6, 36-59.

5. Chinn, K.S.K. and Allen, T.H. (1960). Body fat in men from two skinfolds, weight, height and age. *Report No. 248, U.S. Army Medical Research and Nutrition Laboratory,* Denver, 1-9.

6. Defrise-Gussenhoven, E. (1954). *Croissance et débilité.* Bruxelles : Institute Royal des Sciences Nuturelles de Belgique. Memoire no. 128, 1-70.

7. Dixon, W.J. (1968). *BMD Biomedical Computer Programmes.* Berkeley and Los Angeles : University of California Press.

8. Doll, R. and Bukatsch, J. (1952). An experimental factor analysis of cancer mortality in England and Wales, 1921-30. *J. Hyg., Lond.,* 50, 384-393.

9. Drion (1961). The intercorrelations between the nutrients consumed by a group of families in the Netherlands. *J. Royal Statist. Soc. A,* 124, 314-335, 361-371.

10. Durnin, J.V.G.A. and Ramahan, M.M. (1967). The assessment

145/.....

of the amount of fat in the human body from measurements of the skinfold thickness. *Brit. J. of Nutr.*, 21, 681-689.

11. Du Plessis, J.P., De Lange, D.J. and Fellingham, S.A. (1965a). Biochemical evaluation of the nutrition status of urban primary school children : Riboflavin status. *S. Afr. Med. J.*, 39, 1176-1180.

12. Du Plessis, J.P., De Lange, D.J. and Fellingham, S.A. (1965b). Biochemical evaluation of the nutrition status of urban primary school children : Protein status. *S. Afr. Med. J.*, 39, 1181-1185.

13. Du Plessis, J.P., De Lange, D.J. and Fellingham, S.A. (1966a). Biochemical investigation of the nutrition status of urban school children aged 12-15 years : Protein status. *S. Afr. Med. J.*, 40, 509-514.

14. Du Plessis, J.P., De Lange,D.J. and Fellingham, S.A. (1966b). Biochemical evaluation of the nutrition status of urban school children of 12-15 years : Riboflavin status. *S. Afr. Med. J.*, 40, 518-520.

15. Du Plessis, J.P., De Lange, D.J. and Neser, M.L. (1966c). Biochemical evaluation of the nutrition status of urban primary school children : Vitamin A status. *S. Afr. Med. J.*, 40, 1093-1097.

16. Du Plessis, J.P. (1967). *An evaluation of biochemical criteria for use in nutrition status surveys.* CSIR Res. Rep. No. 261. Pretoria : CSIR.

17. Du Plessis, J.P., De Lange, D.J. and Vivier, F.S. (1967a). The biochemical evaluation of the nutrition status of urban school children : Nicotinic acid status. *S. Afr. Med. J.*, 41, 1211-1216.

146 /.....

18. Du Plessis, J.P., Vivier, F.S. and De Lange, D.J. (1967b). The biochemical evaluation of the nutrition status of urban school children aged 7-15 years. Serum cholesterol and phospholipid levels and serum and urinary amylase activities. *S. Afr. Med. J.*, 41, 1215-1222.

19. Efroymson, M.A. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, Vol. I. Edited by Ralston, A. and Wilf, H.S.,New York: Wiley, 191-203.

20. Ehrenberg, A.S.C. (1962). Some questions about factor analysis. *The Statistician*, 12, No. 3.

21. Engelman, L. (1972). *Personal communication.* Health Sciences Computing Facility, UCLA.

22. Fellingham, S.A. (1966). Statistical planning of the nutrition status surveys on Pretoria school children. *S. Afr. Med. J.*, 40, 228-234.

23. Fellingham, S.A. (1969). *The selection of criteria in medical research.* Paper read at the Biometrics session of the Cross-Disciplinary Sciences Symposium in Biomedical Research, Johannesburg.

24. Gardiner, A.S. and Jeffers, J.N.R. (1963). Analysis of the collective species Betula alba on the basis of leaf measurements. *Silvae Genetica*, 11, 156-161.

25. Gould, S.J. (1965). Evolutionary patterns in pelycosaurian reptiles : A factor-analytic study. *Evolution*, 21, 385-401.

26. Harman, H.H. (1967). *Modern factor analysis, revised ed.* Chicago : The University of Chicago Press.

147/.....

27. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24, 417-441: 498-520.

28. Hotelling, H. (1936). Simplified calculation of principal components, *Psychometrika*, 1, 27-35.

29. Jeffers, J.N.R. (1967). The study of variation in taxonomic research. *The Statistician*, 17, 29-43.

30. Jolicoeur, P. and Mosiman, J.E. (1960). Size and shape variation in the painted turtle. A principal component analysis. *Growth*, 24, 339-354.

31. Kendall, M.G. and Stuart, A. (1966). *The advanced theory of statistics*. Vol III. New York : Hafner Publishing Co.

32. Khatri, C.G. (1964). Distribution of the 'Generalised' multiple correlation matrix in the dual case. *Ann. Math. Statist.*, 35, 1801-1806.

33. Kraus, B.S. and Choi, S.G. (1958). A factorial analysis of the prenatal growth of the human skeleton. *Growth*, 22, 231-242.

34. Lawley, D.N. and Maxwell, M.A. (1963). *Factor analysis as a statistical method*. London : Butterworths.

35. Levine, E. (1968). *A radiological study of hand and wrist ossification in South African children of four population groups*. Unpublished doctoral thesis : University of the Witwatersrand.

148/.....

36. Louw, M.E.J., Du Plessis, J.P., van den Berg, A.S. and Fellingham, S.A. (1969). Intercorrelation study of dietary and biochemical data from school children in the Pretoria area. *S. Afr. Med. J.*, 43, 1516-1527.

37. Lubbe, A.M. (1968). A survey of the nutritional status of White school children in Pretoria. Description and comparative study of two dietary survey techniques. *S. Afr. Med. J.*, 42, 616-622.

38. Lubbe, A.M. and Pretorius, C.L. (1965). "Nutrient inname van blanke 7-jarige kinders soos bepaal deur twee metodes". *S. Afr. Med. J.*, 39, 1147-1148.

39. Lubran, M. (1966). Paper electrophoresis. *J. Amer. Med. Ass.*, 197, 360-361.

40. Metheny, E. (1939). Some differences in bodily proportions between American Negro and White male college students as related to athletic performance. *Res. Q. Am. Ass. Hlth. Phys. Educ.*, 10, p.42.

41. National Research Council (1968). *Recommended dietary allowances*. 7th ed. Publ. No. 1694. Washington, D.C.: National Academy of Sciences.

42. Neser, M.L. (1968a). The leukocyte picture in White, Bantu, Coloured and Indian school children of 6-15 years as observed during the Pretoria nutrition status surveys of 1962-65. *S. Afr. Med. J.*, 42, 444-450.

43. Neser, M.L. (1968b). The sedimentation rate in White, Bantu, Coloured and Indian school children. *S. Afr. Med. J.*, 42, 1128-1137.

149/.....

44. Oberman, J.W., Gregory, K.O., Burke, F.G., Rose, S. and Rice, E.C. (1956). Electrophoretic analysis of serum proteins in infants and children. I. Normal values from birth to adolescence. *New Engl. J. Med.*, 255, 743-755.

45. Oudkerk, A.C.F. (1965). Eating habits of urban Bantu, with special reference to the school going child. *S. Afr. Med. J.*, 39, 1148-1150.

46. Ouelette, R.P. and Qadri, S.U. (1966). Principal component analysis and pattern of growth in *Cristivomer namaycush*. *Growth*, 30, 285-293.

47. Parizkova, Jana (1961). Total body fat and skinfold thickness in children. *Metabolism*, 10, 794-807.

48. Pearce, S.C. and Holland, D.A. (1960). Some applications of multivariate methods in botany. *Applied Statistics*, 9, 1-7.

49. Pearson, K. (1901). On line and planes of closest fit to a system of points in space. *Phil. Mag. II, 6th series*, 557-572.

50. Pillai, K.C.S. (1960). *Statistical tables for tests of multivariate hypothesis*. Manila: The Statistical Centre, University of the Philippines.

51. Potgieter, J.F. (1965). "Inkomste en voedingspeil". *S. Afr. Med. J.*, 39, 1151-1154.

52. Potgieter, J.F. and Fellingham, S.A. (1967). Assessment of methods for dietary surveys. *S. Afr. Med. J.*, 41, 886-890.

150/.....

53. Rees, J.W. (1969). Morphologic variation in the mandible of the white-tailed deer (*Odocoileus virginianus*): A study of population skeletal variation by principal component and canonical analysis. *J. Morphology*, 128, 113-130.

54. Rohlf, F.J. and Sokal, R.R. (1962). The description of taxonomic relationships by factor analysis. *Systematic Zoology*, 11, 1-16.

55. Seal, H.L. (1964). *Multivariate statistical analysis for biologists*, London : Methuen and Co., Ltd.

56. Smit, P.J. (1965). The importance of arm position when measuring upper arm circumference for the evaluation of nutritional status. *S. Afr. Med. J.*, 39, 1185-1186.

57. Smit, P.J. (1968). *Anthropometric, motor performance and physiological studies on South African children involved in a nutritional status survey.* Pretoria: CSIR Research Report No. 273, 1-470.

58. Smit, P.J., Potgieter, J.F. and Fellingham, S.A. (1967a). Body measurements of school children of four racial groups in Pretoria. *S. Afr. Med. J.*, 41, 868-886.

59. Smit, P.J., Potgieter, J.F., Neser, M.L. and Fellingham, S.A. (1967b). Sex, age and race variations in the body measurements of White, Bantu, Coloured and Indian children aged 7-15 years. *S. Afr. Med. J.*, 41, 422-426.

60. Sokal, R.R. and Daly, H.V. (1961). An application of factor analysis to insect behaviour. *The University of Kansas Science Bulletin*, 42, 1067-1095.

61. Spearman, C. (1904). General intelligence objectively determined and measured. *Amer. J. Psychol.*, 15, 201-293.

151/.....

62. Spearman, C. (1926). *The abilities of man.* London : Macmillan.

63. Stead, R.M. (1968). Assessment of methods for dietary appraisal : A comparison of two different modifications of Burke's dietary history method. *S. Afr. Med. J.*, 42, 961-962.

64. Sukhatme, P.V. (1954). *Sampling theory of surveys with applications, Ames, Iowa,* U.S.A. : The Iowa State College Press. 40-42.

65. Troskie, C.G. (1969). The generalised multiple correlation matrix. *S. Afr. Statist. J.*, 3, 109-121.

66. Troskie, C.G. (1971). Regression and correlation. *Proceedings of the third Symposium on Mathmatical Statistics*, ed. by N.F. Laubscher, Pretoria: CSIR Special Report WISK 89, 35-47.

67. Troskie, C.G. (1972). *Personal Communication.* Department of Mathematical Statistics, University of Cape Town.

68. Van der Merwe, A. le R., Potgieter, J.F. and Neser, M.L. (1965). The planning and execution of the 1962 Pretoria nutrition survey. *S. Afr. Med. J.*, 39, 220-221.

69. Wallace, J.T. and Bader, R.S. (1967). Factor analysis in morphometric traits of the house mouse. *Systematic Zoology*, 16, 144.

70. Warburton, F.W. (1962). The practical value of factor analysis in education. *The Statistician*, 12, 172-188.

152/.....

71. Watts, H.L. (1967). *The poverty datum line in three cities and four towns in the Republic of South Africa.* Natal: Institute for Social Research, University of Natal, Fact Paper No. 1.

72. Wittmann, W., Moodie, A.D., Fellingham, S.A. and Hansen, J.D.L. (1967). An evaluation of the relationship between nutritional status and infection by means of a field study. *S. Afr. Med. J.*, 41, 664-682.

/EDA.