STA 303 | 1002

Note Title 3/12/2012

Assignment 2 is ready!

The TA will be in RW 107/109 to help with the assignment on Mon. March 19 2:00 - 4:00 pm
Fri. March 23 2:00 - 4:00 p.m

Note: The Cquest server has moved. If you logon remotely (ssh) you may get an error about spoofing. Follow the instructions in the announcement on Blackboard.

Framingham example

Data:

Chol (mg/dL)	CVD present	CVD absent	Total
High: ≥ 260	41	245	286 $= n_H$
Low: < 260	51	992	1043 $= n_L$
Total	92	1237	1329

$H_0: \pi_H = \pi_L$ where π_H is prob. of developing CVD in high cholesterol group

$H_a: \pi_H \neq \pi_L$

Analysis based on binomial distribution

Assumptions: 286, 1043 are fixed, & people independent

$\hat{\pi}_H = 41/286$, $\text{Var}(\hat{\pi}_H) = \frac{\pi_H(1-\pi_H)}{n_H}$

$\hat{\pi}_L = 51/1043$, $\text{Var}(\hat{\pi}_L) = \frac{\pi_L(1-\pi_L)}{n_L}$

Test statistic: $\frac{\hat{\pi}_H - \hat{\pi}_L}{\text{s.e. of } (\hat{\pi}_H - \hat{\pi}_L)}$

$\text{Var}(\hat{\pi}_H - \hat{\pi}_L) = \text{Var}(\hat{\pi}_H) + \text{Var}(\hat{\pi}_L)$

Under H_0, $\pi_H = \pi_L$ so estimate them by

$$\hat{\pi}_{combined} = 92/1329$$

Under H_0, $S.E.\left(\hat{\pi}_H - \hat{\pi}_L\right) = \sqrt{\frac{92/1329\,(1-92/1329)}{286} + \frac{92/1329\,(1-92/1329)}{1043}}$

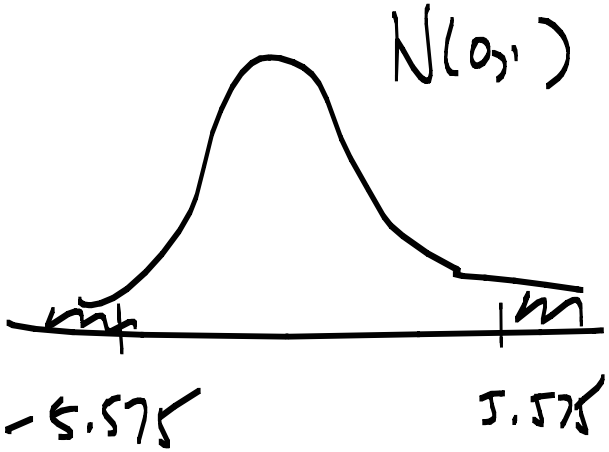
For large samples, proportions are approximately normally distributed (CLT)

So test statistic is approx N. dist'd under H_0

For these data, the test statistic is 5.575

The p-value is very small

Very strong evidence that probability of developing CVD is different for high vs low cholesterol groups



Underlying probability distribution: Binomial

"Product binomial" or "Binomial sampling" approach

ANALYSIS 2 Assume sampled $n = 1329$ people (considered fixed),

classified them 2 ways

Cholesterol status (H or L)

CVD status (present or absent)

Let C denote Cholesterol status
D denote disease status

} Both categorical r.v.'s (with 2 categories)

In general, row factor with I levels
column factor with J levels

π_{ij} = probability an observation falls into row i, column j

$= Pr(C = i, D = j)$

gives joint distribution of C & D

Marginal distributions

$P(C = i) = \pi_{i\cdot}$ = probability an observation falls into row i

$P(D = j) = \pi_{\cdot j}$ = " " " column j

If there is no relationship between C and D, C, D are independent

That is, $\pi_{ij} = \pi_{i\cdot}\,\pi_{\cdot j}$

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

vs $H_a: \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$

Let y_{ij} = count in row i, column j

$y_{i\cdot}$ = total count in row i, $y_{i\cdot} = \sum_{j=1}^{J} y_{ij}$

$y_{\cdot j}$ = " " " column j $= \sum_{i=1}^{I} y_{ij}$

$n = \sum_j \sum_i y_{ij}$ = grand total

Under H_0, $\hat{\mu}_{ij}$ = estimate of expected count in cell (i,j)

$$\hat{\mu}_{ij} = n\,\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}$$

$$= n \frac{y_{i\cdot}}{n} \frac{y_{\cdot j}}{n} = \frac{y_{i\cdot} y_{\cdot j}}{n}$$

Test statistic

$$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

For large samples, under H_0 X^2 has approximately a chi-square distribution with df $(I-1)(J-1)$

Why $(I-1)(J-1)$ df?

IJ observations

Restrictions based on number of estimates we

make to calculate test statistic

To Est $\hat{\mu}_{ij}$, need $y_{i\cdot}$, I-1 estimates (I rows, must sum to n)

$y_{\cdot j}$, J-1 estimates

n fixed.

Restriction on df

$$df = IJ - 1 - (I-1) - (J-1)$$
$$= (I-1)(J-1)$$

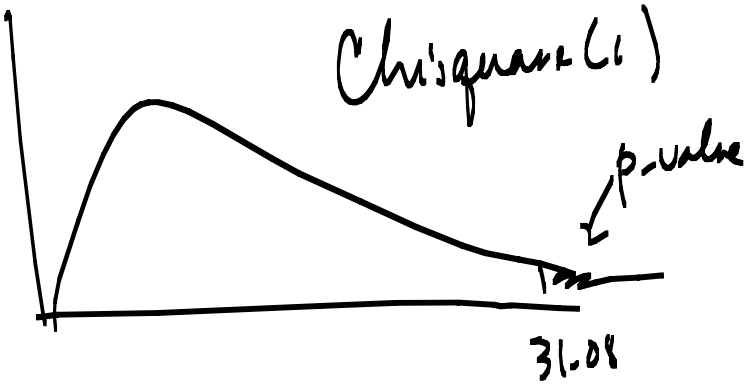
SAS: proc freq; ("freq" short for frequency)

For this analysis: Chi-square statistic

For example $df = 1$ $(J = I = 2)$

$X^2 = 31.08$

pvalue $< .0001$



Strong evidence that C & D are not independent

So CVD status depends on cholesterol status

Exercise: Show X^2 equal to square of test statistic in Analysis 1 when $I = J = 2$

Analysis 2b $n = 1329$ observations (fixed)
classified 2 ways

More formal approach based on MLE & LRTs:

Let Y_{ij} be a r.v. representing the number of observations in row i, column j of table

Observe y_{ij}

For $I = J = 2$ $\underset{\sim}{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22})$

Multinomial distribution:

$$P(\underset{\sim}{Y} = \underset{\sim}{y}) = \frac{n!}{y_{11}! y_{12}! y_{21}! y_{22}!} \pi_{11}^{y_{11}} \pi_{12}^{y_{12}} \pi_{21}^{y_{21}} \pi_{22}^{y_{22}}$$

Underlying probability distribution: Multinomial

Number of ways of distributing n observations so that y_{11} are in row 1, col 1, ...

with $y_{11} + y_{12} + y_{21} + y_{22} = n$

and $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$

log-likelihood: $\log L = \sum_{j=1}^{J} \sum_{i=1}^{I} y_{ij} \log(\pi_{ij}) + \log$ combinatorial term

SAS would call log-likelihood

SAS would call full log likelihood

Maximum likelihood estimates: $\hat{\pi}_{ij} = \frac{y_{ij}}{n}$

(found by maximizing $\log L$ w.r.t. $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ subject to constraint $\sum\sum \pi_{ij} = 1$)

Under assumption of independence: $\pi_{ij} = \pi_{i.}\pi_{.j}$

Substitute $\pi_{ij} = \pi_{i.}\pi_{.j}$ into $\log L$ and maximize wrt $\pi_{1.}, \pi_{2.}, \pi_{.1}, \pi_{.2}$

Subject to constraints $\pi_{1.} + \pi_{2.} = 1$ and $\pi_{.1} + \pi_{.2} = 1$

Solution MLEs under assumption of independence:

$$\hat{\pi}_{1\cdot} = \frac{y_{1\cdot}}{n}, \quad \hat{\pi}_{2\cdot} = \frac{y_{2\cdot}}{n}$$

$$\hat{\pi}_{\cdot 1} = \frac{y_{\cdot 1}}{n}, \quad \hat{\pi}_{\cdot 2} = y_{\cdot 2}/n$$

Then $\hat{\pi}_{12} = \hat{\pi}_{1\cdot}\hat{\pi}_{\cdot 2}$, etc. $\frac{}{n}$,

Leads to same expected counts as X^2

Likelihood ratio test to compare multinomial model under assumption of independence to model without this assumption:

↑ "reduced model"

↑ "full model"

Test stat: $G^2 = -2\log\left(\frac{L_R}{L_F}\right)$

$= 2\log(L_F) - 2\log(L_R)$ ← $\hat{\pi}_{ij}$ for full model

$= 2\left\{\sum_j \sum_i y_{ij} \log\left(\frac{y_{ij}}{n}\right)\right.$

$\left. - \sum_j \sum_i y_{ij} \log\left(\frac{y_{i\cdot}}{n}\frac{y_{\cdot j}}{n}\right)\right\}$

↑ $\hat{\pi}_{ij}$ for reduced model

$= 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right)$

where $\hat{\mu}_{ij}$ is expected count in cell i,j under assumption of independence

Under H_0: ($H_0: \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ vs $H_a: \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$, $i=1,\dots,I$; $j=1,\dots,J$)

G^2 has a chisquare distribution with $df = (I-1)(J-1)$

Why $(I-1)(J-1)$ df?

Unrestricted (full) model: number of parameters (π_{ij}'s) is $IJ-1$ (lose 1 df because $\sum\sum \pi_{ij} = 1$)

reduced model (independence): number of parameter

(parameters are $\pi_{i\cdot}$, $\pi_{\cdot j}$'s)

((lose 2 df because $\sum_i \pi_{i\cdot} = 1$, $\sum_j \pi_{\cdot j} = 1$)

so # of parameters is $I + J - 2$

df for test statistic in LRT is

$$IJ - 1 - (I + J - 2) = (I-1)(J-1)$$

SAS proc freq, G^2 = 'likelihood ratio chi-square'

For our example, $G^2 = 26.4298$

$df = 1$

p-value < 0.0001

Strong evidence that row and column variable (cholesterol & disease status) are not independent

[Not responsible for numbers from proc freq output below likelihood ratio chi-square]

What is the relationship between CVD & cholesterol?
For people with high cholesterol, 14.3% developed CVD
For people with low cholesterol, 4.89% developed CVD

Analysis 3 No fixed counts

Treat the IJ counts as realizations of independent Poisson random variables.

The joint of counts is $P(\underset{\sim}{Y} = \underset{\sim}{y}) = \prod_j \prod_i \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!}$

Log-likelihood function:

$$\log L = \sum_j \sum_i \left(y_{ij} \log(\mu_{ij}) - \mu_{ij} - \log(y_{ij}!) \right)$$

"full log-likelihood"

Use Poisson Regression (proc genmod in SAS)

"Log-linear" models