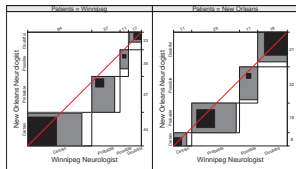
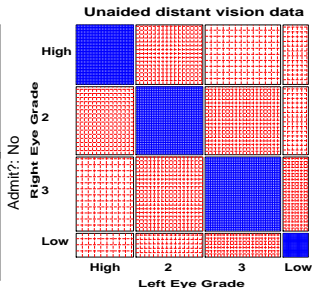
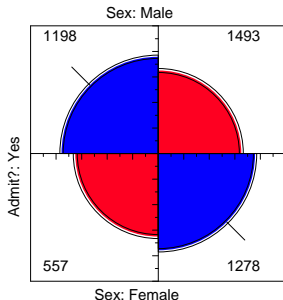


# Two-way tables: Independence and association

Michael Friendly

Psych 6136

September 21, 2017



# Two-way tables: Overview

Two-way contingency tables are a convenient and compact way to represent a data set cross-classified by two discrete variables,  $A$  and  $B$ .

## Special cases:

- $2 \times 2$  tables: two binary factors (e.g., gender, admitted?, died?, ...)
- $2 \times 2 \times k$  tables: a collection of  $2 \times 2$ s, stratified by another variable
- $r \times c$  tables
- $r \times c$  tables, with **ordered** factors

## Questions:

- Are  $A$  and  $B$  statistically **independent**? (vs. **associated**)
- If associated, what is the **strength** of association?
- Measures:  $2 \times 2$ — odds ratio;  $r \times c$ — Pearson  $\chi^2$ , LR  $G^2$
- How to understand the **pattern** or **nature** of association?

## Two-way tables: Examples

$2 \times 2$  table: Admissions to graduate programs at U. C. Berkeley

Table: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total	% Admit	Odds(Admit)
Males	1198	1493	2691	44.52	0.802
Females	557	1278	1835	30.35	0.437
Total	1755	2771	4526	38.78	0.633

Males were nearly twice as likely to be admitted.

- Association between gender and admission?
- If so, is this evidence for gender bias?
- How do characterise strength of association?
- How to test for significance?
- How to visualize?

## 2 × 2 tables: UCB data

In R, the data is contained in `UCBAdmissions`, a  $2 \times 2 \times 6$  table for 6 departments. Collapse over department:

```
data(UCBAdmissions)
UCB <- margin.table(UCBAdmissions, 2:1)
UCB
```

##		Admit	
##	Gender	Admitted	Rejected
##	Male	1198	1493
##	Female	557	1278

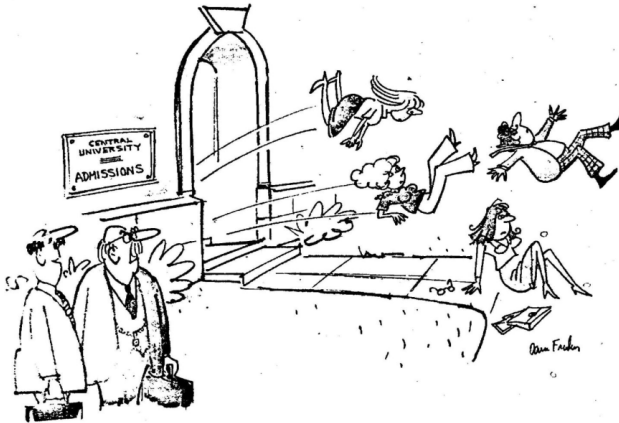
Association between gender and admit can be measured by the **odds ratio**, the ratio of odds of admission for males vs. females. Details later.

```
oddsratio(UCB, log=FALSE)
```

##	odds ratios for Gender and Admit	
##		
##	[1]	1.8411

```
confint(oddsratio(UCB, log=FALSE))
```

##		2.5 %	97.5 %
##	Male:Female/Admitted:Rejected	1.6244	2.0867



"YES, ON THE SURFACE IT WOULD APPEAR TO BE SEX-BIAS  
BUT LET US ASK THE FOLLOWING QUESTIONS..."

- How to analyse these data?
- How to visualize & interpret the results?
- Does it matter that we collapsed over Department?

## Two-way tables: Examples

$r \times c$  table: Hair color and eye color— Students in a large statistics class.

Table: Hair-color eye-color data

Eye Color	Hair Color				Total
	Black	Brown	Red	Blond	
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

- Association between hair color and eye color?
- How do characterise strength of association?
- How to test for significance?
- How to visualize?
- How to interpret the **pattern** of association?

## $r \times c$ tables: HEC data

In R, the data is contained in `HairEyeColor`, a  $4 \times 4 \times 2$  table for males and females. Collapse over gender:

```
data(HairEyeColor)
HEC <- margin.table(HairEyeColor, 2:1)
```

Association between hair and eye color can be tested by the standard Pearson  $\chi^2$  test. Details later.

```
chisq.test(HEC)

##
##  Pearson's Chi-squared test
##
## data:  HEC
## X-squared = 138, df = 9, p-value <2e-16
```

## Two-way tables: Examples

$r \times c$  table with ordered categories: Mental health and parents' SES

Table: Mental impairment and parents' SES

SES	Mental impairment			
	Well	Mild	Moderate	Impaired
1	64	94	58	46
2	57	94	54	40
3	57	105	65	60
4	72	141	77	94
5	36	97	54	78
6	21	71	54	71

- Mental impairment is the response, SES is the predictor
- How do characterise strength of association?
- How to interpret the **pattern** of association?
- How to take **ordinal** nature of the variables into account?



## ordered $r \times c$ tables: Mental data I

In R, the data is contained in `Mental` in `vcdExtra`, a frequency data frame.

```
data(Mental, package="vcdExtra")
str(Mental)

## 'data.frame': 24 obs. of 3 variables:
## $ ses : Ord.factor w/ 6 levels "1"<"2"<"3"<"4"<...: 1 1 1 1 2
## $ mental: Ord.factor w/ 4 levels "Well"<"Mild"<...: 1 2 3 4 1 2
## $ Freq : int 64 94 58 46 57 94 54 40 57 105 ...
```

Convert to a contingency table using `xtabs()`, and test association:

```
mental.tab <- xtabs(Freq ~ ses + mental, data=Mental)
chisq.test(mental.tab)

##
## Pearson's Chi-squared test
##
## data: mental.tab
## X-squared = 46, df = 15, p-value = 5.3e-05
```

## ordered $r \times c$ tables: Mental data II

For ordinal factors, more powerful tests are available with Cochran-Mantel-Haenszel tests:

```
CMHtest (mental.tab)
```

```
## Cochran-Mantel-Haenszel Statistics for ses by mental
##
##               AltHypothesis  Chisq Df    Prob
## cor             Nonzero correlation   37.2   1 1.09e-09
## rmeans   Row mean scores differ   40.3   5 1.30e-07
## cmeans   Col mean scores differ   40.7   3 7.70e-09
## general      General association   46.0  15 5.40e-05
```

Details later, but  $\chi^2/df$  gives a useful comparison.

```
##      cor  rmeans  cmeans  general
##  37.16   8.06   13.56    3.06
```

## 2 by 2 tables: Notation

Row	Column		Total
	1	2	
1	$n_{11}$	$n_{12}$	$n_{1+}$
2	$n_{21}$	$n_{22}$	$n_{2+}$
Total	$n_{+1}$	$n_{+2}$	$n_{++}$

Gender	Admit	Reject	Tot
Male	1198	1493	2691
Female	557	1278	1835
Total	1755	2771	4526

- $\mathbf{N} = \{n_{ij}\}$  are the **observed** frequencies.
- + subscript means **sum over**: row sums:  $n_{i+}$ ; col sums:  $n_{+j}$ ; total sample size:  $n_{++} \equiv n$
- Similar notation for:
  - Cell joint **population** probabilities:  $\pi_{ij}$ ; also use  $\pi_1 = \pi_{1+}$  and  $\pi_2 = \pi_{2+}$
  - Population **marginal** probabilities:  $\pi_{i+}$  (rows),  $\pi_{+j}$  (cols)
  - Sample **proportions**: use  $p_{ij} = n_{ij}/n$ , etc.

# Independence

Two categorical variables,  $A$  and  $B$  are **statistically independent** when:

- The **conditional distributions** of  $B$  given  $A$  are the same for all levels of  $A$

$$\pi_{1j} = \pi_{2j} = \cdots = \pi_{rj}$$

- Joint cell probabilities are the product of the marginal probabilities

$$\pi_{ij} = \pi_{i+} \pi_{+j}$$

For  $2 \times 2$  tables, this gives rise to tests and measures based on

- Difference in row marginal probabilities: test  $H_0 : \pi_1 = \pi_2$
- Odds ratio
- Standard  $\chi^2$  tests also apply for large  $n$
- Fisher's exact test or simulation required in small samples.

# Independence: Example I

In the Arthritis data, people are classified by `Sex`, `Treatment` and `Improved`. Are `Treatment` and `Improved` independent?

- row proportions are the same for Treated and Placebo
- cell frequencies  $\sim$  row total  $\times$  column total

```
data(Arthritis, package="vcd")
arth.tab <- xtabs( ~ Treatment + Improved, data=Arthritis)
round(prop.table(arth.tab, 1), 3)
```

```
##           Improved
## Treatment  None  Some Marked
##   Placebo 0.674 0.163  0.163
##   Treated 0.317 0.171  0.512
```

More people given the Placebo show No improvement; more people Treated show Marked improvement

## Independence: Example II

Frequencies, if **Treatment** and **Improved** were independent:

```
row.totals <- margin.table(arth.tab, 1)
col.totals <- margin.table(arth.tab, 2)
round(outer(row.totals, col.totals) / sum(arth.tab), 1)
```

```
##           Improved
## Treatment None Some Marked
##   Placebo 21.5   7.2   14.3
##   Treated 20.5   6.8   13.7
```

These are the **expected** frequencies, under independence.

```
chisq.test(arth.tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  arth.tab
## X-squared = 13.1, df = 2, p-value = 0.0015
```

# Sampling models: Poisson, Binomial, Multinomial

Some subtle distinctions arise concerning whether the row and/or column marginal totals of a contingency table are **fixed** by the sampling design or **random**.

- **Poisson**: each  $n_{ij}$  is regarded as an independent Poisson variate; nothing fixed
- **Binomial**: each row (or col) is regarded as an independent binomial distribution, with one fixed margin (group total), other random (response)
- **Multinomial**: only the total sample size,  $n_{++}$ , is fixed; frequencies  $n_{ij}$  are classified by  $A$  and  $B$
- These make a difference in how hypothesis tests are derived, justified and explained.
- Happily, for most inferential methods, the same results arise under Poisson, binomial and multinomial sampling

Q: What is an appropriate sampling model for the UCB admissions data? For the Hair-Eye color data? For the Mental impairment data?

## Odds and odds ratios

For a binary response where  $\pi = \text{Pr}(\text{success})$ , the **odds** of a success is

$$\text{odds} = \frac{\pi}{1 - \pi} .$$

- Odds vary **multiplicatively** around 1 (“even odds”,  $\pi = \frac{1}{2}$ )
- Taking logs, the  $\log(\text{odds})$ , or **logit** varies symmetrically around 0,

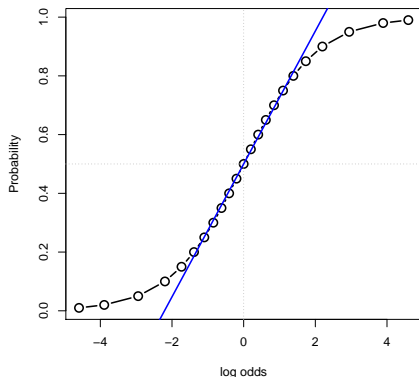
$$\text{logit}(\pi) \equiv \log(\text{odds}) = \log\left(\frac{\pi}{1 - \pi}\right) .$$

```
p <- c(.1, .25, .50, .75, .9)
odds <- p / (1-p)
logodds <- log(odds)
(odds.df <- data.frame(p, odds, logodds))
```

```
##      p  odds logodds
## 1 0.10 0.111   -2.2
## 2 0.25 0.333   -1.1
## 3 0.50 1.000    0.0
## 4 0.75 3.000    1.1
## 5 0.90 9.000    2.2
```



# Log odds



## Log odds:

- Symmetric around  $\pi = \frac{1}{2}$ :  
 $\text{logit}(\pi) = -\text{logit}(1 - \pi)$
- Fairly linear in the middle,  
 $0.2 \leq \pi \leq 0.8$
- The logit transformation of probability provides the basis for logistic regression

# Odds ratio

For two groups, with probabilities of success  $\pi_1, \pi_2$ , the **odds ratio**,  $\theta$ , is the ratio of the odds for the two groups:

$$\text{odds ratio} \equiv \theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

- $\theta = 1 \implies \pi_1 = \pi_2 \implies$  independence, no association
- Same value when we interchange rows and columns (transpose)
- Sample value,  $\hat{\theta}$  obtained using  $n_{ij}$ .

More convenient to characterize association by **log odds ratio**,  $\psi = \log(\theta)$  which is symmetric about 0:

$$\log \text{ odds ratio} \equiv \psi = \log(\theta) = \log \left[ \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \right] = \text{logit}(\pi_1) - \text{logit}(\pi_2) .$$

# Odds ratio: Inference and hypothesis tests

Symmetry of the distribution of the log odds ratio  $\psi = \log(\theta)$  makes it more convenient to carry out tests independence as tests of  $H_0 : \psi = \log(\theta) = 0$  rather than  $H_0 : \theta = 1$

- $z = \log(\hat{\theta}) / SE(\log(\theta)) \sim N(0, 1)$

`oddsratio()` in `vcd` uses  $\log(\theta)$  by default

```
oddsratio(UCB)
```

```
## log odds ratios for Gender and Admit
##
## [1] 0.61035
```

```
summary(oddsratio(UCB))
```

```
##
## z test of coefficients:
##
##               Estimate Std. Error z value Pr(>|z|)
## Male:Female/Admitted:Rejected    0.6104    0.0639    9.55    <2e-16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Odds ratio: Inference and hypothesis tests

Or, in terms of odds ratios directly:

```
oddsratio(UCB, log=FALSE)

## odds ratios for Gender and Admit
##
## [1] 1.8411

confint(oddsratio(UCB, log=FALSE))

##                2.5 % 97.5 %
## Male:Female/Admitted:Rejected 1.6244 2.0867
```

Males 1.84 times as likely to be admitted, with 95% CI of  $1.62 \leq \theta \leq 2.09$ .

**chisq.test()** just tests association:

```
chisq.test(UCB)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  UCB
## X-squared = 91.6, df = 1, p-value <2e-16
```

## Small sample size

- Pearson  $\chi^2$  and LR  $G^2$  tests are valid only when most expected frequencies  $\geq 5$
- Otherwise, use Fisher's exact test or simulated  $p$ -values

### Example

Is there a relation between high cholesterol in diet and heart disease?

```
fat <- matrix(c(6, 2, 4, 11), 2, 2)
dimnames(fat) <- list(cholesterol=c("low", "high"),
                      disease=c("no", "yes"))
```

fat

```
##           disease
## cholesterol no  yes
##           low   6   4
##           high  2  11
```

# Small sample size

The standard Pearson  $\chi^2$  is not significant:

```
chisq.test(fat)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  fat
## X-squared = 3.19, df = 1, p-value = 0.074
```

We get a warning message:

```
In chisq.test(fat) : Chi-squared approximation may be incorrect
```

## Small sample size

Using Monte Carlo simulation to calculate the  $p$ -value:

```
chisq.test(fat, simulate=TRUE)

##
##  Pearson's Chi-squared test with simulated p-value (based on
##  2000 replicates)
##
## data:  fat
## X-squared = 4.96, df = NA, p-value = 0.04
```

This method repeatedly samples cell frequencies from tables with the same margins, and calculates a  $\chi^2$  for each.

The  $\chi^2$  test is now significant

## Small sample size

Fisher's exact test: calculates probability for all  $2 \times 2$  tables as or more extreme than the data.

```
fisher.test(fat)

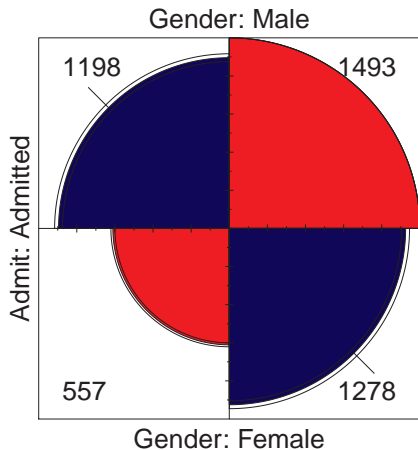
##
##  Fisher's Exact Test for Count Data
##
## data:  fat
## p-value = 0.039
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.86774 105.56694
## sample estimates:
## odds ratio
##    7.4019
```

The  $p$ -value is similar to the result using simulation.



# Visualizing: Fourfold plots

```
fourfold(UCB, std="ind.max") # maximum frequency
```

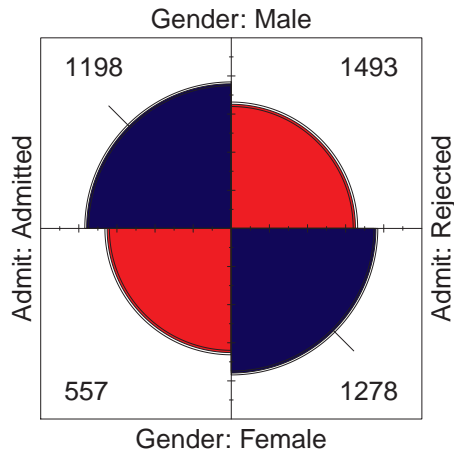


Friendly (1994a):

- Fourfold display: area  $\sim$  frequency,  $n_{ij}$
- Color: blue (+), red(-)
- This version: Unstandardized
- Odds ratio: ratio of products of blue / red cells

# Visualizing: Fourfold plots

```
fourfold(UCB)      #standardize both margins
```

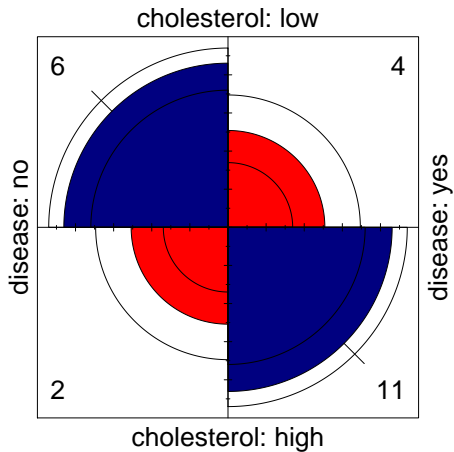


Better version:

- Standardize to equal row, col margins
- Preserves the odds ratio
- Confidence bands: significance of odds ratio
- If **don't** overlap  $\Rightarrow \theta \neq 1$

# Cholesterol data

```
fourfold(fat)
```



## Stratified $2 \times 2 \times k$ tables

The UC Berkeley data was collected for 6 graduate departments:

```
fable(addmargins(UCBAdmissions, 3))
```

##		Dept	A	B	C	D	E	F	Sum
##	Admit	Gender							
##	Admitted	Male	512	353	120	138	53	22	1198
##		Female	89	17	202	131	94	24	557
##	Rejected	Male	313	207	205	279	138	351	1493
##		Female	19	8	391	244	299	317	1278

### Questions:

- Does the overall association between gender and admission apply in each department?
- Do men and women apply equally to all departments?
- Do departments differ in their rates of admission?

**Stratified analysis** tests association between a main factor and a response *within* the levels of control variable(s)

# Stratified $2 \times 2 \times k$ tables

Odds ratios by department:

```
summary(oddsratio(UCBAdmissions))
```

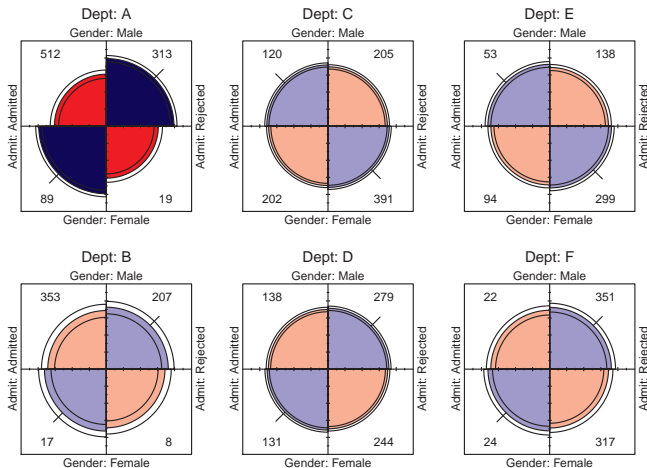
```
##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## A      -1.052      0.263   -4.00  6.2e-05 ***
## B      -0.220      0.438   -0.50    0.62
## C       0.125      0.144    0.87    0.39
## D      -0.082      0.150   -0.55    0.59
## E       0.200      0.200    1.00    0.32
## F      -0.189      0.305   -0.62    0.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Odds ratio only significant,  $\log(\theta) \neq 0$  for department A
- For department A, men are only  $\exp(-1.05) = .35$  times as likely to be admitted as women
- The overall analysis ignoring department is misleading: falsely assumes no associations of admission with department and gender with department.

# Stratified $2 \times 2 \times k$ tables

Fourfold plots by department (intense shading where significant):

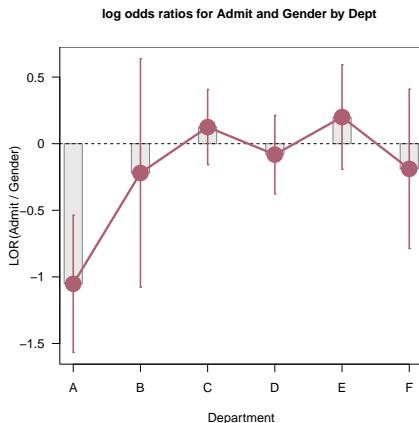
```
fourfold(UCBAdmissions)
```



# Stratified $2 \times 2 \times k$ tables

Or plot odds ratios directly:

```
plot(oddsratio(UCBAdmissions), cex=1.5, xlab="Department")
```



# Stratified tables: Homogeneity of odds ratios

## Related questions:

- Are the  $k$  odds ratios all equal,  $\theta_1 = \theta_2, \dots, \theta_k$ ? (Woolf's test: `woolf_test()`)
- (This is equivalent to the hypothesis of no three-way association)
- *If* homogeneous, is the common odds ratio different from 1? (Mantel-Haenszel test: `mantelhaen.test()`)

```
woolf_test(UCBAdmissions)

##
## Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
##
## data:  UCBAdmissions
## X-squared = 17.9, df = 5, p-value = 0.0031
```

Odds ratios differ across departments, so no sense in testing their common value.



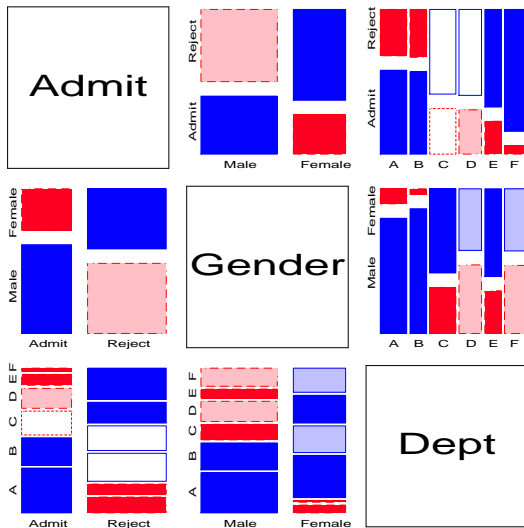
# Exegesis: What happened at UC Berkeley?

Why do the results *collapsed over* department disagree with the results *by* department?

## Simpson's paradox

- Aggregate data are misleading because they falsely assume men and women apply *equally* in each field.
- But:
  - Large differences in admission rates across departments.
  - Men and women apply to these departments differentially.
  - Women applied in large numbers to departments with low admission rates.
- Other graphical methods can show these effects.
- (This ignores possibility of *structural bias* against women: differential funding of fields to which women are more likely to apply.)

Mosaic matrix shows all pairwise associations:



## $r \times c$ tables: Overall analysis

- **Overall tests** of association: `assocstats()` : Pearson chi-square and LR  $G^2$
- **Strength** of association:  $\phi$  coefficient, contingency coefficient (C), Cramer's V ( $0 \leq V \leq 1$ )

$$\phi^2 = \frac{\chi^2}{n}, \quad C = \sqrt{\frac{\chi^2}{n + \chi^2}}, \quad V = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}}$$

- For a  $2 \times 2$  table,  $V = \phi$ .
- (If the data table was collapsed from a 3+ way table, the two-way analysis may be misleading)

```
assocstats(HEC)
```

```
##               X^2 df P(> X^2)
## Likelihood Ratio 146.44  9      0
## Pearson          138.29  9      0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.435
## Cramer's V        : 0.279
```

## $r \times c$ tables: Overall analysis and residuals

- The Pearson  $X^2$  and LR  $G^2$  statistics have the following forms:

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \quad G^2 = \sum_{ij} n_{ij} \log \left( \frac{n_{ij}}{\hat{m}_{ij}} \right)$$

- Expected (fitted) frequencies under independence:  $\hat{m}_{ij} = n_{i+}n_{+j}/n_{++}$
- Each of these is a sum-of-squares of corresponding **residuals**
- Degrees of freedom:  $df = (r - 1)(c - 1)$  — # independent residuals

Can get residuals from **loglm()** in **MASS**:

```
library(MASS)
mod <- loglm(~Hair + Eye, data=HEC, fitted=TRUE)
mod

## Call:
## loglm(formula = ~Hair + Eye, data = HEC, fitted = TRUE)
##
## Statistics:
##              X^2 df P(> X^2)
## Likelihood Ratio 146.44   9      0
## Pearson          138.29   9      0
```

## Extract residuals:

```
res.P <- residuals(mod, type="pearson")
res.LR <- residuals(mod, type="deviance") # default
res.P
```

```
##           Hair
## Eye      Black  Brown    Red   Blond
## Brown  4.398  1.233 -0.075 -5.851
## Blue  -3.069 -1.949 -1.730  7.050
## Hazel -0.477  1.353  0.852 -2.228
## Green -1.954 -0.345  2.283  0.613
```

## Demonstrate SSQ property:

```
unlist(mod[c("pearson", "deviance", "df")])
```

```
##  pearson deviance      df
##   138.29   146.44    9.00
```

```
sum(res.P^2)           # Pearson chisq
```

```
## [1] 138.29
```

```
sum(res.LR^2)          # LR chisq
```

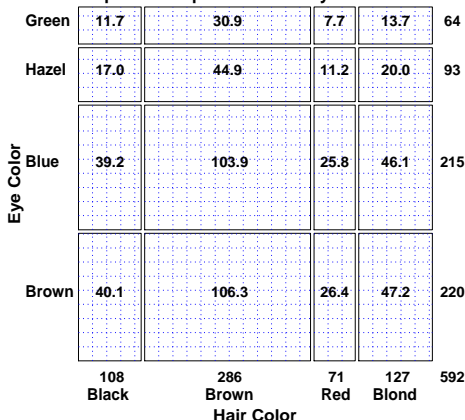
```
## [1] 146.44
```

# Visualizing association: Sieve diagrams

Visual metaphor: **count**  $\sim$  **area**

- When row/col variables are independent,  $n_{ij} \approx \hat{m}_{ij} \sim n_{i+}n_{+j}$
- $\Rightarrow$  each cell can be represented as a rectangle, with area = height  $\times$  width  $\sim$  frequency,  $n_{ij}$  (under independence)

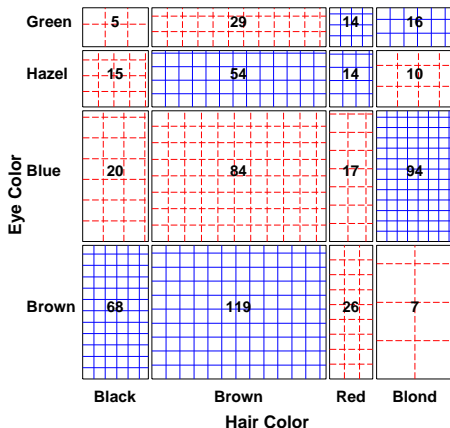
Expected frequencies: Hair Eye Color Data



- This display shows **expected frequencies**, assuming independence, as # boxes within each cell
- The boxes are all of the same size (equal density)
- Real sieve diagrams use # boxes = **observed frequencies**,  $n_{ij}$

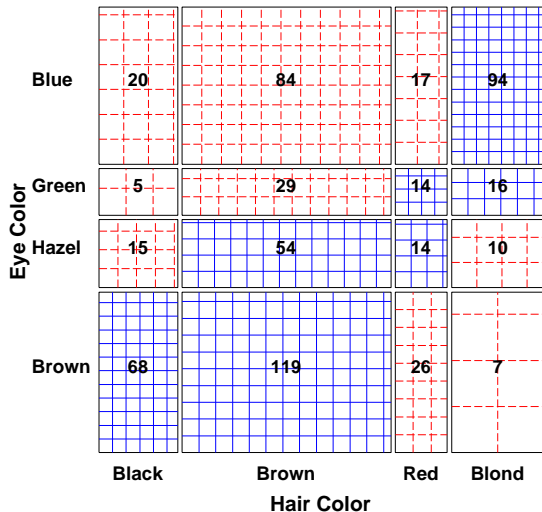
# Sieve diagrams

- Height, width  $\sim$  marginal frequencies,  $n_{i+}$ ,  $n_{+j}$
- $\implies$  Area  $\sim$  expected frequency,  $\hat{m}_{ij} \sim n_{i+}n_{+j}$
- Shading  $\sim$  observed frequency,  $n_{ij}$ , color:  $\text{sign}(n_{ij} - \hat{m}_{ij})$ .
- $\implies$  Independence: Shown when density of shading is uniform.



# Sieve diagrams

**Effect ordering:** Reorder rows/cols to make the pattern coherent

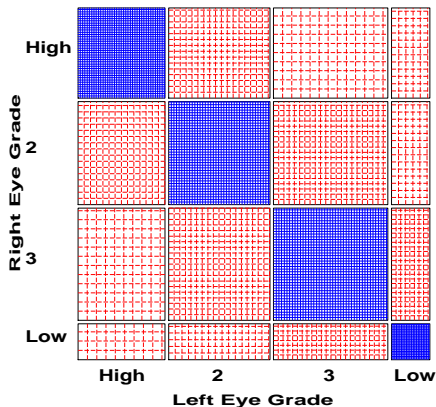




# Sieve diagrams

Vision classification data for 7477 women: visual acuity in left, right eyes

**Unaided distant vision data**



- The obvious association is apparent on the diagonal cells
- A more subtle pattern appears on the off-diagonal cells
- Analysis methods for **square tables** (later) allow testing hypotheses of **symmetry**, **quasi-symmetry**, etc.

# Ordinal factors

The Pearson  $\chi^2$  and LR  $G^2$  give tests of **general association**, with  $(r - 1)(c - 1)$  df.

## More powerful CMH tests

- When either the row or column levels are **ordered**, more specific CMH (Cochran–Mantel–Haentzel) tests which take order into account have greater **power** to detect ordered relations.
- This is similar to testing for **linear trends** in ANOVA
- Essentially, these assign scores to the categories, and test for differences in row / column means, or non-zero correlation.

# CMH tests for ordinal variables

Three types of CMH tests:

## Non-zero correlation

- Use when *both* row and column variables are ordinal.
- CMH  $\chi^2 = (N - 1)r^2$ , assigning scores (1, 2, 3, ...)
- most powerful for *linear* association

## Row/Col Mean Scores Differ

- Use when only *one* variable is ordinal
- Analogous to the Kruskal-Wallis non-parametric test (ANOVA on rank scores)

## General Association

- Use when *both* row and column variables are nominal.
- Similar to overall Pearson  $\chi^2$  and Likelihood Ratio  $G^2$ .

# Sample CMH Profiles

**Only general association:**

	b1	b2	b3	b4	b5	Total	Mean
a1	0	15	25	15	0	55	3.0
a2	5	20	5	20	5	55	3.0
a3	20	5	5	5	20	55	3.0
Total	25	40	35	40	25	165	

**Output:**

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.000	1.000
2	Row Mean Scores Differ	2	0.000	1.000
3	General Association	8	91.797	<b>0.000</b>

# Sample CMH Profiles

## Linear Association:

	b1	b2	b3	b4	b5	Total	Mean
a1	2	5	8	8	8	31	3.48
a2	2	8	8	8	5	31	3.19
a3	5	8	8	8	2	31	2.81
a4	8	8	8	5	2	31	2.52
Total	17	29	32	29	17	124	

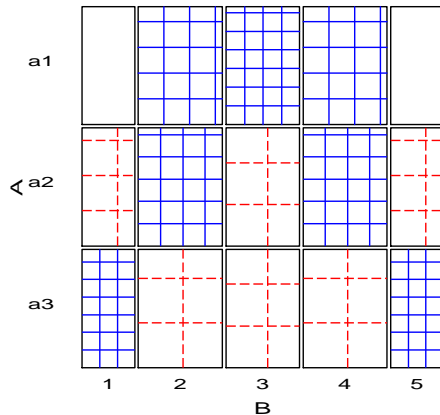
## Output:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	10.639	<b>0.001</b>
2	Row Mean Scores Differ	3	10.676	<b>0.014</b>
3	General Association	12	13.400	0.341

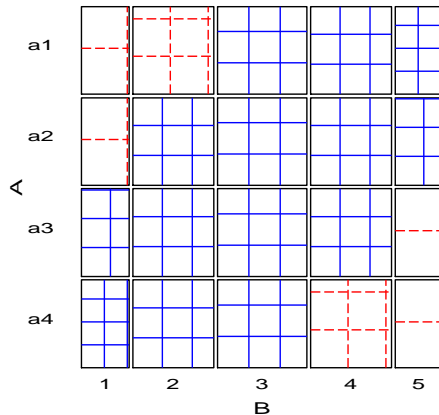
# Sample CMH Profiles

## Visualizing Association: Sieve diagrams

General Association



Linear Association



## Example: Mental health data

- In R, these tests are provided by `CMHtest()` in the `vcdExtra` package
- For the mental health data, both factors are ordinal
- All tests are significant
- The nonzero correlation test, with 1 df, has the smallest  $p$ -value, largest  $\chi^2/df$

```
mental.tab <- xtabs(Freq ~ ses + mental, data=Mental)
CMHtest(mental.tab)

## Cochran-Mantel-Haenszel Statistics for ses by mental
##
##               AltHypothesis  Chisq  Df    Prob
## cor             Nonzero correlation   37.2   1 1.09e-09
## rmeans   Row mean scores differ   40.3   5 1.30e-07
## cmeans   Col mean scores differ   40.7   3 7.70e-09
## general      General association   46.0  15 5.40e-05
```

# Observer Agreement

- **Inter-observer agreement** often used as to assess reliability of a subjective classification or assessment procedure
  - → square table, Rater 1 x Rater 2
  - Levels: diagnostic categories (normal, mildly impaired, severely impaired)
- **Agreement vs. Association:** Ratings can be strongly associated without strong agreement
- **Marginal homogeneity:** Different frequencies of category use by raters affects measures of agreement
- **Measures of Agreement:**
  - Intraclass correlation: ANOVA framework— multiple raters!
  - Cohen's  $\kappa$ : compares the observed agreement,  $P_o = \sum p_{ii}$ , to agreement expected by chance if the two observer's ratings were independent,  $P_c = \sum p_{i+} p_{+i}$ .

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$



# Cohen's $\kappa$

## Properties of Cohen's $\kappa$ :

- perfect agreement:  $\kappa = 1$
- minimum  $\kappa$  may be  $< 0$ ; lower bound depends on marginal totals
- Unweighted  $\kappa$ : counts only diagonal cells (same category assigned by both observers).
- Weighted  $\kappa$ : allows partial credit for near agreement. (Makes sense only when the categories are *ordered*.)

## Weights:

- Cicchetti-Alison (inverse integer spacing)
- Fleiss-Cohen (inverse square spacing)

### Integer Weights

1	2/3	1/3	0
2/3	1	2/3	1/3
1/3	2/3	1	2/3
0	1/3	2/3	1

### Fleiss-Cohen Weights

1	8/9	5/9	0
8/9	1	8/9	5/9
5/9	8/9	1	8/9
0	5/9	8/9	1

# Cohen's $\kappa$ : Example

The table below summarizes responses of 91 married couples to a questionnaire item,

*Sex is fun for me and my partner (a) Never or occasionally, (b) fairly often, (c) very often, (d) almost always.*

Husband's Rating	----- Wife's Rating -----					SUM
	Never fun	Fairly often	Very Often	Almost always		
Never fun	<b>7</b>	7	2	3		19
Fairly often	2	<b>8</b>	3	7		20
Very often	1	5	<b>4</b>	9		19
Almost always	2	8	9	<b>14</b>		33
SUM	12	28	18	33		91

## Cohen's $\kappa$ : Example

The `Kappa()` function in `vcd` calculates unweighted and weighted  $\kappa$ , using equal-spacing weights by default.

```
data(SexualFun, package="vcd")
Kappa(SexualFun)
```

	value	ASE	z	Pr(> z )
## Unweighted	0.129	0.0686	1.89	0.05939
## Weighted	0.237	0.0783	3.03	0.00244

```
Kappa(SexualFun, weights="Fleiss-Cohen")
```

	value	ASE	z	Pr(> z )
## Unweighted	0.129	0.0686	1.89	0.059387
## Weighted	0.332	0.0973	3.41	0.000643

Unweighted  $\kappa$  is not significant, but both weighted versions are. You can obtain confidence intervals with the `confint()` method

# Observer agreement: Multiple strata

When the individuals rated fall into multiple groups, one can test for:

- Agreement within each group
- Overall agreement (controlling for group)
- Homogeneity: Equal agreement across groups

## Example: Diagnostic Classification of MS patients

Patients in Winnipeg and New Orleans were each classified by a neurologist in each city

NO rater:	Winnipeg patients				New Orleans patients			
	Cert	Prob	Pos	Doubt	Cert	Prob	Pos	Doubt
-----								
Winnipeg rater:								
Certain MS	38	5	0	1	5	3	0	0
Probable	33	11	3	0	3	11	4	0
Possible	10	14	5	6	2	13	3	4
Doubtful MS	3	7	3	10	1	2	4	14

# Observer agreement: Multiple strata

Here, simply assess agreement between the two raters in each stratum separately

```
data(MSPatients, package="vcd")
```

```
Kappa(MSPatients[, , 1])
```

```
##              value      ASE      z Pr(>|z|)
## Unweighted 0.208 0.0505 4.12 3.77e-05
## Weighted   0.380 0.0517 7.35 1.99e-13
```

```
Kappa(MSPatients[, , 2])
```

```
##              value      ASE      z Pr(>|z|)
## Unweighted 0.297 0.0785 3.78 1.59e-04
## Weighted   0.477 0.0730 6.54 6.35e-11
```

The **irr** package (inter-rater reliability) provides ICC and other measures, and handles the case of  $k > 2$  raters.

# Bangdiwala's Observer Agreement Chart

The observer agreement chart Bangdiwala (1987) provides

- a simple graphic representation of the strength of agreement, and
- a measure of strength of agreement with an intuitive interpretation.

## Construction:

- $n \times n$  square,  $n$ =total sample size
- Black squares, each of size  $n_{ij} \times n_{ij} \rightarrow$  observed agreement
- Positioned within larger rectangles, each of size  $n_{i+} \times n_{+i} \rightarrow$  maximum possible agreement
- $\Rightarrow$  visual impression of the strength of agreement is  $B$ :

$$B = \frac{\text{area of dark squares}}{\text{area of rectangles}} = \frac{\sum_i^k n_{ij}^2}{\sum_i^k n_{i+} n_{+i}}$$

- $\Rightarrow$  Perfect agreement:  $B = 1$ , all rectangles are completely filled.

## Weighted Agreement Chart: Partial agreement

Partial agreement: include weighted contribution from off-diagonal cells,  $b$  steps from the main diagonal, using weights  $1 > w_1 > w_2 > \dots$ .

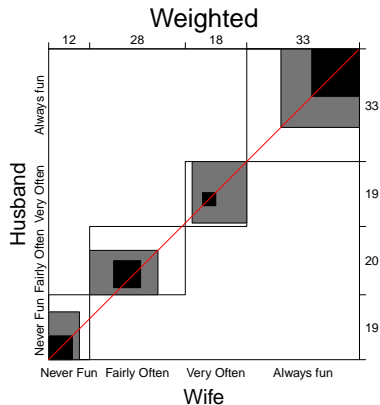
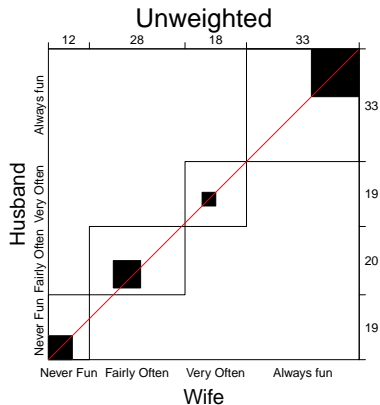
$$\begin{array}{ccccccccccc}
 & & n_{i-b,i} & & & & & & w_2 & & \\
 & & \vdots & & & & & & w_1 & & \\
 n_{i,i-b} & \cdots & n_{i,i} & \cdots & n_{i,i+b} & & w_2 & w_1 & 1 & w_1 & w_2 \\
 & & \vdots & & & & & & w_1 & & \\
 & & n_{i-b,i} & & & & & & w_2 & & 
 \end{array}$$

- Add shaded rectangles, size  $\sim$  sum of frequencies,  $A_{bi}$ , within  $b$  steps of main diagonal
- $\Rightarrow$  weighted measure of agreement,

$$B^w = \frac{\text{weighted sum of agreement}}{\text{area of rectangles}} = 1 - \frac{\sum_i^k [n_{i+} n_{+i} - n_{ii}^2 - \sum_{b=1}^q w_b A_{bi}]}{\sum_i^k n_{i+} n_{+i}}$$

Husbands and wives:  $B = 0.146$ ,  $B^w = 0.498$

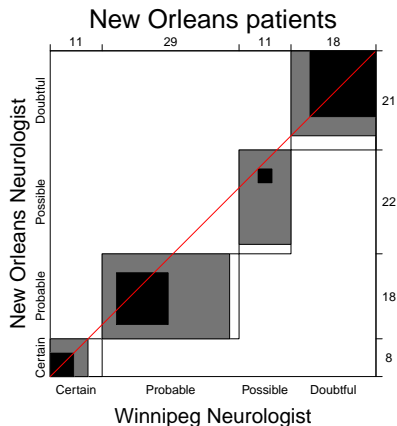
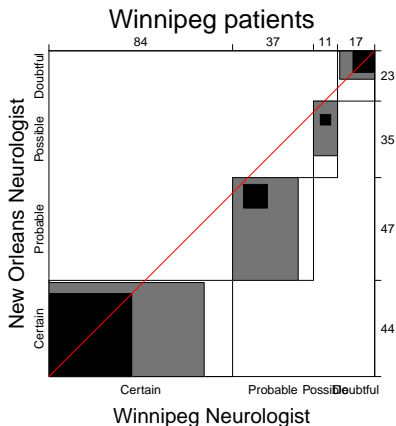
```
agreementplot(SexualFun, main="Unweighted", weights=1)
agreementplot(SexualFun, main="Weighted")
```





# Marginal homogeneity and Observer bias

- Different raters may consistently use higher or lower response categories
- Test– **marginal homogeneity**:  $H_0 : n_{i+} = n_{+i}$
- Shows as departures of the squares from the diagonal line



- Winnipeg neurologist tends to use more severe categories

# Looking ahead

## Loglinear models

Loglinear models generalize the Pearson  $\chi^2$  and LR  $G^2$  tests of association to 3-way and larger tables.

- Allows a range of models from **mutual independence** ( $[A][B][C]$ ) to the **saturated model** ( $[ABC]$ )
- Intermediate models address questions of **conditional independence**, controlling for some factors
- Can test associations in 2-way, 3-way terms analogously to tests of interactions in ANOVA

## Example: UC Berkeley data

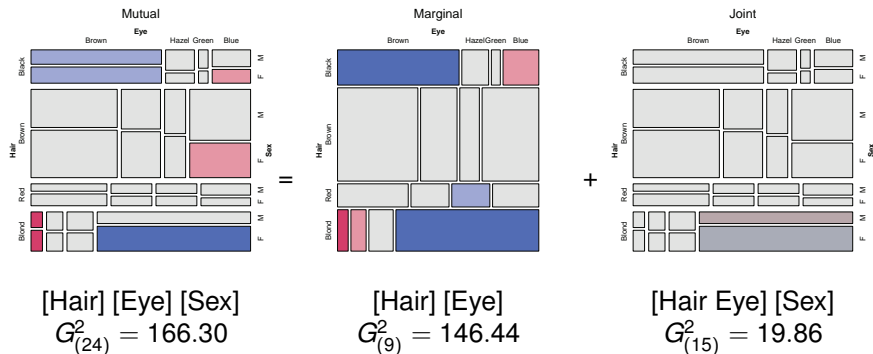
- Mutual independence:  $[Admit][Gender][Dept]$
- Joint independence:  $[Admit][Gender*Dept]$
- Conditional independence:  $[Admit*Dept][Admit*Gender]$ : A specific test for absence of gender bias, controlling for department

# Looking ahead

## Mosaic displays

Mosaic plots provide visualizations of associations in 2+ way tables.

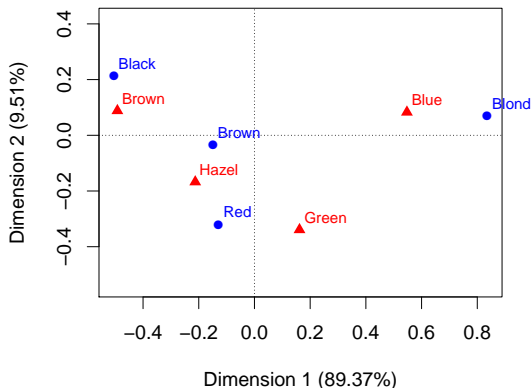
- Tiles:  $\sim$  frequency
- Fit loglinear model
- Shading:  $\sim$  residuals



# Looking ahead

## Correspondence analysis

- Account for max. % of  $\chi^2$  in few (2-3) dimensions
- Find scores for row and column categories
- Plot of row and column scores shows associations



# References I

- Bangdiwala, S. I. Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS User's Group International Conference*, 12:1083–1088, 1987.
- Friendly, M. A fourfold display for 2 by 2 by K tables. Technical Report 217, York University, Psychology Dept, 1994a.
- Friendly, M. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89:190–200, 1994b.