Needed for Q8.23

TABLE 8.1: Party Identification (ID) and Gender, for GSS Data

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	573	516	422	1511
Males	386	475	399	1260
Total	959	991	821	2771

it has two rows and three columns. The row totals and the column totals are called the ***marginal distributions***. The sample marginal distribution for party identification, for instance, is the set of marginal frequencies (959, 991, 821). ■

Percentage Comparisons

Constructing a contingency table from a data file is the first step in investigating an association between two categorical variables. To study how party identification depends on gender, we convert the frequencies to percentages within each row, as Table 8.2 shows. For example, a proportion of 573/1511 = 0.38, or 38% in percentage terms, identify themselves as Democrat. The percentage of males who identify themselves as Democrat equals 31% (386 out of 1260). It seems that females are more likely than males to identify as Democrats.

TABLE 8.2: Party Identification and Gender: Percentages Computed within Rows of Table 8.1

	Party Identification				
Gender	Democrat	Independent	Republican	Total	*n*
Females	38%	34%	28%	100%	1511
Males	31%	38%	32%	101%	1260

The two sets of percentages for females and males are called the ***conditional distributions*** on party identification. They refer to the sample data distribution of party ID, *conditional* on gender. The females' conditional distribution on party ID is the set of percentages (38, 34, 28) for (Democrat, Independent, Republican). The percentages sum to 100 in each row, except possibly for rounding. Figure 8.1 portrays graphically the two conditional distributions.

In a similar way, we could compute conditional distributions on gender for each party ID. The first column would indicate that 60% of the Democrats are females and 40% are males. In practice, it is standard to form the conditional distribution for the response variable, within categories of the explanatory variable. In this example, party ID is a response variable, so Table 8.2 reports percentages within rows, which tells us the percentage of (Democrats, Independents, Republicans) for each gender.

Another way to report percentages provides a single set for all cells in the table, using the total sample size as the base. To illustrate, in Table 8.1, of the 2771 subjects, 573 or 21% fall in the cell (Female, Democrat), 386 or 14% fall in the cell (Male, Democrat), and so forth. This percentage distribution is called the sample ***joint distribution***. It is useful for comparing relative frequencies of occurrences for combinations of variable levels. When we distinguish between response and explanatory variables, though, conditional distributions are more informative than the joint distribution.

8.14. Table 8.24 shows SPSS analyses with the 2004 GSS, for variables party ID and race.

(a) Report the expected frequency for the first cell, and show how SPSS obtained it.

(b) Test the hypothesis of independence between party ID and race. Report the test statistic and *P*-value and interpret.

(c) Use the standardized residuals (labelled ADJ RES here for "adjusted residuals" to describe the pattern of association.

TABLE 8.24

	Count / Exp Val / Adj Res	PARTY_ID			
		democr	indep	repub	Row Total
RACE	black	250	106	17	373
		129.1	129.0	114.9	
		14.2	-2.7	-11.9	
	white	640	783	1775	2198
		760.9	760.0	677.1	
		-14.2	2.7	11.9	
	Column Total	890	889	792	2571

Chi-Square	Value	DF	Significance
Pearson	234.73	2	0.0000

8.15. For a 2 × 4 cross classification of gender and religiosity (very, moderately, slightly, not at all) for recent GSS data, the standardized residual was 3.2 for females who are very religious, −3.2 for males who are very religious, −3.5 for females who are not at all religious, and 3.5 for males who are not at all religious. All other standardized residuals fell between −1.1 and 1.1. Interpret.

8.16. Table 8.25 is from the 2006 General Social Survey, cross-classifying happiness (HAPPY) and marital status (MARITAL).

TABLE 8.25

Marital Status	Very Happy	Pretty Happy	Not Too Happy
Married	600 (13.1)	720 (−5.4)	93 (−10.0)
Widowed	63 (−2.2)	142 (−0.2)	51 (3.4)
Divorced	93 (−6.1)	304 (3.2)	88 (3.6)
Separated	19 (−2.7)	51 (−1.2)	31 (5.3)
Never Married	144 (−7.4)	459 (4.2)	127 (4.0)

(a) Software reports that $\chi^2 = 236.4$. Interpret.

(b) Table 8.25 also shows, in parentheses, the standardized residuals. Summarize the association by indicating which marital statuses have strong evidence of (i) more, (ii) fewer people in the population in the *very happy* category than if the variables were independent.

(c) Compare the married and divorced groups by the difference in proportions in the *very happy* category.

8.17. In a *USA Today*/Gallup poll in July 2006, 82% of Republicans approved of President George W. Bush's performance, whereas 9% of Democrats approved. Would you characterize the association between political party affiliation and opinion about Bush's performance as weak, or strong? Explain why.

8.18. In a recent GSS, the death penalty for subjects convicted of murder was favored by 74% of whites and 43% of blacks. It was favored by 75% of males and 63% of females. In this sample, which variable was more strongly associated with death penalty opinion—race or gender? Explain why.

8.19. Refer to Exercise 8.10, on alcohol use and cigarette use.

(a) Describe the strength of association using the difference between users and nonusers of alcohol in the proportions who have used cigarettes. Interpret.

(b) Describe the strength of association using the difference between users and nonusers of cigarettes in the proportions who have used alcohol. Interpret.

(c) Describe the strength of association using the odds ratio. Interpret. Does the odds ratio value depend on your choice of response variable?

8.20. Table 8.26 cross-classifies 68,694 passengers in autos and light trucks involved in accidents in the state of Maine by whether they were wearing a seat belt and by whether they were injured or killed. Describe the association using

(a) The difference between two proportions, treating whether injured or killed as the response variable.

(b) The odds ratio.

TABLE 8.26

		Injury	
		Yes	No
Seat Belt	Yes	2409	35,383
	No	3865	27,037

Source: Thanks to Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, for supplying these data.

8.21. According to the Substance Abuse and Mental Health Archive, a 2003 national household survey on drug abuse indicated that for Americans aged

26–34, 51% had used marijuana at least once in their lifetime, and 18% had used cocaine at least once.

(a) Find the odds of having used (i) marijuana, (ii) cocaine. Interpret.

(b) Find the odds ratio comparing marijuana use to cocaine use Interpret.

8.22. According to the U.S. Department of Justice, in 2004 the incarceration rate in the nation's prisons was 1 per 109 male residents, 1 per 1563 female residents, 1694 per 100,000 black residents, and 252 per 100,000 white residents (*Source*: www.ojp.usdoj.gov/bjs).

(a) Find the odds ratio between whether incarcerated and (i) gender, (ii) race. Interpret.

(b) According to the odds ratio, which has the stronger association with whether incarcerated, gender or race? Explain.

8.23. Refer to Table 8.1 (page 222) on political party ID and gender. Find and interpret the odds ratio for each 2×2 subtable. Explain why this analysis suggests that the last two columns show essentially no association.

8.24. For college freshmen in 2004, the percent who agreed that homosexual relationships should be legally prohibited was 38.0% of males and 23.4% of females (www.gseis.ucla.edu/heri/american_freshman.html).

(a) The odds ratio is 2.01. Explain what is wrong with the interpretation, "The probability of a yes response for males is 2.01 times the probability of a yes response for females." Give the correct interpretation.

(b) The odds of a yes response equaled 0.613 for males. Estimate the probability of a yes response for males.

(c) Based on the odds of 0.613 for males and the odds ratio of 2.01, show how to estimate the probability of a yes response for females.

8.25. Table 8.27 cross-classifies happiness with family income for the subsample of the 2004 GSS that identified themselves as Jewish.

(a) Find the number of (i) concordant pairs, (ii) discordant pairs.

(b) Find gamma and interpret.

TABLE 8.27

		HAPPY		
		Not_too	Pretty	Very
INCOME	Below	1	2	1
	Average	0	5	2
	Above	2	4	0

(c) Show how to express gamma as a difference between two proportions.

8.26. For the 2006 GSS, $\hat{\gamma} = 0.22$ for the relationship between job satisfaction (SATJOB; categories very dissatisfied, little dissatisfied, moderately satisfied, very satisfied) and family income (FINREL: below average, average, above average).

(a) Would you consider this a very strong or relatively weak association? Explain.

(b) Of the pairs that are concordant or discordant, what proportion are concordant? Discordant

(c) Is this a stronger or a weaker association than the one between job satisfaction and happiness (variable HAPPY), which has $\hat{\gamma} = 0.4$? Explain.

8.27. A study on educational aspirations of high school students[2] measured aspirations using the scale (some high school, high school graduate, some college, college graduate) and family income with three ordered categories. Software provides the results shown in Table 8.28.

(a) Use gamma to summarize the association.

(b) Test independence of educational aspiration and family income using the chi-squared test. Interpret.

(c) Find the 95% confidence interval for gamma. Interpret.

(d) Conduct an alternative test of independence that takes category ordering into account. Why are results so different from the chi-squared test?

TABLE 8.28

Statistic	DF	Value	Prob
Chi-Square	6	8.871	0.1

Statistic	Value	ASE
Gamma	0.163	0.1

8.28. Refer to Exercise 8.13, on happiness and income. The analysis there does not take into account the ordinality of the variables. Using software:

(a) Summarize the strength of association by testing and interpreting gamma.

(b) Construct and interpret a 95% confidence interval for the population value of gamma.

Concepts and Applications

8.29. Refer to the "Student survey" data file (Exercise 1.11 on page 8). Using software, create

[2] S. Crysdale, *Intern. J. Compar. Sociol.*, vol. 16, 1975, pp. 19–36.

inal response variables, an extension of logistic regression uses *cumu-gits*, which are logits of *cumulative probabilities*. The model is called ***lative logit model***. The effects of predictors are the same for each ive probability.

inal response variables, an extension of logistic regression forms logits ng each category with a baseline category. Each logit equation has : parameters.

ar models are useful for investigating association patterns among a set of cal response variables. They consider possible conditional independence and use conditional odds ratios to describe association.

dels for contingency tables, Pearson and likelihood-ratio chi-squared ; test the goodness of fit of the model to the data.

on introduced the chi-squared test for bivariate contingency tables models presented in this chapter did not become popular until near he 1900s. They are examples of ***generalized linear models***, which rete as well as continuous response variables. The statistician and eo Goodman is responsible for many of the developments in this scientists now have available a wide variety of tools for analyzing ita.

PROBLEMS

ribes how the prob-
olican candidate in
s on x, the voter's
ands of dollars) in
ion equation for a

$1.00 + 0.02x$.

sign.
lity of voting for the
n (i) income = 10
0 thousand.
he estimated prob-
publican candidate
er than 0.50?
or which $P(y = 1)$
approximation for
lity for an increase
income.
thousand dollar
n the odds of voting

When the explana-
ily income, x_2 =
and s = sex (1 =
on equation is

$+ 0.08x_2 + 0.20s$.

For this sample, x_1 ranges from 6 to 157 with a standard deviation of 25, and x_2 ranges from 7 to 20 with a standard deviation of 3.

(a) Find the estimated probability of voting Republican for (i) a man with 16 years of education and income 30 thousand dollars, (ii) a woman with 16 years of education and income 30 thousand dollars.

(b) Convert the probabilities in (a) to odds, and find the odds ratio, the odds for men divided by the odds for females. Interpret.

(c) Show how the odds ratio in (b) relates to the sex effect in the prediction equation.

(d) Holding the other variables constant, find the estimated effect on the odds of voting Republican of
i) A standard deviation change in x_2
ii) A standard deviation change in x_1
Which predictor has the larger standardized effect? Interpret.

15.3. A sample of 54 elderly men are given a psychiatric examination to determine whether symptoms of senility are present. A subtest of the Wechsler Adult Intelligence Scale (WAIS) is the explanatory variable. The WAIS scores range from 4 to 20, with a mean of 11.6. Higher values indicate more effective intellectual functioning. Table 15.18 shows results.

(a) Show (i) $\hat{P}(y = 1) = 0.50$ at $x = 7.2$, (ii) $\hat{P}(y = 1) < 0.50$ for $x > 7.2$.

(b) Estimate the probability of senility at $x = 20$.

Variable	B	s
INTERCEPT	2.0429	
WAIS	-0.2821	

(c) The fit of the linear probability mo
$\hat{P}(y = 1) = 0.847 - 0.051x$. Estima
probability of senility at $x = 20$. Doe
make sense?

(d) Test H_0: $\beta = 0$ against H_a : $\beta \neq 0$. Repo
interpret the P-value.

15.4. Refer to the previous exercise. One of the
subtests, called *Picture completion*, asks que
about 20 pictures that have one vital detail mi
It is considered a test of attention to fine
The observations for 20 subjects on (x, y),
x = picture completion score and y = sym
of senility (1 = yes), are

(7,1), (5,1), (3,1), (8,1)(1,1), (2,1)(9,1),
(6,1), (4,1), (6,0), (9,0), (7,0), (7,0), (10,
(12,0), (14,0), (8,0), (8,0), (11,0).

(a) Using software, estimate the logistic
sion equation.

(b) Estimate the probability that sympto
senility are present when (i) $x = 0$, (ii) x

(c) Over what range of x-scores is the esti
probability of senility greater than 0.50?

(d) Estimate the effect of a one-unit increa
on the odds that senility symptoms exist

15.5. The final subsection of Section 15.2 used esti
probabilities to describe the effect of husb
earnings on the decision to buy a home. Perf
similar analysis to describe the effect of wife's
ings, when husband's earnings = \$50,000,
married = 3, the wife is working in two years,
ber of children = 0, add child in two years
head's education = 16 years, and parents'
ownership = 0. Interpret.

15.6. Table 12.1 in Chapter 12 reported GSS da
political ideology by party affiliation of

	1	2	3	4	5	6
Democrat	9	20	17	36	4	5
Republican	0	2	7	23	23	17

Use logistic regression to describe the eff
ideology on the probability of being a Demo

(a) Report the prediction equation, and
mate the probability of Democratic affil
at ideology level (i) 1 = extremely li
(ii) 7 = extremely conservative.

[3]M. Kalmin, *American Sociological Review*, vol. 59

TABLE 15.18

Variable	B	Std. Error	Wald Chi-square	Sig.
INTERCEPT	2.0429	1.0717	3.6338	0.0566
WAIS	-0.2821	0.1007	7.8487	0.0051

(c) The fit of the linear probability model is $\hat{P}(y = 1) = 0.847 - 0.051x$. Estimate the probability of senility at $x = 20$. Does this make sense?

(d) Test H_0: $\beta = 0$ against H_a : $\beta \neq 0$. Report and interpret the P-value.

15.4. Refer to the previous exercise. One of the WAIS subtests, called *Picture completion*, asks questions about 20 pictures that have one vital detail missing. It is considered a test of attention to fine detail. The observations for 20 subjects on (x, y), where x = picture completion score and y = symptoms of senility (1 = yes), are

$(7,1),(5,1),(3,1),(8,1)(1,1),(2,1)(9,1),(3,1),$
$(6,1),(4,1),(6,0),(9,0),(7,0),(7,0),(10,0),$
$(12,0),(14,0),(8,0),(8,0),(11,0).$

(a) Using software, estimate the logistic regression equation.

(b) Estimate the probability that symptoms of senility are present when (i) $x = 0$, (ii) $x = 20$.

(c) Over what range of x-scores is the estimated probability of senility greater than 0.50?

(d) Estimate the effect of a one-unit increase in x on the odds that senility symptoms exist.

15.5. The final subsection of Section 15.2 used estimated probabilities to describe the effect of husband's earnings on the decision to buy a home. Perform a similar analysis to describe the effect of wife's earnings, when husband's earnings = \$50,000, years married = 3, the wife is working in two years, number of children = 0, add child in two years = 0, head's education = 16 years, and parents' home ownership = 0. Interpret.

15.6. Table 12.1 in Chapter 12 reported GSS data on political ideology by party affiliation of

	1	2	3	4	5	6	7
Democrat	9	20	17	36	4	5	0
Republican	0	2	7	23	23	17	2

Use logistic regression to describe the effect of ideology on the probability of being a Democrat.

(a) Report the prediction equation, and estimate the probability of Democratic affiliation at ideology level (i) 1 = extremely liberal, (ii) 7 = extremely conservative.

(b) Use the model to test whether the variables are independent. Report the test statistic and P-value, and interpret.

(c) Use the odds ratio to describe the effect on party affiliation of a change in ideology from (i) 1 = extremely liberal to 2 = liberal, (ii) 1 = extremely liberal to 7 = extremely conservative.

(d) Construct and interpret a 95% confidence interval for the population odds ratio in (c), case (i).

15.7. A study of mother's occupational status and children's schooling[3] reported the prediction equation

$$\text{logit}[\hat{P}(y = 1)] = 0.75 + 0.35b + 0.13f + 0.09m + 0.30fo + 0.21mo - 0.92me - 0.16s,$$

where $y = 1$ if the child obtains a high school degree, b = respondent's year of birth, f = father's education, m = mother's education (0 to 17), fo = father's occupational level, mo = mother's occupational level (1 to 9), me = whether mother employed (1 = yes), s = number of siblings. All effects were significant at the 0.01 level.

(a) Interpret the coefficient of mother's education.

(b) Interpret the coefficient of whether mother employed.

(c) The author reported that a one-point increase in mother's occupational level is associated with a 23% increase in the odds of a high school diploma. Explain how he made this interpretation.

15.8. Let $P(y = 1)$ denote the probability that a randomly selected respondent supports current laws legalizing abortion, estimated using sex of respondent ($s = 0$, male; $s = 1$, female), religious affiliation ($r_1 = 1$, Protestant, 0 otherwise; $r_2 = 1$, Catholic, 0 otherwise; $r_1 = r_2 = 0$, Jewish), and political party affiliation ($p_1 = 1$, Democrat, 0 otherwise; $p_2 = 1$, Republican, 0 otherwise, $p_1 = p_2 = 0$, Independent). The logistic model with main effects has prediction equation

$$\text{logit}[\hat{P}(y = 1)] = 0.11 + 0.16s - 0.57r_1 - 0.66r_2 + 0.47p_1.$$

- 1.67P2

[3] M. Kalmin, *American Sociological Review*, vol. 59, 1994, p. 257.

(a) Give the effect of sex on the odds of supporting legalized abortion; that is, if the odds of support for females equal θ times the odds of support for males, report $\hat{\theta}$.

(b) Give the effect of being Democrat instead of Independent on the estimated odds of support for legalized abortion.

(c) Give the effect of being Democrat instead of Republican on the estimated odds of support for legalized abortion.

(d) Find the estimated probability of supporting legalized abortion, for (i) female Jewish Democrats, (ii) male Catholic Republicans.

15.9. Table 15.19 shows results of a study on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. The response is whether they developed AIDS symptoms during the three-year study. Software reports the fit of the logistic model with main effects:

Parameter	Estimate	Std. Error	Wald Chi-Square	Sig
Intercept	-1.0736	0.2629	16.6705	.0001
[azt=yes]	-0.7195	0.2790	6.6507	.0099
[race=black]	0.0555	0.2886	0.0370	.8476

TABLE 15.19

Race	AZT Use	Symptoms: Yes	Symptoms: No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

Source: *The New York Times*, February 15, 1991.

(a) Set up dummy variables, and report the prediction equation.

(b) Interpret the signs of the azt and race estimates.

(c) For white veter
use, estimate th
toms.

(d) Find the estim
between AZT
symptoms. Inter

(e) Test for the effe

15.10. For Table 15.11 on
a SAS printout for a
juana use as the resp

(a) Set up dummy v
tion equation. In
of alcohol use ar

(b) Why are the esti
at the second ca

(c) Estimate the pro
juana (i) for tho
or cigarettes, (ii)
alcohol and ciga

(d) Show how to co
hol use to an esti

15.11. A sample of inma
Rhode Island Depa
asked whether they
tested for hepatitis
ber who reported i
the 887 men who tes
3044 men who teste
197 women who test
the 288 women who
authors[4] concluded
may be underestimat
reported injecting dru

(a) Report the resul

(b) Define dummy v
for which the odd
and whether inj
the population fo

(c) Fit the model in
based estimate o

15.12. Table 15.21 refers to
admission into gradu
of California in Berk

TABLE 15.20

Parameter		DF	Estimate	Std Err	ChiSquare	Pr
INTERCEPT		1	-5.309	0.4752	124.820	0.
ALCOHOL	yes	1	2.986	0.4647	41.293	0.
ALCOHOL	no	0	0.000	0.0000	0.	
CIGARETT	yes	1	2.848	0.1638	302.141	0.
CIGARETT	no	0	0.000	0.0000	0.	

[4]G. Macolino et al., *American Journal of Public Health*, vol. 95, 2005, pp. 1739–1740.

[5]From D. Freedman, R. Pisani, and R. Purves, *Statistics*, W. W. Norton, 1978, p. 14.