Study Factors at More Than Two Levels

The preceding discussion on the analysis of retrospective data has been in terms of the test factor under study taking only two values. This framework has sufficed for discussion of the underlying statistical ideas and issues. In practice, the study factor will frequently take on more than two, perhaps many, potential values. When the number of study factor values is large, grouping can reduce them to manageable proportions.

The need to consider only a limited number of classes for the study factor stems from the fact that, when an association is anticipated, most of the significant information about the association will come from the results for the more extreme values of the study factor. While it is efficient to concentrate attention on the test factor classes expected to show the greatest differences in association with the disease, it is also profitable to consider intermediate values for the test factor to seek evidence for a consistent pattern of association. For example, in table 1, a highly significant difference between nonsmokers and women currently smoking more than 1 pack of cigarettes daily was illustrated. Inclusion of data for smokers of 1 pack or less a day showing results intermediate between the other classes would have added little, if anything, to the statistical significance of the results, and might actually lower it, if one made an over-all test of the differences among the three smoking classes. However, the observation that the intermediate smoking class does, in fact, show an intermediate relative risk contributes to an orderly pattern and increases our confidence in the conclusions suggested by the data for the remaining two classes.

For any two particular test-factor levels, the relative risk for one over the other may be calculated using only the data pertaining to those two levels or by using the results for all test levels. In the formulas previously given for R, R_1, R_2, R_3, and R_4, the difference between the two calculating procedures is simply one of setting the values of N_{1i}, N_{2i}, and $T_i = N_{1i} + N_{2i}$ in terms of number of cases and controls occurring at the two study-factor levels only, or defining them in terms of total number of cases and controls in the entire study. When total cases and controls are used in defining N_{1i}, N_{2i}, and T_i, it can be shown that for R_1, R_2, R_3, and R_4 the various relative risks will be internally consistent with each other. If the relative risk for the first level is twice that for the second level, which in turn is twice that for the third level, then the relative risk for the first level will be four times that of the third. These exact relationships do not hold for R as an estimator of relative risk, and a somewhat sophisticated extension of the formula for R would be required to secure this property.

The problem of obtaining a summary chi square when the study factor is at more than two levels is complicated by the fact that the deviations from expectation at the various study-factor levels are intercorrelated. When there are but two levels, the two deviations will have perfect negative correlation, and attention need be directed to only one of the deviations. Irrespective of the number of levels, at any one level the deviation from expectation among diseased persons will be equal, but opposite in sign, to the deviation from expectation among controls, so that attention can be confined to the deviations for diseased persons.

The problem can be stated as one of reducing a set of correlated deviations into a summary chi square. Table 2 applies this process for obtaining a summary chi square to the study of the association of epidermoid and undifferentiated pulmonary carcinoma in women and maximum cigarette-smoking rate, classified into three levels, after adjustment for age and occupation.

The general expressions for the expectations and variances of the number of cases at a particular test-factor level are given in the lower right section of table 2. Also shown is the expression for the covariance between the number of cases at two different test-factor levels. Since the total of all the deviations is zero, one would in general need the variances of, and covariances between, the number of cases at all but one of the levels. The number of covariance terms will rise sharply as the number of test levels are increased. At 3 test levels, there are 2 variance terms and 1 covariance term, while at 10 test levels, there would be 9 variances and 36 covariance terms of interest.

For the general case the burden of computation could be heavy. After all the necessary computation for the deviations, their variances and covariances, there would still remain the problem of converting these, presumably by matrix methods, into a summary chi square. Since the retrospective problem will normally involve only a limited number of test-factor levels, precise procedures will be given only for the three-level situation, and approximate procedures outlined for the general case.

The exact computation procedure for the three-level case is detailed in table 2. Lines (1), (2), and (4) show the total observed and expected frequencies and variances of the number of cases (and controls) at each of the three smoking-rate levels, after adjusting for age and occupation. These are the summary totals over each subclassification obtained by application of the formulas appearing in table 2.

Lines (5) and (6) give the chi squares corresponding to the total deviation from expectation at each of the smoking-rate levels. The chi squares in line (5) are corrected for continuity. They relate to the difference of the particular level to which they apply, from the two other levels combined. Following the usual practice of making no continuity corrections when chi squares with more than 1 degree of freedom are under consideration, line (6) shows the uncorrected chi squares.

The computing procedure of table 2 takes advantage of the fact that, since the sum of the deviations from expectation is zero, the variance of the third deviation must equal the sum of the other two variances plus twice the covariance for the first two deviations. The covariance of the first two deviations is readily obtained as illustrated and is used in calculating the summary chi square. The summary chi square is obtained as the sum of squares of two orthogonal deviates, with each

TABLE 2.—*Illustrative computation of summary chi square, when there are 3 levels for study factor. The data relate to the association of epidermoid and undifferentiated pulmonary carcinoma in women with smoking history*

	1+ Pack cigarettes daily			1 Pack or less of cigarettes daily			Occasional or nonsmokers			Total		
	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_1)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_2)	Epidermoid-undifferentiated pulmonary carcinoma	Controls	Total (ΣM_3)	Epidermoid-undifferentiated pulmonary carcinoma (ΣN_1)	Controls (ΣN_2)	Total ($=T$)
(1) Total observed frequencies	19	17	36	32	71	103	51	251	302	102	339	441
(2) Total expected frequencies, adjusted for age and occupation	9.09	26.91	36	23.76	79.24	103	69.15	232.85	302	102	339	441
(3) Total deviation from expectation (1)−(2)	$+9.91 = Y_1$			$+8.24 = Y_2$			$-18.15 = Y_3$					
(4) Variance of total observed frequencies, subject to fixed marginal totals in each age and occupation group	5.9163 = V_1			12.2900 = V_2			14.0723 = V_3					
(5) Individual corrected chi squares $(\lvert Y\rvert-0.5)^2/V$	14.97 = X^2_{1c}			4.88 = X^2_{2c}			22.15 = X^2_{3c}					
(6) Individual uncorrected chi squares Y^2/V	16.60 = X^2_1			5.53 = X^2_2			23.42 = X^2_3					
(7) Covariance (Y_1, Y_2)				−2.0670								
$(Y_3-V_1-V_2)/2$				11.70								
(8) Adjusted Y_2												
$Y_3-(7)Y_1/V_1$				11.5678								
(9) Adjusted V_2												
$V_2-(7)^2/V_1$				11.83 = X^2_2(ad.)								
(10) Adjusted X^2_2												
(8)²/(9)												
(11) Summary chi square (2 degrees of freedom)				16.60 + 11.83 = 28.43								
$X^2_1 + X^2_2$(ad.)												

For the general situation the total expected case frequency at the jth level of a test factor ...
$\Sigma N_{ij}M_{ij}/T_i$

The variance of the total case frequency is
$V_j=\sum_i \frac{N_{ij}N_{2j}M_{ij}(T_i-M_{ij})}{T_i^2(T_i-1)}$

The covariance of the total case frequencies at test levels j and k is
$-\sum_i \frac{N_{ij}N_{ik}M_{ij}M_{ik}}{T_i^2(T_i-1)}$

The index of summation, i, represents the various subclassifications into which the results are divided

For 3 test levels only, since $Y_3=-(Y_1+Y_2)$, it follows that $V_3=V_1+V_2+2$ Covariance (Y_1, Y_2)

square adjusted for its own variance. The first deviate squared is simply the uncorrected chi square at the first level in line (6)—the variance of the deviate remaining as initially calculated. The second deviate is the deviation at the second level adjusted for its correlation with the first deviation [adjusted $Y_2 = Y_2 - b_{21}Y_1$; b_{21} = covariance (Y_1, Y_2)/variance Y_1)]. The variance of the adjusted second deviate is the initial value reduced by that portion of the variation accounted for by the first deviation [Var. (adjusted Y_2) = variance Y_2 − covariance$^2(Y_1, Y_2)$/variance Y_1)].

In the present instance the summary chi square with 2 degrees of freedom is 28.43 [line (11)]. This presumably is close to the chi square with 1 degree of freedom which would have obtained had only the two most extreme smoking classes been compared. If one examines the individual uncorrected chi squares [line (6)], their total is found to be 45.55, the maximum individual figure being 23.42. *It will necessarily be true that the summary chi-square value will lie between the largest of the three chi squares and their total. At almost any reasonable probability level these limits would be sufficient to establish statistical significance without further calculation.* In our companion paper (*27*) this rule sufficed in almost all instances to separate the significant from the nonsignificant results.

Comments on Extensions to More Than Three Factors

Two procedures can be suggested for getting approximate summary chi squares, when there are a large number of levels for the test factors, without the burden of computation that the exact method would entail. Both methods calculate the approximate summary chi square as a sum of squares of approximately orthogonal standardized deviates.

In the first method one computes an uncorrected chi square with 1 degree of freedom for the difference of the first level from all the remaining levels combined (the same first step as in the illustration for the three-level case). Discarding the data from the first level, a second chi square is computed for the difference between the second test-factor level and the remaining levels combined. This is done successively up to and including the last two remaining levels. The approximate summary chi square is then the sum of the separate chi squares with the number of degrees of freedom being one less than the number of test levels.

Exactly orthogonal standardized deviates would be obtained if, in the summary analysis, as each successive total deviation from expectation were evaluated, it was adjusted for its multiple regression on the preceding deviations, and then standardized by the adjusted variance. This, of course, would no longer be a simplified approximate procedure. However, it can be shown that for a single classification, in the multiple regression of any deviation from expectation on any subset of deviations, the regression coefficients will all be equal; the multiple regression on the set of deviations will be the same as the simple regression on their sum. The equality of regression coefficients, while holding true exactly for deviations in the separate subclassifications, will hold only approximately for the total

deviations from expectation (it would hold exactly if equal numbers of individuals were observed from level to level at each subclassification). Nevertheless, this result suggests that approximately orthogonal deviates would be obtained if, in evaluating each successive total deviation, it were adjusted for the cumulative total of deviations already evaluated. Computing procedures to accomplish this can readily be devised.

Both approximate chi-square procedures just outlined, which may have merit when more than three groups are being compared simultaneously, should, in theory, yield linear combinations of independent chi squares. While testing the chi-square values obtained as though they were exact is not likely to be too inappropriate, it may be more correct to obtain a modified number of degrees of freedom, along the lines suggested by Satterthwaite (*47*) for problems involving such linear combinations. What the modified number of degrees of freedom would be has not been investigated by us, and it may prove as easy to apply the exact chi-square procedure, indicated later, as to determine the appropriate degrees of freedom for the approximate chi square.

It is of interest that a somewhat similar task of obtaining an appropriate summary chi square appears in the birth-order problems described by Halperin (*48*). There, it was necessary to compare a set of total observations (across family sizes) with a set of total expectations, one for each birth order. Halperin described a matrix-inversion procedure for reducing the set of correlated deviations into a summary chi square. In that problem it can be shown that all the regression coefficients are equal in the multiple regression of the deviation at a particular birth order on the set of deviations at all succeeding birth orders. The second approximate method described previously for the present problem could thus be used exactly for the birth-order problem, permitting simplified computation of chi square. The procedure indicated by Halperin has the advantage of generality and could be applied to the current and related problems, if one obtained all the necessary variances and covariances and inverted the resulting matrix.

References

(*1*) Snow, J.: On the mode of communication of cholera. *In* Snow on Cholera. New York, The Commonwealth Fund, 1936, pp. 1–139.

(*2*) Holmes, O. W.: The contagiousness of puerperal fever. *In* Medical Classics. Baltimore, Williams & Wilkins Co., vol. 1, 1936, pp. 211–243.

(*3*) Stern, R.: Nota sulle ricerche del dottore Tanchon intorno la frequenza del cancro. Annali Universali di Medicina 110: 484–503, 1844.

(*4*) Stocks, P., and Campbell, J. M.: Lung cancer death rates among non-smokers and pipe and cigarette smokers. Brit. M. J. 2: 923–929, 1955.

(*5*) Wynder, E. L., and Cornfield, J.: Cancer of the lung in physicians. New England J. Med. 248: 441–444, 1953.

(*6*) Lane-Claypon, J. E.: A further report on cancer of the breast, with special reference to its associated antecedent conditions. Rept. Publ. Health & M. Subj., No. 32, 1926, pp. 1–189.

(*7*) Clemmesen, J., Lockwood, K., and Nielsen, A.: Smoking habits of patients with papilloma of urinary bladder. Danish M. Bull. 5: 123–128, 1958.

(*8*) Denoix, P. R., and Schwartz, D.: Tobacco and cancer of the bladder. (Bulletin de L'Association francaise pour l'étude du Cancer.) Cancer 43: 387–393, 1956.

(*9*) Lilienfeld, A. M., Levin, M. L., and Moore, G. E.: The association of smoking with cancer of the urinary bladder in humans. A.M.A. Arch. Int. Med., 1956.

(*10*) Mustacchi, P., and Shimkin, M. B.: Cancer of the bladder and infestation with *Schistosoma hematobium*. J. Nat. Cancer Inst. 20: 825–842, 1958.

(*11*) Lilienfeld, A. M.: The relationship of cancer of the female breast to artificial menopause and marital status. Cancer 9: 927–934, 1956.

(*12*) Lilienfeld, A. M., and Levin, M. L.: Some factors involved in the incidence of breast cancer. *In* Proc. Third National Cancer Conference. Philadelphia, J. B. Lippincott Co., 1957, pp. 105–112.

(*13*) Segi, M., Fukushima, I., Fujisaku, S., Kurihara, M., Saito S., Asano, K., and Kamoi, M.: An epidemiological study on cancer in Japan. Gann Supp. 48, 1957.

(*14*) Dunham, L. J., Thomas, L. B., Edgcomb, J. H., and Stewart, H. L.: Some environmental factors and the development of uterine cancers in Israel and New York City. To be published in Acta Unio internat. contra cancrum.

(*15*) Stocks, P.: Cancer of the uterine cervix and social conditions. Brit. J. Cancer 9: 487–494, 1955.

(*16*) Wynder, E. L., Cornfield, J., Schroff, P. D., and Doraiswami, K. R.: A study of environmental factors in carcinoma of the cervix. Am. J. Obst. & Gynec. 68: 1016–1052, 1954.

(*17*) Mills, C. A., and Porter, M. M.: Tobacco smoking habits and cancer of the mouth and respiratory system. Cancer Res. 10: 539–542, 1950.

(*18*) Wynder, E. L., Bross, I. J., and Day, E.: A study of environmental factors in cancer of the larynx. Cancer 9: 86–110, 1956.

(*19*) Manning, M. D., and Carroll, B. E.: Some epidemiological aspects of leukemia in children. J. Nat. Cancer Inst. 19: 1087–1094, 1957.

(*20*) Breslow, L., Hoaglin, L., Rasmussen, G., and Abrams, H. K.: Occupations and cigarette smoking as factors in lung cancer. Am. J. Pub. Health 44: 171–181, 1954.

(*21*) Doll, R., and Hill, A. B.: A study of the aetiology of carcinoma of the lung. Brit. M. J. 2: 1271–1286, 1952.

(*22*) Levin, M. L.: Etiology of lung cancer; present status. New York J. Med. 54: 769–777, 1954.

(*23*) Sadowsky, D. A., Gilliam, A. G., and Cornfield, J.: The statistical association between smoking and carcinoma of the lung. J. Nat. Cancer Inst. 13: 1237–1258, 1953.

(*24*) Watson, W. L., and Conte, A. J.: Lung cancer and smoking. Am. J. Surg. 89: 447–456, 1955.

(*25*) Wynder, E. L., and Graham, E. A.: Tobacco smoking as possible etiologic factor in bronchiogenic carcinoma. J.A.M.A. 143: 329–336, 1950.

(*26*) Wynder, E. L., Bross, I. J., Cornfield, J., and O'Donnell, W. E.: Lung cancer in women. New England J. Med. 255: 1111–1121, 1956.

(*27*) Haenszel, W., Shimkin, M. B., and Mantel, N.: A retrospective study of lung cancer in women. J. Nat. Cancer Inst. 21: 825–842, 1958.

(*28*) Aird, I., Bentall, H. H., and Roberts, J. A. F.: A relationship between cancer of stomach and the ABO blood groups. Brit. M. J. 1: 799–801, 1953.

(*29*) Buckwalter, J. A., Wohlwend, C. B., Colter, D. C., Tidrick, R. T., and Knowler, L. A.: The association of the ABO blood groups to gastric carcinoma. Surg. Gynec. & Obst. 104: 176–179, 1957.

(*30*) Kraus, A. S., Levin, M. L., and Gerhardt, P. R.: A study of occupational associations with gastric cancer. Am. J. Pub. Health 47: 961–970, 1957.

(*31*) Cornfield, J.: A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. J. Nat. Cancer Inst. 11: 1269–1275, 1951.

(*32*) Dorn, H. F.: Some applications of biometry in the collection and evaluation of medical data. J. Chron. Dis. 1: 638–664, 1955.

(33) NEYMAN, J.: Statistics—servants of all sciences. Science 122: 3166, 1955.
(34) BERKSON, J.: Limitations of the application of fourfo'd table analysis to hospital data. Biometrics Bull. 2: 47–53, 1946.
(35) WHITE, C.: Sampling in medical research. Brit. M. J. 2: 1284–1288, 1953.
(36) GREENWOOD, M., and YULE, G. U.: On the determination of size of family and of the distribution of characters in order of birth from samples taken through members of the sibships. Roy. Stat. Soc. J. 77: 179–197, 1914.
(37) HAENSZEL, W.: Variation in incidence of and mortality from stomach cancer with particular reference to the United States. J. Nat. Cancer Inst. 21: 213–262, 1958.
(38) VIDEBAEK, A., and MOSBECH, J.: The aetiology of gastric carcinoma elucidated by a study of 302 pedigrees. Acta med. scandinav. 149: 137–159, 1954.
(39) WHELPTON, P. K., and FREEDMAN, R.: A study of the growth of American families. Am. J. Sociol. 61: 595–601, 1956.
(40) LEVIN, M. L., GOLDSTEIN, H., and GERHARDT, P. R.: Cancer and tobacco smoking. J.A.M.A. 143: 336–338, 1950.
(41) LEVIN, M. L., KRAUS, A. S., GOLDBERG, I. D., and GERHARDT, P. R.: Problems in the study of occupation and smoking in relation to lung cancer. Cancer 8: 932–936, 1955.
(42) LILIENFELD, A. M.: Possible existence of predisposing factors in the etiology of selected cancers of nonsexual sites in females. A preliminary inquiry. Cancer 9: 111–122, 1956.
(43) WINKELSTEIN, W., JR., STENCHEVER, M. A., and LILIENFELD, A. M.: Occurrence of pregnancy, abortion and artificial menopause among women with coronary artery disease: a preliminary study. J. Chron. Dis. 7: 273–286, 1958.
(44) BILLINGTON, B. P.: Gastric cancer—relationships between ABO blood-groups, site, and epidemiology. Lancet 2: 859–862, 1956.
(45) SCHWARTZ, D., and ANGUERA, G.: Une cause de biais dans certaines enquêtes médicales: le temps de séjour des malades a l'hôpital. Communication à l'Institut International de Statistique, 30ème Session. Stockholm, 1957.
(46) CORNFIELD, J.: A statistical problem arising from retrospective studies. Proc. Third Berkeley Symposium on Mathematical Statistics and Probability 4: 135–148, 1956.
(47) SATTERTHWAITE, F. E.: Synthesis of variance. Psychometrika 6: 309–316, 1941.
(48) HALPERIN, M.: The use of X^2 in testing effect of birth order. Ann. Eugenics 18: 99–106, 1953.