STA 303 / 1002

Note Title 3/15/2012

$I \times J$ Contingency table analysis:

Framingham example: 2×2 table

Analysis 1: Only for 2×2 tables
Assume row totals fixed
Underlying probability distribution: Binomial
2 sample test for proportions
$H_0: \pi_1 = \pi_2$

Not a linear model

<u>Analysis 2</u> Assume overall total fixed

$H_0: \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$, $i = 1, \dots, I$, $j = 1, \dots, J$

2a) "Pearson's" Chisquare test

2b) LRT based on underlying probability distribution: multinomial

Not a linear model

<u>Analysis 3</u> No totals fixed:

Underlying probability distribution: Poisson

Use Poisson regression

Is A (log) <u>linear</u> model

H_0: independence of row & column variables is equivalent to whether Poisson model without the interaction term fits as well as the model with the interaction term

If interaction is needed:
The mean count of people with CVD depends on their cholesterol status

Loglinear Model of Independence

$I \times J$ contingency table, cross-classifies observations ~~into 2 categories~~ 2 ways.

If ways that they are being classified are independent; probability of being in cell (i, j)

is $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$, $i = 1, \dots, I$, $j = 1, \dots, J$

Expected number of observations in each cell is:

$$\mu_{ij} = n\pi_{ij} = \underbrace{n\pi_{i\cdot}\pi_{\cdot j}}_{\text{if independence}}$$

Take log of both sides:

$$\log(\mu_{ij}) = \underbrace{\log n + \log \pi_{i\cdot} + \log \pi_{\cdot j}}_{\text{additive}}$$

Distinguishing feature of log-linear models, row & column variables are treated equally; It's not the case that one is the response and the other is the explanatory variable

Framingham example

Model with interaction: "Full" model, "Saturated" model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 I_{[chol=H]} + \beta_2 I_{[CVD=absent]} + \beta_3 I_{[chol=H]} * I_{[CVD=absent]}$$

Fits data perfectly because # parameters (β's) = # observed counts

Model without interaction "Reduced" model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 I_{[\text{chol}=H]} + \beta_2 I_{[\text{CVD}=\text{absent}]}$$

Use LRT to compare these 2 models

Test statistic = 26.43
$= G^2$

3 ways to get this:

1. Deviance for reduced model (since full model is saturated)
2. Subtract full log likelihood and multiply by 2

(3) Subtract log likelihood and multiply by 2

H_0: $\beta_3 = 0$
Chol & CVD are independent

H_a: $\beta_3 \neq 0$
Chol & CVD not independent

Under H_0, G^2 is an observation from chi-square (1) distribution

$df = 1 = 4 - 3$ (diff in # of parameters in 2 models.)

p-value <u>very</u> small

Strong evidence that chol & CVD are not independent

Forms of independence of 3 events

- Events A, B are jointly independent of C if

$$P(AB \mid C) = P(AB)$$

(joint distribution of A & B is unaffected by C)

- A, B, C are mutually independent iff

$$\text{all of}\left\{\begin{array}{l}\left.\begin{array}{l}P(AB) = P(A)P(B)\\ P(AC) = P(A)P(C)\\ P(BC) = P(B)P(C)\end{array}\right\}\text{pairwise independence}\\ P(ABC) = P(A)P(B)P(C)\end{array}\right.$$

- A, B are <u>conditionally independent given</u> C if

$$P(AB \mid C) = P(A \mid C)\,P(B \mid C)$$