## Loglinear Models and Other Approaches

• Many tests for contingency tables use the "Pearson's Chi-square Statistic":

• An alternative approach uses the "Likelihood Ratio Chi-square Statistic":

• The LR statistic also has an asymptotic $\chi^2$ distribution, with the same degrees of freedom as Pearson's statistic.

• An advantage of the Pearson test statistic is that its asymptotic $\chi^2$ distribution tends to be valid with smaller sample sizes (i.e., when ___________ ) than the $\chi^2$ approximation for the LR statistic (which holds well when ______________ ).

## Loglinear Models

• This is a common method of analyzing contingency tables of more than two dimensions.

• In a 2 × 2 table, the null hypothesis of independence between dimensions is equivalent to

where $\pi_{i+}$ =

and $\pi_{+j}$ =

• **Taking logarithms of both sides, we get:**

**which is a ____________________ model.**

**<u>Recall</u>: Our expected cell count under independence is**

**where $n_{i+}$ =**

**and $n_{+j}$ =**

• **Thus for a 2 × 2 table,**

**and so we have**

• **This fraction is called the <u>odds ratio</u>.**

**It is defined as**

• **Now, if we instead have <u>dependence</u> between dimensions, that implies:**

• Writing the loglinear model in terms of the cell counts rather than cell probabilities, we have:

under independence

under dependence

• These model parameters are estimated using software via iterative methods.

• Using the estimates, we can get fitted values for each cell.

• We then use either the Pearson statistic or the LR statistic to determine (with a $\chi^2$ test) whether the model provides a good fit. $H_0$:

## Three-Way Tables

• This is most useful in cases where the data are classified according to three categorical variables.

Example 1 (2 × 2 × 2 table):

Possible loglinear models for 2 × 2 × 2 tables:

**Example 1: Let $i$ = 1, 2 be the level of Cigarette Use (Yes/No); let $j$ = 1, 2 be the level of Marijuana Use; let $k$ = 1, 2 be the level of Alcohol Use.**

**• The model that includes all possible parameters is called the __________________ model.**

**• The `loglm` function in the `MASS` library in R estimates the parameters of any of these models, calculates the fitted values, and performs the $\chi^2$ tests for fit.**
**• In addition, the `step` function evaluates these possible models based on Akaike's Information Criterion (AIC).**

**Example 1 Possible Questions of Interest:**

**• Do the odds of a cigarette smoker using marijuana differ from the odds of a cigarette non-smoker using marijuana? →**

**• Does the value of this odds ratio depend on alcohol use? →**

**<u>Analysis in R</u>:**

**• The best model appears to be**

**• Example of fitted value calculation using estimated coefficients:**

**• Interpretation of results is best done using odds ratios:**

**Example 2 (2 × 2 × 2 table):**

## Example 2 Possible Questions of Interest:

**• Do the odds of an early plant surviving differ from the odds of a late plant surviving? →**

**• Does the value of this odds ratio depend on the cutting length? →**

**<u>Analysis in R</u>:**
**• The search for the best model:**

**• Interpretation of results via odds ratios:**

**Example 3 ($2 \times 2 \times 6$ table): A study classified UC-Berkeley graduate school applicants according to Admission Status (Admitted/Rejected), Sex (Male/Female), and Department (A/B/C/D/E/F). We adapt a built-in R data set.**

**Example 3 Possible Questions of Interest:**

**• Do the odds of a female being admitted differ from the odds of a male being admitted? →**

**• Does the value of this odds ratio depend on the department to which the applicant applies? →**

**<u>Analysis in R</u>:**

**• The search for the best model:**

**• Interpretation of results via odds ratios:**

**• This example illustrates something similar to what is known as Simpson's Paradox:**

**• Simpson's Paradox occurs when an association between two categorical variables is evident when the data are aggregated over some third categorical variable, but this association is reversed when the data are examined separately at each category of the third variable.**

**• In the Berkeley data, the aggregated data indicate that males are more likely to be admitted than females.**

**• But within each department, females are roughly as likely (or more likely) to be admitted as males.**

**Another example (Accident victim data):**

**Aggregated Data:**

**• Based on this, helicopter transports yield a 32% death rate, while road transports yield a 24% death rate. Which mode of transportation is better?**

**Separated Data:**

**• For serious accidents: helicopter transports yield a 48% death rate, while road transports yield a 60% death rate.**

**• For less serious accidents: helicopter transports yield a 16% death rate, while road transports yield a 20% death rate.**

**Which mode of transportation is better?**

**• In the aggregated data set, "seriousness of injury" is a <u>lurking variable</u>, which can conceal the actual association.**