

## Two-way tables: tests of independence

Situation: *one* multinomial sample of size  $n$ , two categorical variables recorded from each sample member, sample members cross-tabulated according to responses

ex. General Social Survey 1982

		death penalty opinion		
		F	O	
gun registration opinion	F	784	236	1020
	O	311	66	377
		1095	302	1397

In general, the categorical variables might have more than two possible responses. The data can be summarized in a two-way table with  $r$  rows and  $c$  columns:

		Variable B				
		1	2	...	$c$	
Variable A	1	$y_{11}$	$y_{12}$	...	$y_{1c}$	$n_{1.}$
	2	$y_{21}$	$y_{22}$	...	$y_{2c}$	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	$r$	$y_{r1}$	$y_{r2}$	...	$y_{rc}$	$n_{r.}$
		$n_{.1}$	$n_{.2}$	...	$n_{.c}$	$n$

Model:  $Y_{ij}$ 's have a multinomial distribution, with sample size  $n$  and probabilities  $\pi_{ij}$ , where

$$\sum_{i=1}^r \sum_{j=1}^c Y_{ij} = n \quad \text{and} \quad \sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$$

$H_0$ :  $\pi_{ij} = \phi_i \psi_j$  (var A and var B are **independent**)

where  $\phi_1, \phi_2, \dots, \phi_r$  are the marginal probabilities for variable A, and  $\psi_1, \psi_2, \dots, \psi_c$  are the marginal probabilities for variable B. The  $\phi_i$ 's sum to 1, and the  $\psi_j$ 's sum to one, so that the number of unknown parameters in  $H_0$  is  $(r - 1) + (c - 1)$ .

ML estimates of parameters under the null hypothesis:

$$\hat{\phi}_i = \frac{n_{i\cdot}}{n}, \quad \hat{\psi}_j = \frac{n_{\cdot j}}{n}$$

$$\hat{\pi}_{ij} = \hat{\phi}_i \hat{\psi}_j = \left(\frac{n_{i\cdot}}{n}\right) \left(\frac{n_{\cdot j}}{n}\right)$$

ML estimates of expected values of the  $Y_{ij}$ 's under  $H_0$ :

$$\hat{E}_{ij} = n \hat{\pi}_{ij} = n \left(\frac{n_{i\cdot}}{n}\right) \left(\frac{n_{\cdot j}}{n}\right) = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

$H_a$ : all  $rc - 1$  parameters ( $\pi_{ij}$ 's) are needed to describe the data (var A and var B are **dependent** or **associated**)

Test statistic:

$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^c y_{ij} \log_e \left( \frac{y_{ij}}{\hat{E}_{ij}} \right) \quad \text{or}$$

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(y_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

Rejection region:

reject  $H_0$  if  $G^2$  (or  $X^2$ )  $\geq \chi_{\alpha}^2$ , where the chi-square distribution has

$$rc - 1 - (r - 1 + c - 1) = (r - 1)(c - 1)$$

degrees of freedom

Note: in LR test statistic, if  $y_{ij} = 0$ , then that term in the sum equals zero:

$$y_{ij} \log_e \left( \frac{y_{ij}}{\hat{E}_{ij}} \right) = 0$$

Note: many measures of amount or degree of association have been devised ( $\lambda$ , etc), which attempt to summarize numerically how strongly related are var A and var B (like a correlation coefficient for quantitative variables)

## **Multinomial models in SAS**

1. ML estimates & goodness of fit for a reduced parameter multinomial: PROC NLIN (using the *iteratively reweighted least squares algorithm*)
2. Two-way tables and tests of independence: PROC FREQ
3. Product-multinomial models (i.e. more than one multinomial sample) & tests of homogeneity of proportions:  
PROC CATMOD
4. Multiple categorical variables (loglinear models, logit models): PROC CATMOD or PROC GENMOD

## **Learning more about categorical data analysis**

Stat 420 (WSU) Statistical analysis of qualitative data

Stat 514 (ID) Nonparametric statistics (final weeks)