**BIOS 6244 Analysis of Categorical Data**
**September 28, 2005 Lecture**

**Probability Structures for Two-Way Contingency Tables (Section 2.1, pp. 16-19)**

Example

In an animal toxicology study, a comparison was made of a natural vs. synthetic chemical.

| | | Disease: Tumor | Disease: No Tumor |
|---|---|---|---|
| Exposure | Natural | 37 | 40 |
| | Synthetic | 212 | 138 |

Also see Table 2.1, p. 17.

Data arranged in this way are often called a *(two-way) contingency* (or *2x2*) *table*. There are several approaches that can be used to analyze data of this type.

Notation and Terminology

Let X and Y denote the 2 categorical variables we are interested in and suppose X has I categories (or levels) and Y has J categories (or levels).

Then there are I J possible combinations of categories and we display these in a rectangular array having I rows and J columns. Each combination of a row and column is called a *cell*, so there are I J cells in the rectangular array.

In general we refer to the rectangular array described above as an *IxJ contingency table*.

Probability Distributions for Contingency Tables

Suppose that each subject in the sample that is to be cross-classified is randomly chosen from some population of interest.

Let $\pi_{ij} = Pr(X = i, Y = j)$ denote the probability that (X,Y) falls into cell (i,j).

The collection of probabilities $\{\pi_{ij} \mid i = 1, \ldots, I; j = 1, \ldots, J\}$ form the joint *distribution* of X & Y. They must satisfy

$$\sum_{j=1}^{J}\sum_{i=1}^{I}\pi_{ij}=1$$

The marginal distribution of X & Y are defined by the row and column totals, respectively, of the joint probabilities:

$$\pi_{i+}=\sum_{j=1}^{J}\pi_{ij} \text{ and}$$
$$\pi_{+j}=\sum_{i=1}^{I}\pi_{ij} .$$

That is, we use a plus sign "+" to indicate that we have "summed out" the other variable.

For a 2x2 table:

| | | Y = 1 | Y = 2 | |
|---|---|---|---|---|
| X | 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
| | 2 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
| | | $\pi_{+1}$ | $\pi_{+2}$ | $\pi_{++} = 1$ |

For a sample of N cross-classified subjects, we replace π by p.

<u>cell proportions</u>

| | | Y = 1 | Y = 2 | |
|---|---|---|---|---|
| X | 1 | $p_{11}$ | $p_{12}$ | $p_{1+}$ |
| | 2 | $p_{21}$ | $p_{22}$ | $p_{2+}$ |
| | | $p_{+1}$ | $p_{+2}$ | $p_{++} = 1$ |

cell counts

| | | Y = 1 | Y = 2 | |
|---|---|---|---|---|
| X | 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| | | $n_{+1}$ | $n_{+2}$ | $n_{++} = N$ |

Note that

$$p_{ij} = \frac{n_{ij}}{n}$$

$$p_{+1} = \frac{n_{+1}}{n}, \text{ etc.}$$

We can also construct the *conditional distribution* for Y, given X (assuming Y is the outcome and X is the predictor). In other words, there is a separate distribution for Y corresponding to each value of X.

Example, revisited

| | | Disease: Tumor | Disease: No Tumor | |
|---|---|---|---|---|
| Exposure | Natural | 37 | 40 | 77 |
| | Synthetic | 212 | 138 | 350 |
| | | 249 | 178 | 427 |

The estimated *marginal probability* of getting a tumor is:

$$p_{21} = \frac{249}{427} = .583.$$

The estimated *joint probability* of being exposed to the synthetic chemical and getting a tumor is

$$p_{21} = \frac{212}{427} = .496.$$

The estimated *conditional probability* of getting a tumor, given exposure to the synthetic chemical is

$$\Pr(Y = 1 \mid X = 2) = p_{1|2} = \frac{212}{350} = .606.$$

The estimated *conditional probability* of getting a tumor, given no exposure to the synthetic chemical is

$$\Pr(Y = 1 \mid X = 1) = p_{2|1} = \frac{37}{77} = .481.$$

Note that, in our sample at least, the risk of getting a tumor does appear to be related to exposure to the synthetic chemical since .606 ≠ .481.

Also see Table 2.2, p. 18.

The categorical variables X & Y are said to be *independent* if the conditional distributions of Y are the same for all values of X, i.e., for fixed j, Pr(Y = j | X = i) is the same for all i.

Equivalently,

$\pi_{ij} = \pi_{i+} \pi_{+j}$ for all i = 1, …, I and j = 1, …, J.

where $\pi_{ij}$ = probability for cell (i,j)

$\pi_{i+}$ = probability for row i

$\pi_{+j}$ = probability for column j.

Use of Probability Models in Contingency Tables

Poisson: Each of the 4 cell counts $n_{11}$, $n_{12}$, $n_{21}$, $n_{22}$ are assumed to be independent Poisson random variables.

Binomial: Different probability models are used depending on whether the marginal totals are assumed to be *fixed* or *random*. For purposes of illustration,

suppose that the rows of the 2x2 table refer to different exposure groups (e.g., natural chemical vs. synthetic chemical).

(1) Fixed Marginal Totals

Suppose that the 427 animals in the example were randomly assigned *a priori* to be exposed to either the natural or the synthetic chemical. In this type of study, the row totals are treated as if they are *fixed*, since they were known in advance of determining tumor status.

We assume a binomial distribution in each row of the table, i.e.,

$$n_{11} \sim \text{bin}(n_{1+}, \pi_{1+})$$

$$n_{21} \sim \text{bin}(n_{2+}, \pi_{2+})$$

where bin(N, $\pi$) denotes a binomial distribution with # of trials = N and probability of success = $\pi$.

If the rows can be assumed to be independent (which they would be in a randomized design), we call this *independent binomial sampling.*

(2) Random Marginal Totals

Suppose now that the data in our example were based on a cross-classification at a single point in time (i.e., as in a cross-sectional study). In this case, the numbers of animals in each row were *not* known prior to constructing the 2x2 table. Now, a multinomial sampling model applies, in which the cell counts are all assumed to be random, i.e.,

$$n_{11}, n_{12}, n_{21}, n_{22} \sim \text{MN}(N, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}).$$

However, by a property of the multinomial distribution, the conditional distributions for the rows follow a binomial distribution:

$$n_{11}|n_{1+} \sim b(n_{1+}, \pi_{1+})$$

$$n_{21}|n_{2+} \sim b(n_{2+}, \pi_{2+}).$$

For most of the statistical procedures covered in our text, it does not matter whether we assume fixed or random margins since they yield the same results (p-values, CI's, etc.), regardless of which sampling model we assume. However, there are instances when the choice of an appropriate method of statistical inference does depend on the sampling model.