BIOS 6244 Analysis of Categorical Data
October 24, 2005 Lecture

Three-Way Contingency Tables (Chapter 3)

Controlling for one or more confounding variables is an important component of examining the E-D association in an Epi study. (Note: Agresti often refers to confounders as *control variables*.)

Example (passive smoking)

Suppose we wish to examine the association between *passive smoking* (i.e., exposure to second-hand tobacco smoke) and lung cancer (CA) using a cross-sectional design. A representative sample of subjects is selected and a 2x2 contingency table is constructed.

		Lung Cancer	
		Yes	No
Passive Smoker	Yes		
	No		

Fo purposes of this study, "passive smoker" is defined to be a non-smoker whose spouse smokes.

Suppose that we find a significant association between passive smoking and lung CA. We should then ask the question: "Can this association be attributed to some other factor besides passive smoking?" This third categorical variable is referred to as a *confounder*. It is a factor that is associated with both the exposure and disease that obscures or distorts the true E-D association.

For example, suppose that spouses of non-smokers tend to be younger than spouses of smokers. Since age is also a risk factor for lung cancer, we expect there to be fewer cases of lung cancer among the younger spouses. Therefore, the lower proportion of lung CA cases that we find among spouses of non-smokers could be attributable to the fact they are younger, not to the fact that they are married to a non-smoker (as opposed to being married to a smoker). In this case, age would be a confounder of the true E-D association between passive smoking and lung CA.

(Note that, in addition to age, ethnic origin and gender should generally be considered as potential confounders in any Epi study.)

One approach to controlling for a confounding variable is to add a third dimension to the Exposure vs. Disease contingency table. (Sometimes this dimension is referred to as the "layers.") We construct a separate 2-way table for each level of the confounding variable (which we assume to be categorical for purposes of this chapter). This process is called *stratification*

and enables us to examine the *interaction* between the risk factor of interest (e.g., passive smoking) and the confounding variable (e.g., age) and to thereby properly account for the effect of the confounding variable.

Partial Association (Sec. 3.1)

Suppose we are interested in the association between X and Y, after controlling for the effect of a confounding variable Z, where all 3 variables are binary. The 2x2 tables formed at each of the 2 levels of Z are called *partial tables.* These partial tables remove the confounding effect of Z by holding the level of Z constant.

The 2-way contingency table formed by combining these partial tables into one table is called the X-Y *marginal table*. Each cell count is obtained by summing over the levels of Z. Doing this ignores the potential effect of Z and the resulting marginal table contains no information whatsoever about Z.

The associations in partial tables are called *conditional associations* because they refer to the association between X and Y *conditional* on the value of Z (i.e., given the value of Z). These conditional associations can be quite different from the association in the marginal table.

Example (ulcer drugs)

Suppose that a study based on a chart review was performed to compare the effectiveness of two different drugs in treating stomach ulcers. Ulcer patients' records were retrospectively reviewed to determine which of the two drugs they were treated with, and whether or not their ulcer eventually healed. The study was performed at 2 different treatment sites of a large health-care provider. The following marginal table was obtained:

	Cured	Not Cured
Drug A	125	175
Drug B	330	270

Thus, the cure rate with Drug A is $\frac{125}{125+175} = 41.7\%$ and the cure rate with Drug B is $\frac{330}{330+270} = 55.0\%$. Thus, it appears that Drug B is more effective than Drug A.

However, the 2 different treatment sites have not been taken into account. If we stratify by site, we obtain the following partial tables:

		Cured	Not Cured
Site 1	Drug A	40	160
	Drug B	30	170

		Cured	Not Cured
Site 2	Drug A	85	15
	Drug B	300	100

At Site 1, the cure rates are 20% and 15% for Drugs A and B, respectively. At Site 2, the rates are 85% and 75%. Thus, at both sites, Drug A appears to be better than Drug B. This is an example of *Simpson's paradox*, which occurs when the association in the marginal table is in the opposite direction from the all of the conditional associations. (This can also happen with continuous variables.) For these data, the disproportionately large sample size in the group treated with Drug B at Site 2, along with the high cure rate in this group (75%) resulted in the "paradoxical" results. It would be extremely important in this study to try to discover a reason why the cure rates were so different at the two sites – only 17.5% of ulcer patients overall were cured at Site 1, whereas 77% were cured at Site 2. Perhaps the patients at Site 1 were more seriously ill than those at Site 2 and should not be combined with them in the overall analysis. In Section 3.2, we will discuss statistical methods that can be used to adjust for the effects of Simpson's paradox.

This example further demonstrates the importance of examining the impact of *all* potential confounding variables when examining a proposed E-D association in an Epi study. An interesting example of Simpson's paradox dealing with the death penalty is discussed by Agresti on pp. 54-57 of the text.

<u>Conditional and Marginal Odds Ratios (Sec. 3.1.3)</u>

We typically use the odds ratio to describe the marginal and conditional associations in a 3-way table. For a 2x2xK table, where K denotes the # of levels of a categorical confounding variable, we define the *conditional odds ratio* for the k'th partial table to be

$$OR_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}},$$

where μ_{ijk} denotes the expected cell frequency in cell (i,j,k).

If $\{n_{ijk}\}$ are the observed cell frequencies in a sample of size n, then

$$\widehat{OR}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}. \qquad (6)$$

The estimated OR's calculated using Equation (6) are called the *X-Y conditional odds ratios*.

These conditional OR's can be quite different from the marginal OR, in which the confounding variable Z is ignored rather than controlled for. The X-Y marginal OR is given by

$$OR_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}},$$

where $$\mu_{ij+} = \sum_{k=1}^{K} \mu_{ijk} \text{ for all i,j}$$

The sample estimate of the marginal OR is given by

$$\widehat{OR} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}.$$

Example (ulcer drugs, cont.)

The observed *X-Y conditional odds ratio* at Sites 1 and 2 are

$$\widehat{OR}_{XY(1)} = \frac{(40)(170)}{(30)(160)} = 1.42 \text{ and}$$

$$\widehat{OR}_{XY(2)} = \frac{(85)(100)}{(300)(15)} = 1.89.$$

The marginal OR is

$$\widehat{OR} = \frac{(125)(270)}{(330)(175)} = .58.$$

This illustrates the problems one can encounter with data exhibiting Simpson's paradox. In Section 3.2, we will discuss an alternative method of estimating the overall OR for stratified tables.

Marginal vs. Conditional Independence (Sec. 3.1.4)

Consider the true relationship between X and Y, controlling for Z. If X and Y are independent in each partial table, then X and Y are said to be *conditionally independent*, given Z. All conditional OR's between X and Y are then equal to 1. Note that conditional independence of X and Y, given Z, does not imply marginal independence of X and Y. That is, even if $OR_{XY(k)} = 1$ at every level of Z, the marginal OR may be different from 1.

Example (Table 3.2, p. 58)

Table 3.2 Conditional Independence Does Not Imply Marginal Independence

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

The expected cell frequencies in this table show a hypothetical relationship among 3 variables: Y = response (success/failure), X = drug treatment (A or B), and Z = clinic (1 or 2). The conditional associations between X and Y for each of the 2 levels of Z are given by:

$$OR_{XY(1)} = \frac{(18)(8)}{(12)(12)} = 1.0 \text{ and}$$

$$OR_{XY(2)} = \frac{(2)(32)}{(8)(8)} = 1.0 .$$

Thus, for each clinic, response and treatment are conditionally independent. However, the OR for the marginal table is

$$OR_{XY(2)} = \frac{(20)(40)}{(20)(20)} = 2.0 .$$

So, X and Y are *not* marginally independent.

How can we explain the results given above for the data in Table 3.2?

If a 3-way table exhibits unexpected results such as these for the X-Y association, it is also beneficial to examine the conditional OR's for X-Z and Y-Z. In order to calculate $OR_{XZ(j)}$ for j = 1, 2, it is helpful to rearrange Table 3.2 as follows:

		Clinic	
Response	Treatment	1	2
Success	A	18	2
	B	12	8
Failure	A	12	8
	B	8	32

So,

$$OR_{XZ(1)} = \frac{(18)(8)}{(12)(2)} = 6.0 \text{ and}$$

$$OR_{XZ(2)} = \frac{(12)(32)}{(8)(8)} = 6.0 .$$

In other words, the conditional odds (given the response) of receiving Treatment A are 6 times higher at Clinic 1 than at Clinic 2.

It can also be shown that, for these data, $OR_{YZ(1)} = OR_{YZ(2)} = 6.0$, i.e., the conditional odds of success (given the treatment) are 6 times higher at Clinic 1 than at Clinic 2.

These conditional odds tells us that Clinic 1 tends to use Treatment A more often and Clinic 1 also tends to have more successes than Clinic 2. So, for example, if patients who attend Clinic 1 tend to be in better health or tend to be younger than those who go to Clinic 2, perhaps they have a better success rate than subjects who go to Clinic 2, regardless of the treatment received.

Therefore, it is misleading to examine only the marginal table since the marginal odds ratio of 2.0 implies that Treatment A is better than Treatment B. Patients at a particular clinic tend to be more homogeneous than those in the overall sample, and, after controlling for clinic site using the conditional OR's, we see that Treatment A and Treatment B are equivalent in terms of response.

Homogeneous Association (Sec. 3.1.5)

We say that there is *homogeneous X-Y association* in a 2x2xK table when

$$OR_{XY(1)} = OR_{XY(2)} = \ \ldots = OR_{XY(K)}, \text{ i.e.,} \qquad (7)$$

the conditional OR between X and Y is identical at each level of Z.

Conditional independence occurs when each of the conditional OR's in Equation (7) is equal to 1.

For a general IxJxK table, homogeneous X-Y association means that any conditional OR calculated from a 2x2 table consisting of any 2 levels of X and any 2 levels of Y is the same, regardless of the level of Z.

Note that if the X-Y conditional OR's are identical at each level of Z, then the same property holds for the other conditional associations. So, for example, $OR_{XZ(j)}$ will be the same for j = 1, 2, …, J.

Thus, homogeneous association is a *symmetric* property in the sense that it holds regardless of which pair of variables are examined across the levels of the remaining variable. When homogeneous association is present, we say there is no *interaction* between any 2 of the variables in terms of their association with the remaining variable. In an Epi study, we would say that "there is no interaction between the exposure and the confounder with regard to outcome."

When homogeneous association is not present, the conditional OR for any pair of variables changes across the levels of the remaining variable.

Example

Suppose X = smoking (Yes/No),
Y = lung CA (Yes/No),
Z = age (< 45, 45-65, > 65).

Suppose $OR_{XY(1)} = 1.2$, $OR_{XY(2)} = 2.8$, $OR_{XY(3)} = 6.2$.

Clearly, homogeneous association is not present. Smoking is weakly associated with lung CA for younger subjects and this association strengthens considerably with increasing age.

In general, of course, the expected frequencies in the IxJxK table will be unknown. How do we determine from a sample of size n if homogeneous association is likely to be present in the parent population?