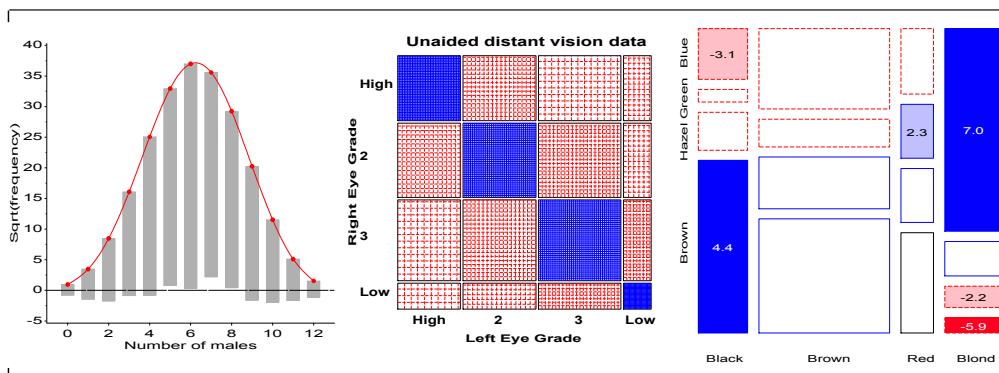


Michael Friendly

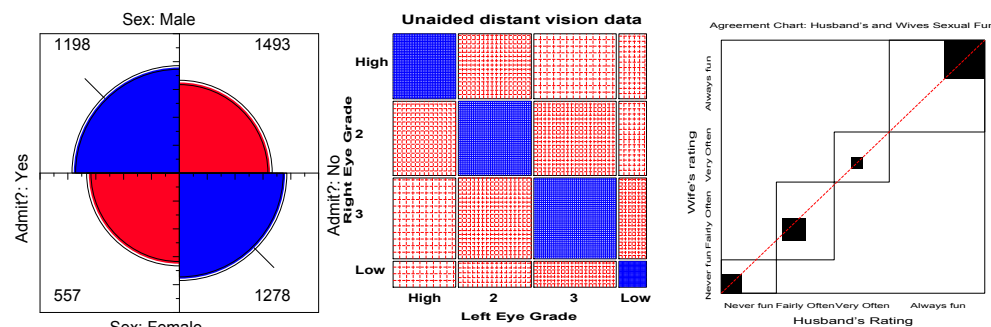
York University

SCS Short Course, 2016

Web notes: datavis.ca/courses/VCD/



Part 2: Visualizing two-way and n -way tables



Topics:

- 2×2 tables and fourfold displays
- Sieve diagrams
- Observer agreement
- Correspondence analysis

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:

- 2×2 tables: two binary factors (e.g., gender, admitted?, died?, ...)
- $2 \times 2 \times k$ tables: a collection of 2×2 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: 2×2 — odds ratio; $r \times c$ — Pearson χ^2 , LR G^2
- How to understand the **pattern** or **nature** of association?

Two-way tables: Examples

2×2 table: Admissions to graduate programs at U. C. Berkeley

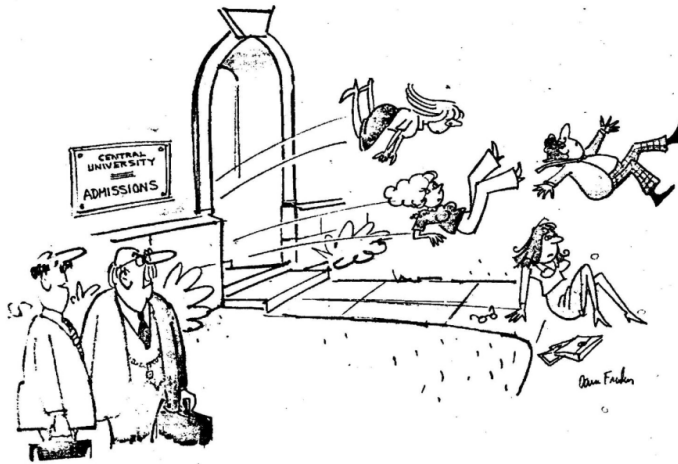
Table: Admissions to Berkeley graduate programs

	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?

Is this evidence for gender bias in admission?



"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?

2 × 2 tables: UCB data

In R, the data is contained in UCBAAdmissions, a 2 × 2 × 6 table for 6 departments. Collapse over department:

```
data(UCBAAdmissions)
UCB <- margin.table(UCBAAdmissions, 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)
```

```
## [1] 1.8411
```

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

```
##          lwr    upr
## [1,] 1.6244 2.0867
```

6 / 56

Standard analysis: PROC FREQ

PROC FREQ gives the standard Pearson and LR χ^2 tests:

```
1 proc freq data=berkeley;
2   weight freq;
3   tables gender*admit / chisq;
```

Output:

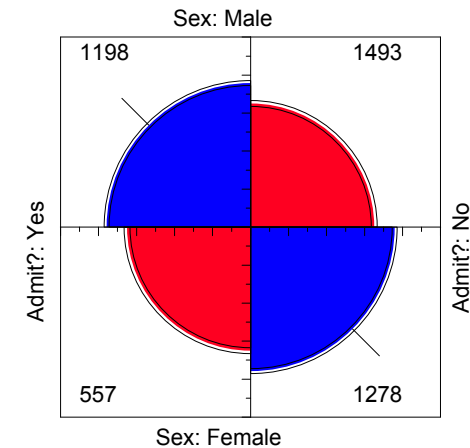
Statistics for Table of gender by admit

Statistic	DF	Value	Prob
Chi-Square	1	92.2053	<.0001
Likelihood Ratio Chi-Square	1	93.4494	<.0001
Continuity Adj. Chi-Square	1	91.6096	<.0001
Mantel-Haenszel Chi-Square	1	92.1849	<.0001
Phi Coefficient		0.1427	

How to visualize and interpret?

Fourfold displays for 2 × 2 tables

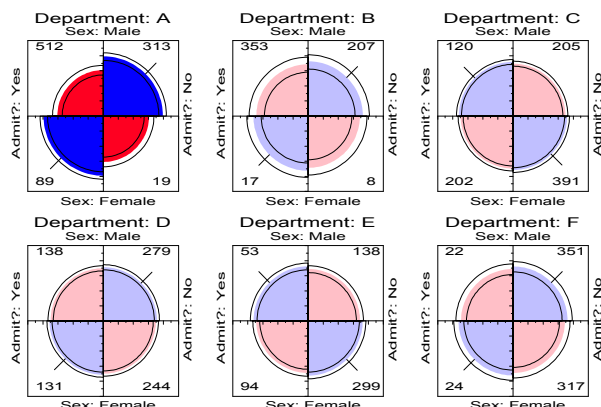
- **Quarter circles:** radius $\sim \sqrt{n_{ij}} \Rightarrow$ area \sim frequency
- **Independence:** Adjoining quadrants \approx align
- **Odds ratio:** ratio of areas of diagonally opposite cells
- **Confidence rings:** Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining rings overlap



- Confidence rings do not overlap: $\theta \neq 1$ (reject H_0)

Fourfold displays for $2 \times 2 \times k$ tables

- Data in Table 2 had been pooled over departments
- Stratified analysis: one fourfold display for each department
- Each 2×2 table standardized to equate marginal frequencies
- Shading: highlight departments for which $H_a : \theta_i \neq 1$



- Only one department (A) shows association; $\theta_A = 0.349 \rightarrow$ women $(0.349)^{-1} = 2.86$ times as likely as men to be admitted.

9/56

What happened here?

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.)

10/56

The FOURFOLD program and the FFOLD macro

- The **FOURFOLD** program is written in SAS/IML.
- The **FFOLD** macro provides a simpler interface.
- Printed output: (a) significance tests for individual odds ratios, (b) tests of homogeneity of association (here, over departments) and (c) conditional association (controlling for department).

Plot by department:

`berk4f.sas`

```

1 %include catdata(berkeley);
2
3 %ffold(data=berkeley,
4   var=Admit Gender,      /* panel variables */
5   by=Dept,               /* stratify by dept */
6   down=2, across=3,     /* panel arrangement */
7   htext=2);             /* font size */

```

Aggregate data: first sum over departments, using the **TABLE** macro:

```

8 %table(data=berkeley, out=berk2,
9   var=Admit Gender,      /* omit dept */
10  weight=count,          /* frequency variable */
11  order=data);
12 %ffold(data=berk2, var=Admit Gender);

```

11/56

Odds ratio plots

```

> library(vcd)
> oddsratio(UCBAdmissions, log=FALSE)

```

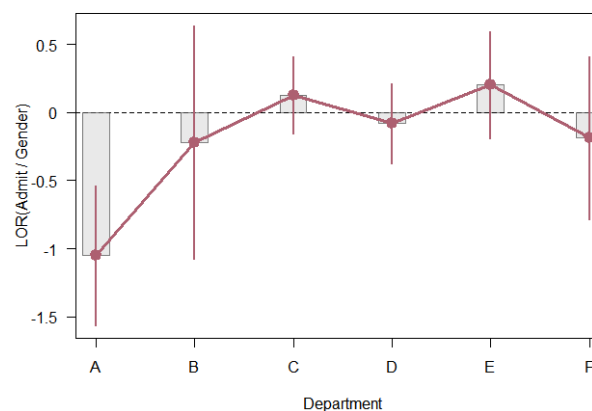
	A	B	C	D	E	F
0.349 0.803 1.133 0.921 1.222 0.828						

```

> lor <- oddsratio(UCBAdmissions) # capture log odds ratios
> plot(lor)

```

log odds ratios for Admit and Gender by Dept



12/56

Two-way frequency tables

Table: Hair-color eye-color data

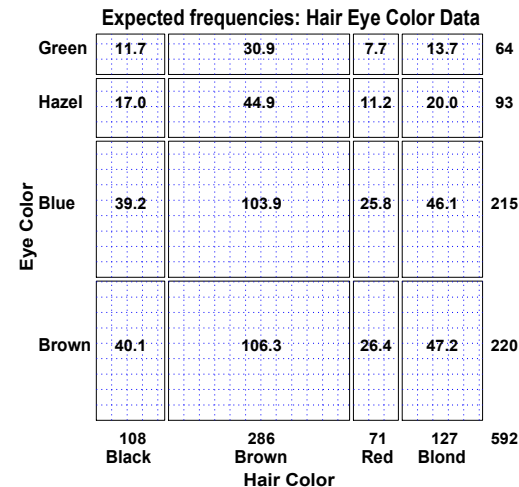
Eye Color	Hair Color				Total
	Black	Brown	Red	Blond	
Green	5	29	14	16	64
Hazel	15	54	14	10	93
Blue	20	84	17	94	215
Brown	68	119	26	7	220
Total	108	286	71	127	592

- With a χ^2 test (PROC FREQ) we can tell that hair-color and eye-color are associated.
- The more important problem is to understand **how** they are associated.
- Some graphical methods:
 - Sieve diagrams
 - Agreement charts (for square tables)
 - Mosaic displays

Two-way frequency tables: Sieve diagrams

• 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)



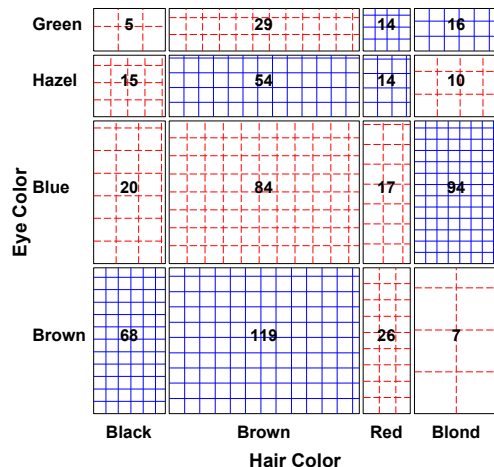
- This display shows **expected frequencies**, assuming independence, as # boxes within each cell
- The boxes are all of the same size (equal density)
- Real sieve diagrams use # boxes = **observed frequencies**, n_{ij}

13 / 56

14 / 56

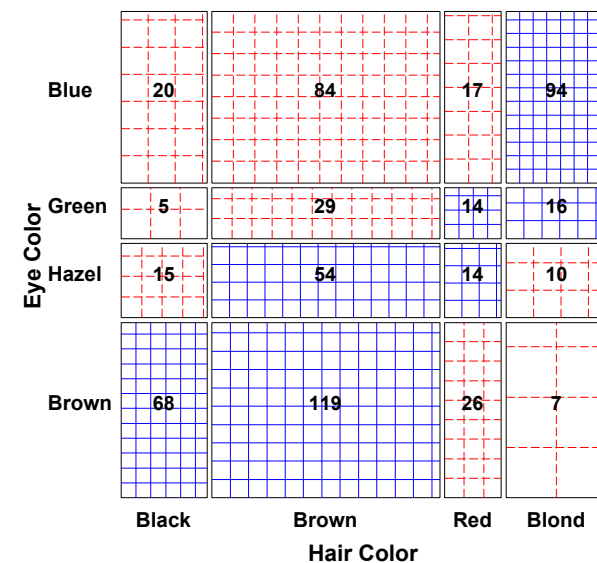
Sieve diagrams

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



Sieve diagrams

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

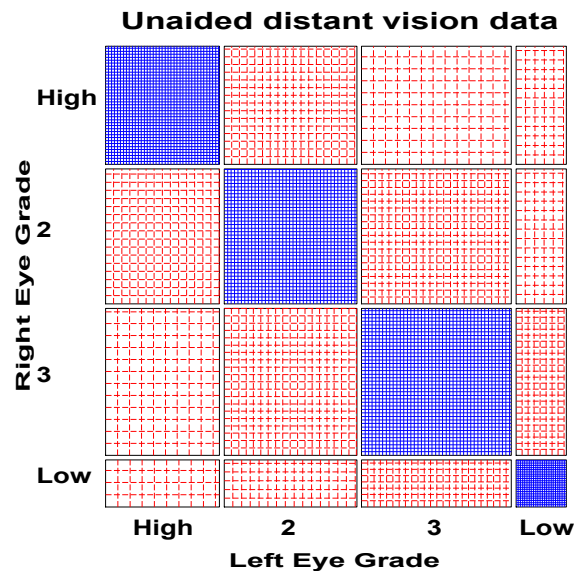


15 / 56

16 / 56

Sieve diagrams

- Vision classification data for 7477 women



17 / 56

Sieve diagrams: SAS Example

sievem.sas

```

1 data vision;
2   do Left='High', '2', '3', 'Low';
3     do Right='High', '2', '3', 'Low';
4       input count @@; output;
5     end;
6   end;
7   label left='Left Eye Grade' right='Right Eye Grade';
8 datalines;
9   1520  266  124  66
10  234  1512  432  78
11  117  362  1772  205
12   36   82  179  492
13 ;
14 %sieveplot(data=vision, var=Left Right,
15            title=Unaided distant vision data);

```

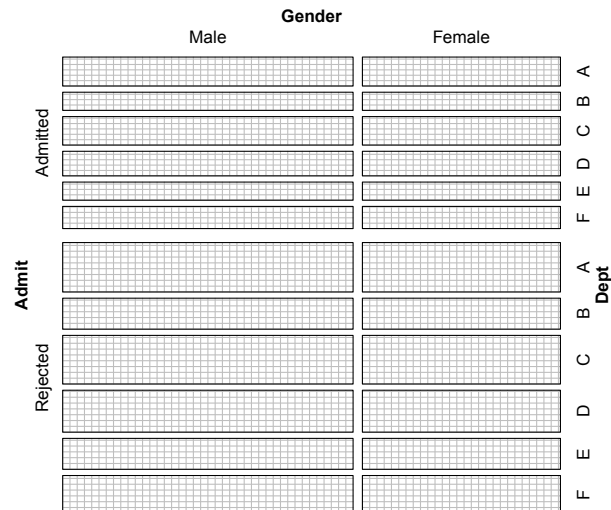
Online weblet: <http://datavis.ca/online/sieve/>

18 / 56

Sieve diagrams: n-way tables in R

```
> sieve(UCBAdmissions, sievetype='expected')
```

Berkeley Data: Mutual Independence (exp)

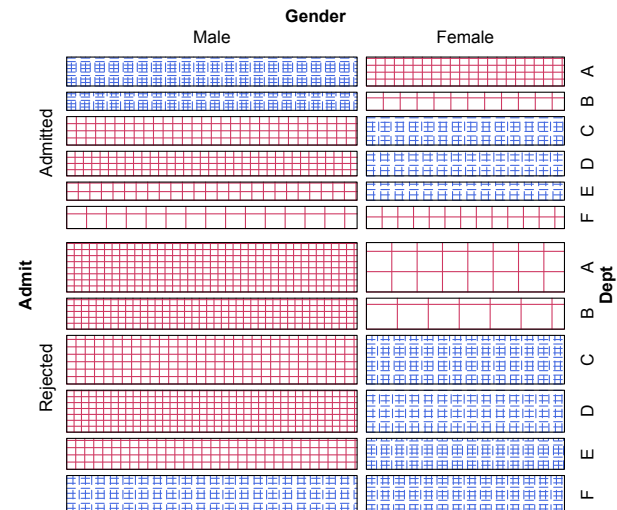


19 / 56

Sieve diagrams: n-way tables in R

```
> sieve(UCBAdmissions, shade=TRUE)
```

Berkeley data: Mutual independence (obs)



20 / 56

Observer Agreement

- **Inter-observer agreement** often used as to assess reliability of a subjective classification or assessment procedure
 - → square table, Rater 1 × 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 κ : compares the observed agreement, $P_o = \sum p_{ii}$, to agreement expected by chance if the two observer's ratings were independent, $P_c = \sum p_{i+} p_{+i}$.

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

21 / 56

Cohen's κ

Properties of Cohen's κ :

- perfect agreement: $\kappa = 1$
- minimum κ may be < 0 ; lower bound depends on marginal totals
- Unweighted κ : counts only diagonal cells (same category assigned by both observers).
- Weighted κ : 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

22 / 56

Cohen's κ : 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	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

23 / 56

Cohen's κ : R Example

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

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

##           value   ASE    z
## Unweighted 0.129 0.0686 1.89
## Weighted   0.237 0.0783 3.03

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

##           value   ASE    z
## Unweighted 0.129 0.0686 1.89
## Weighted   0.332 0.0973 3.41
```

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

24 / 56

Computing κ with SAS

- PROC FREQ: Use AGREE option on TABLES statement
 - Gives both unweighted and weighted κ (default: CA weights)
 - AGREE (wt=FC) uses Fleiss-Cohen weights
 - Bowker's (Bowker, 1948) test of symmetry: $H_0 : p_{ij} = p_{ji}$

kappa3.sas

```

1 title 'Kappa for Agreement';
2 data fun;
3   do Husband = 1 to 4;
4     do Wife = 1 to 4;
5       input count @@;
6       output;
7     end; end;
8 datalines;
9 7 7 2 3
10 2 8 3 7
11 1 5 4 9
12 2 8 9 14
13 ;
14 proc freq;
15   weight count;
16   tables Husband * Wife / noprint agree; /* default: CA weights*/
17   tables Husband * Wife / noprint agree(wt=FC);

```

25 / 56

Computing κ with SAS

Output (CA weights):

Statistics for Table of Husband by Wife				
Test of Symmetry				
Statistic (S)	3.8778			
DF	6			
Pr > S	0.6932			
Kappa Statistics				
Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.1293	0.0686	-0.0051	0.2638
Weighted Kappa	0.2374	0.0783	0.0839	0.3909
Sample Size = 91				

Using Fleiss-Cohen weights:

Weighted Kappa	0.3320	0.0973	0.1413	0.5227
----------------	--------	--------	--------	--------

26 / 56

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 multiple sclerosis by two neurologists, for two populations (Landis and Koch, 1977)

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

Analysis:

```

proc freq;
  tables strata * rater1 * rater2 / agree;

```

27 / 56

Observer agreement: Multiple strata

msdiag.sas

```

1 data msdiag;
2   do patients='Winnipeg ', 'New Orleans';
3     do N_rating = 1 to 4;
4       do W_rating = 1 to 4;
5         input count @;
6         output;
7       end;
8     end;
9   end;
10  label N_rating = 'New Orleans neurologist'
11      W_rating = 'Winnipeg neurologist';
12  datalines;
13 38 5 0 1
14 33 11 3 0
15 10 14 5 6
16 3 7 3 10
17 5 3 0 0
18 3 11 4 0
19 2 13 3 4
20 1 2 4 14
21 ;
22
23 *-- Agreement, separately, and controlling for Patients;
24 proc freq data=msdiag;
25   weight count;
26   tables patients * N_rating * W_rating / norow nocol nopct agree;

```

28 / 56

Observer agreement: Multiple strata

Output, strata 1: (New Orleans patients):

Statistics for Table 1 of N_rating by W_rating Controlling for patients=New Orleans				
Test of Symmetry				
Statistic (S)	9.7647			
DF	6			
Pr > S	0.1349			
Kappa Statistics				
Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2965	0.0785	0.1427	0.4504
Weighted Kappa	0.4773	0.0730	0.3341	0.6204
Sample Size = 69				

29 / 56

Observer agreement: Multiple strata

Output, strata 2: (Winnipeg patients):

Statistics for Table 2 of N_rating by W_rating Controlling for patients=Winnipeg				
Test of Symmetry				
Statistic (S)	46.7492			
DF	6			
Pr > S	<.0001			
Kappa Statistics				
Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2079	0.0505	0.1091	0.3068
Weighted Kappa	0.3797	0.0517	0.2785	0.4810
Sample Size = 149				

30 / 56

Observer agreement: Multiple strata

Overall test:

Summary Statistics for N_rating by W_rating Controlling for patients				
Overall Kappa Coefficients				
Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.2338	0.0424	0.1506	0.3170
Weighted Kappa	0.4123	0.0422	0.3296	0.4949

Homogeneity test: $H_0 : \kappa_1 = \kappa_2 = \dots = \kappa_k$

Tests for Equal Kappa Coefficients			
Statistic	Chi-Square	DF	Pr > ChiSq
Simple Kappa	0.9009	1	0.3425
Weighted Kappa	1.1889	1	0.2756
Total Sample Size = 218			

31 / 56

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 :

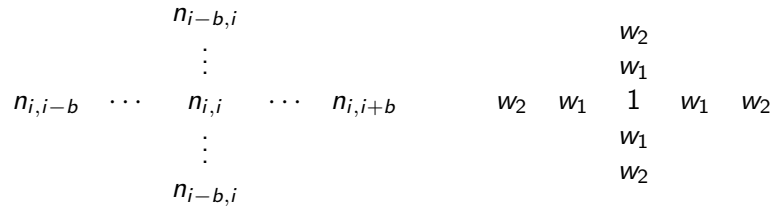
$$B = \frac{\text{area of dark squares}}{\text{area of rectangles}} = \frac{\sum_i^k n_{ii}^2}{\sum_i^k n_{i+} n_{+i}}$$

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

32 / 56

Weighted Agreement Chart: Partial agreement

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

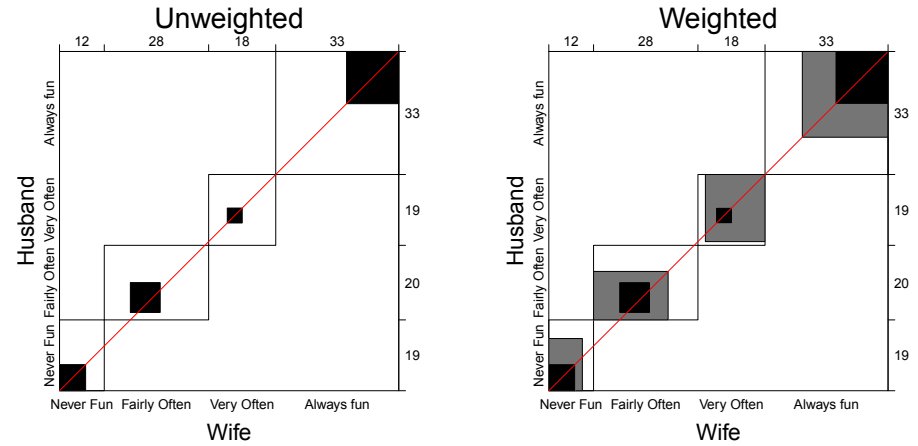


- Add shaded rectangles, size \sim sum of frequencies, A_{bi} , within b steps of main diagonal
- \Rightarrow weighted measure of agreement,

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

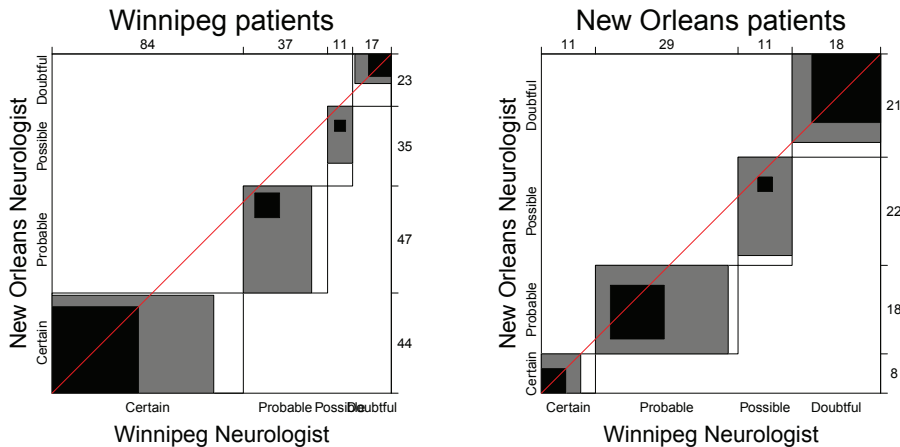
Husbands and wives: $B = 0.146$, $B^w = 0.498$

```
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

Correspondence analysis

Correspondence analysis (CA)

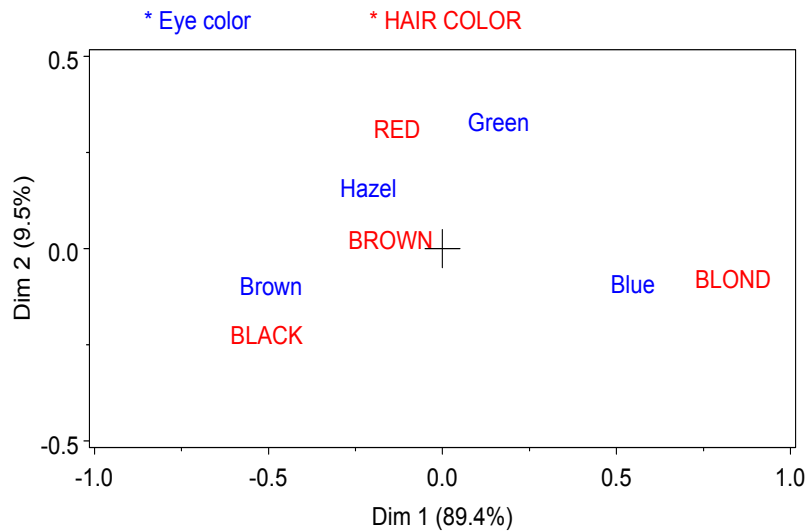
Analog of PCA for frequency data:

- account for maximum % of χ^2 in few (2-3) dimensions
- finds scores for row (x_{im}) and column (y_{jm}) categories on these dimensions
- uses Singular Value Decomposition of residuals from independence, $d_{ij} = (n_{ij} - \hat{m}_{ij}) / \sqrt{\hat{m}_{ij}}$

$$\frac{d_{ij}}{\sqrt{n}} = \sum_{m=1}^M \lambda_m x_{im} y_{jm}$$

- **optimal scaling**: each pair of scores for rows (x_{im}) and columns (y_{jm}) have highest possible correlation ($= \lambda_m$).
- plots of the row (x_{im}) and column (y_{jm}) scores show associations

Hair color, Eye color data:



- Interpretation: row/column points “near” each other are positively associated
- Dim 1: 89.4% of χ^2 (dark \leftrightarrow light)
- Dim 2: 9.5% of χ^2 (RED/Green vs. others)

37 / 56

PROC CORRESP and the CORRESP macro

- Two forms of input dataset:
 - dataset in *contingency table* form – column variables are levels of one factor, observations (rows) are levels of the other.

Obs	Eye	BLACK	BROWN	RED	BLOND
1	Brown	68	119	26	7
2	Blue	20	84	17	94
3	Hazel	15	54	14	10
4	Green	5	29	14	16

- Raw category responses (*case form*), or cell frequencies (*frequency form*), classified by 2 or more factors (e.g., output from PROC FREQ)

Obs	Eye	HAIR	Count
1	Brown	BLACK	68
2	Brown	BROWN	119
3	Brown	RED	26
4	Brown	BLOND	7
...			
15	Green	RED	14
16	Green	BLOND	16

38 / 56

Software: PROC CORRESP, CORRESP macro & R

• PROC CORRESP

- Handles 2-way CA, extensions to n -way tables, and MCA
- Many options for scaling row/column coordinates and output statistics
- OUTC= option \rightarrow output dataset for plotting
- SAS V9.1+: PROC CORRESP uses ODS Graphics

• CORRESP macro

- Uses PROC CORRESP for analysis
- Produces labeled plots of the category points in either 2 or 3 dimensions
- Many graphic options; can equate axes automatically
- See: <http://datavis.ca/sasmac/corresp.html>

• R

- The `ca` package provides 2-way CA, MCA and more
- `plot(ca(data))` gives reasonably useful plots
- Other R packages: `vegan`, `ade4`, `FactoMiner`, ...

39 / 56

Example: Hair and Eye Color

- **Input the data** in contingency table form

```

1 data haireye;
2   input  EYE $ BLACK BROWN RED BLOND ;
3   datalines;
4     Brown    68    119    26    7
5     Blue     20     84    17   94
6     Hazel    15     54    14   10
7     Green     5     29    14   16
8 ;

```

40 / 56

Example: Hair and Eye Color

- Using **PROC CORRESP** directly— ODS graphics (V9.1+)

```
ods rtf; /* ODS destination: rtf, html, latex, ... */
ods graphics on;
proc corresp data=haireye short;
  id eye; /* row variable */
  var black brown red blond; /* col variables */
ods graphics off;
ods rtf close;
```

- Using the **CORRESP** macro— labeled high-res plot

```
%corresp (data=haireye,
  id=eye, /* row variable */
  var=black brown red blond, /* col variables */
  dimlab=Dim); /* options */
```

41 / 56

Example: Hair and Eye Color

Printed output:

```
The Correspondence Analysis Procedure
      Inertia and Chi-Square Decomposition

Singular  Principal  Chi-
Values    Inertias  Squares  Percents  18  36  54  72  90
-----
0.45692  0.20877  123.593  89.37%  *****
0.14909  0.02223   13.158   9.51%  ***
0.05097  0.00260    1.538   1.11%
-----
0.23360  138.29 (Degrees of Freedom = 9)

      Row Coordinates
                Dim1          Dim2
Brown          -.492158        -.088322
Blue           0.547414        -.082954
Hazel          -.212597        0.167391
Green          0.161753        0.339040

      Column Coordinates
                Dim1          Dim2
BLACK          -.504562        -.214820
BROWN          -.148253        0.032666
RED            -.129523        0.319642
BLOND          0.835348        -.069579
```

42 / 56

Example: Hair and Eye Color

Output dataset(selected variables):

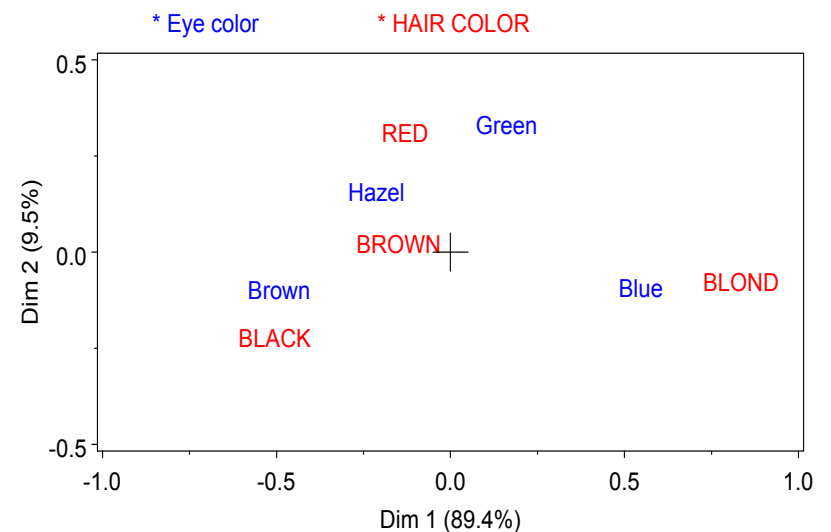
Obs	_TYPE_	EYE	DIM1	DIM2
1	INERTIA	.	.	.
2	OBS	Brown	-0.49216	-0.08832
3	OBS	Blue	0.54741	-0.08295
4	OBS	Hazel	-0.21260	0.16739
5	OBS	Green	0.16175	0.33904
6	VAR	BLACK	-0.50456	-0.21482
7	VAR	BROWN	-0.14825	0.03267
8	VAR	RED	-0.12952	0.31964
9	VAR	BLOND	0.83535	-0.06958

Row and column points are distinguished by the `_TYPE_` variable: OBS vs. VAR

43 / 56

Example: Hair and Eye Color

Graphic output from **CORRESP** macro:



44 / 56

CA in R: the ca

```
> HairEye <- margin.table(HairEyeColor, c(1, 2))
> library(ca)
> ca(HairEye)
```

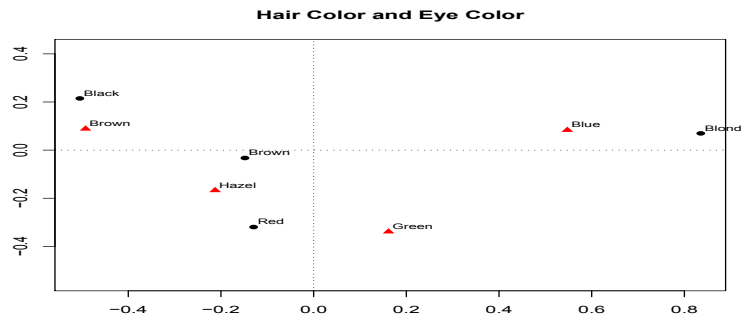
Principal inertias (eigenvalues):

Value	1	2	3
0.208773	0.208773	0.022227	0.002598
Percentage	89.37%	9.52%	1.11%

...

Plot the ca object:

```
> plot(ca(HairEye), main="Hair Color and Eye Color")
```



45 / 56

Multi-way tables

Correspondence analysis can be extended to n -way tables in several ways:

- **Multiple correspondence analysis (MCA)**

- Extends CA to n -way tables
- only uses bivariate associations

- **Stacking approach**

- n -way table flattened to a 2-way table, combining several variables "interactively"
- Each way of stacking corresponds to a *loglinear model*
- Ordinary CA of the flattened table \rightarrow visualization of that model
- Associations among stacked variables are *not visualized*

- Here, I only describe the stacking approach, and only with SAS

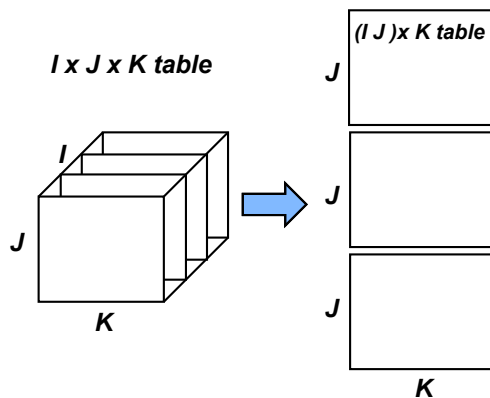
- In SAS 9.3, the MCA option with PROC CORRESP provides some reasonable plots.
- For R, see the `ca`— the `mjca()` function is much more general

46 / 56

Multi-way tables: Stacking

- **Stacking approach:** van der Heijden and de Leeuw (1985)—

- three-way table, of size $I \times J \times K$ can be sliced and stacked as a two-way table, of size $(I \times J) \times K$



- The variables combined are treated "interactively"
- Each way of stacking corresponds to a loglinear model
 - $(I \times J) \times K \rightarrow [AB][C]$
 - $I \times (J \times K) \rightarrow [A][BC]$
 - $J \times (I \times K) \rightarrow [B][AC]$
- Only the associations in separate $[[$ terms are analyzed and displayed

47 / 56

Multi-way tables: Stacking

- **PROC CORRESP:** Use TABLES statement and option CROSS=ROW or CROSS=COL. E.g., for model $[A B] [C]$,

```
proc corresp cross=row;
  tables A B, C;
  weight count;
```

- **CORRESP macro:** Can use / instead of ,

```
%corresp(
  options=cross=row,
  tables=A B/ C,
  weight count);
```

48 / 56

Example: Suicide Rates

Suicide rates in West Germany, by Age, Sex and Method of suicide

Sex	Age	POISON	GAS	HANG	DROWN	GUN	JUMP
M	10-20	1160	335	1524	67	512	189
M	25-35	2823	883	2751	213	852	366
M	40-50	2465	625	3936	247	875	244
M	55-65	1531	201	3581	207	477	273
M	70-90	938	45	2948	212	229	268
F	10-20	921	40	212	30	25	131
F	25-35	1672	113	575	139	64	276
F	40-50	2224	91	1481	354	52	327
F	55-65	2283	45	2014	679	29	388
F	70-90	1548	29	1355	501	3	383

- CA of the [Age Sex] by [Method] table:
 - Shows associations between the Age-Sex combinations and Method
 - Ignores association between Age and Sex

Example: Suicide Rates

suicide5.sas ...

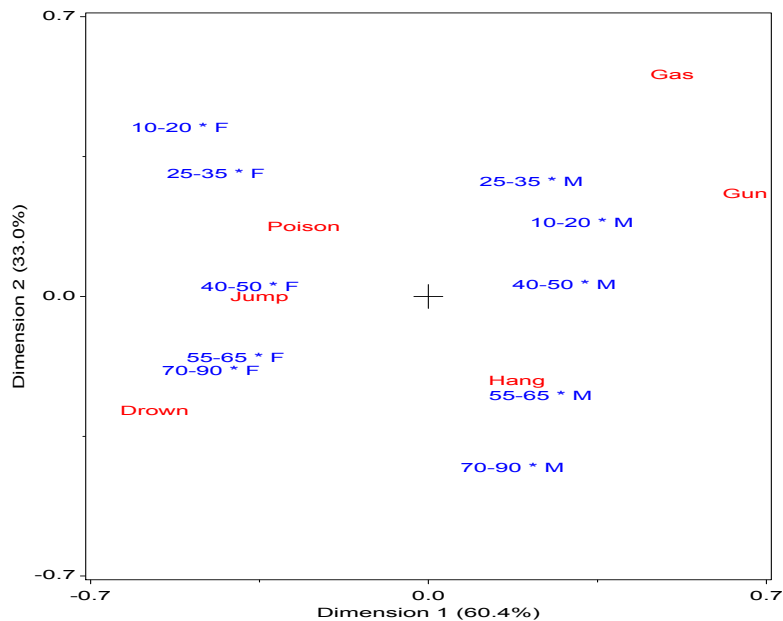
```

1 %include catdata(suicide);
2 *-- equate axes!;
3 axis1 order=(-.7 to .7 by .7) length=6.5 in label=(a=90 r=0);
4 axis2 order=(-.7 to .7 by .7) length=6.5 in;
5 %corresp(data=suicide, weight=count,
6   tables=%str(age sex, method),
7   options=cross=row short,
8   vaxis=axis1, haxis=axis2);
    
```

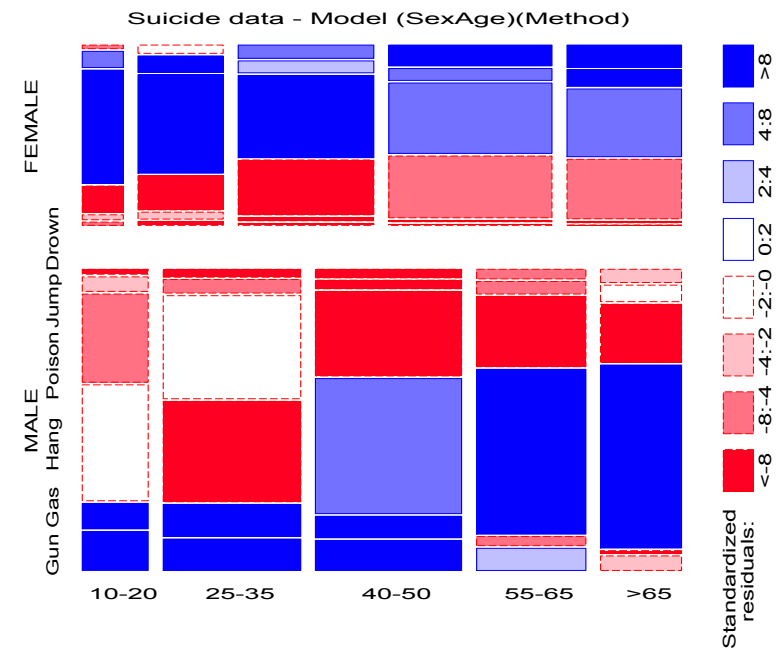
Output:

Inertia and Chi-Square Decomposition								
Singular Values	Principal Inertias	Chi-Squares	Percents	12	24	36	48	60
0.32138	0.10328	5056.91	60.41%	*****				
0.23736	0.05634	2758.41	32.95%	*****				
0.09378	0.00879	430.55	5.14%	**				
0.04171	0.00174	85.17	1.02%					
0.02867	0.00082	40.24	0.48%					
	0.17098	8371.28	(Degrees of Freedom = 45)					

CA Graph:



Looking forward— Loglinear models and mosaic displays:



Summary: Part 2

• Fourfold displays

- Odds ratio: ratio of areas of diagonally opposite quadrants
- Confidence rings: visual test of $H_0 : \theta = 1$
- Shading: highlight strata for which $H_a : \theta \neq 1$

• Sieve diagrams

- Rows and columns \sim marginal frequencies \rightarrow area \sim expected
- Shading \sim observed frequencies
- Simple visualization of pattern of association
- SAS: `sieveplot` macro; R: `sieve()`

• Agreement

- Cohen's κ : strength of agreement
- Agreement chart: visualize weighted & unweighted agreement, marginal homogeneity
- SAS: `agreeplot` macro; R: `agreementplot()`

• Correspondence analysis

- Decompose χ^2 for association into 1 or more dimensions
- \rightarrow scores for row/col categories
- CA plots: Interpretation of *how* the variables are related
- SAS: `corresp` macro; R: `ca()`

53 / 56

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.
- Bickel, P. J., Hammel, J. W., and O'Connell, J. W. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187:398–403, 1975.
- Bowker, A. H. Bowker's test for symmetry. *Journal of the American Statistical Association*, 43:572–574, 1948.
- Friendly, M. Conceptual and visual models for categorical data. *The American Statistician*, 49:153–160, 1995.
- Friendly, M. Multidimensional arrays in SAS/IML. In *Proceedings of the SAS User's Group International Conference*, volume 25, pp. 1420–1427. SAS Institute, 2000.
- Friendly, M. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, 2002.
- Friendly, M. and Kwan, E. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4):509–539, 2003.

54 / 56

References II

- Hoaglin, D. C. and Tukey, J. W. Checking the shape of discrete distributions. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Exploring Data Tables, Trends and Shapes*, chapter 9. John Wiley and Sons, New York, 1985.
- Koch, G. and Edwards, S. Clinical efficiency trials with categorical data. In Peace, K. E., editor, *Biopharmaceutical Statistics for Drug Development*, pp. 403–451. Marcel Dekker, New York, 1988.
- Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174., 1977.
- Mosteller, F. and Wallace, D. L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, New York, NY, 1984.
- Ord, J. K. Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A*, 130:232–238, 1967.
- Tufte, E. R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.
- Tukey, J. W. Some graphic and semigraphic displays. In Bancroft, T. A., editor, *Statistical Papers in Honor of George W. Snedecor*, pp. 292–316. Iowa State University Press, Ames, IA, 1972.

55 / 56

References III

- Tukey, J. W. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, 1977.
- van der Heijden, P. G. M. and de Leeuw, J. Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50:429–447, 1985.

56 / 56