

# Categorical Data Analysis

Richard L. Scheaffer  
University of Florida

The reference material and many examples for this section are based on Chapter 8, Analyzing Association Between Categorical Variables, from *Statistical Methods for the Social Sciences, 3rd edition* (Alan Agresti and Barbara Finlay, Prentice Hall). Alan Agresti, a statistics professor at the University of Florida, is one of the world's leading research authorities on categorical data analysis. Barbara Finlay is a sociologist at Texas A&M University.

The technique for analyzing categorical data that is most frequently taught in introductory statistics courses is the Chi-square test for independence. Consider the following two-way table that displays the results from a survey that Dr. Scheaffer asked his freshman and sophomore students to complete. The responses in the table are answers to the question, "Do you currently have a job?" Columns show the yes/no answers; rows show whether students were freshmen or sophomores.

## Do you currently have a job?

	No	Yes
Freshmen	25	12
Sophomores	11	14
	36	26

Table 1

The  $\chi^2$ -value is 3.4, and the P-value is 0.065. What information do these numbers provide? Since the P-value is relatively small, many researchers would conclude that there is *some* association between job status and class. Does the chi-square value of 3.4 tell us anything more? Does the chi-square statistic tell us what we really want to know? Most likely what we really want to know about is the *nature* of the association. The chi-square value does not tell us anything at all about the nature of the association. In fact, the chi-square value tells us very little and is considered by many statisticians to be overused as a technique for analyzing categorical data. In this situation we can look directly at the data and conclude that sophomores have jobs at a higher rate than freshmen, but we do not get this information from the test results. The test tells us only that there is a relationship somewhere in the data. There are other techniques for analyzing categorical data that give us more information. In this situation, for example, we would like to look at the proportions to get additional information.

What we would like to have for categorical data analysis are techniques that give us useful information like we get from correlation and regression analysis for continuous data. Recall that the regression equation, in particular the slope from the regression equation, and the correlation value give us information we want to know. Once we have these values, we could

throw away the actual data and still have lots of useful information, especially if we know whether or not the slope is significantly different from zero. We would like to have some additional measures for categorical data that tell us

1. Is there evidence of association?
2. If so, how strong is it?

Consider the following two-way table that categorizes a sample of people in the work force by income level (high or low) and educational level (end after high school or end after college).

		Income level	
		Low	High
Education Level	High School	a	b
	College	c	d

Table 2

If there were a positive association in these categorical variables, how would  $a$ ,  $b$ ,  $c$ , and  $d$  be related? If there is a positive association, we would expect low income to be associated with low education and high income to be associated with high education. Thus we would expect the values of  $a$  and  $d$  to be high and the values of  $b$  and  $c$  to be low. If there were a negative association between these variables, we would expect low education to be associated with high income and high education to be associated with low income. Thus we would expect the values of  $b$  and  $c$  to be high and the values of  $a$  and  $d$  to be low. Note that these questions of association are meaningful only when we have ordered categories. In this situation our categories can be ordered high and low, but this will not be the case with all categorical variables (for example gender could not be ordered). Finally, we can think of several situations that would support no association between the variables. For example,  $a$  and  $b$  could be large with  $c$  and  $d$  small, or these values could be switched, or all cell counts could be approximately equal.

In the example above, the two-way table provides very much the same information about the association between variables as we get by examining a scatterplot. In fact we can often do everything with data in a two-way table that we can do with data in a scatterplot. We can do logistic regression, we can examine residuals, and we can create a measure similar to correlation. The important point is that we can do far more categorical analysis than we generally do just with chi-square tests.

Consider the following quote made in 1965 by Sir Austin Bradford Hill. “Like fire, the chi-square statistic is an excellent servant and a bad master.” Hill was a British statistician, epidemiologist, and medical researcher. He is credited with practically inventing in the 1940’s the randomized comparative experiments that are now known as clinical trials. Hill saw the need for systematically designed experiments, and he advocated completely randomized

designs. It is noteworthy that in the United States, the National Institutes of Health did not begin requiring well-designed clinical trials for medical researchers until 1965. In fact, the requirement today is not that there be a randomized design, but rather that a procedure be employed that is “appropriate for the investigation.” Prior to that time, researchers were free to do pretty much what they wanted to do. As a result, some experiments were designed well and others were flawed.

Table 3 contains data provided in *Statistical Methods for the Social Sciences*, page 280. The data were obtained from the 1991 General Social Survey. Subjects were asked whether they identify more strongly with the Democratic party, the Republican party, or with independents.

<b>Gender</b>	<b>Party Identification</b>			<b>Total</b>
	<b>Democrat</b>	<b>Independent</b>	<b>Republican</b>	
Females	279	73	225	577
Males	165	47	191	403
Total	444	120	416	980

Table 3

The table classifies 980 respondents according to gender and party identification. Looking at the numbers in the table, we see that females tend to be Democrats and males tend to be Republicans. So it appears that there is an association between these two variables. The P-value from the chi-square test equals 0.03, confirming that there is an association. Is the association strong or weak? This is a harder question, but we can use residual analysis to help us find the answer.

In the context of a two-way table, a residual is defined as the difference in the observed frequency and the expected frequency:  $residual = f_o - f_e$ . To more easily make comparisons, statisticians generally prefer to use standardized or adjusted residuals. Standardized residuals are calculated by dividing the residual value by the standard error of the residual.

$$adjusted\ residual = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}}$$

The adjusted residuals associated with the gender and party preference data are included in parentheses next to the observed frequencies in Table 4.

<b>Gender</b>	<b>Party Identification</b>		
	<b>Democrat</b>	<b>Independent</b>	<b>Republican</b>
Females	279 (2.3)	73 (0.5)	225 (−2.6)
Males	165 (−2.3)	47 (−0.5)	191 (2.6)

Table 4

The sign of the adjusted residual is positive when the observed frequency is higher than the expected frequency. In Table 4, we can see that the observed frequency of Democratic females is 2.3 standard errors higher than would be expected if there were no association between party and gender. Similarly, the observed frequency of Republican females is 2.6 standard errors lower than we would expect if there were no association between party and gender. Standardized residuals allow us to see the direction and strength of the association between categorical variables. A large standardized residual provides evidence of association in that cell.

### **Measuring the Strength of Association**

In the case of continuous variables we use  $r$  and  $r^2$  to measure the strength of the relationship between the variables. Measuring the strength of association between categorical variables is not quite so simple. We'll look at three ways of measuring this strength: proportions, odds and odds ratios, and concordant and discordant pairs.

#### **Comparing Proportions**

Comparing proportions is a simple procedure. Looking back at the job versus university class data in Table 1, we can see that  $25/36 = 0.69$  of the students without jobs were freshmen and that  $12/26 = 0.46$  of the students with jobs were freshmen.

#### **Do you currently have a job?**

	No	Yes
Freshmen	25	12
Sophomores	11	14
	36	26

Table 1

It is interesting to observe that the differences in proportions must range between  $-1$  and  $+1$ . A difference close to one in magnitude indicates a high level of association, while a difference close to zero represents very little association. Note the similarity here to how we interpret values of the correlation coefficient in the continuous case. In this example, the difference in proportions is  $0.69 - 0.46 = 0.23$ , suggesting a low or moderate level of association.

#### **Odds and Odds Ratios**

As an example of using odds and odds ratios to measure the strength of association, we look back at Table 2, the two-way table for income level and education level.

Education Level		Income level	
		Low	High
High School		a	b
College		c	d

Table 2

For high school graduates, the estimated probability of low income is  $\frac{a}{a+b}$  and the estimated probability of high income is  $\frac{b}{a+b}$ . Therefore, for high school graduates, the odds favoring low income are given by

$$\text{odds} = \frac{\text{probability of success}}{\text{probability of failure}} = \frac{\left(\frac{a}{a+b}\right)}{\left(\frac{b}{a+b}\right)} = \frac{a}{b} = \frac{\# \text{ successes}}{\# \text{ failures}}$$

where “success” represents low income. For college graduates, the odds favoring low income are similarly calculated to be  $c/d$ . Generally we look at the ratio of odds from two rows, which we denote by  $\hat{q}$ . In this example,

$$\hat{q} = \text{odds ratio} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

To interpret the odds ratio, suppose  $\hat{q} = 2$ . This would indicate that the odds in favor of having a low income job if a person is a high school graduate are twice the odds in favor of having a low income job if a person is a college graduate.

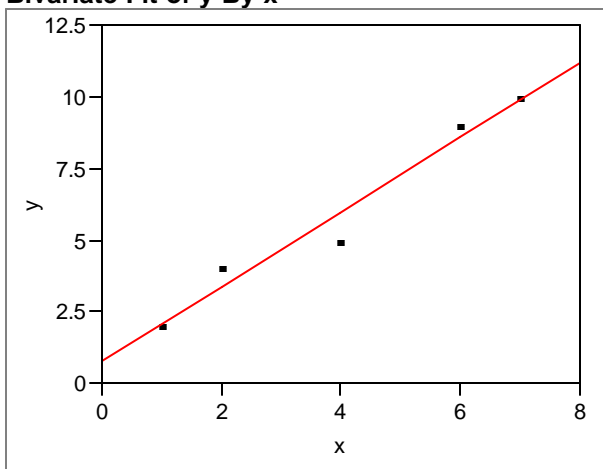
### The Effect of Sample Size

In the continuous case using regression, when we double the sample size without changing the results, what happens to the slope, correlation coefficient, and p-value for a test of significance for the slope? As an example, consider the following two sets of data  $(x, y)$  and  $(x^*, y^*)$ . The ordered pairs  $(x^*, y^*)$  are just  $(x, y)$  repeated.

$x$	1	2	4	6	7
$y$	2	4	5	9	10

$x^*$	1	2	4	6	7	1	2	4	6	7
$y^*$	2	4	5	9	10	2	4	5	9	10

### Bivariate Fit of y By x



— Linear Fit

#### Linear Fit

$$y = 0.7692308 + 1.3076923 x$$

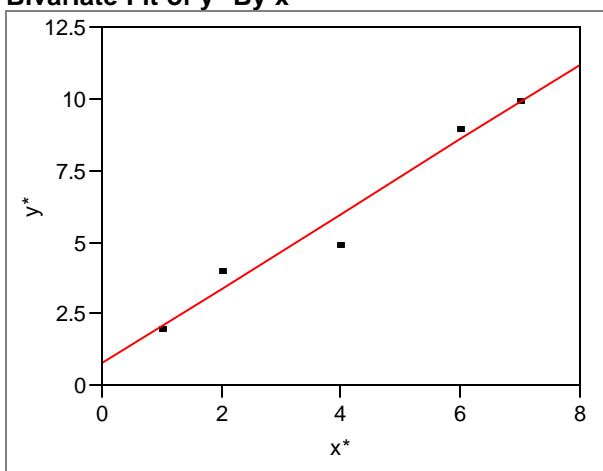
#### Summary of Fit

RSquare	0.966555
RSquare Adj	0.955407
Root Mean Square Error	0.716115
Mean of Response	6
Observations (or Sum Wgts)	5

#### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.7692308	0.646642	1.19	0.3198
x	1.3076923	0.140442	9.31	0.0026

### Bivariate Fit of y\* By x\*



— Linear Fit

#### Linear Fit

$$y^* = 0.7692308 + 1.3076923 x^*$$

### Summary of Fit

RSquare	0.966555
RSquare Adj	0.962375
Root Mean Square Error	0.620174
Mean of Response	6
Observations (or Sum Wgts)	10

### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.7692308	0.395986	1.94	0.0880
$x^*$	1.3076923	0.086003	15.21	<.0001

Notice that the parameter estimates for the two data sets are identical, as are the values of *RSquare*. However, the *t*-ratio is larger and the P-value is smaller for  $(x^*, y^*)$ . The chi-square value and P-value behave similarly to the *t*-ratio and P-value for regression. Do we have a measure that is stable for chi-square in the same way that the slope and correlation are stable for regression relative to sample size?

Consider the following table from *Statistical Methods for the Social Sciences*, page 268. These data show the results by race of responses to a question regarding legalized abortion. In each table, 49% of the whites and 51% of the blacks favor legalized abortion.

	A			B			C		
	Yes	No	Total	Yes	No	Total	Yes	No	Total
White	49	51	100	98	102	200	4900	5100	10000
Black	51	49	100	102	98	200	5100	4900	10000
Total	100	100	200	200	200	400	10000	10000	20000
	$\chi^2 = .08$			$\chi^2 = .16$			$\chi^2 = 8.0$		
	P = .78			P = .69			P = .005		

Table 5

Note that the chi-square value increases from a value that is not statistically significant to a value that is highly significant as the sample size increases from 200 to 400 to 20,000. Since the P-value for this test of association is very sensitive to the sample size, we need other measures to describe the strength of the relationship. In every table above, the difference in proportions is 0.02, and the odds ratio is approximately 0.92. The small difference in proportions and the odds ratio with value close to one indicate that there is not very much going on in any of these situations. The association is very weak. This example illustrates that we should not simply depend on a test of significance. Increasing sample size can always make very small differences in proportions statistically significant. We need to use other measures to get a more complete picture of the strength of association between two categorical variables.

To get information about the strength of association between two categorical variables, we can do inference procedures based on the odds ratio  $\hat{q}$ . First we must know the distribution

of  $\hat{q}$ . We can work out the asymptotic distribution of the odds ratio, but it is much easier to work out the distribution of the natural logarithm of the odds ratio. Note that the values of the odds ratio range from zero to very large positive numbers, that is  $0 \leq \hat{q} < \infty$ , and the values of  $\hat{q}$  are highly skewed to the right. When the odds of success are equal in the two rows being compared,  $\hat{q} = 1$ , indicating no association between the variables. By taking the natural logarithm of  $\hat{q}$ , we pull in the tail. It turns out that the logarithm of the odds ratio, that is  $\ln(\hat{q})$ , is approximately normally distributed for large values of  $n$ . Thus we can perform inference procedures on  $\ln(\hat{q})$  since statisticians know that the standard error of  $\ln(\hat{q})$  is

$$SE[\ln(\hat{q})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}},$$

and the quantity  $\frac{\ln(\hat{q}) - 0}{SE[\ln(\hat{q})]}$  has a distribution that is approximately standard normal for large values of  $n$ .

As an example, reconsider the job data provided in Table 1. For freshmen, the odds favoring not having a job are  $\frac{25}{12} = 2.08\bar{3}$ . For sophomores, the odds favoring not having a job are  $\frac{11}{14} \approx 0.7857$ . So the odds ratio is  $\hat{q} \approx \frac{2.08\bar{3}}{0.7857} \approx 2.65$ . This tells us that the odds in favor of having no job if the student is a freshman are 2.65 times the odds in favor of having no job if the student is a sophomore. Note that by taking the reciprocal of this odds ratio,  $\frac{1}{\hat{q}} \approx 0.38$ , we see that the odds in favor of having a job if the student is a freshman are only 0.38 times the odds in favor of having a job if the student is a sophomore.

We want to determine whether the odds ratio  $\hat{q} = 2.65$  differs significantly from what we would expect if there were no difference in the odds of having a job for the freshmen and sophomore students. Note that if there were no difference in the odds for the two groups, the odds ratio would be exactly equal to one. Thus, under the null hypothesis of no difference, the natural logarithm of the odds ratio would be equal to zero. Based on our sample of students,  $\ln(\hat{q}) = \ln(2.65) \approx 0.974$ . The standard error of  $\ln(\hat{q})$  is given by  $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} = \sqrt{\frac{1}{25} + \frac{1}{12} + \frac{1}{11} + \frac{1}{14}} \approx 0.53$ . Assuming that  $n$  is sufficiently large, which in this case may be questionable, we can do a  $z$ -test using as our test statistic

$$z = \frac{\ln(\hat{q}) - 0}{SE(\ln(\hat{q}))} = \frac{.974 - 0}{.53} \approx 1.84.$$



This  $z$ -value indicates that there is not much evidence of a strong association between job status and class category, even though the odds ratio is 2.65. This is a two-sided test, so  $z$ -values larger than 2 are generally considered significant. If for example the  $z$ -statistic had been approximately equal to 3, we would conclude that the odds ratio is significantly greater than one. This would indicate that the odds in favor of having no job if a student is a freshman are significantly higher than the odds in favor of having no job if a student is a sophomore.

As another example, we can analyze the data in Table 3. Odds ratios are computed on two-by-two tables, so we will ignore Independents and look just at Democrats versus Republicans. In this case  $\hat{q} = 1.435$ . So the odds in favor of being a Democrat if a person is female are 1.435 times as great as the odds in favor of being a Democrat if a person is male. Note that  $\ln(\hat{q}) = 0.361$  and  $SE[\ln(\hat{q})] = 0.139$ , so

$$z = \frac{\ln(\hat{q}) - 0}{SE(\ln(\hat{q}))} = \frac{.361 - 0}{.139} > 2$$

and we conclude that this result is significant. We could now do a similar analysis of Democrats versus Independents by ignoring the Republicans. In fact there are three two-by-two subtables that could be constructed from Table 3. We need only look at two of them (as many subtables as there are degrees of freedom); the third result is a function of the other two.

	Democrats versus Republicans	Democrats versus Independents	Independents versus Republicans
$\hat{q}$	1.435	1.089	1.318
$\ln(\hat{q})$	0.361	0.085	0.276
$SE[\ln(\hat{q})]$	0.139	0.211	0.211
$z$	2.597	0.402	1.308

It is important to realize that this test on the odds ratio does not duplicate the chi-square test. The chi-square test is used to determine whether or not there is any association between the categorical variables. This  $z$ -test is used to determine whether there is a particular association.

Before leaving two-by-two tables, we will interject a historical note about the development of chi-square procedures. The chi-square statistic was invented by Karl Pearson about 1900. Pearson knew what the chi-square distribution looks like, but he was unsure about the degrees of freedom. About 15 years later, Fisher got involved. He and Pearson were unable to agree on the degrees of freedom for the two-by-two table, and they could not settle the issue mathematically. They had no nice way to do simulations, which would be the modern approach, so they gathered lots of data in two-by-two tables where they thought the variables

should be independent. For each table they calculated the chi-square statistic. Recall that the expected value for the chi-square statistic is the degrees of freedom. After collecting many chi-square values, Pearson and Fisher averaged all the values they had obtained and got a result very close to one. Fisher is said to have commented, “The result was embarrassingly close to 1.” This confirmed that there is one degree of freedom for a two-by-two table. Some years later this result was proved mathematically.

### Concordant and Discordant Pairs

Analytical techniques based on concordant and discordant pairs were developed by Maurice Kendall. These methods are useful only when the categories can be ordered. A pair of observations is *concordant* if the subject who is higher on one variable is also higher on the other variable. A pair of observations is *discordant* if the subject who is higher on one variable is lower on the other. If a pair of observations is in the same category of a variable, then it is neither concordant or discordant and is said to be *tied* on that variable.

		Income level	
		Low	High
Education Level	High School	a	b
	College	c	d

Table 2

Looking back at Table 2, we see that the income category is ordered by low and high. Similarly the education category is ordered, with education ending at high school being the low category and education ending at college being the high category. All  $d$  observations, representing individuals in the high income and high education category, “beat” all  $a$  observations, who represent individuals in the low income and low education category. Thus there are  $C = ad$  concordant pairs. All  $b$  observations are higher on the income variable and lower on the education variable, while all  $c$  observations are lower on the income variable and higher on the education variable. Thus there are  $D = bc$  discordant pairs.

The strength of association can be measured by calculating the difference in the proportions of concordant and discordant pairs. This measure, which is called *gamma* and is denoted by  $\hat{g}$ , is defined as

$$\hat{g} = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}.$$

Note that since  $\hat{g}$  represents the difference in two proportions, its value is between  $-1$  and  $1$ , that is,  $-1 \leq \hat{g} \leq 1$ . A positive value of gamma indicates a positive association, while a negative value of gamma indicates a negative association. Similar to the correlation coefficient,  $\hat{g}$ -values close to  $0$  indicate a weak association between variables while values close to  $1$  or

-1 indicate a strong association. To obtain the value  $\hat{g} = 1$ ,  $D$  must equal 0. In order to have  $D = 0$ , either  $b = 0$  or  $c = 0$ . Whenever either  $b$  or  $c$  is close to 0 in value, the value of  $\hat{g}$  will be close to 1. Therefore,  $\hat{g}$  is not necessarily the best measure of the strength of association, however it is a common measure.

A more sensitive measure of association between two ordinal variables is called Kendall's tau- $b$ , denoted  $t_b$ . This measure is more complicated to compute, but it has the advantage of adjusting for ties. The result of adjusting for ties is that the value of  $t_b$  is always a little closer to 0 than the corresponding value of gamma. The  $t_b$  statistic can be used for inference so long as the value of  $C + D$  is large.

Consider again the data given in Table 1 where  $a = 25, b = 12, c = 11, d = 14$ . The number of concordant pairs is  $C = ad = 25 \cdot 14 = 350$ ; the number of discordant pairs is  $D = bc = 12 \cdot 11 = 132$ . In this example,  $\hat{I} = \frac{C - D}{C + D} = 0.45$ , which indicates that the association between having a job and college class is not particularly strong. The  $t_b$ -value is 0.23, which, as expected, is closer to 0; the P-value associated with the test of  $H_0 : t_b = 0$  is 0.07.

Some computer software will provide information about concordant and discordant pairs whenever analysis is performed on a two-way table. If the categories are not ordered, the results may not be meaningful, however. Suppose that the data given in Table 3 were ordered. One might assume an ordering from low conservative to highly conservative for party identification and from low to high testosterone for gender. The number of concordant pairs is given by  $C = 279(47 + 191) + 73(191) = 80,345$ . The number of discordant pairs is given by  $D = 225(165 + 47) + 73(165) = 59,745$ , and  $\hat{g} = \frac{C - D}{C + D} = 0.15$ . Recall that the chi-square test reveals a highly significant chi-square value. Based on  $\hat{g}$ , we can now see that the strength of the association is very modest. Note that the interpretation of  $\hat{g}$  may be questioned in this situation, however, since in a sense we forced the ordering.

The measure of association  $\hat{g}$  is not very sensitive to sample size. Unlike the P-value associated with a chi-square test of association, the measure of  $\hat{g}$  changes very little as sample size increases. So  $\hat{g}$  gives us a measure that is stable for chi-square in the same way that the slope and correlation are stable for regression relative to sample size.

### **Physicians' Health Study Example**

The Physicians' Health Study was a clinical trial performed in the United States in the late 1980's and early 1990's. It involved approximately 22,000 physician subjects over the age

of 40 who were divided into three groups according to their smoking history: current smokers, past smokers, and those who had never smoked. Each subject was assigned randomly to one of two groups. One group of physicians took a small dose of aspirin daily; physicians in the second group took a placebo instead. These physicians were observed over a period of several years; for each subject, a record was kept of whether he had a myocardial infarction (heart attack) during the period of the study.

Data for the 10,919 physicians who had never smoked are provided in the following table.

	Placebo	Aspirin
<b>Had heart attack</b>	96	55
<b>Did not have heart attack</b>	5392	5376
<b>Total</b>	5488	5431

A chi-square test to determine whether there is association between having a heart attack and taking aspirin results in  $\chi^2 = 10.86$  and a  $P$ -value less than 0.001. Thus we can conclude that there is *strong evidence* that aspirin has an effect on heart attacks. (Note that this statement is not synonymous with saying that aspirin has a strong effect on heart attack.) Based partially on the strength of this evidence that aspirin has an effect on heart attack, the Physicians' Health Study was terminated earlier than originally planned so that the information could be shared. In the United States, researchers are ethically bound to stop a study when significance becomes obvious.

The highly significant chi-square value does not give information about the direction or strength of the association. The odds ratio for the group who had never smoked is  $\hat{q} = 1.74$ . Thus these data indicate that the odds in favor of having a heart attack if taking the placebo are 1.74 times the odds in favor of having a heart attack if taking aspirin for the group who had never smoked. Looking at this from the other direction, we can say that the odds in favor of having a heart attack if a physician is taking aspirin are only .57 times the odds in favor of having a heart attack if he is taking a placebo. Note that for this group,  $\ln(\hat{q}) = 0.55$  and  $SE \left[ \ln(\hat{q}) \right] = 0.17$ , so the  $z$ -value will indicate that the odds ratio is significantly greater than zero.

Data for the physicians who were past smokers are provided in the following table.

	Placebo	Aspirin
<b>Had heart attack</b>	105	63
<b>Did not have heart attack</b>	4276	4310

The odds ratio for this group is  $\hat{q} = 1.68$ .

Data for the physicians who were current smokers are provided in the following table.

	Placebo	Aspirin
<b>Had heart attack</b>	37	21
<b>Did not have heart attack</b>	1188	1192

The odds ratio for this group is  $\hat{q} = 1.77$ . Note that the odds ratios are similar for all groups, indicating a favorable effect from aspirin across all groups.

To get more information from all these data, we can look at a logistic regression model for the entire group of 22,000 participating physicians. In this model there were two explanatory variables, aspirin or placebo and smoking status. The model will have the form

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) = b_0 + b_1A + b_2P + b_3C.$$

The response variable was whether or not the subject had a heart attack. All variables are coded as 0 or 1. For example, letting  $A$  represent the aspirin variable,  $A=1$  indicates aspirin and  $A=0$  indicates placebo. There are two variables for smoking status. We let  $P$  represent past smoker and code  $P=1$  to indicate that the subject is a past smoker. If  $P=0$  then the subject is not a past smoker. Similarly  $C=1$  indicates that a subject is a current smoker, while  $C=0$  indicates that the subject is not a current smoker. A subject who has never smoked would be indicated by having the values of  $C$  and  $P$  both zero.

The following linear model is based on a maximum likelihood method:

$$\ln(\text{estimated odds}) = -4.03 - 0.548A + 0.339P + 0.551C.$$

This equation simplifies to

$$\begin{aligned}\text{estimated odds} &= e^{-4.03}(e^{-0.548A})(e^{0.339P})(e^{0.551C}) \\ &= 0.018(0.578^A)(1.403^P)(1.735^C)\end{aligned}$$

Letting  $A = P = C = 0$ , we can predict the odds in favor of having a heart attack for the physicians who had never smoked and were not taking aspirin. According to the model, these odds are .018. Note that in the raw data the odds are  $\frac{96}{5392} \approx 0.018$ . Thus it appears that the model fits very well for this group. Letting  $A = 1$  and  $P = C = 0$ , the model predicts that the odds in favor of having a heart attack for physicians who never smoked and were taking aspirin

are lower:  $0.018 \cdot 0.578 \approx 0.010$ . Note that this prediction is also very close to what we see in the raw data:  $\frac{55}{5376} \approx 0.010$ . Note also that the number we multiply 0.018 by to calculate the odds in favor of having a heart attack for physicians in this group who were taking aspirin, that is 0.578, is very close to the odds ratio observed earlier for this group of physicians. For physicians who are current or past smokers, the model will incorporate a factor of 1.403 or 1.735 and thus predict higher odds than for those who have never smoked. All possible combinations are shown in the table below:

Conditions	A	P	C	Estimated Odds
No Aspirin Never Smoked	0	0	0	0.018
No Aspirin Past Smoker	0	1	0	0.0253
No Aspirin Current Smoker	0	1	1	0.0312

Aspirin Never Smoked	1	0	0	0.0104
Aspirin Past Smoker	1	1	0	0.0146
Aspirin Current Smoker	1	1	1	0.0181

Data are also provided showing outcomes by age group for the participants in this study.

#### Age 40-49

	Placebo	Aspirin
<b>Had heart attack</b>	24	27
<b>Did not have heart attack</b>	4500	4500

For the 40-49 year-olds, the odds ratio  $\hat{q} = 0.89$ ,  $\ln(\hat{q}) = -0.12$ , and  $SE \left[ \ln(\hat{q}) \right] = 0.28$ . So for the younger participants in this study, the odds of having a heart attack if taking the placebo were 0.89 times the odds of having a heart attack if taking aspirin. At first glance this might suggest that aspirin had a negative effect for the younger physicians. Noting the relative size of the standard error to the  $\ln(\hat{q})$ , however, informs us that this ratio is not significantly different from zero ( $z = \frac{-0.12}{0.28} = -0.42$ ).

As the tables below reveal, different results are observed for all groups of older participants.

#### Age 50-59

	Placebo	Aspirin
<b>Had heart attack</b>	87	51
<b>Did not have heart attack</b>	3638	3674

#### Age 60-69

	Placebo	Aspirin
<b>Had heart attack</b>	84	39
<b>Did not have heart attack</b>	1961	2006

#### Age 70-84

	Placebo	Aspirin
<b>Had heart attack</b>	44	22
<b>Did not have heart attack</b>	696	718

Values of  $\hat{q}$  for these groups are 1.72 for the 50-59 year old group, 2.20 for the 60-69 year-old group, and 2.06 for the 70-84 year-olds. These ratios inform us that aspirin is more effective for all of the older groups, though the effectiveness appears to drop off a little for the oldest participants. The odds ratios provide information that would lead us in general to recommend aspirin to a man fifty or older, but not to a man below fifty.

### Turtle Example Revisited

Now we will look again at the data Bob Stephenson provided relating temperature and gender of turtles. This situation is different from the others we have looked at, as the response variable (gender) is categorical but the explanatory variable (temperature) is continuous. We can treat the continuous variable as categorical by separating temperatures into groups, in this case five temperature groups.

The computer output from the earlier logistic regression analysis showed that temperature is significant. The data can be displayed in the two-way table that follows:

	Temperature Group					Total
	1	2	3	4	5	
<b>Male</b>	2	17	26	19	27	91
<b>Female</b>	25	7	4	8	1	45

<b>Total</b>		27	24	30	27	28	136
--------------	--	----	----	----	----	----	-----

The computer output from the logistic regression analysis contains information about concordant and discordant pairs, Kendall's Tau-a, and much more. We will look at the concordant and discordant pairs. Temperatures are ordered in the table from low to high, with the lowest temperatures in group 1 and the highest temperatures in group 5. We will order gender so that male is the high group (since this is what we are looking for), so the table rows are arranged from high to low. Thus the high/high category is the one in the upper right corner with frequency 27. We obtain the number of concordant pairs as follows:

$$C = 27(25+7+4+8) + 19(25+7+4) + 26(25+7) + 17(25) = 3129.$$

Similarly, the number of discordant pairs is computed to be

$$D = 2(20) + 17(13) + 26(9) + 19(1) = 514.$$

The resulting gamma-value is 0.72, indicating a pretty strong association between temperature and maleness. The value of Kendall's Tau-a, which accounts for ties, drops to 0.28. The gamma-value is a "rough" measure of strength of association while Kendall's Tau-a is a stronger, more sensitive measure. In addition to the adjustment for ties, there is another reason that the two values are so different in this situation. Kendall's Tau-a provides a measure of the strength of *linear* association between the variables. Because there is curvature in the relationship between temperature and gender, the more sensitive measure is going to reflect this, thus causing Kendall's Tau-a value to be lower. Note that the proportion of males increases very fast from temperature category 1 to temperature category 2. Increases for higher temperature categories are much smaller, so a graph displaying the proportion of males versus temperature category is increasing but concave down. Because there is no adjustment for ties, the gamma measure is a less sensitive detector of this curvature.

Note again that measures of association for categorical variables are being used in this situation where one of the variables really is continuous. So caution should be exercised in using and interpreting these measures. They are provided by the software, but they may not be so meaningful in this situation.

The two-way table provided in the logistic regression output shows observed and expected frequencies and a corresponding chi-square statistic. The expected frequencies in this table are not calculated according to the usual chi-square method (row total times column total divided by grand total). Rather they are calculated by substituting into the logistic regression model. The logistic model predicts log odds, which can then be converted to an estimate of the odds, which can then be converted to a probability. The probability can be used to determine the expected frequencies for each cell of the table. Note that the logistic regression model has two terms and consequently uses two degrees of freedom. Based on these observed and expected frequencies we can compute the chi-square statistic in the usual way. The information provided by this chi-square test indicates how well the model fits the data.



There is another measure which is denoted by  $G^2$  on the logistic regression output. This  $G^2$  value is the *deviance*, which is another measure of the discrepancy between the observed frequencies and those that would be expected according to the logistic model.  $G^2$  is calculated according to the formula

$$G^2 = 2 \sum f_o \ln \left( \frac{f_o}{f_e} \right).$$

The  $G^2$  statistic is calculated from a maximum likelihood equation. On the other hand, Pearson's chi-square is not a maximum likelihood statistic. Almost always there will be methods that are better than Pearson's chi-square statistic for a given situation; however, in many cases, the chi-square statistic works reasonably well though it may not be best. In some sense,  $G^2$  is a better measure of discrepancy than the more generally used chi-square. Both are interpreted the same way. When the model fits well, we expect both the chi-square and the  $G^2$  values to be small. Large values of these statistics indicate large discrepancies between the observed and expected frequencies and therefore indicate a poor fit. For the turtle example, the chi-square and  $G^2$  values were both large. This would lead us to think that we should continue looking for a better model.

One other issue remains with the output from logistic regression. We are told that the odds ratio equals 9.13. What information does the odds ratio provide in this situation? Recall that we fit a model of the form  $\ln \left( \frac{\hat{p}}{1-\hat{p}} \right) = a + bx$ , where  $x$  is a continuous variable representing temperature. We can get the odds for any particular cell of the two-way table by exponentiating:

$$e^{\ln \left( \frac{\hat{p}}{1-\hat{p}} \right)} = e^{a+bx} = k \cdot c^x, \text{ so } \frac{\hat{p}}{1-\hat{p}} = k \cdot c^x.$$

According to the logistic regression model, the odds grow exponentially as the  $x$ -value

increases. If we calculate the ratio  $\frac{\left( \frac{\hat{p}}{1-\hat{p}} \right)_{x+1}}{\left( \frac{\hat{p}}{1-\hat{p}} \right)_x} = \frac{k \cdot c^{x+1}}{k \cdot c^x} = c = e^b$ , we see that the increase in

odds per one unit increase in the  $x$ -variable is  $e^b = e^{2.2110} = 9.1248$ . This number, which is consistent with the odds ratio reported in the logistic regression computer output, tells us that the ratio of the odds of a turtle being male is 9.13 per increase in one degree of temperature. Note also that we can use the confidence interval reported for  $b$  to create a confidence interval for the odds ratio by computing  $e^{\text{lower}}$  and  $e^{\text{upper}}$ .

### Sensitivity and Specificity

The words "sensitivity" and "specificity" have their origins in screening tests for diseases. When a single test is performed, the person may in fact have the disease or the person may be disease free. The test result may be positive, indicating the presence of disease, or the test result may be negative, indicating the absence of the disease. The table below displays test results in the columns and true status of the person being tested in the rows.

		Test Result (T)	
		Positive (+)	Negative (–)
True Status of Nature (S)	Disease (+)	a	b
	No Disease (–)	c	d

Though these tests are generally quite accurate, they still make errors that we need to account for.

**Sensitivity:** We define sensitivity as the probability that the test says a person has the disease when in fact they do have the disease. This is  $P(T^+ | S^+) = \frac{a}{a+b}$ . Sensitivity is a measure of how likely it is for a test to pick up the presence of a disease in a person who has it.

**Specificity:** We define specificity as the probability that the test says a person does not have the disease when in fact they are disease free. This is  $P(T^- | S^-) = \frac{d}{c+d}$ .

Ideally, a test should have high sensitivity and high specificity. Sometimes there are tradeoffs in terms of sensitivity and specificity. For example, we can make a test have very high sensitivity, but this sometimes results in low specificity. Generally we are able to keep both sensitivity and specificity high in screening tests, but we still get false positives and false negatives.

**False Positive:** A false positive occurs when the test reports a positive result for a person who is disease free. The false positive rate is given by  $P(S^- | T^+) = \frac{c}{a+c}$ . Ideally we would like the value of  $c$  to be zero; however, this is generally impossible to achieve in a screening test involving a large population.

**False Negative:** A false negative occurs when the test reports a negative result for a person who actually has the disease. The false negative rate is given by:  $P(S^+ | T^-) = \frac{b}{b+d}$ .

Which false result is the more serious depends on the situation. But we generally worry more about false positives in screening tests. We don't want to tell someone that they have a serious disease when they do not really have it.

The following example is take from *Statistical Methods for the Social Sciences, 3rd edition* (Alan Agresti and Barbara Finlay, Prentice Hall, page 287). The data are provided for the results from a screening test for HIV that was performed on a group of 100,000 people. Note that the prevalence rate of HIV in this group was very low.

		Test Result		
		Positive (+)	Negative (–)	
HIV Status	Positive (+)	475	25	500
	Negative (–)	4975	94525	99500
		5450	94550	100000

Based on the results above, the sensitivity of this test was  $\frac{475}{500} = 0.95$  and the specificity was  $\frac{94525}{99500} = 0.95$ . This test appears to be pretty good. What are the false positive and false negative rates for this test?

$$\text{False positive rate} = \frac{4975}{5450} = 0.91$$

$$\text{False negative rate} = \frac{25}{94550} = 0.0003$$

Even when a test has a high sensitivity and specificity, you are still going to get a high false positive rate if you are screening a large population where the prevalence of the disease is low. False positive rate is not just a function of sensitivity and specificity. It is also a function of the prevalence rate of the disease in the population you are testing. Thus there is danger in indiscriminately applying a screening test to a large population where the prevalence rate of the disease is very low.

In a significance test, sensitivity and specificity relate to correct decisions and false positives and false negatives correspond to Type I and Type II errors. Consider the columns in the table below. A positive test result corresponds to a statistically significant test result and thus leads to rejection of the null hypothesis. A negative test result corresponds to a test result that is not statistically significant; thus we decide not to reject  $H_0$ . Now consider the rows in this table. If the true status of nature is positive (that is, if something is going on and what we are testing for is really true), then the alternative hypothesis is true. if what we are testing for is not true (and nothing is going on), then the null hypothesis is correct.

		Test Result (T)	
		Positive (+)	Negative (–)
True Status of Nature (S)	$H_a$ True (+)	a	b (Type II)
	$H_0$ “True” (–)	c (Type I)	d

Note that the upper left entry in the table corresponds to the correct decision to reject the null hypothesis when the alternative is really true. A high value of  $a$  results in a high probability of detecting that a null hypothesis is false and corresponds to a test that is very sensitive. The lower right entry corresponds to the correct decision not to reject the null hypothesis when it should not be rejected. A high value of  $d$  results in a high probability of not rejecting a "true" null hypothesis and corresponds to a test with high specificity. The lower left entry corresponds to a test result that causes us to reject the null hypothesis when it is really "true;" this entry reflects Type I error and corresponds to a false positive. The upper right entry corresponds to a test result that causes us not to reject the null hypothesis when it is not true; this entry reflects Type II error and corresponds to a false negative. Note that a small value of  $c$  is consistent with low probability of Type I error and a low value of  $b$  is consistent with low probability of Type II error.