Describing Contingency tables

Today’s topics:

1. Probability structure for contingency tables (distributions, sensitivity/specificity, sampling schemes).

2. Comparing two proportions (relative risk, odds ratios).

3. $2 \times 2$-tables with covariates (marginal and conditional independence).

4. $I \times J$ tables (summary measures of association)

Sections skipped: none
Let $X$ and $Y$ be two categorical variables, $X$ with $I$ categories and $Y$ with $J$ categories. Classifications of subjects on both variables have $IJ$ combinations. The responses $(X, Y)$ of a random subject have a probability distribution.

A rectangular table having $I$ rows and $J$ columns displays the distribution. The cells of the table represent the $IJ$ outcomes. When the cells contain frequency counts, the table is called a contingency table or cross-classification table.

Example:

<table>
<thead>
<tr>
<th>Number of Classes</th>
<th>Solution Accuracy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Error</td>
</tr>
<tr>
<td>One Classes</td>
<td>80</td>
<td>251</td>
</tr>
<tr>
<td>Multiple Classes</td>
<td>84</td>
<td>66</td>
</tr>
<tr>
<td>Total</td>
<td>164</td>
<td>317</td>
</tr>
</tbody>
</table>
Let $\pi_{ij}$ denote the probability that $(X, Y)$ occurs in the cell in row $i$ and column $j$. The probability distribution \{\pi_{ij}\} is the joint distribution of $X$ and $Y$.

The marginal distribution are the row and column totals that result from summing the joint probabilities, denoted by \{\pi_{i+}\}.

Probabilities sum to one:

$$\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$$

When $X$ is an explanatory variable (a fixed variable) and $Y$ a response, the notion of a joint distribution is no longer meaningful. Given that a subject is in row $i$ of $X$, $\pi_{j|i}$ denotes the probability of classification in column $j$ of $Y$. The probabilities \{\pi_{1|i}, \ldots, \pi_{J|i}\} form the conditional distribution of $Y$ at category $i$ of $X$. Note that

$$\sum_j \pi_{j|i} = 1$$
Sensitivity & Specificity

Example:

<table>
<thead>
<tr>
<th>Disease</th>
<th>Diagnosis/Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.82</td>
<td>1.0</td>
</tr>
<tr>
<td>No</td>
<td>0.01</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity of a test is the conditional probability that a diagnostic test is positive given that the subject has the disease.

Specificity of a test is the conditional probability that a diagnostic test is negative given that the subject does not have the disease.
**Independence**

When both variables are response variables, descriptions of the association between the two can use the joint as well as the conditional distribution. In this case $\pi_{j|i} = \pi_{ij}/\pi_{i+}$.

Two categorical variables are called independent if the joint probabilities equal the product of marginal distribution

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

or when the conditional distributions are equal for different rows

$$\pi_{j|1} = \pi_{j|2} = \ldots, \pi_{j|I}, \text{ for } j = 1, \ldots, J$$

often called homogeneity of the conditional distributions.
Sampling distributions

The probability distributions introduced last week extend to contingency tables.

1. Poisson distribution treats cell counts as independent Poisson random variables with parameters $\mu_{ij}$. The joint probability mass function is then the product of the Poisson probabilities $P(Y_{ij} = n_{ij})$.

2. When the total sample size is fixed, but not the row or column totals, the multinomial sampling model applies.

3. When the row (or column) totals are fixed, or when the $X$ variable is explanatory and $Y$ response each row follows a multinomial model.

   When the different rows are independent, the joint probability distribution for the entire data set is the product multinomial sampling or independent multinomial sampling.
Types of study

1. **Case-control studies** use a retrospective design (look-into-the-past design). In this case $Y$ is fixed and we should condition on $\pi_j$.

2. **Clinical trials**: Subjects are assigned to treatments. $X$ is fixed. This is an experimental study (the others are observational studies).

3. **Cohort studies**: Subjects choose their own treatment and are followed over time.

4. **Cross-sectional design**: Subjects are classified simultaneously on both variables. Only the total sample size is fixed.
Comparing two proportions

Comparison of groups on a binary response variable, like success/failure, dead/alive,....

We confine ourselves to two groups, i.e. the table is of size $2 \times 2$. The conditional probability on success $\pi_{1|i}$ equals $1 - \pi_{2|i}$ and $\pi_i$ (the probability on success in group $i$) will be used instead of $\pi_{1|i}$.

1. **Difference of proportions**: $\Delta \pi = \pi_1 - \pi_2$ is a basic comparison of the rows. Possible values of $\Delta \pi$ are $-1 \leq \Delta \pi \leq 1$ and when $\Delta \pi = 0$ the response is statistically independent of the rows.

2. **Relative risk** $\pi_1/\pi_2$ can be any positive number. A relative risk of 1 indicates independence. This measure is useful in case a fixed difference of proportions is relatively more important for low or high $\pi$. 

Describing Contingency tables 2.8
Comparing two proportions- cont.

3 Odds ratio. For a probability of success $\pi_i$, the odds are defined by

$$\Omega = \pi_i/(1 - \pi_i).$$

The odds are nonnegative with a $\Omega > 1$ indicating that success is more likely. The ratio of odds

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11} \times \pi_{22}}{\pi_{12} \times \pi_{21}}$$

sometimes called cross-product ratio.
Properties of the odds ratio

1. The condition $\Omega_1 = \Omega_2$ and thus $\theta = 1$ refers to independence. When $\theta > 1$ subjects in row 1 are more likely to have success than are subjects in row 2.

2. The odds ratio is multiplicatively symmetric around 1, i.e. a value of 4 indicates the same amount of association as a value of 1/4. The log of the odds ratio is symmetric around 0, i.e. a value of -4 refers to the same amount of association as a value of 4. The log odds ratio is also convenient for inference (next week).

3. The odds ratio does not change when an entire row is multiplied by a constant.

4. The odds ratio is equally valid for prospective, retrospective or cross-sectional designs.

5. For cell counts $n_{ij}$ the sample odds ratio is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$
Example

Frequencies & Conditional proportions:

<table>
<thead>
<tr>
<th>Number of Classes</th>
<th>Solution Accuracy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Error</td>
</tr>
<tr>
<td>One</td>
<td>80</td>
<td>251</td>
</tr>
<tr>
<td>Multiple</td>
<td>84</td>
<td>66</td>
</tr>
<tr>
<td>Total</td>
<td>164</td>
<td>317</td>
</tr>
</tbody>
</table>

1. Difference of proportions $\Delta \pi = .24 - .56 = -.32$

2. Relative risk $\frac{.24}{.56} = .43$. The probability of success for those who did multiple classes is about two times as high as for those who followed one class.

3. Odds ratio

$$\hat{\theta} = \frac{80 \times 66}{84 \times 251} = .25$$

The odds are 4 times as high in the multiple class group compared to single class group.

Describing Contingency tables 2.11
Multiple 2 × 2-tables

An important part of most studies is the choice of control variables. In studying the relationship between $X$ and $Y$ one should control for any covariate that can influence this relationship. (think about suppression and spurious correlation)

We will study the relationship between $X$ and $Y$ while controlling for $Z$. This can be done by studying the relationship at fixed levels of $Z$. This is a single slice (two-way table) from a three-way table, called a partial table.

The table obtained by summing the partial tables is called the marginal table. In the marginal table $Z$ is ignored.

The associations in partial tables is called conditional association, since it is the association for fixed values of $Z$.

Conditional association can have a different direction than marginal association. This phenomenon is called Simpson's paradox.
Conditional and Marginal odds ratios

Let $\mu_{ijk}$ denote expected frequencies for some sampling model. We assume both $X$ and $Y$ are binary.

Within a fixed level $k$ of variable $Z$ the odds ratio

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

is called the conditional odds ratio.

Define $\mu_{ij+} = \sum_k \mu_{ijk}$, then

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

is called the marginal odds ratio.
Conditional and Marginal Independence

If \( X \) and \( Y \) are independent in the partial table \( k \), then \( X \) and \( Y \) are called conditionally independent at level \( k \) of \( Z \). When \( Y \) is a response:

\[
P(Y = j | X = i, Z = k) = P(Y = j | Z = k), \text{ for all } i, j
\]

When this is true for all \( k \), \( X \) and \( Y \) are called conditionally independent given \( Z \).

Suppose a multinomial sampling scheme applies with joint probabilities \( \pi_{ijk} = P(X = i, Y = j, Z = k) \).

Conditional independence is equal to

\[
\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}
\]

Marginal independence is equal to

\[
\pi_{ij+} = \pi_{i++}\pi_{+j+}
\]

Conditional independence does not imply marginal independence, nor the other way around!
Homogeneous Association

A $2 \times 2 \times K$ table has homogeneous association when

$$\theta_{XY}(1) = \theta_{XY}(2) = \cdots = \theta_{XY}(K)$$

(Conditional independence is a special case.)

If

$$\theta_{XY}(1) = \theta_{XY}(2) = \cdots = \theta_{XY}(K)$$

then also

$$\theta_{XZ}(1) = \theta_{XZ}(2) = \cdots = \theta_{XZ}(J)$$

Homogeneous association is a symmetric property. When it occurs it is said that there is no interaction between two variables in their effects on the other variable.

If an interaction exists, the conditional odds ratio is different for each level of the covariate and the covariate is said to be an effect modifier.
I × J tables

For 2 × 2 tables a single odds ratio can summarize the association. For I × J tables this is not possible, however a set of odds ratios or another summary index can describe features of the association.

For rows a and b and columns c and d the odds ratio \( \frac{\pi_{ac}\pi_{bd}}{\pi_{ad}\pi_{bc}} \) uses four cells in a rectangular pattern. For I × J tables there are many of such odds ratios. However, this set contains much redundant information.

**Basic sets**

The local odds ratios defined by

\[
\theta_{ij}^{(l)} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{ij+1}\pi_{i+1,j}} , \quad 1 \leq i \leq (I - 1), \quad 1 \leq j \leq (J - 1)
\]

The spanning cell odds ratios (here with the cell \( i = 1, j = 1 \) as spanning cell)

\[
\theta_{ij}^{(s)} = \frac{\pi_{11}\pi_{ij}}{\pi_{1j}\pi_{i1}} , \quad 2 \leq i \leq I, \quad 2 \leq j \leq J.
\]
Summary measures

These measures summarize the association in a single number.

1. Summary measures of association. Various forms exist but in general it is difficult to understand them.

2. Ordinal trends: Concordant and discordant pairs. A pair of subjects is concordant if the subject ranked higher on $X$ also ranks higher on $Y$. The pair is discordant if the subject ranked higher on $X$ ranks lower on $Y$. Such measures are only valid for ordinal data

3. Ordinal trends: gamma