exist to describe an $r \times c$ table well by a single number. Nominal measures based on predictive power (called *tau* and *lambda*) and gamma for ordinal data were defined in 1954 by two prominent statistician–social scientists, Leo Goodman and William Kruskal. Most software for analyzing contingency tables prints their measures and several others. Some nominal measures, such as the contingency coefficient and Cramer's V, are difficult to interpret (other than larger values representing stronger association) and, in our view, not especially useful.

We do not present the summary nominal measures in this text. We believe you get a better feel for the association by making percentage comparisons of conditional distributions, by viewing the pattern of standardized residuals in the cells of the table, by constructing odds ratios in 2×2 subtables, and by building models such as those presented in Chapter 15. These methods become even more highly preferred to summary measures of association when the analysis is multivariate rather than bivariate.

8.5 ASSOCIATION BETWEEN ORDINAL VARIABLES*

We now turn our attention to other analyses of contingency tables that apply when the variables are ordinal. The categories of ordinal variables are ordered. Statistical analyses for ordinal data take this ordering into account. This section introduces a popular ordinal measure of association, and Section 8.6 presents related methods of inference.

EXAMPLE 8.7 How Strongly Associated Are Income and Happiness?

Table 8.15 is a contingency table with ordinal variables. These data, from the 2004 GSS, refer to the relation between family income (FINRELA) and happiness (HAPPY). This table shows results for black Americans, and Exercise 8.13 analyzes data for white Americans.

Let's first get a feel for the data by studying the conditional distributions on happiness. Table 8.15 shows these in parentheses. For instance, the conditional distribution (24%, 54%, 22%) displays the percentages in the happiness categories for subjects with family income below average. Only 22% are very happy, whereas 36% of the subjects at the highest income level are very happy. Conversely, a lower percentage (9%) of the high-income group are not too happy compared to those in the lowest income group (24%). The odds ratio for the four corner cells is $(16 \times 8)/(15 \times 2) = 4.3$. It seems that subjects with higher incomes tended to have greater happiness. ■

Needed for Q8.28

TABLE 8.15: Family Income and Happiness for a GSS Sample

	Happiness			
Family Income	Not Too Happy	Pretty Happy	Very Happy	Total
Below average	16 (24%)	36 (54%)	15 (22%)	67 (100.0%)
Average	11 (16%)	36 (53%)	21 (31%)	68 (100.0%)
Above average	2 (9%)	12 (55%)	8 (36%)	22 (100.0%)
Total	29	84	44	157

Ordinal data exhibit two primary types of association between variables x and y—*positive* and *negative*. Positive association results when subjects at the high end of the scale on x tend also to be high on y, and those who are low on x tend to be low on y. For example, a positive association exists between income and happiness if those with low incomes tend to have lower happiness, and those with high incomes tend

were slain by whites. Let y denote race of victim and x denote race of murderer.

(a) Which conditional distributions do these statistics refer to, those of y at given levels of x or those of x at given levels of y? Set up a contingency table showing these distributions.

(b) Are x and y independent or dependent? Explain.

How large a χ^2 value provides a P-value of 0.05 for testing independence for the following table dimensions?

a) 2×2 **b)** 3×3 **c)** 2×5 **d)** 5×5 **e)** 3×9

Show that the contingency table in Table 8.21 has four degrees of freedom by showing how the four cell counts given determine the others.

TABLE 8.21

10	20		60
30	40		100
			40
50	80	70	

In 2000 the GSS asked whether a subject is willing to accept cuts in the standard of living to help the environment (GRNSOL), with categories (very willing, fairly willing, neither willing nor unwilling, not very willing, not at all willing). When this was cross-tabulated with sex, $\chi^2 = 8.0$.

(a) What are the hypotheses for the test to which refers?

(b) Report the df value on which χ^2 is based.

(c) What conclusion would you make, using a significance level of (i) 0.05, (ii) 0.10? State your conclusion in the context of this study.

Table 8.22 refers to a survey of senior high school students in Dayton, Ohio.

(a) Construct conditional distributions that treat cigarette smoking as the response variable. Interpret.

(b) Test whether cigarette use and alcohol use are statistically independent. Report the P-value and interpret.

TABLE 8.22

		Cigarette Use	
		Yes	No
Alcohol Use	Yes	1449	500
	No	46	281

Source: Thanks to Professor Harry Khamis for providing these data.

8.11. Are people happier who believe in life after death? Go to the GSS Web site sda.berkeley.edu/GSS and download the contingency table for the 2006 survey relating happiness and whether you believe in life after death (variables HAPPY and POSTLIFE, with YEAR(2006) in the 'selection filter').

(a) State a research question that could be addressed with the output.

(b) Report the conditional distributions, using happiness as the response variable, and interpret.

(c) Report the χ^2 value and its P-value. (You can get this by checking 'Statistics'.) Interpret.

(d) Interpret the standardized residuals. (You can get them by checking 'z-statistic'.)

8.12. In the GSS, subjects who were married were asked the happiness of their marriage, the variable coded as HAPMAR.

(a) Go to sda.berkeley.edu/GSS/ and construct a contingency table for 2006 relating HAPMAR to family income measured as (above average, average, below average), by entering FINRELA(r: 1-2; 3; 4-5) as the row variable and YEAR(2006) in the selection filter. Use a table or graph with conditional distributions to describe the association.

(b) By checking 'Statistics,' you request the chi-squared statistic. Report it and its df and P-value, and interpret.

8.13. The sample in Table 8.15 is 157 black Americans. Table 8.23 shows cell counts and standardized residuals for income and happiness for white subjects in the 2004 GSS.

(a) Explain how to interpret the Pearson chi-squared statistic and its associated P-value.

(b) Explain how to interpret the standardized residuals in the four corner cells.

TABLE 8.23

Rows: income Columns: happiness

	not	pretty	very	All
below	62	187	45	294
	5.34	3.43	-7.40	
average	47	270	181	498
	-2.73	-0.57	2.53	
above	22	127	118	267
	-2.37	-2.88	4.73	
All	131	584	131	1069

Cell Contents: Count
Standardized residual

Pearson Chi-Square = 72.15, DF = 4, P-Value = 0.000

Referred to by Q8.28

5–34, 51% had used marijuana at least once in
eir lifetime, and 18% had used cocaine at least
nce.

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

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

According to the U.S. Department of Justice, in 004 the incarceration rate in the nation's pris-ns was 1 per 109 male residents, 1 per 1563 male residents, 1694 per 100,000 black resi-ents, and 252 per 100,000 white residents (*Source*: ww.ojp.usdoj.gov/bjs).

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

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

efer to Table 8.1 (page 222) on political party D and gender. Find and interpret the odds ratio or each 2 × 2 subtable. Explain why this analysis uggests that the last two columns show essentially o association.

or college freshmen in 2004, the percent who greed that homosexual relationships should be gally prohibited was 38.0% of males and 23.4% f females (www.gseis.ucla.edu/heri/american_ reshman.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 proba-bility 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.

able 8.27 cross-classifies happiness with family ncome for the subsample of the 2004 GSS that dentified 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

rysdale, *Intern. J. Compar. Sociol.*, vol. 16, 1975, pp. 19–36.

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 (FINRELA; 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 $\hat{\gamma} = 0.40$? Explain.

First question

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 aspirations 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	0.181

Statistic	Value	ASE
Gamma	0.163	0.080

Second question

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 finding 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 and

Concepts and Applications

10.25. Refer to the "Student survey" data file (Exercise 1.11 on page 7). Construct partial tables relating opinion about abortion to opinion about life after death, controlling for attendance at religious services, measured using the two categories, (Never or occasionally, Most weeks or every week). Prepare a report (a) posing and interpreting a possible arrow diagram, before you analyze the data, for relationships among the variables; (b) interpreting the sample associations in the bivariate table and the partial tables; (c) revising, if necessary, your arrow diagram based on the evidence in the sample data.

10.26. For the student survey data (Exercise 1.11), are there any pairs of variables for which you expect the association to disappear under control for a third variable? Explain.

10.27. Using the most recent GSS, construct a contingency table relating gender (GSS variable SEX) and party identification (PARTYID). Is there still a gender gap? Control for political ideology (POLVIEWS) by forming partial tables for the most conservative and the most liberal subjects. Does the association seem to persist for these subjects?

10.28. Suppose that X_1 = father's education is positively associated with Y = son's income at age 40. However, for the regression analysis conducted separately at fixed levels of X_2 = son's education, the correlation does not differ significantly from zero. Do you thing this is more likely to reflect a chain relationship or a spurious relationship? Explain.

10.29. Table 10.9 shows the mean number of children in Canadian families, classified by whether the family was English speaking or French speaking and by whether the family lived in Quebec or in another province. Let Y = number of children in family, X_1 = primary language of family, and X_2 = province (Quebec, others).

(a) Describe the association between Y and X_1, based on the overall means in this table.

(b) Describe the association between Y and X_1, controlling for X_2.

TABLE 10.9

Province	English	French
Quebec	1.64	1.80
Other	1.97	2.14
Overall	1.95	1.85

(c) Explain how it is possible that for each level of province the mean is higher for French speaking families yet overall the mean is higher for English speaking families. (This illustrates *Simpson's paradox*. See Exercise 10.14.)

10.30. Eighth-grade math scores on the National Assessment of Educational Progress had means of 277 in Nebraska and 271 in New Jersey. For white students the means were 281 in Nebraska and 283 in New Jersey. For black students, the means were 236 in Nebraska and 242 in New Jersey. For other nonwhite students, the means were 259 in Nebraska and 260 in New Jersey.[6]

(a) Identify the group variable specifying the two states as the explanatory variable. What is the response variable and the control variable?

(b) Explain how it is possible for New Jersey to have the higher mean for each race yet for Nebraska to have the higher mean when the data are combined. (This illustrates *Simpson's paradox*.).

10.31. Example 7.1 (page 187) discussed a study that found that prayer did not reduce the incidence of complications for coronary surgery patients.

(a) Just as association does not imply causality, so does a lack of association not imply a lack of causality, because there may be an alternative explanation. Illustrate this using this study.

(b) A summary of this study in *Time Magazine* (December 4, 2006, p. 87) noted that "the prayers said by strangers were provided by the clergy and were all identical. Maybe that prevented them from being truly heartfelt. In short, the possible confounding factors in this study made it extraordinary limited." Explain what the "possible confounding" means, in the context of this study.

10.32. A study observes that subjects who say they exercise regularly reported only half as many serious illnesses per year, on the average, as those who say they do not exercise regularly. The results section in the article states, "We next analyzed whether age was a confounding variable affecting this association." Explain what this sentence means and how age could potentially explain the association between exercising and illnesses.

10.33. A research study funded by Wobegon Springs Mineral Water, Inc., discovers that the probability that a newborn child has a birth defect is lower for families that regularly buy bottled water than for families that do not. Does this association reflect a causal link between drinking bottled water and a reduction in birth defects? Why or why not?

10.34. The percentage of women who get breast cancer is higher now than at the beginning of this century. Suppose that cancer incidence tends to increase with age, and suppose that women tend to live

[6] H. Wainer and L. Brown, *American Statistician*, vol. 58, 2004.

Third question, 10.29

Do "use http://teaching.sociology.ul.ie/ws/floridacrime" for Q11.8

TABLE 11.12

	B	Std. Error	t	Sig
(Constant)	-11.526	16.834	-0.685	0.4960
INCOME	2.609	0.675	3.866	0.0003

	B	Std. Error	t	Sig
(Constant)	40.261	16.365	2.460	0.0168
INCOME	-0.809	0.805	-1.005	0.3189
URBAN	0.646	0.111	5.811	0.0001

Fourth question.

8. Refer to the previous exercise. Using software with the "Florida crime" data file at the text website:

(a) Construct box plots for each variable and scatterplots and partial regression plots between y and each of x_1 and x_2. Interpret these plots.

(b) Find the prediction equations for the (i) bivariate effects of x_1 and of x_2, (ii) multiple regression model. Interpret.

(c) Find R^2 for the multiple regression model, and show that it is not much larger than r^2 for the model using urbanization alone as the predictor. Interpret.

9. Recent UN data from several nations on y = crude birth rate (number of births per 1000 population size), x_1 = women's economic activity (female labor force as percentage of male), and x_2 = GNP (per capita, in thousands of dollars) has prediction equation $\hat{y} = 34.53 - 0.13x_1 - 0.64x_2$.

(a) Interpret the coefficient of x_1.

(b) Sketch on a single graph the relationship between y and x_1 when $x_2 = 0$, $x_2 = 10$, and $x_2 = 20$. Interpret the results.

(c) The bivariate prediction equation with x_1 is $\hat{y} = 37.65 - 0.31x_1$. The correlations are $r_{yx_1} = -0.58$, $r_{yx_2} = -0.72$, and $r_{x_1x_2} = 0.58$. Explain why the coefficient of x_1 in the bivariate equation is quite different from in the multiple predictor equation.

10. For recent UN data for several nations, a regression of carbon dioxide use (CO_2, a measure of air pollution) on gross domestic product (GDP) has a correlation of 0.786. With life expectancy as a second explanatory variable, the multiple correlation is 0.787.

(a) Explain how to interpret the multiple correlation.

(b) For predicting CO_2, did it help much to add life expectancy to the model? Does this mean that life expectancy is very weakly correlated with CO_2? Explain.

11. Table 11.13 shows a printout from fitting the multiple regression model to recent statewide data, excluding D.C., on y = violent crime rate (per 100,000 people), x_1 = poverty rate (percentage with income below the poverty level), and x_2 = percent living in urban areas.

Fifth question

TABLE 11.13

	Sum of Squares	DF	Mean Square	F	Sig
Regression	2468368.07	2	1224184.04	31.249	0.0001
Residual	1841257.15	47	39175.68		
Total	4289625.22	49			

R	R Square	Std Error of the Estimate
.7555	.5708	197.928

	B	Std. Error	t	Sig
(Constant)	-498.683	140.988	-3.537	0.0009
POVERTY	32.622	6.677	4.885	0.0001
URBAN	9.112	1.321	6.900	0.0001

Correlations

	VIOLENT	POVERTY	URBAN
VIOLENT	1.0000	.3688	.5940
POVERTY	.3688	1.0000	-.1556
URBAN	.5940	-.1556	1.0000

(a) Report the prediction equation.

(b) Massachusetts had $y = 805$, $x_1 = 10.7$, and $x_2 = 96.2$. Find its predicted violent crime rate. Find the residual, and interpret.

(c) Interpret the fit by showing the prediction equation relating $\hat{y}$ and x_1 for states with (i) $x_2 = 0$, (ii) $x_2 = 50$, (iii) $x_2 = 100$. Interpret.

(d) Interpret the correlation matrix.

(e) Report R^2 and the multiple correlation, and interpret.

11.12. Refer to the previous exercise. Sixth question

(a) Report the F statistic for testing H_0: $\beta_1 = \beta_2 = 0$, report its df values and P-value, and interpret.

(b) Show how to construct the t statistic for testing H_0: $\beta_1 = 0$, report its df and P-value for H_a: $\beta_1 \neq 0$, and interpret.

(c) Construct a 95% confidence interval for β_1, and interpret.

(d) Since these analyses use data for all the states, what relevance, if any, do the inferences have in (a)–(c)?

11.13. Refer to the previous two exercises. When we add x_3 = percentage of single-parent families to the model, we get the results in Table 11.14.

(a) Report the prediction equation and interpret the coefficient of poverty rate.

TABLE 11.14

Variable	Coefficient	Std. Error
Intercept	−1197.538	
Poverty	18.283	6.136
Urban	7.712	1.109
Single parent	89.401	17.836
R^2	0.722	
n	50	

TABLE 11.17

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	----	---	----	----	----	----
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0600	----	----

the percentage of adults owning homes, controlling for the other variables. Interpret.

(b) Find a 95% confidence interval for the change in the mean of y for a 50-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.

Seventh question

11.19. Use software with the "house selling price" data file at the text Web site to conduct a multiple regression analysis of y = selling price of home (dollars), x_1 = size of home (square feet), x_2 = number of bedrooms, x_3 = number of bathrooms.

(a) Use scatterplots to display the effects of the predictors on y. Interpret, and explain how the highly discrete nature of x_2 and x_3 affects the plots.

(b) Report the prediction equation and interpret the estimated partial effect of size of home.

(c) Inspect the correlation matrix, and report the variable having the (i) strongest association with y, (ii) weakest association with y.

(d) Report R^2 for this model and r^2 for the simpler model using x_1 alone as the predictor. Interpret.

11.20. Refer to the previous exercise.

(a) Test the partial effect of number of bathrooms, and interpret.

(b) Find the partial correlation between selling price and number of bathrooms, controlling for number of bedrooms. Compare it to the correlation, and interpret.

(c) Find the estimated standardized regression coefficients for the model, and interpret.

(d) Write the prediction equation using standardized variables. Interpret.

11.21. Exercise 11.11 showed a regression analysis for statewide data on y = violent crime rate, x_1 = poverty rate, and x_2 = percent living in urban areas. When we add an int
the prediction equation $\hat{y}$ =
$1.29x_2 + 0.76x_1x_2$.

(a) As the percentage li
increases, does the effec
to increase or decrease?

(b) Show how to interpret th
by finding how it simpli
and 100.

11.22. A study analyzes relations
percentage vote for Democ
percentage of registered vo
crats, and x_2 = percentage of
vote in the election, for seve
tions in 2006. The researche
since they expect a higher sl
at larger values of x_2 than at
obtain the prediction equatio
$0.05x_2 + 0.005x_1x_2$. Does
the direction of their predicti

11.23. Use software with the "hou
file to allow interaction betv
rooms and number of bathro
selling price.

(a) Interpret the fit by sh
equation relating $\hat{y}$ and
for homes with (i) two
bathrooms.

(b) Test the significance of
Interpret.

11.24. A multiple regression analysi
tionship between y = colle
explanatory variables, using
195 students at Slippery R
high school GPA and total S
into the model. The sum
SSE = 20. Next, parents' ed
income are added, to deter

3753.3 --- ---

Parameter Estimate	Standard Error	t	Sig
70.0000			
0.1000	0.0450	----	----
-0.1500	0.0750	----	----
0.1000	0.2000	----	----
-0.0400	0.0500	----	----
0.1200	0.0500	----	----

ılts owning homes, con-
variables. Interpret.
e interval for the change
a 50-unit increase in the
owning homes, control-
ables. Interpret.
ouse selling price" data
onduct a multiple regres-
price of home (dollars),
e feet), x_2 = number of
of bathrooms.
lisplay the effects of the
rpret, and explain how
ıture of x_2 and x_3 affects

on equation and inter-
partial effect of size of

on matrix, and report the
(i) strongest association
ssociation with y.
ıodel and r^2 for the sim-
alone as the predictor.

rcise.
of number of bathrooms,

relation between selling
f bathrooms, controlling
ooms. Compare it to the
rpret.
standardized regression
ıodel, and interpret.
equation using standard-
ret.
a regression analysis for
violent crime rate, x_1 =
percent living in urban

areas. When we add an interaction term, we get the prediction equation $\hat{y} = 158.9 - 14.72x_1 - 1.29x_2 + 0.76x_1x_2$.

(a) As the percentage living in urban areas increases, does the effect of poverty rate tend to increase or decrease? Explain.

(b) Show how to interpret the prediction equation by finding how it simplifies when $x_2 = 0$, 50, and 100.

11.22. A study analyzes relationships among y = percentage vote for Democratic candidate, x_1 = percentage of registered voters who are Democrats, and x_2 = percentage of registered voters who vote in the election, for several congressional elections in 2006. The researchers expect interaction, since they expect a higher slope between y and x_1 at larger values of x_2 than at smaller values. They obtain the prediction equation $\hat{y} = 20 + 0.30x_1 + 0.05x_2 + 0.005x_1x_2$. Does this equation support the direction of their prediction? Explain.

11.23. Use software with the "house selling price" data file to allow interaction between number of bedrooms and number of bathrooms in their effects on selling price.

8th question

(a) Interpret the fit by showing the prediction equation relating $\hat{y}$ and number of bedrooms for homes with (i) two bathrooms, (ii) three bathrooms.

(b) Test the significance of the interaction term. Interpret.

11.24. A multiple regression analysis investigates the relationship between y = college GPA and several explanatory variables, using a random sample of 195 students at Slippery Rock University. First, high school GPA and total SAT score are entered into the model. The sum of squared errors is SSE = 20. Next, parents' education and parents' income are added, to determine if they have an