INSTRUCTOR'S MANUAL
CATEGORICAL DATA ANALYSIS
Solutions to Selected Problems

Alan Agresti

This manual contains solutions and hints to solutions for many of the exercises in *Categorical Data Analysis*, second edition, by Alan Agresti (John Wiley, & Sons, 2002). It is prepared for instructors who use this book as a course text. It should not be distributed elsewhere without permission of the author or of John Wiley & Sons.

Please report any errors in these solutions to the author (Dept. of Statistics, University of Florida, Gainesville, Florida 32611-8545, e-mail AA@STAT.UFL.EDU), so they can be corrected in future revisions of the manual.

Chapter 1

1. a. nominal, b. ordinal, c. interval, d. nominal, e. ordinal, f. nominal, g. ordinal.

2. a. Binomial, $n = 100$, $\pi = .25$.
b. The mean is $n\pi = 25$ and the standard deviation is $\sqrt{n\pi(1-\pi)} = 4.33$. Yes, 50 correct responses would be surprising, since 50 is $z = (50 - 25)/4.33 = 5.8$ standard deviations above the mean of a distribution that is approximately normal.
c. Multinomial, $n = 100$, $\pi_1 = \cdots = \pi_4 = 0.25$.
d. $E(n_j) = n\pi_j = 25$, $\text{Var}(n_j) = n\pi_j(1-\pi_j) = 100(.25)(.75) = 18.75$, $\text{Cov}(n_j, n_k) = -n\pi_j\pi_k = -100(.25)(.25) = -6.25$, $\text{Corr}(n_j, n_k) = -6.25/\sqrt{(18.75)(18.75)} = -.333$.

3. π varies from batch to batch, so the counts come from a mixture of binomials rather than a single $bin(n, \pi)$. $\text{Var}(Y) = E[\text{Var}(Y \mid \pi)] + \text{Var}[E(Y \mid \pi)] > E[\text{Var}(Y \mid \pi)] = E[n\pi(1-\pi)]$.

4. a. The geometric probability, $(5/6)^6$.
b. Note that $Y = y$ when there are $y-1$ successes and then a failure. The probability of a sequence of independent events is the product of the probabilities of the separate events. Thus, $p(y) = (5/6)^{y-1}(1/6)$, $y = 1, 2, \ldots$.

5. $\hat{\pi} = 842/1824 = .462$, so $z = (.462 - .5)/\sqrt{.5(.5)/1824} = -3.28$, for which $P = .001$ for H_a: $\pi \neq .5$. The 95% Wald CI is $.462 \pm 1.96\sqrt{.462(.538)/1824} = .462 \pm .023$, or (.439, .485). The 95% score CI is also (.439, .485).

6.a. Expected frequency = 12.5 for each category. For the count of 0, $0 \log(0/12.5) = 0$, so result follows.
b. Score statistic $z = (0 - .5)/\sqrt{.5(.5)/25} = -5.0$, so $z^2 = 25.0$.
c. $z = (0 - .5)/\sqrt{0(1.0)/25} = -\infty$.

7. a. $\ell(\pi) = \pi^{20}$, so $\hat{\pi} = 1.0$.
b. Wald statistic $z = (1.0 - .5)/\sqrt{1.0(0)/20} = \infty$. Wald CI is $1.0 \pm 1.96\sqrt{1.0(0)/20} = 1.0 \pm 0.0$, or (1.0, 1.0).
c. $z = (1.0 - .5)/\sqrt{.5(.5)/20} = 4.47$, $P < .0001$. Score CI is (0.839, 1.000).
d. Test statistic $2(20)\log(20/10) = 27.7, df = 1$. From problem 1.25a, the CI is $(\exp(1.96^2/40), 1) = (0.908, 1.0)$.
e. P-value $= 2(.5)^{20} = .00000191$. Clopper-Pearson CI is (0.832, 1.000). CI using Blaker method is (0.840, 1.000).

f. $n = 1.96^2(.9)(.1)/(.05)^2 = 138$.

8. The chi-squared goodness-of-fit test of the null hypothesis that the binomial proportions equal (.75, .25) has expected frequencies (827.25, 275.75), and $X^2 = 3.46$ based on $df = 1$. The P-value is 0.063, giving moderate evidence against the null.

9. The sample mean is 0.61. Fitted probabilities for the truncated distribution are 0.543, 0.332, 0.102, 0.021, 0.003. The estimated expected frequencies are 108.5, 66.4, 20.3, 4.1, and 0.6, and the Pearson $X^2 = 0.7$ with $df = 3$ (0.3 with $df = 2$ if one truncates at 3 and above).

11. $\text{Var}(\hat{\pi}) = \pi(1-\pi)/n$ decreases as π moves toward 0 or 1 from 0.5.

12. a. $\text{Var}(Y) = n\pi(1-\pi)$, binomial.

b. $\text{Var}(Y) = \sum \text{Var}(Y_i) + 2\sum_{i<j} \text{Cov}(Y_i, Y_j) = n\pi(1-\pi) + 2\rho\pi(1-\pi)\binom{n}{2} > n\pi(1-\pi)$.

c. $\text{Var}(Y) = E[\text{Var}(Y|\pi)] + \text{Var}[E(Y|\pi)] = E[n\pi(1-\pi)] + \text{Var}(n\pi) = n\bar{\pi} - nE(\pi^2) + [n^2E(\pi^2) - n^2\bar{\pi}^2] = n\rho + (n^2-n)[E(\pi^2) - \rho^2] - n\rho^2 = n\rho(1-\rho) + (n^2-n)Var(\pi) > n\rho(1-\rho)$.

d. Conditionally, Y is the sum of non-identical Bernoulli trials, so is not binomial. Conditionally, the probability of a particular sequence is $\prod \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. Since the responses are independent, the unconditional probability of that sequence is $\prod (E\pi_i)^{y_i}(1-E(\pi_i))^{1-y_i}$, which corresponds to a sequence of identical, independent trials.

13. This is the binomial probability of y successes and $k-1$ failures in $y+k-1$ trials times the probability of a failure at the next trial.

14. Using results shown in Sec. 14.1.4, $\text{Cov}(n_j, n_k)/\sqrt{\text{Var}(n_j)\text{Var}(n_k)} = -n\pi_j\pi_k/\sqrt{n\pi_j(1-\pi_j)n\pi_k(1-\pi_k)}$. When $c = 2$, $\pi_1 = 1 - \pi_2$ and correlation simplifies to -1.

15. For binomial, $m(t) = E(e^{tY}) = \sum_y \binom{n}{y}(\pi e^t)^y(1-\pi)^{n-y} = (1 - \pi + \pi e^t)^n$, so $m'(0) = n\pi$.

16. $t_o = -2\log[(\text{prob. under } H_0)/(\text{prob. under } H_a)]$, so $(\text{prob. under } H_0)/(\text{prob. under } H_a) = \exp(-t_o/2)$.

17. a. $\ell(\mu) = \exp(-n\mu)\mu^{\sum y_i}$, so $L(\mu) = -n\mu + (\sum y_i)\log(\mu)$ and $L'(\mu) = -n + (\sum y_i)/\mu = 0$ yields $\hat{\mu} = (\sum y_i)/n$.
b. (i) $z_w = (\bar{y} - \mu_0)\sqrt{\bar{y}/n}$, (ii) $z_s = (\bar{y} - \mu_0)\sqrt{\mu_0/n}$, (iii) $-2[-n\mu_0 + (\sum y_i)\log(\mu_0) + n\bar{y} - (\sum y_i)\log(\bar{y})]$.
c. (i) $\bar{y} \pm z_{\alpha/2}\sqrt{\bar{y}/n}$, (ii) all μ_0 such that $|z_s| \le z_{\alpha/2}$, (iii) all μ_0 such that LR statistic $\le \chi^2_1(\alpha)$.

18. Conditional on $n = y_1 + y_2$, y_1 has a $bin(n, \pi)$ distribution with $\pi = \mu_1/(\mu_1 + \mu_2)$, which is .5 under H_0. The large sample score test uses $z = (y_1/n - .5)/\sqrt{.5(.5)/n}$. If (ℓ, u) denotes a CI for π (e.g., the score CI), then the CI for $\pi/(1-\pi) = \mu_1/\mu_2$ is $[\ell/(1-\ell), u/(1-u)]$.

19. a. No outcome can give $P \le .05$, and hence one never rejects H_0.
b. When $T = 2$, mid P-value = .04 and one rejects H_0. Thus, P(Type I error) = P($T = 2$) = .08.
c. P-values of the two tests are .04 and .02; P(Type I error) = P($T = 2$) = .04 with both tests.
d. P(Type I error) = E[P(Type I error | T)] = (5/8)(.08) = .05.

20. a. all π_0 such that $|\hat{\pi} - \pi_0|/\sqrt{\hat{\pi}(1-\hat{\pi})/n} \le z_{\alpha/2}$, which yields $\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$.
b. all π_0 such that $|\hat{\pi} - \pi_0|/\sqrt{\pi_0(1-\pi_0)/n} \le z_{\alpha/2}$. Squaring and solving quadratic equation for π_0 gives result.

21. a. With the binomial test the smallest possible P-value, from $y = 0$ or $y = 5$, is $2(1/2)^5 = 1/16$. Since this exceeds .05, it is impossible to reject H_0, and thus P(Type I error) = 0. With the large-sample score test, $y = 0$ and $y = 5$ are the only outcomes to give $P \le .05$ (e.g., with $y = 5$, $z = (1.0 - .5)/\sqrt{.5(.5)/5} = 2.24$ and $P = .025$). Thus, for that test, P(Type I error) = $P(Y = 0) + P(Y = 5) = 1/16$.
b. For every possible outcome the Clopper-Pearson CI contains .5. e.g., when $y = 5$, the CI is (.478, 1.0), since for $\pi_0 = .478$ the binomial probability of $y = 5$ is $.478^5 = .025$.

22. P(CI contains π) $\le P(1 \le Y \le n-1) = 1 - P(Y = n) - P(Y = 0) = 1 - \pi^n - (1-\pi)^n$. This converges to 0 as $\pi \to 0$ or as $\pi \to n$.

23. For π just below $.18/n$, $P(\text{CI contains } \pi) = P(Y = 0) = (1-\pi)^n = (1-.18/n)^n \approx \exp(-.18) = 0.84$.

24. $g(\pi) = \pi(1-\pi)/n^*$ is a concave function of π, so if π is random, $g(E\pi) \geq Eg(\pi)$ by Jensen's inequality. Now $\tilde{\pi}$ is the expected value of π for a distribution putting probability $n/(n+z^2_{\alpha/2})$ at $\hat{\pi}$ and probability $z^2_{\alpha/2}/(n+z^2_{\alpha/2})$ at $1/2$.

25. a. The likelihood-ratio (LR) CI is the set of π_0 for testing H_0: $\pi = \pi_0$ such that LR statistic $= -2\log[(1-\pi_0)^n/(1-\hat{\pi})^n] \leq z^2_{\alpha/2}$, with $\hat{\pi} = 0.0$. Solving for π_0, $n\log(1-\pi_0) \geq -z^2_{\alpha/2}/2$, or $(1-\pi_0) \geq \exp(-z^2_{\alpha/2}/2n)$, or $\pi_0 \leq 1-\exp(-z^2_{\alpha/2}/2n)$. Using $\exp(x) = 1+x+...$ for small x, the upper bound is roughly $1-(1-z^2_{.025}/2n) = z^2_{.025}/2n = 1.96^2/2n \approx 2^2/2n = 2/n$.
b. Solve for $(0-\pi)/\sqrt{\pi(1-\pi)/n} = -z_{\alpha/2}$.
c. Upper endpoint is solution to $\pi_0^0(1-\pi_0)^n = \alpha/2$, or $(1-\pi_0) = (\alpha/2)^{1/n}$, or $\pi_0 = 1-(\alpha/2)^{1/n}$. Using the expansion $\exp(x) \approx 1+x$ for x close to 0, $(\alpha/2)^{1/n} = \exp\{\log[(\alpha/2)^{1/n}]\} \approx 1+\log[(\alpha/2)^{1/n}]$, so the upper endpoint is $\approx 1-\{1+\log[(\alpha/2)^{1/n}]\} = -\log(\alpha/2)^{1/n} = -\log(.025)/n = 3.69/n$.
d. The mid P-value when $y = 0$ is half the probability of that outcome, so the upper bound for this CI sets $(1/2)\pi_0^0(1-\pi_0)^n = \alpha/2$, or $\pi_0 = 1-\alpha^{1/n}$.

26. The *cdf* is $F(y) = 1-\pi^{y+1}$. Equating it to $\alpha/2$ and solving for π yields the upper endpoint, $(1-\alpha/2)^{1/(y+1)}$. Setting $P(Y \geq y) = 1-F(y-1) = \pi^y = \alpha/2$ yields the lower endpoint, $(\alpha/2)^{1/y}$. For $y = 0$ the upper bound is $(1-\alpha/2)$, and the upper bound is larger than this for $y > 1$. Thus, all π between 0 and $1-\alpha/2$ *never* fall above a confidence interval, and thus they can be excluded only by falling below the interval.

28. If we form the P-value using the right tail, then mid P-value $= \pi_j/2 + \pi_{j+1} + \cdots$. Thus, E(mid P-value) $= \sum_j \pi_j(\pi_j/2 + \pi_{j+1} + \cdots) = (\sum_j \pi_j)^2/2 = 1/2$.

29. The right-tail mid P-value equals $P(T > t_o) + (1/2)p(t_o) = 1-P(T \leq t_o)+(1/2)p(t_o) = 1-F_{mid}(t_o)$.

30. a. The kernel of the log likelihood is $L(\theta) = n_1\log\theta^2 + n_2\log[2\theta(1-\theta)] + n_3\log(1-\theta)^2$. Take $\partial L/\partial\theta = 2n_1/\theta + n_2/\theta - n_2/(1-\theta) - 2n_3/(1-\theta) = 0$ and solve for θ.
b. Find the expectation using $E(n_1) = n\theta^2$, etc. Then, the asymptotic variance is the inverse information $= \theta(1-\theta)/2n$, and thus the estimated $SE = \sqrt{\hat{\theta}(1-\hat{\theta})/2n}$.
c. The estimated expected counts are $[n\hat{\theta}^2,\ 2n\hat{\theta}(1-\hat{\theta}),\ n(1-\hat{\theta})^2]$. Compare these to the observed counts $(n_1,\ n_2,\ n_3)$ using X^2 or G^2, with $df = (3-1)-1 = 1$, since 1 parameter is estimated.

31. Since $\partial^2 L/\partial\pi^2 = -(2n_{11}/\pi^2) - n_{12}/\pi^2 - n_{12}/(1-\pi)^2 - n_{22}/(1-\pi)^2$,
the information is its expected value, which is
$-2n\pi^2/\pi^2 - n\pi(1-\pi)/\pi^2 - n\pi(1-\pi)/(1-\pi)^2 - n(1-\pi)/(1-\pi)^2$,
which simplifies to $-n(1+\pi)/\pi(1-\pi)$. The asymptotic standard error is the square root of the inverse information, or $\sqrt{\pi(1-\pi)/n(1+\pi)}$.

33. c. Let $\hat{\pi} = n_1/n$, and $(1-\hat{\pi}) = n_2/n$, and denote the null probabilities in the two categories by π_0 and $(1-\pi_0)$. Then, $X^2 = (n_1 - n\pi_0)^2/n\pi_0 + (n_2 - n(1-\pi_2))^2/n(1-\pi_0)$
$= n[(\hat{\pi}-\pi_0)^2(1-\pi_0) + ((1-\hat{\pi})-(1-\pi_0))^2\pi_0]/\pi_0(1-\pi_0)$,
which equals $(\hat{\pi}-\pi_0)^2/[\pi_0(1-\pi_0)/n] = z_S^2$.

34. Let X be a random variable that equals $\pi_{j0}/\hat{\pi}_j$ with probability $\hat{\pi}_j$. By Jensen's inequality, since the negative log function is convex, $E(-\log X) \geq -\log(EX)$. Hence, $E(-\log X) = \sum\hat{\pi}_j\log(\hat{\pi}_j/p_{j0}) \geq -\log[\sum\hat{\pi}_j(\pi_{j0}/\hat{\pi}_j)] = -\log(\sum\pi_{j0}) = -\log(1) = 0$. Thus $G^2 = 2nE(-\log X) \geq 0$.

35. If Y_1 is χ^2 with $df = \nu_1$ and if Y_2 is independent χ^2 with $df = \nu_2$, then the *mgf* of Y_1+Y_2 is the product of the *mgf*s, which is $m(t) = (1-2t)^{-(\nu_1+\nu_2)/2}$, which is the *mgf* of a χ^2 with $df = \nu_1+\nu_2$.

36. a. By the Bonferroni inequality, if the probability of an event (an error, the CI not containing the difference) is α/c, then the probability of the union of the c events (i.e., at least one error) is no greater than the sum of these probabilities, or α.
b. Follows again from the Bonferroni inequality.

Chapter 2

1. $P(-|C) = 1/4$ and $P(+|\bar{C}) = 2/3$. Sensitivity $= P(+|C) = 1 - P(-|C) = 3/4$. Specificity $= P(-|\bar{C}) = 1 - P(+|\bar{C}) = 1/3$.

2. $\theta = (.8/.2)/(.2/.8) = 16$.

3. The odds ratio is $\hat{\theta} = 7.965$; the relative risk of fatality for 'none' is 7.897 times that for 'seat belt'; difference of proportions = .0085. The proportion of fatal injuries is close to zero for each row, so the odds ratio is similar to the relative risk.

4. a. Relative risk.
b. (i) $\pi_1 = .55\pi_2$, so $\pi_1/\pi_2 = .55$.
(ii) $1/.55 = 1.82$.

5. Relative risks are 3.3, 5.4, 11.5, 34.7; e.g., 1994 probability of gun-related death in U.S. was 34.7 times that in England and Wales.

6. Prob. of winning = odds/(odds+1), which equals 11/21 for Italy and 3/13 for Bulgaria. These probabilities do not sum to 1.

7. a. .0012, 10.78; relative risk, since difference of proportions makes it appear there is no association.
b. (.001304/.998696)/(.000121/.999879) = 10.79; this happens when the proportion in the first category is close to zero.

8. a. The quoted interpretation is that of the relative risk. Should substitute *odds* for *probability*.
b. For females, probability $= 2.9/(1 + 2.9) = .744$. Odds for males $= 2.9/11.4 = .25$, so probability $= .25/(1 + .25) = .20$.

9. X given Y. Applying Bayes theorem, $P(V = w|M = w) = P(M = w|V = w)P(V = w)/[P(M = w|V = w)P(V = w) + P(M = w|V = b)P(V = b) = .83$ P(V=w)/[.83 P(V=w) + .06 P(V=b)]. We need to know the relative numbers of victims who were white and black. Odds ratio = (.94/.06)/(.17/.83) = 76.5.

10. a. (.847/.153)/(.906/.094) = .574.
b. This is interpretation for relative risk, not the odds ratio. The actual relative risk = .847/.906 = .935; i.e., 60% should have been 93.5%.

11. a. Relative risk: Lung cancer, 14.00; Heart disease, 1.62. (Cigarette smoking seems more highly associated with lung cancer)
Difference of proportions: Lung cancer, .00130; Heart disease, .00256. (Cigarette smoking seems more highly associated with heart disease)
Odds ratio: Lung cancer, 14.02; Heart disease, 1.62. *e.g.*, the odds of dying from lung cancer for smokers are estimated to be 14.02 times those for nonsmokers. (Note similarity to relative risks.)
b. Difference of proportions describes excess deaths due to smoking. That is, if N = no. smokers in population, we predict there would be $.00130N$ fewer deaths per year from lung cancer if they had never smoked, and $.00256N$ fewer deaths per year from heart disease. Thus elimination of cigarette smoking would have biggest impact on deaths due to heart disease.

12. Marginal odds ratio = 1.84, but most conditional odds ratios are close to 1.0 except in Department A where odds ratio = .35. Note that males tend to apply in greater numbers to Departments A and B, in which admissions rates are relatively high, and females tend to aply in greater numbers to Departments C, D, E, F, in which admissions rates are relatively low. This results in the marginal association whereby the odds of admission for males are 84% higher than those for females.

14. a. 0.18 for males and 0.32 for females; e.g., for male children, the odds that a white was a murder victim were 0.18 times the odds that a nonwhite was a murder victim.
b. 0.21.

15. The age distribution is relatively higher in Maine.

16. Kentucky: Counts are (31, 360 / 7, 50) when victim was white and (0, 18 / 2, 106) when victim was black. Conditional odds ratios are 0.62 and 0.0, whereas marginal odds ratio is 1.42. Simpson's paradox occurs. Whites tend to kill whites and blacks tend to kill blacks, and killing a white is more likely to result in the death penalty.

17. The odds of carcinoma for the various smoking levels satisfy:
(Odds for high smokers)/(Odds for low smokers) $= \frac{\text{(Odds for high smokers)/(Odds for nonsmokers)}}{\text{(Odds for low smokers)/(Odds for nonsmokers)}} = 26.1/11.7 = 2.2$.

19. gamma = .360 ($C = 1508$, $D = 709$); of the untied pairs, the difference between the proportion of concordant pairs and the proportion of discordant pairs equals .360. There is a tendency for wife's rating to be higher when husband's rating is higher.

21. a. Let "pos" denote positive diagnosis, "dis" denote subject has disease.

$$P(dis|pos) = \frac{P(pos|dis)P(dis)}{P(pos|dis)P(dis) + P(pos|no\ dis)P(no\ dis)}$$

b. .95(.005)/[.95(.005) + .05(.995)] = .087.

		Test		
		+	−	Total
Reality	+	.00475	.00025	.005
	−	.04975	.94525	.995

Nearly all (99.5%) subjects do not have AIDS. The 5% errors for them swamp (in frequency) the 95% correct cases for subjects truly having AIDS. The odds ratio = 361; i.e., the odds of a positive test result are 361 times higher for those having AIDS than for those not having AIDS.

23. a. The numerator is the extra proportion that got the disease above and beyond what the proportion would be if no one had been exposed (which is $P(D \mid \bar{E})$).
b. Use Bayes Theorem and result that $RR = P(D \mid E)/P(D \mid \bar{E})$.

24. a. For instance, if first row becomes first column and second row becomes second column, the table entries become

$$\begin{matrix} n_{11} & n_{21} \\ n_{12} & n_{22} \end{matrix}$$

The odds ratio is the same as before. The difference of proportions and relative risk are only invariant to multiplication of cell counts within rows by a constant.

25. Suppose $\pi_1 > \pi_2$. Then, $1 - \pi_1 < 1 - \pi_2$, and $\theta = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)] > \pi_1/\pi_2 > 1$. If $\pi_1 < \pi_2$, then $1 - \pi_1 > 1 - \pi_2$, and $\theta = [\pi_1/(1-\pi_1)]/[\pi_2/(1-\pi_2)] < \pi_1/\pi_2 < 1$.

26. Let E_1 = classification in level 1 of Y (column var.), E_2 = classification in level 1 of X (row 1), E_3 = classification in stratum 1 of Z. Then $P(E_1|E_2) > P(E_1|\bar{E}_2)$ is equivalent to $\pi_1 > \pi_2$ in the XY marginal table, or $\theta > 1$. The other two orderings given in the problem refer to corresponding conditional probabilities in the partial tables relating X and Y at the two levels of Z. The orderings imply the odds ratio is less than 1 in each partial table. It follows that Simpson's paradox says that the marginal XY odds ratio can exceed 1 even if both the conditional XY odds ratios at the two levels of Z are less than 1.

27. This simply states that ordinary independence for a two-way table holds in each partial table.

28.

$$\theta_{11(1)}/\theta_{11(2)} = \frac{[(\mu_{111}\mu_{222})/(\mu_{121}\mu_{211})]}{[(\mu_{112}\mu_{222})/(\mu_{122}\mu_{212})]} = \frac{[(\mu_{111}\mu_{212})/(\mu_{211}\mu_{112})]}{[(\mu_{121}\mu_{222})/(\mu_{221}\mu_{122})]} = \theta_{1(1)1}/\theta_{1(2)1}.$$

Hence, $\theta_{11(1)} = \theta_{11(2)}$ iff $\theta_{1(1)1} = \theta_{1(2)}$, and equality of the XY conditional odds ratios is equivalent to equality of the XZ conditional odds ratios, and likewise equality of the YZ conditional odds ratios. An argument for defining "no three-factor interaction" in three-way tables as equality of conditional odds ratios is that the odds ratio exhibits this symmetry, unlike measures of association that are not functions of the odds ratio.

29. Yes, this would be an occurrence of Simpson's paradox. One could display the data as a $2 \times 2 \times K$ table, where rows = (Smith, Jones), columns = (hit, out) response for each time at bat, layers = (year 1, ..., year K). This could happen if Jones tends to have relatively more observations (i.e., "at bats") for years in which his average is high.

30. a.

.05 .10	.20 .15
.15 .20	.10 .05
$Z = 1$	$Z = 2$

b.

.15 .10	.10 .15
.10 .15	.15 .10
$Z = 1$	$Z = 2$

33. This condition is equivalent to the conditional distributions of Y in the first $I-1$ rows being identical to the one in row I. Equality of the I conditional distributions is equivalent to independence.

34. a. $\log \theta_j \geq 0$ is equivalent to $F_{j|2} \leq F_{j|1}$.

35. Use an argument similar to that in Sec. 1.2.5. Since Y_{i+} is sum of independent Poissons, it is Poisson. In the denominator for the calculation of the conditional probability, the distribution of $\{Y_{i+}\}$ is a product of Poissons with means $\{\mu_{i+}\}$. The multinomial distributions are obtained by identifying $\pi_{j|i}$ with μ_{ij}/μ_{i+}.

36. a. This follows since $\Pi_c = 2\pi_{11}\pi_{22}$, $\Pi_d = 2\pi_{12}\pi_{21}$.
b. This follows since it is a difference of proportions.
c. $Q = 1$ iff $\pi_{12}\pi_{21} = 0$, so if either $\pi_{12} = 0$ or $\pi_{21} = 0$.
d. Divide numerator and denominator of Q by $\pi_{12}\pi_{21}$.

37. a. Note that ties on X and Y are counted both in T_X and T_Y, and so T_{XY} must be subtracted. $T_X = \sum_i n_{i+}(n_{i+} - 1)/2$, $T_Y = \sum_j n_{+j}(n_{+j} - 1)/2$, $T_{XY} = \sum_i \sum_j n_{ij}(n_{ij} - 1)/2$.
c. The denominator is the number of pairs that are untied on X.

39. If in each row the maximum probability falls in the same column, say column 1, then $E[V(Y \mid X)] = \sum_i \pi_{i+}(1 - \pi_{1|i}) = 1 - \pi_{+1} = 1 - \max\{\pi_{+j}\}$, so $\lambda = 0$. Since the maximum being the same in each row does not imply independence, $\lambda = 0$ can occur even when the variables are not independent.

Chapter 3

4. a. $G^2 = 90.3$, $df = 2$; very strong evidence of association ($P < .0001$).
c. $G^2 = 7.2$ for comparing races on (Democrat, Independent) choice, and $G^2 = 83.2$ for comparing races on (Dem. + Indep., Republican) choice; extremely strong evidence that whites are more likely than blacks to be Republicans. (To get independent components, combine the two groups compared in the first analysis and compare them to the other group in the second analysis.)

5. The values $X^2 = 7.01$ and $G^2 = 7.00$ ($df = 2$) show considerable evidence against the hypothesis of independence (P-value = .03). The standardized Pearson residuals show that the number of female Democrats and Male Republicans is significantly greater than expected under independence, and the

number of female Republicans and Male Democrats is significantly less than expected under independence. e.g., there were 279 female Democrats, the estimated expected frequency under independence is 261.4, and the difference between the observed count and fitted value is 2.23 standard errors.

6. Use the percentages to reconstruct frequencies in the 3×2 contingency table, and compute X^2 (df = 2) and find the P-value.

7. $G^2 = 27.59$, $df = 2$, so $P < .001$. For first two columns, $G^2 = 2.22$ ($df = 1$), for those columns combined and compared to column three, $G^2 = 25.37$ ($df = 1$). The main evidence of association relates to whether one suffered a heart attack.

8. It is not necessary for each row to have the same number of observations. This adjustment procedure is improper, and the test should have been conducted on the original observations. 9. b. Compare rows 1 and 2 ($G^2 = .76$, $df = 1$, no evidence of difference), rows 3 and 4 ($G^2 = .02$, $df = 1$, no evidence of difference), and the 3×2 table consisting of rows 1 and 2 combined, rows 3 and 4 combined, and row 5 ($G^2 = 95.74$, $df = 2$, strong evidences of differences).

10. The X^2 statistic has $df = 9$ and is designed for the general alternative, ignoring the ordering of rows and columns. The M^2 statistic uses the ordering through the scores and has $df = 1$. It focuses the statistic on a narrower alternative and yields a smaller P-value.

11.a. $X^2 = 8.9$, $df = 6$, $P = 0.18$; test treats variables as nominal and ignores the information on the ordering.
b. Residuals suggest tendency for aspirations to be higher when family income is higher.
c. Ordinal test gives $M^2 = 4.75$, $df = 1$, $P = .03$, and much stronger evidence of an association.

13. a. It is plausible that control of cancer is independent of treatment used. (i) P-value is hypergeometric probability $P(n_{11} = 21 \text{ or } 22 \text{ or } 23) = .3808$, (ii) P-value is sum of probabilities that are no greater than the probability (.2755) of the observed table.
b. The asymptotic CI (.31, 14.15) uses the delta method formula (3.1) for the SE. The 'exact' CI (.21, 27.55) is the Cornfield tail-method interval that guarantees a coverage probability of at least .95.
c. .3808 - .5(.2755) = .243. With this type of P-value, the actual error probability tends to be closer to the nominal value, but it does not guarantee that it is no greater than the nominal value.

14. Table has entries (7,8) in row 1 and (0,15) in row 2. $P = \binom{15}{7}\binom{15}{0}/\binom{30}{7} = 15!23!/8!30! =$ .003; strong evidence of better results for treatment than control.

15. a. entire real line, b. (.618, ∞),

17. $P = 0.164$, $P = 0.0035$ takes into account the positive linear trend information in the sample.

18. a. No. P-value is .22 for Pearson test and .208 for the one-sided Fisher's exact test P-value, and .245 for the two-sided Fisher's exact test P-value based on summing all probabilities no greater than observed.
b. Large-sample CI for odds ratio is (.51, 15.37), and exact based on Cornfield approach is (.39, 31.04).

20. No. Since convergence to normality is faster on the log scale, large-sample intervals use it. The CI for the log odds ratio is centered at the sample log odds ratio. After exponentiating, the odds ratio is not at the center of the resulting CI, since the exponential function is nonlinear.

21. For proportions π and $1-\pi$ in the two categories for a given sample, the contribution to the asymptotic variance is $[1/n\pi + 1/n(1-\pi)]$. The derivative of this with respect to π is $1/n(1-\pi)^2 - 1/n\pi^2$, which is less than 0 for $\pi < 0.5$ and greater than 0 for $\pi > 0.5$. Thus, the minimum is with proportions (.5, .5) in the two categories.

22. Note the delta method was used to derive the standard error of the sample logit in Sec. 3.1.6. If the endpoints of the CI for the logit are (ℓ, u), then the corresponding endpoints of the CI for π are its inverse, $[e^\ell/(1+e^\ell), e^u/(1+e^u)]$.