J. chron. Dis. 1967, Vol. 20, pp. 511–524. Pergamon Press Ltd. Printed in Great Britain

A MULTIVARIATE ANALYSIS OF THE RISK OF CORONARY HEART DISEASE IN FRAMINGHAM

JEANNE TRUETT*, JEROME CORNFIELD† and WILLIAM KANNEL, M.D.‡

National Heart Institute, National Institutes of Health, Bethesda, Maryland

(*Received* 30 *December* 1966; *in revised form* 16 *February* 1967)

INTRODUCTION

IT IS the function of longitudinal studies, like that of coronary heart disease in Framingham, [1] to investigate the effects of a large variety of variables, both singly and jointly on the risk of developing a disease. The traditional analytic method of the epidemiologist, multiple cross-classification, quickly becomes impracticable as the number of variables to be investigated increases. Thus, if 10 variables are under consideration, and each variable is to be studied at only three levels, e.g. serum cholesterols of less than 225 mg/100 ml, 225–274, and 275 and over, there would be 59,049 cells in the multiple cross-classification. Even with only 10 cases for the denominator of the rate for each cell, a cohort of approximately 600,000 persons would be required.

Study populations of this size are not often available and one is consequently led to seek a more powerful form of analysis than inspection of the results of a multiple cross-classification. One such method was suggested by CORNFIELD. [2] He considered the case of k variables, say $x_1, x_2 \ldots x_k$ and assumed that the multivariate frequency distributions of those who would (CHD) and those who would not (NCHD) develop the disease could be represented by two known mathematical functions, say $f_1(x_1 \ldots x_k)$ and $f_0(x_1 \ldots x_k)$. In that case the probability $P(x_1 \ldots x_k)$, that an individual characterized by the variable values $x_1 \ldots x_k$ would develop the disease is given by

$$\mathrm{P}(x_1 \ldots x_k) = 1 \Big/ \left[1 + \frac{1-p}{p}\,\frac{f_0(x_1 \ldots x_k)}{f_1(x_1 \ldots x_k)}\right], \tag{1}$$

where p is the unconditional probability of developing the disease. In particular if the frequency distributions f_0 and f_1 are multivariate normal, with different means, but the same variances and covariances

$$\frac{1-p}{p}\,\frac{f_0(x_1 \ldots x_k)}{f_1(x_1 \ldots x_k)} = e^{-[\alpha + \sum_{i=1}^{k} \beta_i x_i]}, \tag{2}$$

*Statistician, Biometrics Research Branch.
†Chief, Biometrics Research Branch.
‡Director, Framingham Heart Study.

where the β_i are the coefficients of the linear discriminant function. [3] When the frequency distributions are multivariate normal but with different variances and covariances, the expression in the exponent is quadratic rather than linear in the x's. When there is only one variable, i.e. with $k=1$, expression (1) with f_0/f_1 given by (2) will be recognized as the logistic function. [4] The general function for k variables is therefore referred to as the multiple logistic.

The multiple logistic function seems to promise a more penetrating analysis than can be achieved by contemplation of cross-classifications. The main limitation to its use has been that the assumption of multivariate normality is rarely satisfied, even approximately. The original application by Cornfield was confined to two variables, serum cholesterol and systolic blood pressure, and in this case it was possible to transform the variables so that the multivariate assumption was satisfied. It has not been possible to find any systematic procedure for extending this transformation to a larger number of variables, however, and in the case of some (e.g. dichotomous variables) it is in principle impossible. This paper investigates the consequences on estimated risk of using the multivariate normal assumption, when departures from it are substantial. Although the primary interest is methodological, some substantive results of interest emerge.

MATERIALS AND METHODS

The analysis is based upon the 12-yr incidence of coronary heart disease of 2187 men and 2669 women, aged 30–62 and found free of coronary disease at first examination in Framingham. "Coronary heart disease" includes all definite myocardial infarction, coronary insufficiency, angina pectoris and death from coronary heart disease. As in all previous publications reporting results in Framingham, the lost to follow-up are treated as NCHD. Seven risk factors measured on the initial examination have been investigated:

Age (yr)
Serum cholesterol (mg/100 ml)
Systolic blood pressure (mm Hg)
Relative weight (100 × actual weight ÷ median for sex–height group)
Hemoglobin (g/100 ml)
Cigarettes per day, coded as
  0 = never smoked
  1 = less than a pack a day
  2 = one pack a day
  3 = more than a pack a day
ECG, coded as
  0 for normal
  1 for definite or possible left ventricular hypertrophy, definite non-specific abnormality and intraventricular block.

Two hundred and seventy-one individuals for whom information on one or more risk factors were missing were excluded.

Analyses have been performed separately for the age groups 30–39, 40–49 and 50–62 for men and 30–49, 50–62 for women as well as for all ages combined. For each group analyzed the pooled variance–covariance matrix was obtained and its inverse computed. We denote the element in the ith row and jth column of the inverse as $\hat{\sigma}^{ij}$. We also compute the means of each of the seven risk factors in the CHD populations, $\bar{x}_{i1}$, and in the NCHD populations, $\bar{x}_{i0}$, $i=1, 2, \ldots 7$ and the differences d_i, where $d_i = \bar{x}_{i1} - \bar{x}_{i0}$. The numbers of individuals in each of the populations are

	Free of CHD NCHD (N_0)		Developed CHD CHD (N_1)	
	Men	Women	Men	Women
All ages	1929	2540	258	129
30–39	749	1824	40	39
40–49	654		88	
50–62	526	716	130	90

The linear discriminant function coefficients are estimated as

$$\hat{\beta}_i = \sum_{j=1}^{7} d_j \hat{\sigma}^{ij} \qquad i=1, 2 \ldots 7, \tag{3}$$

the constant α is estimated as

$$\hat{\alpha} = -\tfrac{1}{2}\Sigma\hat{\beta}_i(\bar{x}_{i0}+\bar{x}_{i1}) - \log_e \frac{N_0}{N_1}\ . \tag{4}$$

For the variance of the discriminant function coefficients we have taken

$$\mathrm{Var}\,(\hat{\beta}_i) = \hat{\sigma}^{ii}\left(\frac{1}{N_0}+\frac{1}{N_1}\right). \tag{5}$$

This is exact when the variances and covariances are known, in which case the $\hat{\beta}_i$ can be treated as normal variables. In view of the 600 or more degrees of freedom available for estimating the pooled variances and covariances, the error of this assumption would not appear to be large.

The value of the risk for each individual was computed as

$$\hat{\mathrm{P}} = 1/[1+e^{-(\hat{\alpha}+\Sigma\hat{\beta}_i x_i)}]. \tag{6}$$

(Risk, probability and 12-yr incidence are used interchangeably.) For each group a frequency distribution of the values of $\hat{\mathrm{P}}$ was obtained, the deciles of the distributions determined, the observed numbers of cases of coronary disease tallied by decile and the number of cases expected in each decile determined by summing the calculated risks given by (6) for all the subjects in the decile (Tables 1, 2 and 3). The expected number of cases in Table 4 was similarly computed.

The frequency distributions of Fig. 2 were obtained by separating the combined populations into CHD and NCHD distributions and determining deciles of risk and extreme percentiles for each. The ordinates, for all but the two end tenths, are plotted as 0.1 over the difference between the two deciles defining the tenth. The tails of the distributions were defined in more detail by using the 2nd, 5th, 95th and 98th percentile as well and plotting as ordinate 0.02/difference, 0.03/difference, 0.05/difference, etc.

RESULTS

Expected and observed numbers of cases of CHD and 12-yr incidence for each of the deciles of risk are shown in Table 1 for the risk functions based on all age groups for men and for women. Table 2 shows the same comparison for men and Table 3 for women for the risk function computed separately for the different age groups. A similar comparison is shown for each of six individual risk factors in Table 4, separately for men and women and by age group.

TABLE 1. EXPECTED AND OBSERVED NUMBER OF CASES OF CHD AND OBSERVED INCIDENCE IN 12 YR OF FOLLOW-UP AT FRAMINGHAM OF MEN AND WOMEN AGED 30–62 YR AND FREE OF CHD AT ORIGINAL EXAMINATION, BY DECILE OF RISK

Decile of risk	2187 Men		Observed 12-yr incidence (no. of cases per 100)	2669 Women		Observed 12-yr incidence (no. of cases per 100)
	Number of cases			Number of cases		
	Expected	Observed		Expected	Observed	
10	90.5	82	37.5	70.4	54	20.2
9	47.1	44	20.1	24.7	23	8.6
8	32.6	31	14.2	15.0	21	7.9
7	25.0	33	15.1	9.8	14	5.2
6	19.7	22	10.1	6.5	5	1.9
5	15.0	20	9.1	4.4	6	2.2
4	11.5	13	5.9	3.2	2	0.7
3	8.6	10	4.6	2.3	0	0.0
2	6.0	3	1.4	1.7	3	1.1
1	3.4	0	0.0	1.1	1	0.4
Total	259.4	258	11.8	139.1	129	4.8

Two conclusions emerge from inspection of these results. (a) Despite the markedly non-multivariate normal nature of the distribution the agreement between observation and expectation is quite good both for the deciles of risk and for the individual risk factors. There is some tendency, however, for the expected to exceed observed at both the highest and lowest deciles of risk and to fall below in the middle. (b) The separation in incidence between lowest and highest decile is pronounced and is considerably greater than that achieved by the traditional classification by number of risk factors present. The total number of cases expected tends to be somewhat larger than observed as a result of the positive skew in the distributions of well persons, particularly women, illustrated in Fig. 2. One-half the men and more than one-fourth the women 50–62 in the highest decile of risk developed CHD in the 12 yr of follow-up. The highest 12-yr probability of disease computed for any individual for each of the groups and the actual 12 yr follow-up experience are

	Men	Women
30–39	0.986 (event)	0.838 (event)
40–49	0.742 (event)	
50–62	0.770 (no event)	0.773 (event)

While the percentage excess in incidence at the highest decile of risk is much greater among younger men, the arithmetic excess increases with age.

TABLE 2. EXPECTED AND OBSERVED NUMBER OF CASES OF CHD AND OBSERVED INCIDENCE IN 12 YR OF FOLLOW-UP AT FRAMINGHAM IN MEN, BY AGE AT FIRST EXAMINATION AND DECILE OF RISK

Decile of risk	789 aged 30–39		Observed 12-yr incidence (no. of cases per 100)	742 aged 40–49		Observed 12-yr incidence (no. of cases per 100)	656 aged 50–62		Observed 12-yr incidence (no. of cases per 100)
	Number of cases			Number of cases			Number of cases		
	Expected	Observed		Expected	Observed		Expected	Observed	
10	21.9	18	22.8	26.1	26	35.0	31.8	32	48.8
9	7.7	8	10.1	15.4	15	20.2	21.3	17	25.9
8	4.1	4	5.1	11.4	6	8.1	16.9	23	35.1
7	2.6	4	5.1	8.9	11	14.8	13.6	8	12.2
6	1.8	4	5.1	7.1	9	12.1	11.4	10	15.2
5	1.3	1	1.3	5.9	4	5.4	9.8	14	21.3
4	0.9	1	1.3	4.8	9	12.1	8.3	7	10.7
3	0.6	0	0.0	3.9	3	4.0	7.1	10	15.2
2	0.4	0	0.0	3.0	4	5.4	5.7	3	4.6
1	0.2	0	0.0	1.9	1	1.3	4.0	6	9.1
Total	41.5	40	5.1	88.4	88	11.9	129.9	130	19.8

TABLE 3. EXPECTED AND OBSERVED NUMBER OF CASES OF CHD AND OBSERVED INCIDENCE IN 12 YR OF FOLLOW-UP AT FRAMINGHAM IN WOMEN, BY AGE AT FIRST EXAMINATION AND DECILE OF RISK

Decile of risk	1863 aged 30–49		Observed 12-yr incidence (no. of cases per 100)	806 aged 50–62		Observed 12-yr incidence (no. of cases per 100)
	Number of cases			Number of cases		
	Expected	Observed		Expected	Observed	
10	24.1	19	10.2	27.0	23	28.5
9	7.4	4	2.1	14.0	11	13.6
8	4.2	8	4.3	10.6	13	16.1
7	2.8	2	1.1	8.6	12	14.9
6	2.0	1	0.5	7.4	10	12.4
5	1.4	1	0.5	6.4	4	5.0
4	1.0	1	0.5	5.4	5	6.2
3	0.6	0	0.0	4.7	5	6.2
2	0.4	1	0.5	3.8	5	6.2
1	0.2	2	1.1	2.8	2	2.5
Total	44.1	39	2.1	90.7	90	11.2

Estimated linear discriminant function coefficients and their approximate standard errors are shown in Table 5 for men and Table 6 for women. Each value indicates the amount by which the logit of risk increases for unit increase in the risk factor, where the logit of risk is $\log_e P/(1-P)$ and P is the 12-yr probability of developing CHD. The relation between logit of risk and risk is illustrated in Fig. 1. For example, the logit of risk increases by 0.708 for every 10 yr of age. Thus, if one starts with a risk of 0.05 at, say age 35, the logit is increased from $\log_e (0.05/0.95)$, or -2.944 to -2.236 at age 45. Since $P=0.0966$ when the logit of risk$=-2.236$, the absolute risk 10 yr later is 0.0966. If one started with a risk of 0.20 at, say age 45, the logit is increased by the same absolute amount or from -1.386 to 0.678. This corresponds to a risk of 0.3367 10 yr later. Since $0.0966-0.05$ does not equal $0.3367-0.20$, a constant increase in the logit of risk does not imply a constant increase in risk.

With the exception of hemoglobin, all coefficients for men are well in excess of their standard errors for the age groups combined. Although the individual age group coefficients are not as well determined they agree reasonably well with those for all ages. Each of the seven coefficients is smaller for the age group 50–62 than for the 30–39 group, thus quantifying the decreased effect of each risk factor on logit of risk at higher ages. The coefficient for women is clearly smaller than that for men for cigarettes smoked and suggestively smaller for cholesterol and relative weight. There is no clear-cut tendency in women for the effect of the risk factors to decrease with age.

The data in Tables 5 and 6 are in natural units, and comparisons between the values of the coefficients for different risk factors must take this into account. Thus, the appropriate interpretation of the male coefficients for age, 0.0708, and serum cholesterol, 0.0105, is not that age is a more "important risk factor," but that 7 mg % of cholesterol is equivalent to 1 yr of age in its effect on risk. The effect of an ECG abnormality on risk in both men and women is equivalent to a 15–20 yr difference in age ($1.046 \div 0.0708$ for men and $1.434 \div 0.0765$ for women).

TABLE 4. EXPECTED AND OBSERVED MARGINAL DISTRIBUTIONS OF CHD BY INDIVIDUAL RISK FACTORS

Variable	Men									Women					
	30–39			40–49			50–62			30–49			50–62		
	Frequency	No. of cases		Frequency	No. of cases		Frequency	No. of cases		Frequency	No. of cases		Frequency	No. of cases	
		Obs.	Exp.		Obs.	Exp.		Obs.	Exp.		Obs.	Exp.		Obs.	Exp.
Cholesterol															
< 190	212	0	2	128	13	9	123	13	14	542	6	3	58	9	6
190–219	210	5	5	198	13	18	176	33	29	552	5	6	135	12	14
220–249	199	12	11	222	28	27	174	35	34	412	10	10	218	21	23
≥ 250	168	23	24	194	34	34	183	49	52	357	18	25	395	48	48
Systolic blood pressure															
< 129	351	10	9	270	24	24	189	24	22	1009	13	12	152	7	8
130–139	214	11	10	169	20	17	128	20	22	362	3	8	130	7	9
140–149	116	6	8	150	19	19	116	26	23	202	6	7	127	11	10
≥ 150	108	13	15	153	25	28	223	60	63	290	17	18	397	65	64
Relative weight															
< 89	168	7	4	143	8	9	103	21	18	541	8	6	110	7	9
90–99	233	6	8	192	16	19	181	30	31	535	6	11	168	14	16
100–109	180	9	11	212	34	25	191	39	38	385	13	11	183	19	18
≥ 110	208	18	19	195	30	35	181	40	42	402	12	16	345	50	48
Hemoglobin															
< 120	85	1	3	80	9	9	103	26	22	430	7	7	137	11	12
120–129										697	13	18	277	26	28
130–139	194	12	10	211	26	24	177	41	39	533	14	14	283	38	32
140–149	252	10	11	234	24	27	215	37	40	203	5	6	109	15	18
≥ 150	258	17	18	217	29	28	161	26	30						
Cigarettes smoked															
0	229	2	5	252	15	18	296	50	52	966	27	30	596	65	66
1	131	4	4	119	15	13	122	29	24	568	5	9	149	16	16
2	232	15	12	194	28	23	149	27	32	255	6	4	46	8	6
3	197	19	20	177	30	34	89	24	22	74	1	1	15	1	3
ECG abnormality															
0	753	36	37	695	77	75	589	106	104	1806	36	39	733	74	71
1	36	4	5	47	11	14	67	24	26	57	3	5	73	16	20

TABLE 5. LINEAR DISCRIMINANT FUNCTION COEFFICIENTS AND CONSTANT TERM, MEN (IN NATURAL UNITS)

Risk factors	Age groups			
	Combined ages	30–39	40–49	50–62
Constant ($\hat{\alpha}$)	−10.8986	−17.6355	−13.6995	−8.6035
Age (yr)	0.0708	0.0920	0.1201	0.0724
Cholesterol (mg %)	0.0105	0.0231	0.0074	0.0091
Systolic blood pressure (mm Hg)	0.0166	0.0219	0.0086	0.0158
Relative weight	0.0138	0.0139	0.0269	0.0077
Hemoglobin (g %)	−0.0837	0.0257	−0.0109	−0.1697
Cigarettes smoked (see code)	0.3610	0.5981	0.4336	0.2723
ECG abnormality (0,1)	1.0459	1.2874	1.0525	0.7311
Standard errors of estimated coefficients				
Age	0.0083	0.0628	0.0413	0.0307
Cholesterol	0.0016	0.0040	0.0027	0.0023
Systolic blood pressure	0.0036	0.0011	0.0063	0.0043
Relative weight	0.0051	0.0126	0.0090	0.0076
Hemoglobin	0.0542	0.1361	0.0944	0.0776
Cigarettes/day	0.0587	0.1436	0.0984	0.0922
ECG abnormality	0.2706	0.7994	0.4752	0.3369

TABLE 6. LINEAR DISCRIMINANT FUNCTION COEFFICIENTS AND CONSTANT TERM, WOMEN (IN NATURAL UNITS)

Risk factors	Age groups		
	Combined ages	30–49	50–62
Constant ($\hat{\alpha}$)	−12.5933	−15.1064	−11.6930
Age (yr)	0.0765	0.1365	0.0805
Cholesterol (mg %)	0.0061	0.0173	0.0026
Systolic blood pressure (mm Hg)	0.0221	0.0098	0.0163
Relative weight	0.0053	0.0043	0.0078
Hemoglobin (g %)	0.0355	−0.0272	0.0691
Cigarettes smoked (see code)	0.0766	−0.0859	0.1869
ECG abnormality (0,1)	1.4338	1.2974	0.8957
Standard errors of estimated coefficients			
Age	0.0133	0.0339	0.0352
Cholesterol	0.0021	0.0041	0.0024
Systolic blood pressure	0.0043	0.0093	0.0041
Relative weight	0.0054	0.0100	0.0062
Hemoglobin	0.0844	0.1490	0.1088
Cigarettes/day	0.1158	0.1964	0.1692
ECG abnormality	0.4342	0.9484	0.4100

A common way of obtaining unit-free comparisons is to express each variable as a multiple of its own standard deviation; each coefficient in natural units is multiplied by its own standard deviation to obtain a coefficient in standard units. Such coefficients are shown in Table 7. Each coefficient measures the change in the logit of risk for a change of one standard deviation in a risk factor. Measured this way the most "important" single risk factor for men of all ages combined is age, but for the individual age groups, number of cigarettes smoked, serum cholesterol and systolic blood pressure appear more "important". Standard errors of the coefficients in standard units can be obtained from the standard errors in Tables 5 and 6 by multiplication by the ratio of the coefficient in standard to that in natural units. This is exact under the same conditions that the original standard errors are.

The relative unimportance of weight as a risk factor at the lowest and highest age groups, when all other risk factors are simultaneously considered, is noteworthy. This is not inconsistent with the possibility that a reduction in weight would by virtue of its effect on other risk factors, e.g. cholesterol, have important effects on the risk of CHD. Nevertheless, above age 50 for constant levels of other risk factors the coefficient for relative weight is only one-third that for cigarettes and only one-fourth that for cholesterol or blood pressure. This is an average pattern for all CHD; for specific manifestations, such as angina pectoris, the pattern of relative importance might be somewhat different.

These results describe the relations between risk and risk factors found in the first 12 yr of follow-up. It is pertinent to inquire about the extent to which this description is useful in predicting CHD experience subsequent to the period on which it is based. There have been 77 new male cases of CHD and 43 female cases

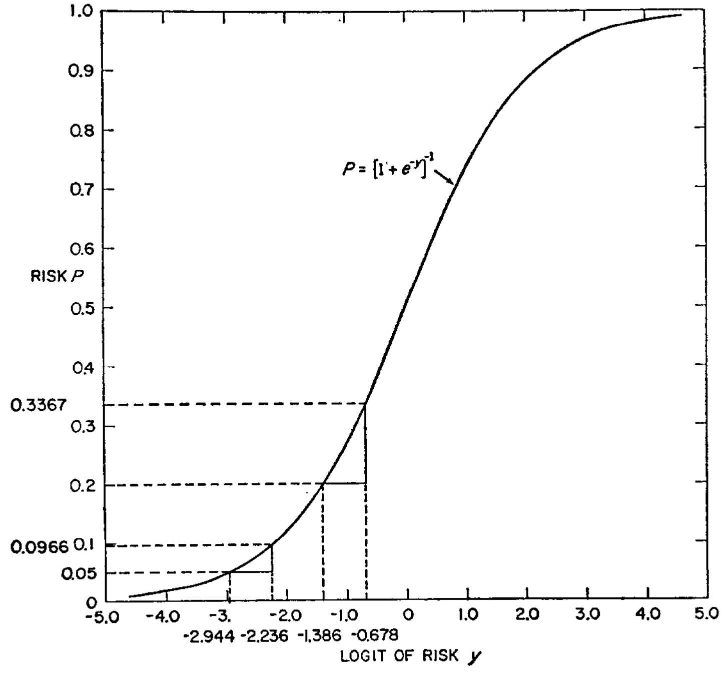

FIG. 1. Relation between logit of risk and risk.

TABLE 7. LINEAR DISCRIMINANT FUNCTION COEFFICIENTS (STANDARD UNITS)

Risk factors	Men Combined ages	Age groups 30–39	 40–49	 50–62
Age	0.5934	0.2394	0.3334	0.2370
Cholesterol	0.4444	0.9613	0.3207	0.3790
Systolic blood pressure	0.3334	0.3427	0.1669	0.3809
Relative weight	0.1890	0.1941	0.3619	0.1036
Hemoglobin	−0.1050	0.0313	−0.0134	−0.2206
Cigarettes smoked	0.4192	0.6823	0.5084	0.3004
ECG abnormality	0.2626	0.2685	0.2556	0.2197

Risk factors	Women Combined ages	Age groups 30–49	 50–62
Age	0.6259	0.7325	0.2600
Cholesterol	0.2844	0.7322	0.1207
Systolic blood pressure	0.5556	0.1947	0.4776
Relative weight	0.0975	0.0751	0.1481
Hemoglobin	0.0392	−0.0304	0.0734
Cigarettes smoked	0.0625	−0.0731	0.1262
ECG abnormality	0.3048	0.2234	0.2526

subsequent to the first 12 yr. Table 8 classifies these cases by whether they fell in the upper or lower half of the risk scale computed from the data for the first 12 yr. For all ages combined there is a two-fold difference in incidence for men and six-fold for women. This largely reflects the fact that older people are developing more disease than younger people. For the individual age groups only males aged 30–39 and females aged 30–49 show any excess incidence in the upper half of the risk scale, and even here the relative excess is well below that found for the first 12 yr of experience, as shown in Tables 1, 2 and 3. This form of presentation does not allow for the fact that the number at risk after the 12-yr follow-up was lower at the 5 highest deciles than at the 5 lowest. Adjustment for this tends to equalize the incidence of CHD in the two decile groups above age 40.

TABLE 8. NEW CASES OF CHD SUBSEQUENT TO THE 12-YR FOLLOW-UP, BY AGE, SEX AND DECILE GROUP OF RISK

	Total new cases	New cases by decile of 12-yr risk Five highest deciles	 Five lowest deciles
Men			
Combined ages	77	51	26
30–39	10	8	2
40–49	33	16	17
50–62	34	14	20
Women			
Combined ages	43	37	6
30–39	15	10	5
50–62	28	13	15

DISCUSSION

For the multiple logistic function to provide an exact description of the relation between risk and risk factors it is sufficient that the underlying distributions be multivariate normal. It is by no means necessary, however. In fact a much weaker condition is sufficient, namely that the linear compound of risk factors, $y=\hat{\alpha}+\Sigma\hat{\beta}_i x_i$, be univariate normal. The circumstances under which a linear compound of independent variables will be normal are given by the central limit theorem, [5] and of dependent variables by Bernstein's theorem. [6] These theorems are asymptotic. That univariate normality holds approximately even for the present compound of seven non-normal risk factors is indicated by the relative frequency histograms shown in Fig. 2. Although none of them can be called exactly normal, and that for women free of CHD displays a rather pronounced skew, it is clear that the linear compound is much more nearly normal than many of its components, and that this explains the perhaps unexpected agreement between observation and theoretical expectation.

It does not necessarily follow, however, that the application of multivariate normal theory to non-normal data is without consequence. *Any* normally distributed compound of risk factors can be used to obtain a risk function leading to agreement

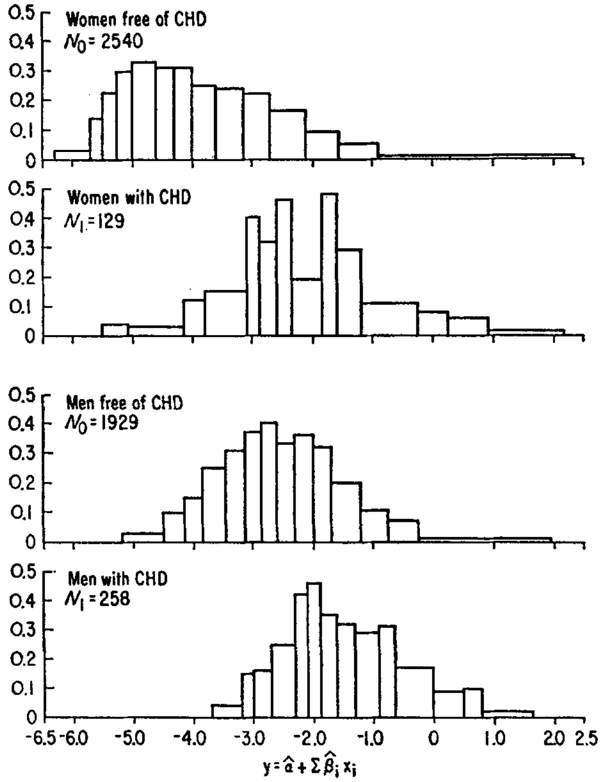

FIG. 2. Relative frequency histograms for values of the exponent in risk functions computed for combined age groups.

between observation and expectation. We naturally prefer that function which leads to the sharpest gradient of risk. It is shown in an Appendix note that if the assumption of multivariate normality is not satisfied the linear compound implied by that assumption need not lead to the sharpest gradient of risk, even though it does satisfactorily reproduce the observations.

As an example of the consequences of departures from assumptions, consider the question of interactions. If the assumptions are correct, the effect of any one risk factor is the same no matter what the levels of the other factors, i.e. there is no interaction. But it is clear from Table 5 that there are in fact interactions between age and other variables. Thus, for the age groups 30–39 and 50–62 the discriminant function coefficients for cholesterol are 0.0231 ± 0.0040 and 0.0091 ± 0.0024, so that the effect of cholesterol is clearly less striking at the older ages. It is possible to study the interactions without modifying the assumptions by proceeding as in Tables 5 and 6, where separate discriminant function coefficients have been computed for different values of the risk factor, age. A more complete analysis of first order interactions would involve repeating this analysis for different serum cholesterol groups, different systolic blood pressure levels, etc. An alternative way of studying interactions which we have not yet investigated would involve relaxing the assumption of equal variance–covariance matrices.

The more general question of how much steeper a risk gradient could be obtained by finding a better representation of the data than the multivariate normal distribution is not easily answered. But even though it is probably possible to do better, it seems clear that the multivariate normal assumption, even though untrue, leads to an analysis which (a) is reasonably consistent with the actual data and (b) provides a more informative way of assessing the contributions to risk of combinations of risk factors than other methods now in common use.

We turn now to the question of prediction. It is a common finding that discriminant functions and multiple regression equations describe the data from which they were derived better than they do new data. The formal basis for this has been studied for multiple regression [7] and for discriminant functions. [3] It is known to depend on the relation between the number of variables used, k, and the number of individuals studied, $N_0 + N_1$. It does not seem that the failure to predict outcomes subsequent to the 12 yr as well as the 12-yr experience has been described can be explained on this formal basis. First of all, there is the empirical observation that the prediction was best for the groups in which N_1 was actually smallest relative to k, namely males 30–39 and females 30–49 but failed completely for the older age groups where N_1 is considerably larger. Secondly, it is known [3, equation (7.3)] that an unbiased estimate of the standardized distance between two populations can be obtained by multiplying the standardized distance yielded by the data by the factor $(N_1 + N_0 - k - 3)/(N_1 + N_0 - 2)$ and then subtracting $k\,(1/N_0 + 1/N_1)$. But for males aged 50–62 the first factor is 0.988 and the subtractive term is 0.0096. These two corrections are entirely too small to explain the difference between the 90 to 40 ratio found in the 12-yr experience and the 14 to 20 ratio found subsequent to the 12 yr.

An explanation is more likely to be found, in our opinion, in the decrease in the gradient of risk with increasing age and with the increasing remoteness of the original measurements from the period being predicted. The correctness of this explanation can be investigated in two ways. The first is further analysis of the Framingham experience for sub-divisions of the 12-yr experience. This we plan to do. The second is application of the present risk function to the experience of other prospective studies, which we hope will be undertaken by others.

APPENDIX NOTE

Consider two functions of risk factors, y_1 and y_2. The function y_1 is obtained from the true ratio of multivariate frequency distributions. The function y_2 is a linear function obtained by assuming multivariate normality. We shall assume that the univariate distribution of y_1 is $N(\mu_{11}, \sigma_1^2)$ in the CHD population and $N(\mu_{10}, \sigma_1^2)$ in the NCHD population and that similarity y_2 is $N(\mu_{21}, \sigma_2^2)$ in the CHD population and $N(\mu_{20}, \sigma_2^2)$ in the NCHD population. Then from the argument leading to equation (1) with $k=1$

$$\mathbf{P}_1 = \left[1 + \frac{1-p}{p}\, e^{-\frac{\mu_{11}-\mu_{10}}{\sigma_1^2}\left(y_1 - \frac{\mu_{11}+\mu_{10}}{2}\right)}\right]^{-1} \tag{7}$$

$$\mathbf{P}_2 = \left[1 + \frac{1-p}{p}\, e^{-\frac{\mu_{21}-\mu_{20}}{\sigma_2^2}\left(y_2 - \frac{\mu_{21}+\mu_{20}}{2}\right)}\right]^{-1}. \tag{8}$$

The change in logit of risk per unit change in y_1/σ_1 is by equation (7) $(\mu_{11}-\mu_{10})/\sigma_1$ and per unit change in y_2/σ_2 is $(\mu_{21}-\mu_{20})/\sigma_2$ by equation (8). But it is an immediate consequence of the Neyman–Pearson lemma [8] that $(\mu_{11}-\mu_{10})/\sigma_1 > (\mu_{21}-\mu_{20})/\sigma_2$ and hence that the risk gradient based on the actual theoretical distribution is steeper that that based on the linear compound implied by the mulivariate normal assumption.

SUMMARY AND CONCLUSIONS

The dependence of the 12-yr probability of developing coronary heart disease in Framingham on 7 risk factors has been investigated using discriminant functions. Despite marked departures of the actual distributions from multivariate normality the description provided by the theoretical risk function agrees well with the actual data. This method of analysis appears, therefore, to provide a powerful method of analyzing the simultaneous effects of many risk factors on incidence, even in the absence of multivariate normality.

The combined effect of all risk factors on risk is striking. The difference in incidence between highest and lowest deciles is thirty-fold for men and seventy-fold for women. Relative differences in incidence between highest and lowest deciles of risk are most marked at the younger age groups in both men and women. The most important risk factors, aside from age itself, are cholesterol, cigarette smoking, ECG abnormality and blood pressure. Weight, while also a significant risk factor, has a considerably smaller effect than these four.

Acknowledgement—Mrs. Christine Cole rendered very skillful assistance in programming the required calculations for this analysis.

REFERENCES

1. Dawber, T. R., Kannel, W. B. and Lyell, L. P.: An approach to longitudinal studies in a community: The Framingham Study, *Ann. N.Y. Acad. Sci.* **107,** 539–556, 1963.

2. CORNFIELD, J.: Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis, *Fedn Proc.* **21,** 58–61, 1962.

3. CORNFIELD, J.: Discriminant functions, *Review of the International Statistical Institute,* **35,** 2, 1967.

4. PEARL, R.: *Introduction to Medical Biometry and Statistics.* W. B. Saunders, Philadelphia and London, 1923.

5. MUNROE, M. E.: *Theory of Probability.* McGraw Hill, New York, 1951.

6. LOEVE, M.: Fundamental limit theorems of probability theory, *Ann. math. Statist.* **21,** 321–338, 1950.

7. OLKIN, I. and PRATT, J. W.: Unbiased estimation of certain correlation coefficients, *Ann. math. Statist.* **29,** 201–211, 1958.

8. NEYMAN, J. and PEARSON, E. S.: On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. R. Soc.* London, Series A, **231,** 289–337, 1933.