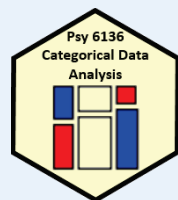
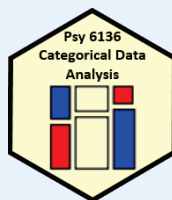


## Two-way tables Independence & association



Michael Friendly  
Psych 6136  
<http://friendly.github.io/psy6136>



## Two-way tables: Overview

Two-way frequency tables are a convenient way to represent a dataset cross-classified by two discrete variables, A & 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?

2

## Methods

- The methods discussed this week are generally simple **non-parametric** or **randomization** methods
- There is no underlying formal **model** with parameters
- Hypothesis tests based on some test statistic:
  - Pearson  $\chi^2$
  - Odds ratio,  $\theta$
  - Cohen's  $\kappa$
  - Friendly's ☹ or ☺      Plenty of room for new stats!
  - Arjun's  $\psi$
- $p$ -values, confidence intervals based on:
  - Large sample theory:  $\chi^2 \sim \chi^2$  as  $N \rightarrow \infty$  (smaller suffices)
  - Permutation or simulation distributions

3

## $2 \times 2$ Example: Berkeley admissions

Table: Admissions to Berkeley graduate programs

	Admitted	Rejected	Total	% Admit	Odds(Admit)	
Males	1198	1493	2691	44.52	0.802	odds ratio ( $\theta$ ) $\approx 1.84$
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

- Is there an association between gender & admission?
- If so, is this evidence for gender bias?
- How to measure **strength** of association?
- How to test for significance?
- How to visualize?

4

# UCBAdmissions data

In R, the data is contained in `UCBAdmissions`, a 2 x 2 x 6 table for 6 departments.  
We collapse over department

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

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

$\text{odds}_M = 1198 / 1493 = 0.802$   
 $\text{odds}_F = 557 / 1278 = 0.437$

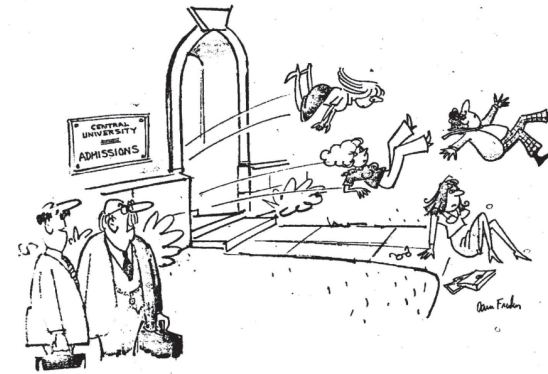
Association in 2 x 2 table can be measured by the odds ratio ( $\theta$ ): odds of admission for males vs. females

```
> oddsratio(UCB, log=FALSE)
odds ratios for Gender and Admit
```

```
[1] 1.84
> confint(oddsratio(UCB, log=FALSE))
                2.5 % 97.5 %
Male:Female/Admitted:Rejected 1.62 2.09
```

Note use of `confint()` for obtaining the CI( $\theta$ )

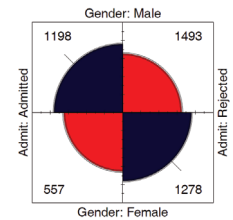
5



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

Questions:

- ❖ How to analyze these results? What tests for odds ratio?
- ❖ How to visualize & interpret?
- ❖ Does it matter that we collapsed over Department?



6

## Table 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

- $N = \{n_{ij}\}$  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.

7

## r x c Example: Hair color, eye color

Data from 592 students in a statistics class: write down your hair and eye color

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

- ❖ Is there an association between hair color and eye color?
- ❖ How to measure **strength** of association?
- ❖ How to test for significance?
- ❖ How to visualize?
- ❖ How to understand the **pattern** (nature) of association?

8

# HairEyeColor data

In R, the dataset is `HairEyeColor`, a 4 x 4 x 2 table: Hair x Eye x Sex.  
For now, collapse over sex.

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

```
> chisq.test(HEC)
```

Pearson's Chi-squared test

```
data: HEC
X-squared = 138, df = 9, p-value <2e-16
```

Association can be tested by the standard Pearson  $\chi^2$  test.  
Details later

```
> MASS::loglm(~Hair + Eye, data=HEC)
```

Statistics:

	X^2	df	P(> X^2)
Likelihood Ratio	146	9	0
Pearson	138	9	0

Or, as a loglinear model for independence  
Formula:  $\sim A + B = A \perp B$

9

# Measures of association

`vcd::assocstats()` collects tests and measures in a convenient summary

```
> 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
```

For 3+ way tables, it gives the results for the strata defined by **all last dimensions**

```
> assocstats(HairEyeColor)
$`Sex:Male`
              X^2 df P(> X^2)
Likelihood Ratio 44.445  9 1.168e-06
Pearson          41.280  9 4.447e-06

Phi-Coefficient   : NA
Contingency Coeff.: 0.359
Cramer's V        : 0.222
```

```
$`Sex:Female`
              X^2 df P(> X^2)
Likelihood Ratio 112.23  9      0
Pearson          106.66  9      0

Phi-Coefficient   : NA
Contingency Coeff.: 0.504
Cramer's V        : 0.337
```

10

# Measures of association

## • 2 x 2 tables

### ▪ Odds ratio

$$\theta = \frac{\text{odds}(B_1 | A_1)}{\text{odds}(B_1 | A_2)} = \frac{n_{11} / n_{12}}{n_{21} / n_{22}}$$

### ▪ Phi coefficient

- Analog of correlation
- $\phi^2 = \% \text{ of variance}$

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1+}n_{2+}n_{+1}n_{+2}} = \pm \sqrt{\chi^2 / n}$$

## • r x c tables

### ▪ Cramer's V – generalization of phi

$$\text{Cramer V} = \sqrt{\frac{\chi^2}{n \min(r-1, c-1)}}$$

### ▪ Pearson contingency coef

$$\text{Pearson C} = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

11

# Example: Bartlett data

2 x 2 x 2 Data on plum root cuttings: Length (short|long), planted (Now|Spring), Survived? (Alive|Dead)

- Does survival depend on time of planting?
- Is there a 3-way association, i.e., does (Alive, Time) differ by Length? ( $\theta_1 = \theta_1$ )

```
> assocstats(Bartlett)
$`Length:Long`
              X^2 df P(> X^2)
Likelihood Ratio 43.87  1 3.50e-11
Pearson          43.20  1 4.94e-11

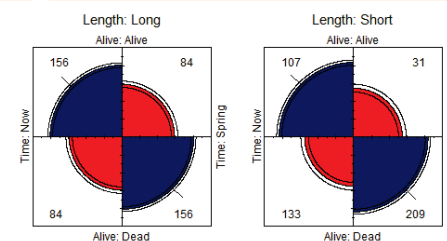
Phi-Coefficient   : 0.3
Contingency Coeff.: 0.287
Cramer's V        : 0.3
```

```
$`Length:Short`
              X^2 df P(> X^2)
Likelihood Ratio 61.31  1 4.88e-15
Pearson          58.74  1 1.80e-14

Phi-Coefficient   : 0.35
Contingency Coeff.: 0.33
Cramer's V        : 0.35
```

```
oddsratio(Bartlett, log=FALSE)
odds ratios for Alive and Time
by Length
```

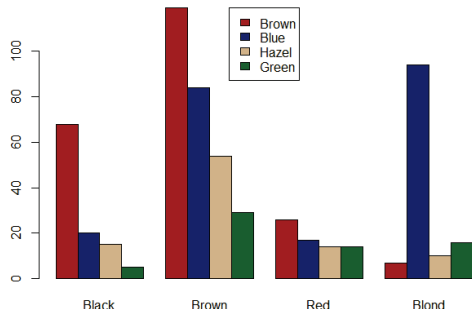
	Long	Short
3.45	5.42	



12

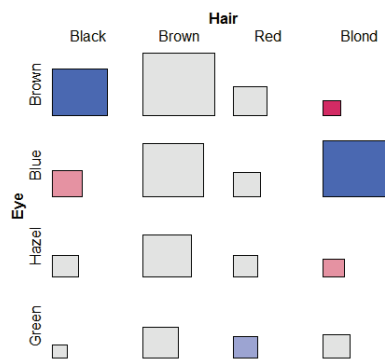
## Simple plots for $r \times c$ tables

```
barplot(HEC, beside=TRUE, ...)
```



But: harder to compare across hair-color groups than within them

```
tile(HEC, shade=TRUE)
```



Neither of these extend to more than 2 variables

13

## Ordered tables

$r \times c$  table with ordered categories: Mental health and Parents' SES categories (1="High" and 6="Low")

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 a **predictor**
- ❖ How to measure **strength** of association?
- ❖ How to understand the **pattern** of association?
- ❖ How to take **ordinal nature** of variables into account?

14

## Mental data: Association

The data is contained in `vcdExtra::Mental`, a frequency data frame, with **ordered** factors

```
> 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 2 2 2 3 ...
 $ mental: Ord.factor w/ 4 levels "Well"<"Mild"<...: 1 2 3 4 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 = 5e-05
```

This  $\chi^2$  test doesn't take ordinality into account. It just tests for **general association**.

15

## Mental data: Ordinal tests

For **ordinal** factors, more powerful (focused) tests are available with Cochran-Mantel-Haenszel tests in `vcdExtra::CMHtest()`

```
> 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

both ordinal  
cols ordinal  
rows ordinal  
neither

$\chi^2$  / df shows why ordered tests are more powerful

```
> xx <- CMHtest(mental.tab)
> xx$table[, "Chisq"] / xx$table[, "Df"]
      cor rmeans cmeans general
37.16   8.06  13.56   3.06
```

Think: more df  $\rightarrow$  more diffuse; less focused; less powerful against  $H_1$

16

# 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  $A$

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

- Joint cell probabilities are the product of the marginal probabilities

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

For 2 x 2 tables, this gives rise to tests and measures based on:

- ❖ Difference in row/col marginal probabilities: Test  $H_0: \pi_1 = \pi_2$
- ❖ Odds ratio,  $\hat{\theta} = (n_{11} / n_{12}) / (n_{21} / n_{22})$ . Test  $H_0: \theta = 1$
- ❖ Standard  $\chi^2$  test, with largish  $n$
- ❖ Small samples: Fisher's exact test, or simulation / permutation tests

17

# Independence: Example

A contrived example, where I generate cell frequencies as the product of row and column marginal totals:  $n_{ij} = n_{i+} \times n_{+j}$

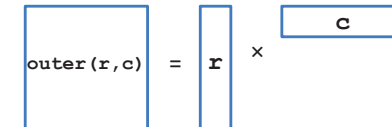
```
> educ <- c(50, 100, 50) # marginal frequencies
> names(educ) <- c("Low", "Med", "High")

> party <- c(20, 50, 30) # marginal frequencies
> names(party) <- c("NDP", "Liberal", "Cons")

> table <- outer(educ, party) / sum(party) # cell = row * col / n
> names(dimnames(table)) <- c("Education", "Party")
> table
```

	Party		
Education	NDP	Liberal	Cons
Low	10	25	15
Med	20	50	30
High	10	25	15

Outer product:



18

# Independence: Example

- The row proportions of party are the same for each educ group
- The col proportions of educ are the same for each party

```
> prop.table(table, 1)
```

	NDP	Liberal	Cons
Low	0.2	0.5	0.3
Med	0.2	0.5	0.3
High	0.2	0.5	0.3

```
> prop.table(table, 2)
```

	NDP	Liberal	Cons
Low	0.25	0.25	0.25
Med	0.50	0.50	0.50
High	0.25	0.25	0.25

So, the  $X^2$  is exactly zero, and measures of strength are zero

```
> vcd::assocstats(table)
```

	$X^2$	df	P(> $X^2$ )
Likelihood Ratio	0	4	1
Pearson	0	4	1

```
Phi-Coefficient : NA
Contingency Coeff.: 0
Cramer's V : 0
```

19

# Independence?: Arthritis data

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

But, more people given the Placebo show no improvement; more people Treated show marked improvement

20

# Independence?: Arthritis data

If Treatment and Improved were independent, frequencies  $\sim$  row x col margins

```
> row.totals <- margin.table(arth.tab, 1)
> col.totals <- margin.table(arth.tab, 2)
> round(outer(row.totals, col.totals)/ sum(arth.tab), 0)
      Improved
Treatment None Some Marked
Placebo    22    7    14
Treated    20    7    14
```

These are the **expected frequencies**, under independence; but for the data:

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

Pearson's Chi-squared test

data:  arth.tab
X-squared = 13.1, df = 2, p-value = 0.0015
```

$$\text{Pearson } \chi^2$$

$$\chi^2_{(r-1) \times (c-1)} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum d_{ij}^2$$

21

# Sampling models: Poisson, Binomial, Multinomial

Subtle distinctions arise concerning whether the row and/or margins are **fixed** by 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 dist<sup>n</sup>, 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
- Makes a difference in how hypothesis tests are justified & explained
- Happily, for most inferential methods,  $\approx$  same results are obtained under the three sampling models

Q: what is an appropriate sampling model for the UCB admissions data? For hair-eye color? For the mental impairment data?

22

# Odds and log(Odds)

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(odds), or **logit** varies symmetrically around 0,

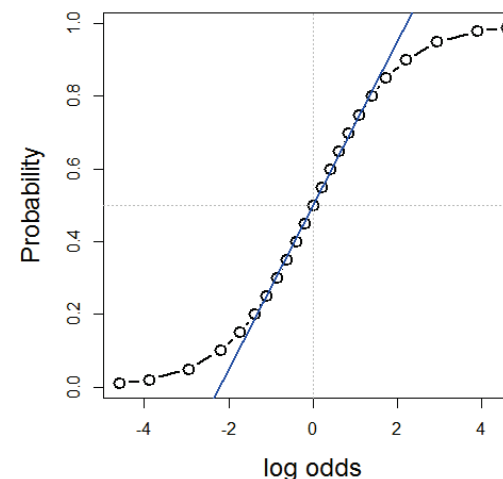
$$\text{logit}(\pi) \equiv \log(\text{odds}) = \log\left(\frac{\pi}{1 - \pi}\right)$$

```
> p <- c( 0.05, .1, .25, .50, .75, .9, .95)
> odds <- p / (1-p)
> logodds <- log(odds)
> (odds.df <- data.frame(p, odds, logodds))
   p    odds logodds
1 0.05 0.0526  -2.94
2 0.10 0.1111  -2.20
3 0.25 0.3333  -1.10
4 0.50 1.0000   0.00
5 0.75 3.0000   1.10
6 0.90 9.0000   2.20
7 0.95 19.0000   2.94
```

23

# Log odds

```
plot(logodds, p, type='b', xlab="log odds", ylab="Probability", ...)
abline(lm(p ~ logodds, subset=(p>=.2 & p<=.8)), col="blue")
```



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 is the basis for **logistic regression**

(An alternative, the cumulative normal,  $\Phi^{-1}(\pi)$ , gives rise to **probit regression**)

24



## 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:

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

- $\theta = 1 \implies \pi_1 = \pi_2 \implies$  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:

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

25

## Odds ratio: Inference & 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  $H_0 : \theta = 1$

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

$$SE(\log(\theta)) = \sqrt{\sum_{ij} n_{ij}^{-1}}$$

**vcd::oddsratio()** has option, `log=, TRUE` by default  
The `summary()` method calculates z tests

```
> 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
```

26

## Odds ratio: Confidence intervals

Results should be reported with confidence intervals, either for the odds ratio,  $\theta$ , or for  $\psi = \log(\theta)$

```
> confint(oddsratio(UCB, log = FALSE))
                2.5 % 97.5 %
Male:Female/Admitted:Rejected 1.624 2.087
> confint(oddsratio(UCB))
                2.5 % 97.5 %
Male:Female/Admitted:Rejected 0.4851 0.7356
```

Summary in words:

For the Berkeley admissions data:

- The Pearson  $\chi^2$  test of association between Gender and Admission was highly significant,  $\chi_1^2 = 91.6$ ,  $p < .0001$
- This corresponded to an odds ratio of admission for Males vs. Females of  $\theta = 1.84$  (CI: 1.62, 2.09), meaning that overall, males were 84% more likely to be admitted
- On the scale of log odds,  $\psi = \log(\theta)$ , the estimate was  $\psi = 0.610$  (CI: 0.485, 0.736), meaning a significant positive association between Gender(Male) and admission.

27

## Small sample size

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

Example: Cholesterol diet and heart disease

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

> fat
              disease
cholesterol no yes
          low  6  4
          high  2 11
```

28

## Small sample size

The standard Pearson  $\chi^2$  test is not significant  
For 2 x 2 tables with small  $n$ , a correction  $|O - E| - \frac{1}{2}$  is standardly applied (Yates)

```
> chisq.test(fat)

Pearson's Chi-squared test with Yates' continuity correction

data: fat
X-squared = 3.19, df = 1, p-value = 0.074
```

Yet, we get a warning. Maybe Friendly's ☹ ?

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

29

## Small sample size: Simulation

A Monte-Carlo method uses simulation to calculate a  $p$ -value, Friendly's ☺

```
> 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  $p$ -value compares the observed  $X^2$  to its' distribution in the simulations.
- The  $\chi^2$  test is now significant; well, barely, but ☺
- The main point is that the test no longer depends on large sample theory ☺ ☺
- Simulation is a **general principle** for testing hypotheses ☺ ☺ ☺

30

## Small sample size: Fisher exact test

Fisher's exact test: calculates probability for all  $2 \times 2$  tables with odds ratio as or more extreme than that in the data, keeping the margins fixed.

```
> fisher.test(fat)

Fisher's Exact Test for Count Data

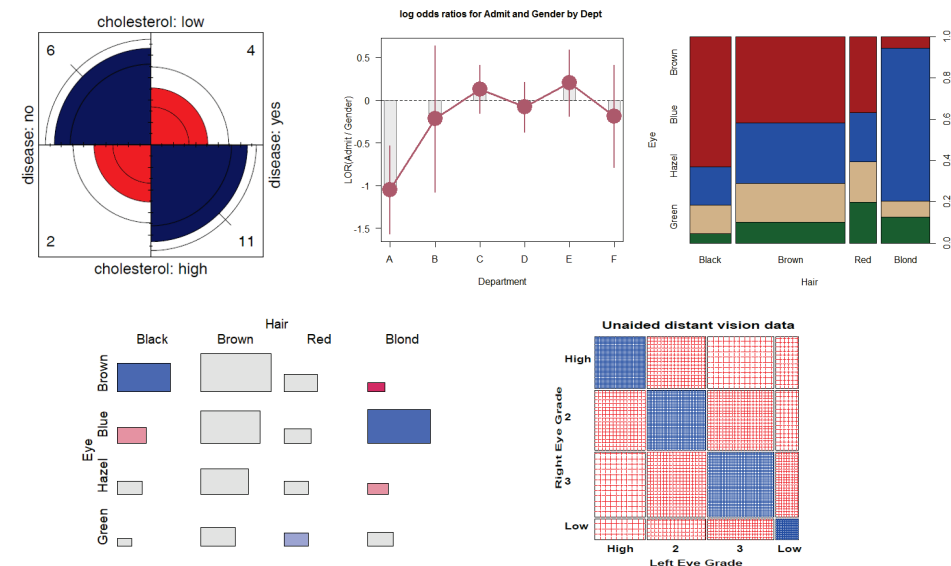
data: fat
p-value = 0.039
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.86774 105.56694
sample estimates:
odds ratio
 7.4019
```

The  $p$ -value is similar to that obtained using simulation.

Fisher's test is available for larger  $r \times c$  tables, but the method gets computationally intensive as  $r * c$  increases

31

## Visualizing association

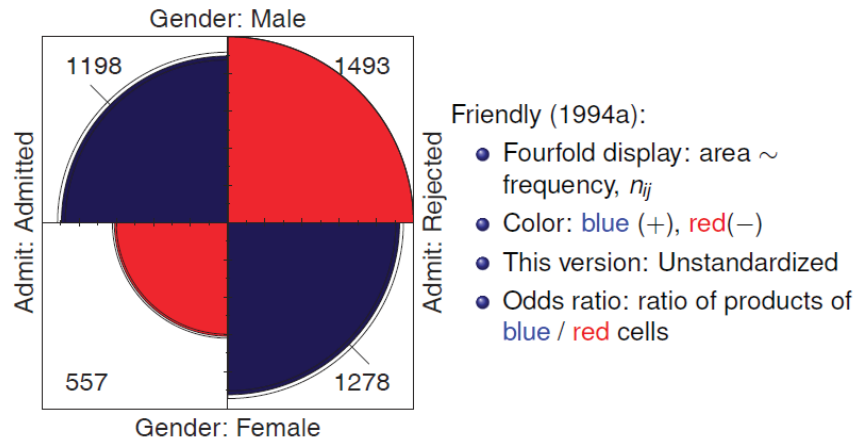


32



## Visualizing: fourfold plots

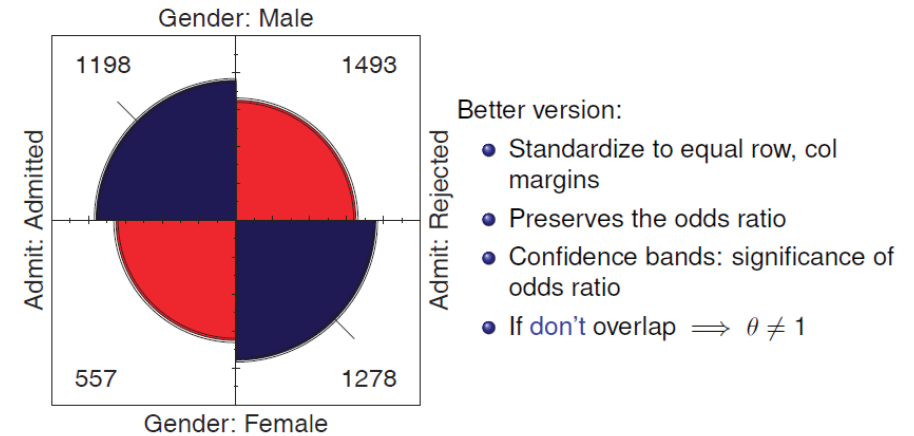
```
fourfold(UCB, std="ind.max") # maximum frequency
```



33

## Visualizing: fourfold plots

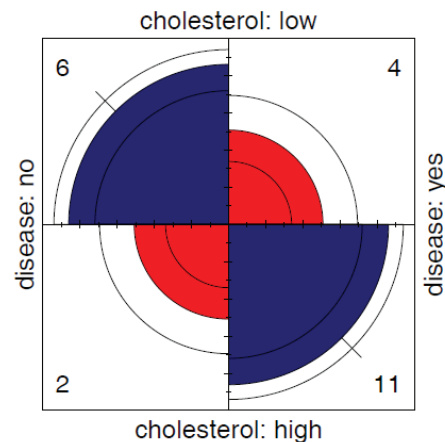
```
fourfold(UCB) #standardize both margins
```



34

## Cholesterol data

```
fourfold(fat)
```



35

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

The UC Berkeley data was obtained from 6 graduate departments

```
> ftable(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)

36

## 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

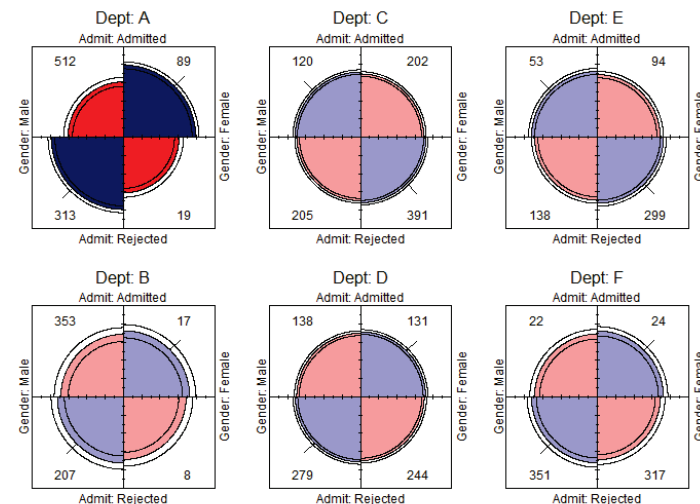
- ❖ Odds ratio only significant,  $\log(\theta) \neq 0$ , for department A
- ❖ For dept. A, men are only  $\exp(-1.05) = .35$  times as likely to be admitted as women
- ❖ The overall analysis (ignoring department) is misleading: falsely assumes no association of {admission, department} and {gender, department}

37

## Stratified fourfold plots

Fourfold plots by department (intense shading where significant)

```
> fourfold(UCBAdmissions)
```

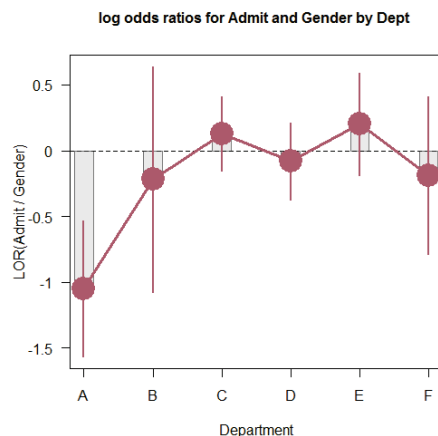


38

## Log odds ratio plot

Plot the log odds ratios with confidence limits

```
> plot(oddsratio(UCBAdmissions), cex=2, xlab="Department")
```



39

## Stratified tables: Homogeneity of association

Questions:

- Are the  $k$  odds ratios all equal,  $\theta_1 = \theta_2 = \dots = \theta_k$ ?
  - Woolf's test: `vcd::woolftest()`
- This is the same as the hypothesis of no three-way association
- If homogeneous, is the common odds ratio different from 1?
  - Mantel-Haenszel test: `stats::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

The odds ratios differ across departments, so no sense testing their common value

40

# What happened at UC Berkeley?

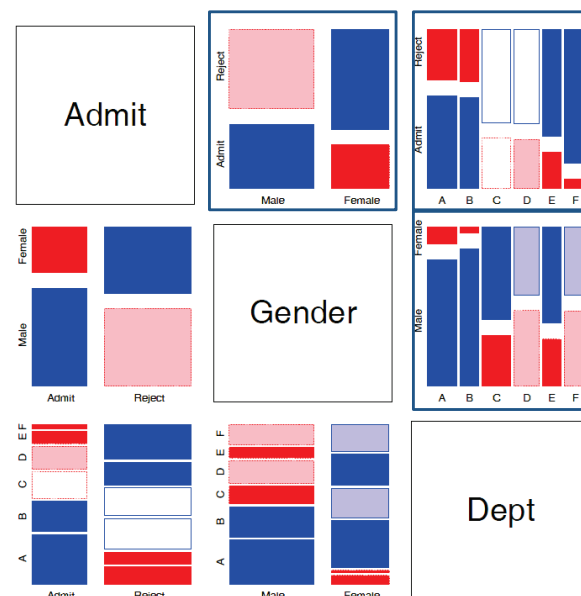
Why do results **collapsed over department** disagree with the results **by department**?

## 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.)

41

# Mosaic matrices



Scatterplot matrix analog for categorical data

All pairwise views  
Small multiples → comparison

The answer: **Simpson's Paradox**

- Depts A, B were easiest
- Applicants to A, B mostly male
- ∴ Males more likely to be admitted **overall**

42

## r × 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
```

43

## r × c tables: Overall analysis

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

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

- 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

Residuals, fitted values, test statistics returned by **MASS::loglm()**

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

44

Residuals and fitted values are obtained with “extractor” methods

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

```
> fitted(mod)
      Hair
Eye      Black  Brown   Red  Blond
Brown  40.1   106.3  26.39  47.2
Blue   39.2   103.9  25.79  46.1
Hazel  17.0   44.9   11.15  20.0
Green  11.7   30.9   7.68   13.7
```

loglm() returns an object (mod) of class “loglm”

Method functions, \*.loglm(), include: residuals(), fitted(), anova(), summary() & various plot methods

Direct calculation of Pearson & LR  $\chi^2$

```
> sum(res.P^2) # Pearson chisq
[1] 138.29
> sum(res.LR^2) # LR chisq
[1] 146.44
```

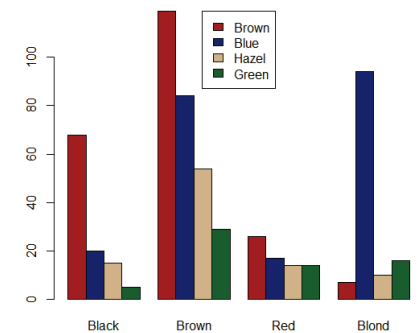
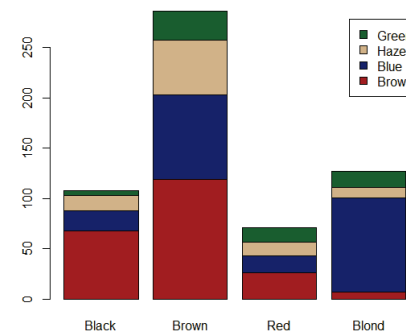
45

## Plots for two-way tables

Barplots are easy, but not often very useful. Why?

```
col <- c("brown", "darkblue", "tan",
  "darkgreen")
barplot(HEC, col = col,
  legend=TRUE)
```

```
barplot(HEC, col = col,
  beside=TRUE, legend=TRUE, ...)
```



46

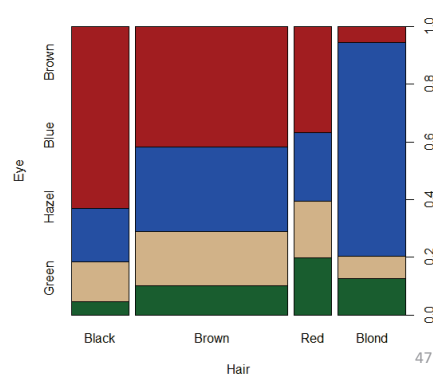
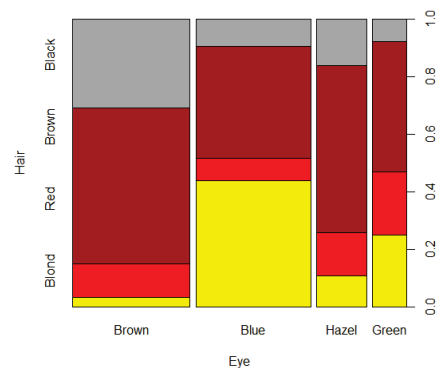
## Spine plots

Spine plots show the **marginal** proportions of one variable, and the **conditional** proportions of the other.

**Independence:** cells align

```
col <- c("darkgrey", "brown", "red",
  "yellow")
spineplot(HEC, col=rev(col))
```

```
col <- c("brown", "blue", "tan",
  "darkgreen")
spineplot(t(HEC), col=rev(col))
```



47

## Tile plots

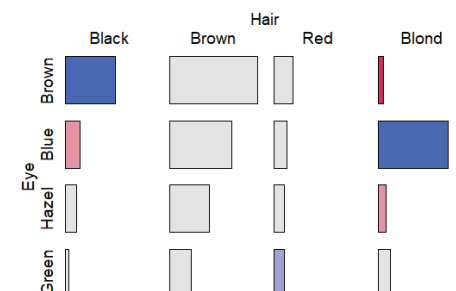
Tile plots show a matrix of rectangular tiles, **area ~ frequency**.

They can be **scaled** to facilitate different types of comparisons: cells, rows, cols

They can be **shaded** to show the sign & magnitude of **residuals** from independence

```
tile(HEC, shade=TRUE, legend=FALSE)
```

```
tile(HEC, tile_type="width", ...)
```

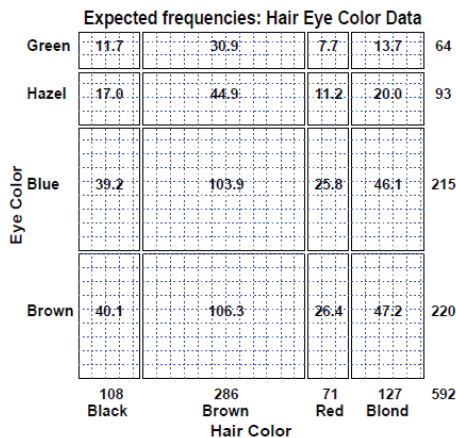


48

# 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)



This display shows **expected** frequencies,  $m_{ij}$ , as # boxes within each cell

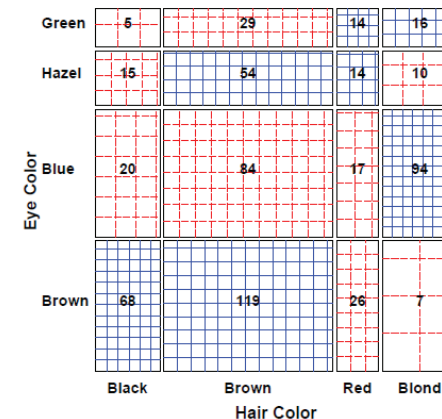
Under independence, boxes all of the same size & equal density

Real sieve diagrams use # boxes = **observed** frequencies,  $n_{ij}$

49

# Sieve diagrams

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



The rectangles have area  $\sim$  expected frequency

# boxes = observed frequency

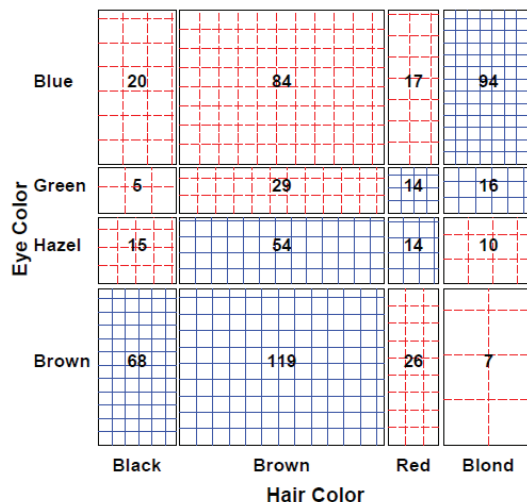
$n_{ij} > m_{ij} \rightarrow$  **greater density**

$n_{ij} < m_{ij} \rightarrow$  **less density**

50

## Sieve diagrams: Effect ordering

Permuting the rows / cols to make the **pattern** more coherent



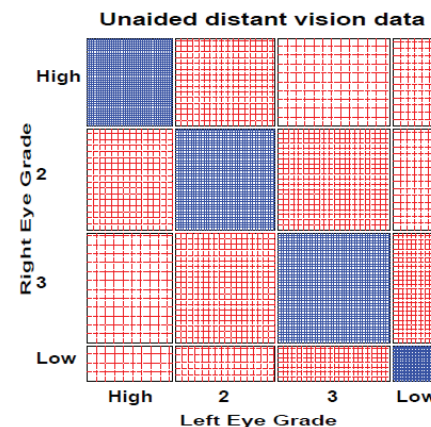
Here, I reordered the eye colors according to lightness

The opposite-corner pattern suggests an explanation for the association

51

## Sieve diagrams: Subtle patterns

Vision classification of 7477 women in Royal Ordnance factories: visual acuity grade in left & right eyes



- The obvious association is apparent in the diagonal cells
- A more subtle pattern appears in the **off-diagonal** cells
- Analysis methods for **square** tables allow testing hypotheses beyond independence
  - **Symmetry**
  - **Quasi-symmetry**, ...

52

## Ordinal factors

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

More powerful CMH tests:

- When either row or col levels are **ordered**, more specific CMH (Cochran–Mantel–Haenszel) tests which take order into account have greater **power** to detect ordered relations.
  - Use fewer df, so ordinal tests are more focused on detecting a particular “signal”
- This is similar to testing for **linear trends** in ANOVA
- Essentially, these assign **scores** to the categories & test for differences in row / col means, or non-zero correlation

53

## CMH tests for ordinal factors

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  $G^2$ .

54

## 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>

55

## 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

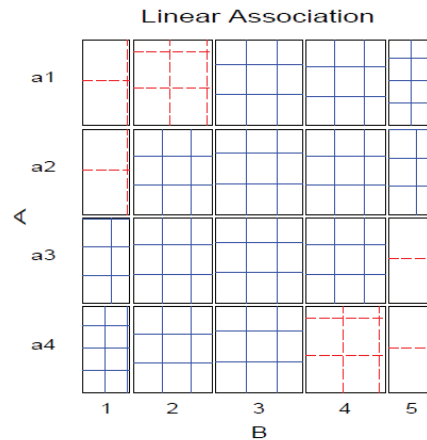
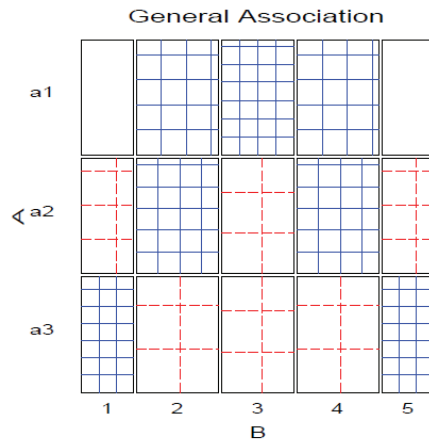
56



# Visualizing the association

The association here is U-shaped  
Only general association detects this

Higher levels of A are associated  
with lower levels of B



57

# Example: Mental health data

For the mental health data, both **ses** and **mental** are ordinal

All tests are significant, but the nonzero correlation test, with 1 df has the smallest p-value & largest  $\chi^2$  / df

```
> 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

both ordinal  
cols ordinal  
rows ordinal  
neither

$\chi^2$  / df shows why ordered tests are more powerful

```
> xx <- CMHtest(mental.tab)
> xx$table[, "Chisq"] / xx$table[, "Df"]
cor rmeans cmeans general
37.16 8.06 13.56 3.06
```

58

# 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,  $P_c = \sum p_{i+} p_{+i}$ .

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

59

# 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				Fleiss-Cohen Weights			
1	2/3	1/3	0	1	8/9	5/9	0
2/3	1	2/3	1/3	8/9	1	8/9	5/9
1/3	2/3	1	2/3	5/9	8/9	1	8/9
0	1/3	2/3	1	0	5/9	8/9	1

60



## Example: Cohen's $\kappa$

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	7	7	2	3		19
Fairly often	2	8	3	7		20
Very often	1	5	4	9		19
Almost always	2	8	9	14		33
SUM	12	28	18	33		91

61

## Example: Cohen's $\kappa$

**vcd::Kappa()** 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

62

## 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

To what extent to the neurologists agree?

Do they agree equally for the patients for the two cities

63

## Observer agreement: Multiple strata

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

```
data(MSPatients, package="vcd")
Kappa(MSPatients[,1])
```

Winnipeg patients

```
##      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])
```

New Orleans patients

```
##      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
```

Somewhat larger agreement for the New Orleans patients

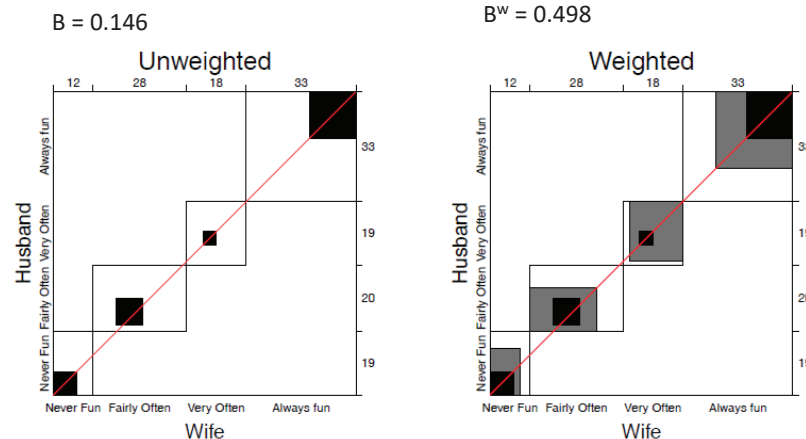
The **irr** package (inter-rater-reliability) provides ICC and other measures; also handles the case of  $k > 2$  raters

64

## Bangdiwala's Observer agreement chart

The observer agreement chart (Bangdiawala, 1987) provides:

- A simple graphic representation of the strength of agreement
- A measure of strength of agreement with an intuitive interpretation



65

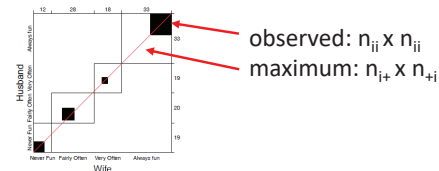
## Bangdiwala's Observer agreement chart

### 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  $B$ :

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

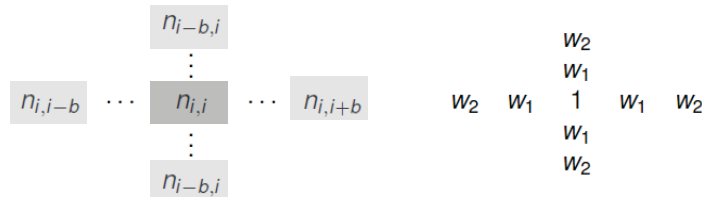
- $\Rightarrow$  Perfect agreement:  $B = 1$ , all rectangles are completely filled.



66

## Weighted agreement chart: Partial agreement

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



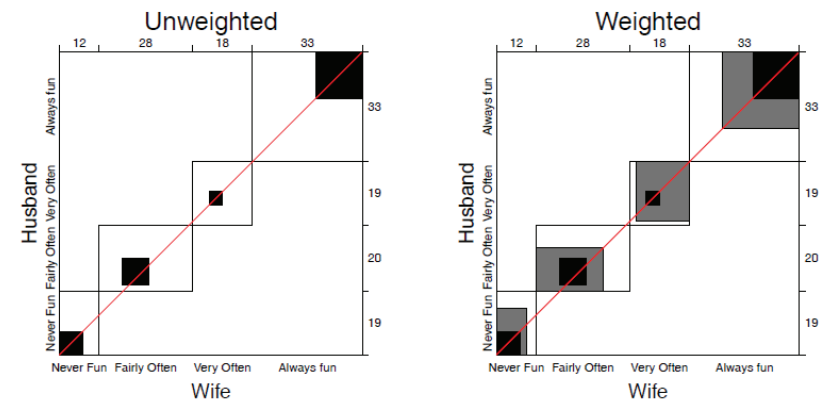
- Add shaded rectangles, size  $\sim$  sum of frequencies,  $A_{bi}$ , within  $b$  steps of main diagonal
- $\Rightarrow$  weighted measure of agreement,

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

67

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

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

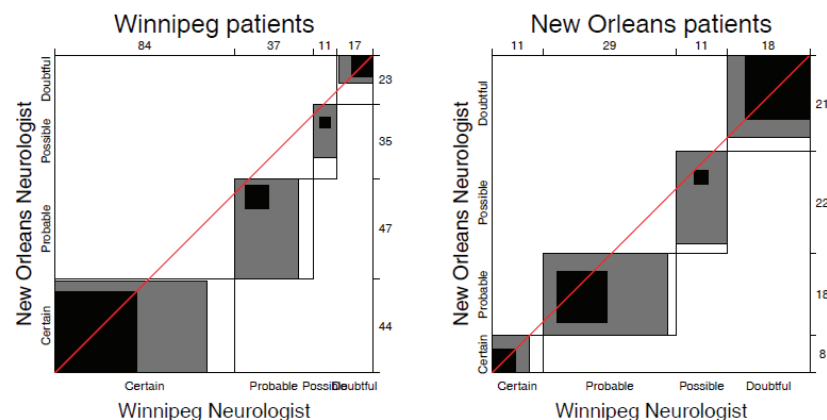


The smallest exact agreement occurs for "very often", but husbands & wives more on this allowing  $\pm 1$  step disagreement

68

## Marginal homogeneity & 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

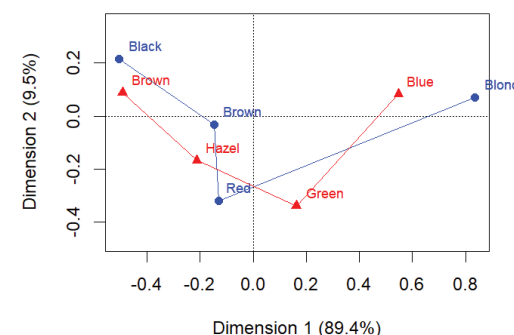
69

## Looking ahead: Correspondence analysis

### Like PCA for categorical data

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

Dim 1: dark to light  
Dim 2: something about red hair, green eyes?



70

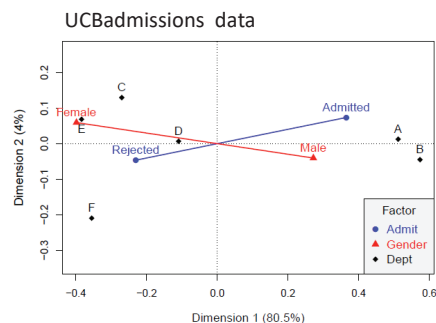
## Looking ahead: Correspondence analysis

### Multiple correspondence analysis extends this to 3+ way tables

- Analyses **all two-way** associations together
- Category points: nearness indicates positive associations

Dim 1: Admission  
Dim 2: ??? (only 4%)

The relations of Dept to Gender and Admit are easy to interpret



71

## Looking ahead: Models

### Loglinear models [loglm()]

- 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

### Generalized linear models [glm()]

- Similar to ordinary  $\text{lm}()$ , but w/ Poisson  $\text{dist}^n$  of counts: family="poisson"
- Formula notation:  $\text{Freq} \sim A + B + C$ ;  $\text{Freq} \sim (A + B + C)^2$
- Familiar diagnostic methods & plots (outliers, influence)

72

## Looking ahead: Models

### Example: UC Berkeley data

- **Mutual** independence:  $[Admit][Gender][Dept] = \sim A + G + D$
- **Joint** independence:  $[Admit][Gender Dept] = \sim A + G * D$
- **Conditional** independence:  $[D Admit][D Gender] = \sim D * (A + G)$ 
  - Specific test of absence of gender bias, **controlling** for department
- **No three-way** association:  $[A G][A D][G D] = \sim (A + D + G)^2$

```
library(MASS)
loglm(~ Admit + Dept + Gender, data=UCBAdmissions) # mutual independence
loglm(~ Admit + Dept * Gender, data=UCBAdmissions) # joint independence
loglm(~ Dept * (Admit + Gender), data=UCBAdmissions) # conditional independence
loglm(~ (Admit + Gender + Dept )^2, data=UCBAdmissions) # all two-way, no three-way
```

#### Bracket notation:

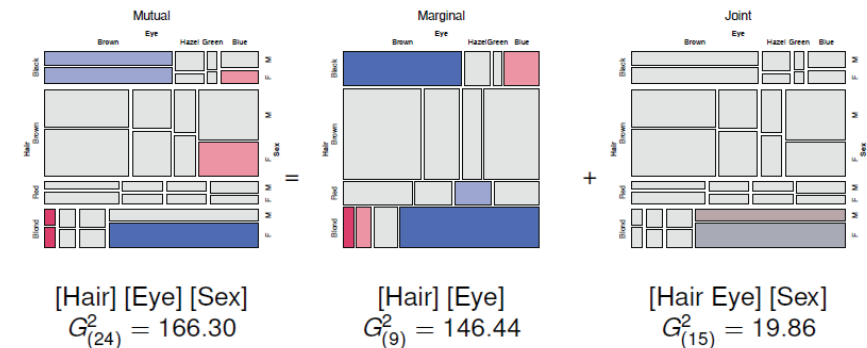
- terms in the **same** bracket are allowed to be **associated**  $[A G] \equiv A * G$
- terms in **separate** brackets are asserted to be **independent**  $[A] [G] \equiv A + G$

73

## Looking ahead: Mosaic plots

Mosaic plots provide visualizations of associations in 2+ way tables

- Tiles  $\sim$  **frequency**; conditioned by A, then B, then C, ...
- Fit: any loglinear model  $[A][B][C]$ ,  $[AB][C]$ ,  $[AB][AC]$ , ...,  $[ABC]$
- Shading:  $\sim$  **residuals**, contributions to  $\chi^2$
- Show: associations **not accounted** for by model



74

## Summary

- Two-way tables summarize frequencies of two categorical factors
  - $2 \times 2$ : a special case, with **odds ratio** as a measure
  - $r \times c$ : factors can be **unordered** or **ordered**
  - $r \times c \times k$ : stratified tables,  $r \times c$  with groups or circumstances
- Tests & measures of association
  - Pearson  $\chi^2$ , LR  $G^2$ : **general association**
  - More powerful **CMH tests** for ordered factors
- Visualization
  - $2 \times 2$ : fourfold plots
  - $r \times c$ : sieve diagrams, tile plots, ...
  - More graphical methods to come ...

75