ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Visualizing categorical data & inference

Applied Multivariate Statistics – Spring 2013

Goals

- Chi-Square test of independence
- R: mosaic plot, cotabplot (with shading)

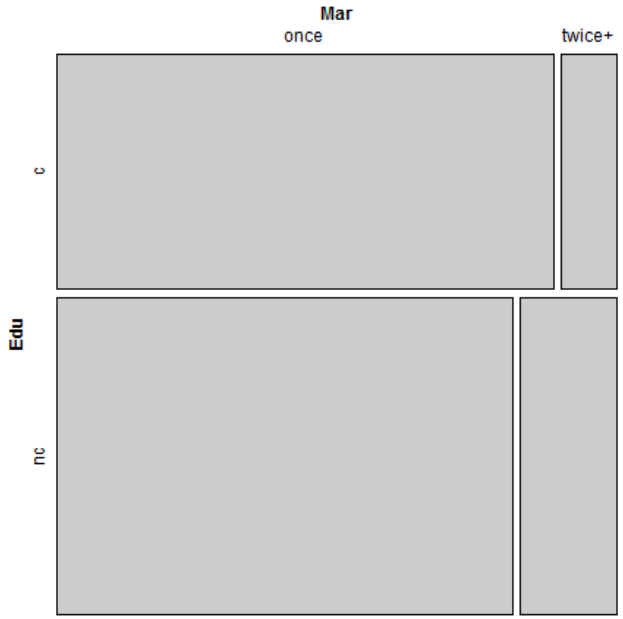
Start simple: Two binary variables

- Education and Marriage (Kiser and Schaefer, 1949)

Education	Married Once	Married More Than Once	Total
College	550	61	611
No College	681	144	825
Total	1231	205	1436

- Two questions:
  - How to visualize (esp. if more than two variables)?
  - Dependence? Why?

Visualizing categorical data: Mosaic Plot



Education	Married Once	Married More Than Once	Total
College	550	61	611
No College	681	144	825
Total	1231	205	1436

Area proportional to table entry

"observed values"
$O_{ij} = n_{ij}$

Chi-Square Test of Independence

	A=1	A=2	Total
B=1	n11	n12	n1*
B=2	n21	n22	n2*
	n*1	n*2	n

H_0: A and B are independent; therefore

$$P(A = i \cap B = j) \overset{=}{} P(A = i) \cdot P(B = j) \approx \hat{P}(A = i) \cdot \hat{P}(B = j) =$$
$$= \frac{n_{\cdot i}}{n} \cdot \frac{n_{j\cdot}}{n} = \hat{\pi}_{ij}$$

Expected values in cells if H_0 is true: $E_{ij} = n \cdot \hat{\pi}_{ij}$

Chi-Square Test of Independence

	A=1	A=2	Total
B=1	n11	n12	n1*
B=2	n21	n22	n2*
	n*1	n*2	n

How different are observed and expected values?
Most popular: ***Pearson*** Chi-Square Statistics

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I} \sum_{j=1}^{J} R_{ij}^2$$

Pearson Residuals

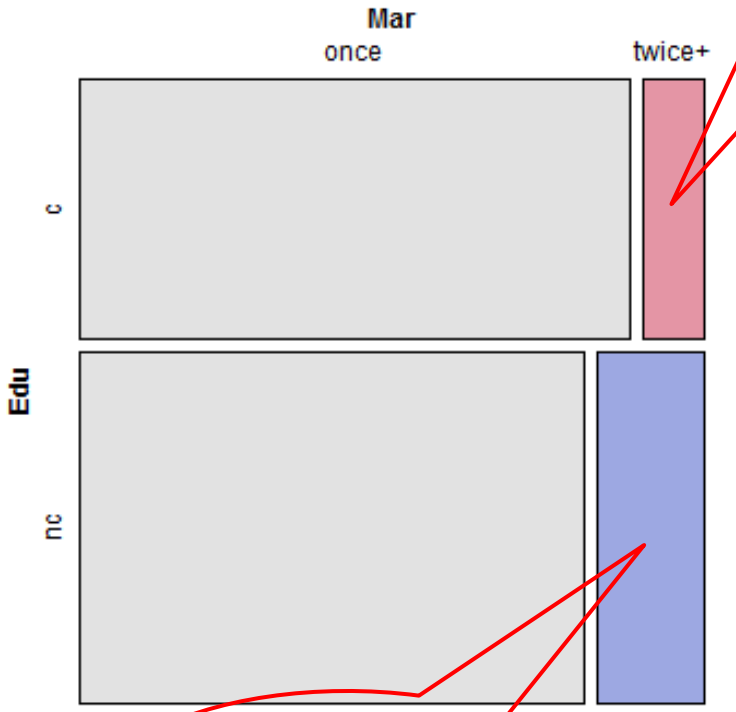
$$R_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

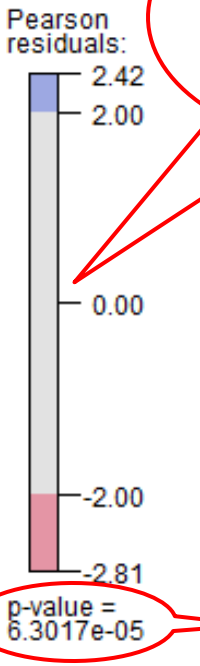
Contribution of each cell to misfit

If H_0 is true, X^2 follows a Chi-Square distribution with (I-1)(J-1) degrees of freedom (if n large and no empty cells)
Thus, can compute p-values.

Alternative: Permutation test; more computer intensive but more precise

Mosaic plot with shading





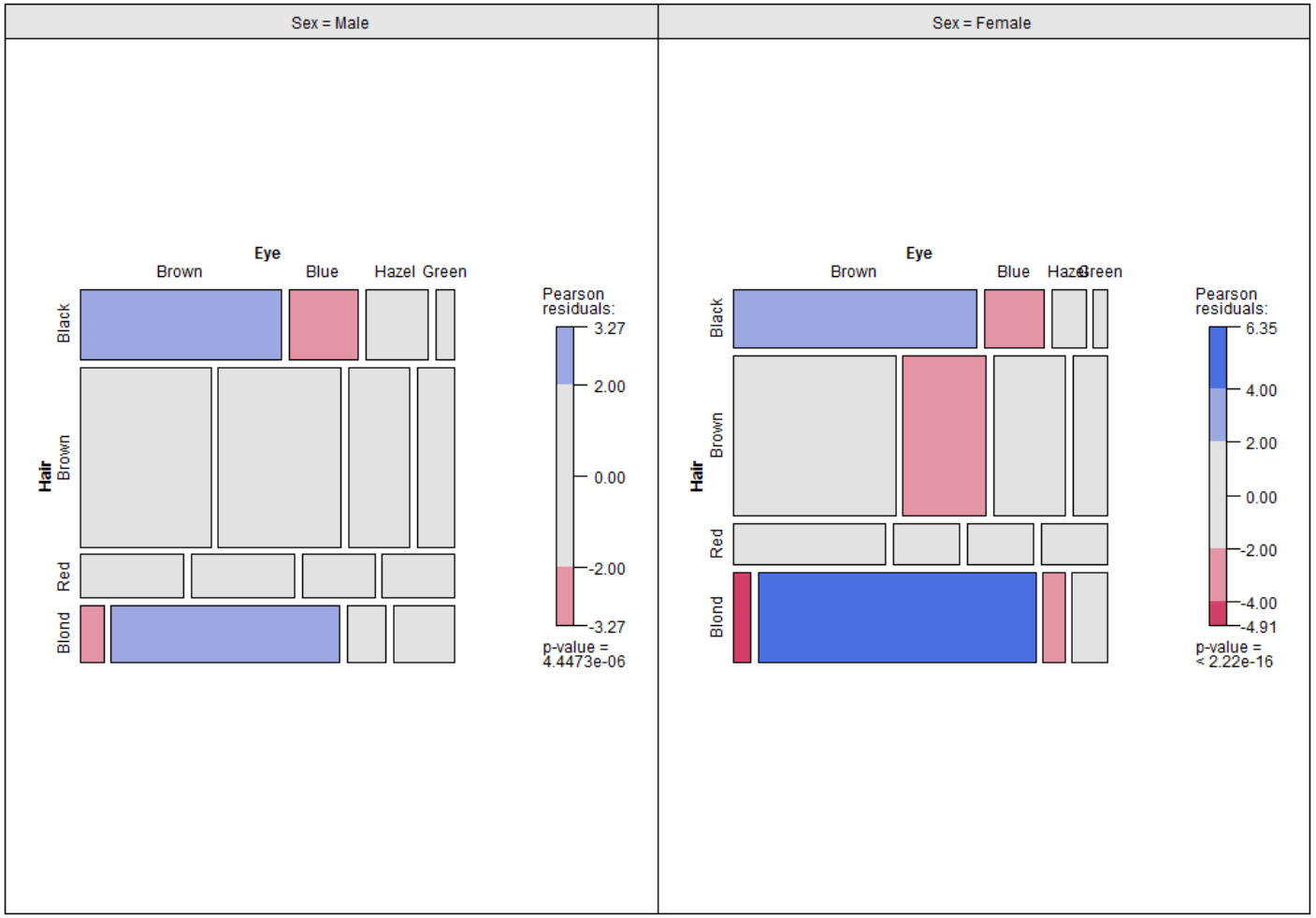
Suprisingly small observed cell count

Use color if Pearson residual is outside [-2, 2]

Suprisingly large observed cell count

p-value of independence test: Highly significant

Conditional plots: Mosaic plot per group



Case study: Admission UC Berkeley

Concepts to know

- Chi-Square test of independence

R commands to know

- mosaic (with shading)
- Cotabplot (with shading)
  (both in package “vcd”)