Extra Exercises for Categorical Data Analysis

This file contains extra exercises that did not fit in the second edition of *Categorical Data Analysis*, by Alan Agresti (John Wiley, & Sons, 2002). Instructors are welcome to use them.

Chapter 1

1. In the following examples, identify the response variable and the explanatory variables.

- **(a)** Attitude toward gun control (favor, oppose), Gender (female, male), Mother's education (high school, college).
- **(b)** Heart disease (yes, no), Blood pressure, Cholesterol level.
- **(c)** Hospital (A, B, C), Chemotherapy treatment (standard, new), Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression).
- **(d)** Race (white, nonwhite), Religion (Catholic, Jewish, Protestant, Other), Vote for President (Democrat, Republican, Other), Annual income.

2. Which measurement scale is most appropriate for attitude toward legalization of abortion (disapprove always, approve in certain cases, approve always).

3. Describe a potential study with a categorical response variable. List explanatory variables that would be important. For each variable, identify the measurement scale, and indicate whether to treat it as continuous or discrete, quantitative or qualitative.

4. In a large city, 50% of the population is black. Prospective jurors for court trials are selected from this population. For each selection of a juror, π denotes the probability that a black person is selected. A supposedly random sampling of 12 prospective jurors contains 1 black person. Using the exact binomial test, find the P-value for testing $H_0 : \pi = 0.5$ against $H_a : \pi \neq 0.5$.

5. A criminologist studies the proportion of U. S. citizens who live in a home in which firearms are available. The 1991 General Social Survey asked respondents, "Do you have in your home any guns or revolvers?" Of the respondents, 393 answered 'yes' and 583 answered 'no.' Analyze these data.

Chapter 2

6. The odds ratio between whether a boy scout (yes, no) and juvenile delinquent behavior (yes, no) is 1.0 at each fixed level of socioeconomic status (SES), but 0.5 marginally. Why is it misleading to claim that scouting leads to lower delinquency rates?

7. Table 1.1 comes from a study that investigated the effect of oral contraceptive use on the likelihood of heart attacks. The 58 subjects in the first column represent married women under 45 years of age treated for myocardial infarction in two hospital regions in England and Wales during 1968–1972. Each case was matched with three control patients in the same hospitals who were not being treated for myocardial infarction. All subjects were then asked whether they had ever used oral contraceptives. Analyze these data.

Chapter 3

8. Refer to Table 1.1. Is there evidence of an association between myocardial infarction and use of oral contraceptives? Use an inferential procedure, and interpret.

9. According to a survey by the European Commission in late 2000 of about 16,000 Europeans (Eurobarometer 54), 1000 in each country, the percent support for a common currency (the euro) was 64% in the Netherlands, 21% in the U.K., and 79% in Italy. Analyze these data.

10. For multinomial sampling in an $I \times J$ table, assuming statistical independence show that the ML estimator $\hat{\pi}_{ij} = n_{i+}n_{+j}/n^2$.

11. Refer to the example in Section 3.3.7. Explain how the range of n for which $\text{MSE}(\{\hat{\pi}_{ij}\}) < \text{MSE}(\{p_{ij}\})$ changes as δ increases.

12. For testing independence in an $I \times 2$ table, show that

$$X^2 = \frac{(\sum_i n_{i1}p_{1|i}) - n_{+1}p_{+1}}{p_{+1}p_{+2}} = \frac{\sum_i n_{i+}(p_{1|i} - p_{+1})^2}{p_{+1}p_{+2}}.$$

Fisher (1934) and Cochran (1954) attributed these formulas to A. E. Brandt and G. W. Snedecor.

Chapter 4

13. Refer to Problem 4.18. When k is unknown show the negative binomial does not have exponential dispersion form (4.14). Jørgensen (1986) argued that a more appropriate form for two-parameter discrete distributions is

$$f(y; \theta, \phi) = \exp\{y\theta - b(\theta)/a(\phi) + c(y, \phi)\}.$$

Show the negative binomial distribution has this form.

Chapter 5

14. A research study used multiple logistic regression to predict the stage of breast cancer (1 = advanced, 0 = local) at diagnosis for a sample of women. A table referring to demographic factors reported the estimated odds ratio for the effect of living arrangement (three categories) as 2.02 for spouse versus alone and 1.71 for others versus alone. Estimate the odds ratios for spouse versus others.

15. Refer to Table 1.2. To investigate the association, one can use logistic regression but interchange roles of response and explanatory variables. Fit (a) the saturated model, treating party identification as nominal, (b) the model that treats party identification as ordinal with equally-spaced scores. Interpret effects for each.

16. Use models to analyze Table 1.3 on smoking habits of students in Arizona high schools.

17. Table 1.4, reported by Clogg and Shockey (1988), comes from the 1982 General Social Survey.

- **(a)** Treating vote as the response, fit logit model (5.11) with nominal main effects. Does there seem to be a trend in the effects at the seven levels of political views?
- **(b)** Fit a logit model that uses the ordinal nature of political views. Carefully interpret parameter estimates for this model.
- **(c)** Test the fit of the models in (a) and (b), and analyze whether the model in (a) gives a significantly better fit.

18. For data from the 1998 General Social Survey on Y = whether one favors the death penalty for persons convicted of murder (1 = yes), x_1 = race (1 = white, 0 = other), and x_2 = opinion about how courts treat criminals (1 = not harsh enough, 2 = about right, 3 = too harshly), $\text{logit}[\hat{P}(Y = 1)] = 1.30 + 1.24x_1 - 0.82x_2$. Interpret the predictor effects. Find the estimate of $P(Y = 1)$ when $x_1 = 0$ and $x_2 = 3$. (Thanks to J. Carter and K. Sodec for this analysis.)

19. Table 1.5 is based on data reported by Cornfield (1962) (See Sec. 6.2.2). Subjects were classified on blood pressure, cholesterol level, and whether they developed coronary heart disease during a

follow-up period. For instance, at the lowest level of both predictors, 2 of 53 cases had heart disease. Plot sample logits or smooth the data to show the trend using cholesterol level alone to predict heart disease. Fit and interpret a logit model that describes the trend.

20. For Table 1.5, fit a logit model that simultaneously describes effects of cholesterol and blood pressure on heart disease. Interpret effects.

21. Refer to the previous exercise. Describe each predictor's effect by estimating (a) the model slope for a standard deviation change in the predictor, (b) the change in the probability of heart disease between the scores for the first and last categories, at the mean score for the other predictor.

22. Table 1.6 refers to the effect of academic achievement on self-esteem among black and white college students. Treating self-esteem as a response variable, analyze these data.

23. Table 1.7, from DiFrancisco and Critelman (1984), refers to effects of nationality and education level on whether one follows politics regularly. Analyze these data.

Chapter 6

24. Table 1.8, given by Chin et al. (1961), classifies 174 poliomyelitis cases in Des Moines, Iowa by age of subject, paralytic status, and by whether the subject had been injected with the Salk vaccine.

- **(a)** Test the hypothesis that severity is independent of whether vaccinated, controlling for age.
- **(b)** Use another procedure for testing this hypothesis, and compare results to those obtained in (a).
- **(c)** Estimate the common odds ratio between severity and whether vaccinated, using (i) the Mantel–Haenszel estimator, (ii) the unconditional ML estimator. Interpret.
- **(d)** If you have appropriate software, obtain the conditional ML estimator. Compare results to those in (c).

25. Refer to Table 6.13. An alternative scenario has P(Nonroutine Care) values equal to (0.50, 0.45, 0.40, 0.45, 0.25, 0.15). Calculate the noncentrality for the likelihood-ratio model-based test of NO partial association. Find the approximate powers for sample sizes 500 and 1000, for a 0.05-level test. How large a sample is needed to achieve power 0.90?

26. Refer to the example in Sec. 6.5.5. Suppose instead we used a linear logit model with equally-spaced scores for categories of N. Calculate the power for the test of conditional independence, with $df = 1$. Find the approximate power for sample size 1000. How do these compare to powers for the additive logit model? Explain the discrepancy.

Chapter 7

27. Refer to the model discussed for Table 7.1 in Sec. 7.1.2. Show that for small alligators in Lake Hancock, the estimated probabilities of primary food choice (fish, invertebrates, reptile, bird, other) are (0.54, 0.09, 0.05, 0.07, 0.25).

28. Table 1.9 shows results of logit modeling of occupational attainment in the U.S. using S = years of schooling, E = labor market experience (calculated as age $-$ years of schooling $-$ 5), R = race (1 = white, 0 = black), and G = gender (1 = male, 0 = female). The categories of occupational attainment are professional (P), white collar (W), blue collar (B), craft (C), and menial (M).

- **(a)** Obtain parameter estimates for modeling $\log(\pi_W/\pi_B)$, and interpret.
- **(b)** Explain why the estimates in the Race column indicate that occupational groups are ordered (W, C, P, B, M) in terms of relative number of white workers, controlling for the other factors.

29. Table 1.10, analyzed by Sugiura and Otake (1974) and Landis et al. (1978), shows the relationship between the deaths from leukemia during 1950–1970 and estimated radiation dosage from atomic bombing at the end of World War II. Subjects are stratified according to their age at time of

bombing. Using the midpoint scores (0, 5, 30, 75, 150, 300) for the levels of dosage, Table 1.11 shows results of CMH tests between dosage and survival status.

(a) Interpret the CMH statistics. Why are two of the statistics identical?

(b) Interpret the effect by fitting a logit model with a linear effect of dose on the probability of death from leukemia.

30. Table 1.12 classifies 1398 children on tonsil size and on whether they are carriers of the virus *Streptococcus pyogenes*. Analyze these data.

31. Table 1.13 is from Bock and Jones (1968), one of the first books to present sophisticated models for categorical data. Using an ordinal scale, subjects indicated their preference for black olives. The sample consisted of independent samples of Armed Forces personnel selected from six combinations of urbanization (urban, rural) and location (NE, MW, SW). Analyze these data.

32. A model for Table 7.5 uses baseline-category logits. What are advantages and disadvantages of this approach compared to cumulative logit modeling?

33. For an $I \times J$ contingency table, show that statistical independence is equivalent to

$$\text{logit}[P(Y \leq j \mid X = i)] = \alpha_j, \quad i = 1, \ldots, I, \; j = 1, \ldots, J-1.$$

Chapter 8

34. Construct a loglinear model for a $2 \times 2 \times K$ table that has homogeneous XY association except for a different association in the first stratum. Derive the likelihood equations and show the first stratum has a perfect fit. Show residual $df = K - 2$.

35. Opposition to the legal availability of abortion is stronger among the religious than the nonreligious, and stronger among those with conservative sexual attitudes than those with more permissive attitudes. Does this imply that the religious are more likely than the nonreligious to have conservative sexual attitudes? Use sample tables in your answer.

36. For a three-way table with binary response Y, give the equivalent loglinear and logit models for which (a) Y is jointly independent of X and Z, (b) no interaction exists between X and Z in their effects on Y.

37. For a 3-way table, the general model between X and Y at level k of Z is

$$\log \mu_{ijk} = \lambda(k) + \lambda_i^X(k) + \lambda_j^Y(k) + \lambda_{ij}^{XY}(k).$$

Show that parameters in model (XYZ) satisfy $\lambda = [\sum \lambda(k)]/K$, $\lambda_i^X = [\sum_k \lambda_i^X(k)/K$, $\lambda_{ij}^{XY} = [\sum_k \lambda_{ij}^{XY}(k)]/K$, $\lambda_k^Z = \lambda(k) - \lambda$, $\lambda_{ik}^{XZ} = \lambda_i^X(k) - \lambda_i^X$, $\lambda_{ijk}^{XYZ} = \lambda_{ij}^{XY}(k) - \lambda_{ij}^{XY}$.

38. Following Problem 8.16(e), define parameters such that

$$\lambda_1^X = \lambda_1^Y = \lambda_1^Z = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = \cdots = \lambda_{ij1}^{XYZ} = 0.$$

Show the two-factor terms are log odds ratios using the cell at the first level of each variable, and a three-factor term is a log of ratios of odds ratios. Illustrate for a $2 \times 2 \times 2$ table, showing $\lambda_{22}^{XY} = \log[\theta_{11(1)}]$ and $\lambda_{222}^{XYZ} = \log[\theta_{11(1)}/\theta_{11(2)}]$. Explain how to set up dummy variables so model fitting yields estimates having these constraints.

39. In a $2 \times 2 \times 2$ table, show $\theta_{11(1)} = \theta_{11(2)}$ implies $\theta_{1(1)1} = \theta_{1(2)1}$ and $\theta_{(1)11} = \theta_{(I)11}$.

40. When $\{n_i\}$ has a multinomial distribution with probabilities $\{\pi_i = \mu_i/(\sum_a \mu_a)\}$, show that the part of the log likelihood involving both the data and parameters is $\sum_i n_i \log(\mu_i)$, the same as for Poisson sampling.

41. Consider loglinear model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta I(ab)$$

where $I(ab) = 1$ in cell (a, b) and equals 0 otherwise.

(a) Find the likelihood equations, and note $\hat{\mu}_{ab} = n_{ab}$.

(b) Show that residual $df = IJ - I - J$.

(c) State an IPF algorithm for finding fitted values that satisfy the model. (Hint: Replace the entry in cell (a, b) by 0. Apply IPF for the independence model, with a starting value of 0 in cell (a, b), to obtain other fitted values.)

42. Show that ML estimates for Poisson loglinear models are identical to those obtained after splitting the sample into several independent multinomial samples. Specifically, suppose a set of Poisson means $\{\mu_{ij}\}$ satisfy

$$\log \mu_{ij} = \alpha_i + \mathbf{x}_{ij}\boldsymbol{\beta}.$$

Decompose the Poisson log likelihood so part refers to the row totals and part refers to the effect of conditioning on those totals.

43. Explain how IPF can standardize a three-way table so that each marginal two-way table has uniform cell frequencies.

Chapter 9

44. Table 1.14 reports the frequency of all reported game-related concussions for players on 49 college football teams, between 1975 and 1982. The total time at risk for these data was 216,690 athlete-games. Suppose the total was identical for offense and defense, the total for blocking was six times that for tackling, and the total for rushing plays was 2.2 times that for passing plays. Find the total time at risk per cell and the sample rates of concussion. Which activity has greatest sample rate? Use loglinear models to summarize these rates.

45. For logit model $\text{logit}[P(Y = 1|X = i, Z = k)] = \alpha + \beta i + \beta_k^Z$, $i = 0, 1$, is $\hat{\beta}$ the same as with model $\text{logit}[P(Y = 1|X = i)] = \alpha + \beta i$ for the table collapsed over Z? Explain.

46. When $I = 2$, explain why the row effects model is equivalent to the linear-by-linear association model.

47. Model (9.11) treats Z as nominal and is not comparable to (9.22), neither being a special case of the other. However, when Z is ordinal and we replace β_k in (9.11) by βk, show that model *is* a special case of (9.22).

48. Express the RC model as a probability function for cell probabilities $\{\pi_{ij}\}$. Demonstrate the similarity of this function to the bivariate normal density having unit standard deviations. Show that β in the RC model corresponds to $\rho/(1-\rho^2)$ for the bivariate normal density, where ρ is the correlation. See Goodman (1981a,b, 1985) and Becker (1989b).

49. Refer to Table 9.5. Use software to fit the row effects model. Check the results in Sec. 9.5.3 for these data. For each pair of rows a and b, construct a 95% confidence interval for $\exp(\mu_b - \mu_a)$, and interpret.

50. Use models discussed in this chapter to analyze Table 2.13.

51. Show that $G^2[(Y, XZ) \mid (XY, XZ)]$ is identical to $G^2[(X, Y)]$ for the XY marginal table.

52. Suppose $\{y_i\}$ are independent Poisson random variables with means $\{\mu_i\}, i = 1, \cdots, N$.

(a) Let $z_i = (y_i - \mu_i)/(\mu_i)^{1/2}$. Show $sum_i \text{Var}(z_i) = N$.

(b) Let $e_i = (y_i - \hat{\mu}_i)/(\hat{\mu}_i)^{1/2}$, where $\{\hat{\mu}_i\}$ are fitted values for a model $\{\mu_i\}$ satisfy. Give a heuristic argument that $\sum_i \text{Var}(e_i)$ asymptotically equals df for testing the model fit.

53. For three dimensions, state a generalization of the RC model for the XY association that is a special case of (XY, XZ, YZ) and contains the homogeneous $L \times L$ model as a special case.

54. Suppose we fit a multiplicative model M to a table, except for certain structural-zero cells where $\mu_a = 0$. The model is $\mu_i = E_i M$, where $E_a = 0$ for those cells and all other $E_i = 1$. Explain how to fit

this using the model-with-offset representation. (In practice, E_a must be a very small constant, such as 10^{-8}, so that its logarithm exists. Some software allows the user to fit this model by assigning zero *weights* to certain cells.)

55. Suppose $n_{11+} = 0$. Do finite ML estimates exist of all parameters for loglinear model (XY, XZ, YZ)? Explain.

Chapter 10

56. Table 1.15 refers to a sample of juveniles convicted of a felony in Florida in 1987. Matched pairs were formed using criteria such as age and the number of prior offenses. For each pair, one subject was handled in the juvenile court and the other was transferred to the adult court. The response of interest was whether the juvenile was rearrested by the end of 1988. Compare the true proportions rearrested for the adult and juvenile court assignments. Interpret.

57. Refer to Table 10.5. For the symmetry model, find and interpret the standardized Pearson residual for the cell in row 1 and column 4.

58. Table 1.16 reports subjects' religious affiliation in 1991 and when their age was 16, for categories (1) Protestant, (2) Catholic, (3) Jewish, (4) None or Other.

(a) The symmetry model has $G^2 = 32.2$ ($df = 6$). Interpret, and use residuals to analyze transition patterns.

(b) The quasi-symmetry model has $G^2 = 2.0$ ($df = 3$). Interpret.

(c) Test marginal homogeneity. Show the small P-value mainly reflects the large sample size, a small decrease in the proportion classified Catholic, and an increase in the proportion classified None or Other.

(d) Fit the quasi-independence model, and interpret.

59. Table 1.17, from Breslow (1982), compares 80 esophageal cancer patients with 80 matched control subjects. The response is the number of beverages reported drunk at "burning hot" temperatures. In analyzing whether cases tended to drink more beverages burning hot than did controls, use X^2 to check model fit, since most cell counts are small.

(a) Fit the symmetry model, and explain how $n_{ab} \leq n_{ba}$ whenever $a < b$ contributes to the lack of fit ($X^2 = 15.1, df = 6$).

(b) Show the quasi-symmetry model has $X^2 = 2.5$ ($df = 3$).

(c) Fit an ordinal model, interpret the effect, and use it to test marginal homogeneity.

60. Treating the response as continuous, use a normal paired-difference procedure to compare cases and controls on the mean number of beverages drunk burning hot. Compare the results to an ordinal test of marginal homogeneity. List assumptions on which each procedure is based.

61. Table 1.18 describes unaided distance vision for a sample of women. Analyze these data.

62. Table 1.19 reports results when subjects were asked their opinion on early teens (age 14-16) having sex and on a man and a woman having sex before marriage. The outcome categories are 1 = always wrong, 2 = almost always wrong, 3 = wrong only sometime, 4 = not wrong at all. Analyze these data.

63. A sample of married couples indicate their candidate preference in a presidential election. Table 1.20 reports the results. Analyze these data.

64. Refer to Table 8.19. The two-way table relating health (as rows) and law enforcement (as columns) has cell counts, by row, (292, 117, 25 / 72, 60, 6 / 14, 12, 9). Analyze these data.

65. Refer to the previous problem. Using the ordinal quasi-symmetry model or a proportional odds model, estimate the marginal effect. Interpret.

66. A wildlife biologist wants to estimate the number of alligators in Lake Lochloosa, Florida. She catches n_{1+} alligators, tags them, and releases them back into the lake. Two weeks later, she catches a second sample of n_{+1} alligators, of which n_{11} were also in the first sample. She cannot observe

n_{22}, the number not caught either time, and hence the population size N. If whether an alligator is captured in the second sample is independent of whether it was captured in the first sample, argue that a reasonable estimator is $\hat{N} = n_{1+}n_{+1}/n_{11}$ (Sekar and Deming 1949).

67. Suppose quasi independence holds. Let $E_{ab} = 1$ for $a \neq b$ and $E_{ab} = \epsilon > 0$ for $a = b$. Show the amended cell counts $\{y^*_{ab}\}$ with zeroes on the main diagonal have expected values $\mu_{ab} = E_{ab}\alpha_a\beta_b$, or $\mu_{ab}/E_{ab} = \alpha_a\beta_b$ for all a and b, as $\epsilon \downarrow 0$. (Thus, fitting a model with structural zeroes corresponds to fitting a general loglinear model of form given in Section 9.7.5, and giving zero weight to certain cells.)

68. Show the ML fit of loglinear model (W, XYZ) for the $6 \times I^3$ table with entries $\{y^*_{1ijk} = y_{ijk}, y^*_{2ijk} = y_{ikj}, y^*_{3ijk} = y_{jik}, y^*_{4ijk} = y_{jki}, y^*_{5ijk} = y_{kij}, y^*_{6ijk} = y_{kji}\}$ has entries $\{\hat{\mu}^*_{hijk}\}$ related to the ML fit $\{\hat{\mu}_{ijk}\}$ for the complete symmetry model by $\{\hat{\mu}_{ijk} = \hat{\mu}^*_{1ijk}\}$.

69. Construct a loglinear model for an I^3 table having the following quasi-independence interpretation: Conditional on the event that the three responses are completely different, the responses are mutually independent. Find the residual *df*.

70. Show the quasi-association model (10.29) is a special case of quasi symmetry.

71. Refer to the previous problem. Specify a logit model for the probability of rearrest, using court assignment as a predictor. Explain how to estimate and interpret the effect of court assignment.

72. Table 1.21 refers to a case-control study investigating a possible relationship between cataracts and the use of head coverings during the summer. Each case reporting to a clinic for care for a cataract was matched with a control of the same sex and similar age not having a cataract. The row and column categories refer to the frequency with which the subject used head coverings. Analyze these data.

73. Table 1.22 refers to matches among five men tennis players during 1989-1990. Analyze these data.

74. Table 1.23 reports respondents' current region of residence and region of residence at age 16. Fit the quasi-independence model. Describe lack of fit. What can you say about the numbers of people who moved from the Northeast to the South and from the Midwest to the West, relative to what quasi independence predicts?

75. Table 1.24 relates mother's education to father's education for a sample of eminent black Americans (defined as persons having biographical sketch in the publication, *Who's Who Among Black Americans*). Analyze these data.

76. For a longitudinal binary matched-pairs study, data are available for some subjects at both times, for others only at the first time or the second time. Of n subjects observed both times, let p_{ab} denote the proportion having outcome a at time 1 and b at time 2. Of n_t subjects observed only at time t, let q_t denote the proportion making the first outcome. Treat n, n_1, and n_2 as fixed, and let $a = n/(n + n_1)$, $b = n/(n + n_2)$, and $\mathbf{p}' = (p_{11}, p_{12}, p_{21}, q_1, q_2)$.

(a) Treating $\{np_{ab}\}$ as a multinomial sample and treating n_1q_1 and n_2q_2 as independent binomials, show $\widehat{Cov}(\mathbf{p}) = \mathbf{S}$ in Table 1.26.

(b) Of all subjects observed at time t, let P_t denote the proportion having the first outcome. Show that $P_t = \mathbf{d}_t'\mathbf{p}$, with $\mathbf{d}_1' = (a, a, 0, 1-a, 0)$ and $\mathbf{d}_2' = (b, 0, b, 0, 1-b)$. Thus, $\widehat{Var}(P_i) = \mathbf{d}_i'\mathbf{S}\mathbf{d}_i$, and $\widehat{Var}(P_1 - P_2) = (\mathbf{d}_1 - \mathbf{d}_2)'\mathbf{S}(\mathbf{d}_1 - \mathbf{d}_2)$.

(c) Table 1.25 is from a study at the Univ. of Florida about drug use in an elderly population. Subjects were asked whether they took tranquilizers. Some were interviewed in 1979, some in 1985, and others both times. Find P_1 and P_2. Assuming $E(p_{1+}) = E(q_1) = \pi_1$ and $E(p_{1+}) = E(q_2) = \pi_2$, construct a 95% confidence interval for $\pi_1 - \pi_2$. (This approach is reasonable when data are *missing completely at random*.)

77. For an $I \times I$ table $\{n_{ab}\}$, construct the $I \times I \times 2$ tables $\{n_{ab1} = n_{ab},\ n_{ab2} = n_{ba}\}$ and $\{\mu_{ab1} = \mu_{ab},\ \mu_{ab2} = \mu_{ba}\}$.

(a) If quasi symmetry holds for $\{\mu_{ab}\}$, show $\theta_{ab(1)}/\theta_{ab(2)} = 1$ for $\{\mu_{abc}\}$, for all a and b.

(b) Show that likelihood equations for the quasi-symmetry model for $\{\mu_{ab}\}$ correspond to like-

lihood equations for loglinear model (XY, XZ, YZ) for $\{\mu_{abc}\}$.

(c) Show that $\{\hat{\mu}_{ab}\}$ for the quasi-symmetry model are identical to $\{\hat{\mu}_{ab1}\}$ for model (XY, XZ, YZ) fitted to $\{n_{abc}\}$ (Bishop et al. 1975, pp. 289–290).

(d) Show that model $\log \mu_{abc} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}$ for $\{\mu_{abc}\}$ corresponds to symmetry for $\{\mu_{ab}\}$.

78. Refer to Table 10.12. Let μ_{abc} denote the expected frequency for outcome c ($c = 1$, win; $c = 2$, lose) for home team a when it plays away team b. Find the loglinear model for $\{\mu_{abc}\}$ that is equivalent to logit model (10.32). Check that residual $df = I(I-2)$, as in the logit model. Show that $\lambda_1^Z - \lambda_2^Z$ in the loglinear model represents the home-team advantage.

79. Consider complete symmetry for $T = 3$ matched observations. Show that

$$\hat{\mu}_{abc} = (n_{abc} + n_{acb} + n_{bac} + n_{bca} + n_{cab} + n_{cba})/6.$$

How does this simplify for $\hat{\mu}_{aaa}$, $a = 1, \ldots, I$?

Chapter 11

80. Use GEE with the cumulative logit model for marginal distributions to model the esophageal cancer data in Table 1.17. Interpret the marginal effect, and show how to use the model to test marginal homogeneity.

Chapter 12

81. Use a GLMM to analyze the esophageal cancer data in Table 1.17.

82. Use a logistic-normal model to analyze the data in Larsen et al. (2000).

Chapter 13

83. Refer to the data in Crowder (1978). Analyze these data using at least two different approaches for overdispersed binary data. Compare results and interpret.

84. In problem 13.28, explain why the more general model in which effects vary by treatment sequence corresponds to a separate quasi-symmetry fit for each treatment sequence. Show the likelihood-ratio statistic comparing the two models is 12.8 ($df = 10$). Show that adding two period effect terms to the simpler model decreases the likelihood-ratio statistic by only 0.7 ($df = 2$).

Chapter 15

85. Show that the loglinear model (8.11) of homogenous association for an $I \times J \times K$ table is specified by $(I-1)(J-1)(K-1)$ constraint equations, such as

$$\log[(\pi_{ijk}\pi_{i+1,j+1,k})/(\pi_{i+1,jk}\pi_{i,j+1,k})]$$
$$-\log[(\pi_{ij,k+1}\pi_{i+1,j+1,k+1})/\pi_{i+1,j,k+1}\pi_{i,j+1,k+1})] = 0.$$

For it, are WLS estimates the same as minimum modified chi-squared estimates?

Table 1.1.

	Myocardial Infarction	
Oral Contraceptive Practice	Yes	No
Used	23	34
Never used	35	132
Total	58	166

Reprinted with permission from Mann et al., *British J. Med.* **2**: 241-245 (1975).

Table 1.2.

	Party Identification		
Race	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

Table 1.3.

	Student Smokes	Student Does Not Smoke
Both parents smoke	400	1380
One parent smokes	416	1823
Neither parent smokes	188	1168

By permission, S.V. Zagona, *Studies and Issues in Smoking Behavior*, Tucson: The University of Arizona Press, Copyright 1967.

Table 1.4.

		1980 Presidential Vote	
Race	Political Views[a]	Reagan	Carter or other
White	1	1	12
	2	13	57
	3	44	71
	4	155	146
	5	92	61
	6	100	41
	7	18	8
Nonwhite	1	0	6
	2	0	16
	3	2	23
	4	1	31
	5	0	8
	6	2	7
	7	0	4

Source: 1982 General Social Survey.
[a]Political views range from 1 = extremely liberal to 7 = extremely conservative.

Table 1.5.

Blood Pressure	Serum Cholesterol (mg/100 ml)						
	<200	200–209	210–219	220–244	245–259	260–284	>284
<177	2/53	0/21	0/15	0/20	0/14	1/22	0/11
117–126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
127–136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
137–146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
147–156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
157–166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
167–186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
>186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source: Reprinted with permission from Cornfield (1962).

Table 1.6.

Gender	Cumulative GPA	Black		White	
		High Self-Esteem	Low Self-Esteem	High Self-Esteem	Low Self-Esteem
Males	High	15	9	17	10
	Low	26	17	22	26
Females	High	13	22	22	32
	Low	24	23	3	17

Source: Reprinted with permission of the Helen Dwight Reid Educational Foundation from D. H. Demo and K. D. Parker, *J. Social Psych.*, **127**: 345–355 (1987). Published by Heldref Publications, copyright ©1987.

Table 1.7.

Follow Politics Regularly	USSR		USA		UK		Italy		Mexico	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Primary	94	84	227	112	356	144	166	526	447	430
Secondary	318	120	371	71	256	76	142	103	78	25
College	473	72	180	8	22	2	47	7	22	2

Source: Reprinted with permission from DiFrancesco and Critelman (1984).

Table 1.8.

Age	Salk Vaccine	Paralysis	
		No	Yes
0–4	Yes	20	14
	No	10	24
5–9	Yes	15	12
	No	3	15
10–14	Yes	3	2
	No	3	2
15–19	Yes	7	4
	No	1	6
20–39	Yes	12	3
	No	7	5
40+	Yes	1	0
	No	3	2

Source: Reprinted with permission, based on data from Chin et al. (1961).

Table 1.9.

Logit	Intercept	Schooling	Experience	Race	Gender
$\log(\pi_B/\pi_M)$	1.056	−0.124	−0.015	0.700	1.252
$\log(\pi_C/\pi_M)$	−3.769	−0.001	−0.008	1.458	3.112
$\log(\pi_W/\pi_M)$	−3.305	0.225	0.003	1.762	−0.523
$\log(\pi_P/\pi_M)$	−5.959	0.429	0.008	0.976	0.656

Source: Reprinted with permission from P. Schmidt and R. P. Strauss, *Intern. Econ. Rev.*, 16, pp. 471-486 (1975).

Table 1.10. Deaths from Leukemia Observed at Atomic Bomb Casualty Commission (1950–1970)

Age	Survival Status[a]	Dose (rad)					
		Not in City	0–9	10–49	50–99	100–199	200+
0–9	LD	0	7	3	1	4	11
	NLD	5015	10752	2989	694	418	387
10–19	LD	5	4	6	1	3	6
	NLD	5973	11811	2620	771	792	820
20–34	LD	2	8	3	1	3	7
	NLD	5669	10828	2798	797	596	624
35–49	LD	3	19	4	2	1	10
	NLD	6158	12645	3566	972	694	608
50+	LD	3	7	3	2	2	6
	NLD	3695	9053	2415	655	393	289

Source: Reprinted from Sugiura and Otake (1974), by courtesy of Marcel Dekker, Inc.
[a]LD = death from leukemia, NLD = nondeath from leukemia.

Table 1.11.

```
---------------------------------------------------------------
            Summary Statistics for survival by dose
                      Controlling for age

   Cochran-Mantel-Haenszel Statistics (Based on Table Scores)
 Statistic    Alternative Hypothesis    DF       Value      Prob
 ---------------------------------------------------------------
     1        Nonzero Correlation        1    397.0091    <.0001
     2        Row Mean Scores Differ     1    397.0091    <.0001
     3        General Association        5    427.0519    <.0001
---------------------------------------------------------------
```

Table 1.12.

	Tonsil Size		
	Not Enlarged	Enlarged	Greatly Enlarged
Noncarriers	497	560	269
Carriers	19	29	24

Source: From M. C. Holmes and R. E. O. Williams, *J. Hyg.*, **52**, 165-179 (1954). Reprinted with permission from Cambridge University Press.

Table 1.13. Preference for Black Olives, by Urbanization and Location[a]

		Preference					
Urbanization	Location	A	B	C	D	E	F
Urban	MW	20	15	12	17	16	28
	NE	18	17	18	18	6	25
	SW	12	9	23	21	19	30
Rural	MW	30	22	21	17	8	12
	NE	23	18	20	18	10	15
	SW	11	9	26	19	17	24

Source: Reprinted with permission from Holden-Day (Bock and Jones 1968, p. 244).
[a] *Key*: A, Dislike extremely; B, dislike very much or moderately; C, dislike slightly or neither like or dislike; D, like slightly; E, like moderately; F, like very much or like extremely.

Table 1.14.

		Activity	
Team	Situation	Tackle	Block
Offense	Rushing	125	129
	Passing	85	31
Defense	Rushing	216	61
	Passing	62	16

Source: Reprinted with permission from Buckley, W. E. (1988), *Amer. J. Sports Med.*, **16**: 51–56.

Table 1.15.

Adult Court	Juvenile Court	
	Rearrest	No Rearrest
Rearrest	158	515
No Rearrest	290	1134

Source: Based on a study at the Univ. of Florida by D. Bishop, C. Frazier, L. Lanza-Kaduce, and L. Winner. Thanks to Dr. Larry Winner for showing me these data.

Table 1.16.

Affiliation at Age 16	Religious Affiliation Now			
	1	2	3	4
1	863	30	1	52
2	50	320	0	33
3	1	1	28	1
4	27	8	0	33

Source: 1991 General Social Survey

Table 1.17.

	Control			
Case	0	1	2	3
0	31	5	5	0
1	12	1	0	0
2	14	1	2	1
3	6	1	1	0

Source: Reprinted with permission from the Biometric Society; data from Breslow (1982).

Table 1.18.

	Left Eye Grade			
Right Eye Grade	Best	Second	Third	Worst
Best	1520	266	124	66
Second	234	1512	432	78
Third	117	362	1772	205
Worst	36	82	179	492

Source: Reprinted with permission from the Biometrika Trustees (Stuart 1955).

Table 1.19.

Teen Sex	Premarital Sex			
	1	2	3	4
1	141	34	72	109
2	4	5	23	38
3	1	0	9	23
4	0	0	1	15

Source: 1989 General Social Survey

Table 1.20.

Husband's Preference	Wife's Preference	
	Democrat	Republican
Democrat	200	25
Republican	75	200

Table 1.21.

Cataract Case	Control			
	Always or Almost Always	Frequently	Occasionally	Never
Always or almost always	29	3	3	4
Frequently	5	0	1	1
Occasionally	9	0	2	0
Never	7	3	1	0

Source: J. M. Dolezal *et al., Amer. J. Epidemiol., 129*: 559-568 (1989).

Table 1.22.

Winner	Loser				
	Edberg	Lendl	Agassi	Sampras	Becker
Edberg	–	5	3	2	4
Lendl	4	–	3	1	2
Agassi	2	0	–	1	3
Sampras	0	1	2	–	0
Becker	6	4	2	1	– –

Source: Reprinted with permission from World Tennis magazine.

Table 1.23.

Residence at Age 16	Residence in 1991			
	Northeast	Midwest	South	West
Northeast	245	16	40	20
Midwest	12	333	31	51
South	14	31	321	16
West	3	51	12	309

Source: 1991 General Social Survey

Table 1.24.

Mother's Education	Father's Education			
	8th Grade or less	Part High School	High School	College
8th Grade or less	81	3	9	11
Part High School	14	8	9	6
High School	43	7	43	18
College	21	6	24	87

Source: Reprinted with permission from E. J. Mullins and P. Sites, *Amer. Sociol. Rev.*, **49**: 672-685 (1984).

Table 1.25.

	1985		
Take Drug	Yes	No	Not Sampled
1979			
Yes	175	190	230
No	139	1518	982
Not Sampled	64	595	

Source: Mary Moore.

Table 1.26.

$$\mathbf{S} = \begin{bmatrix} \frac{p_{11}(1-p_{11})}{n} & \frac{-p_{11}p_{12}}{n} & \frac{-p_{11}p_{21}}{n} & 0 & 0 \\ \frac{-p_{11}p_{12}}{n} & \frac{p_{12}(1-p_{12})}{n} & \frac{-p_{12}p_{21}}{n} & 0 & 0 \\ \frac{-p_{11}p_{21}}{n} & \frac{-p_{12}p_{21}}{n} & \frac{p_{21}(1-p_{21})}{n} & 0 & 0 \\ 0 & 0 & 0 & \frac{q_1(1-q_1)}{n_1} & 0 \\ 0 & 0 & 0 & 0 & \frac{q_2(1-q_2)}{n_2} \end{bmatrix}$$