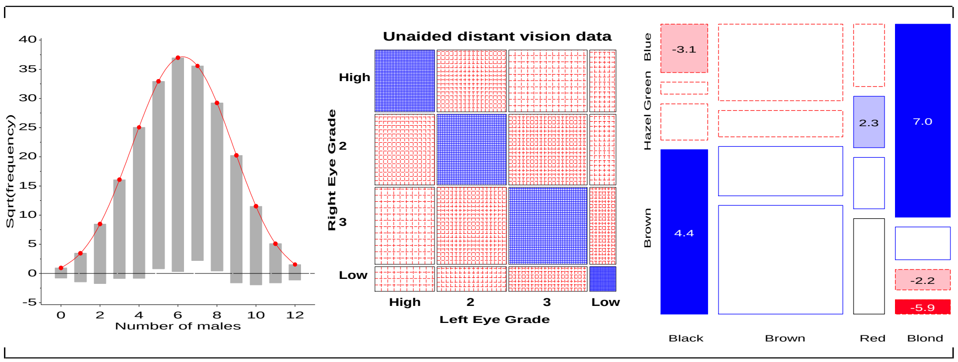
Visualizing Categorical Data with SAS and R

Michael Friendly

York University

Short Course, 2012
Web notes: `datavis.ca/courses/VCD/`



Course goals

Emphasis: visualization methods

- Basic ideas: categorical vs. quantitative data
- Some novel displays: sieve diagrams, fourfold displays, mosaic plots, ...
- Some that extend more familiar ideas to the categorical data setting.

Emphasis: theory ⇒ practice

- Show *what* can be done, in both SAS and R (most in SAS)
- Framework for *thinking* about categorical data analysis in visual terms
- Provide software tools you can *use*

What is included, and what is *not*

- *Some* description of statistical methods— only as necessary
- *Many* software examples— only explained as necessary
- *Too much* material— some skipping may be required

Course structure, Parts 1–3

1. Overview and introduction

- Categorical data? Graphics?
- Discrete distributions
- Testing association

2. Visualizing two-way and n-way tables

- 2 $\times$2 tables; $r \times c$ tables: Fourfold & sieve diagrams
- Observer agreement: Measures and graphs
- Correspondence analysis

3. Mosaic displays and loglinear models

- n-way tables: graphs and models
- Mosaics software
- Structured tables

Course structure, Parts 4–5

4. Logit models and logistic regression

- Logit models; logistic regression models
- Effect plots
- Influence and diagnostic plots

5. Polytomous response models

- Proportional odds models
- Nested dichotomies
- Generalized logits

What is categorical data?

- Simplest case: 1-way frequency distribution
  - Unordered factor

Hair	Black	Brown	Red	Blond
	108	286	71	127

Hair color among 592 students

Party	BQ	Cons	Green	Liberal	NDP	Total
N	104	392	126	404	174	1200
%	8.7	32.6	10.5	33.7	14.5	100

Voting intention in Harris-Decima poll, 8/21/08

  - Ordered, quantitative factor

nMales												
0	1	2	3	4	5	6	7	8	9	10	11	12
3	24	104	286	670	1033	1343	1112	829	478	181	45	7

# of sons in Saxony families with 12 children

What is categorical data?

- Contingency tables ($2 \times 2 \times \ldots$)
  - Two-way

Admit	Gender Male	Female
Admitted	1198	557
Rejected	1493	1278

Admission to graduate programs at UC Berkeley

  - Three-way

... by Department

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

What is categorical data?

- Contingency tables (larger)
  - Two-way

Hair	Eye Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

  - Three-way

Sex	Hair	Eye Brown	Blue	Hazel	Green
Male	Black	32	11	10	3
	Brown	53	50	25	15
	Red	10	10	7	7
	Blond	3	30	5	8
Female	Black	36	9	5	2
	Brown	66	34	29	14
	Red	16	7	7	7
	Blond	4	64	5	8

Table and case-form

- The previous examples were shown in table form
  - # observations = # cells in the table
  - variables: factors + COUNT
- Each has an equivalent representation in case form
  - # observations = total COUNT
  - variables: factors
- Case form is required if there are continuous variables

Sex	Hair	Eye Brown	Blue	Hazel	Green
Male	Black	32	11	10	3
	Brown	53	50	25	15
	Red	10	10	7	7
	Blond	3	30	5	8
Female	Black	36	9	5	2
	Brown	66	34	29	14
	Red	16	7	7	7
	Blond	4	64	5	8

Categorical data: Analysis methods

Methods of analysis for categorical data fall into two main categories:

Non-parametric, randomization-based methods

- Make minimal assumptions
- Useful for hypothesis-testing: Are A and B associated?
- Mostly for two-way tables (possibly stratified)
- SAS: `PROC FREQ`
  - Pearson Chi-square
  - Fisher's exact test (for small expected frequencies)
  - Mantel-Haenszel tests (ordered categories: test for *linear* association)
- R: `chisq.test()`, `mantelhaen.test()`, ...
- SPSS: Crosstabs

Categorical data: Analysis methods

Model-based methods

- Must assume random sample (possibly stratified)
- Useful for estimation purposes: Size of effects (std. errors, confidence intervals)
- More suitable for multi-way tables
- Greater flexibility; fitting specialized models
  - Symmetry, quasi-symmetry, structured associations for square tables
  - Models for ordinal variables
- SAS: `PROC LOGISTIC`, `CATMOD`, `GENMOD` , `INSIGHT` (Fit YX)
  - estimate standard errors, covariances for model parameters
  - confidence intervals for parameters, predicted Pr{response}
- R: `glm()` family, `car` package, `gnm` package, ...
- SPSS: Hiloglinear, Loglinear, Generalized linear models

Categorical data: Response vs. Association models

Response models

- Sometimes, one variable is a natural discrete response.
- Q: How does the response relate to explanatory variables?
  - Admit $\sim$ Gender + Dept
  - Party $\sim$ Age + Education + Urban

$\Rightarrow$ Logit models, logististic regression, generalized linear models

Association models

- Sometimes, the main interest is just association
- Q: Which variables are associated, and how?
  - Berkeley data: [Admit Gender]? [Admit Dept]? [Gender Dept]
  - Hair-eye data: [Hair Eye]? [Hair Sex]? [Eye, Sex]

$\Rightarrow$ Loglinear models

This is similar to the distinction between regression/ANOVA vs. correlation and factor analysis

Graphical methods: Tables and Graphs

If I can't picture it, I can't understand it. Albert Einstein

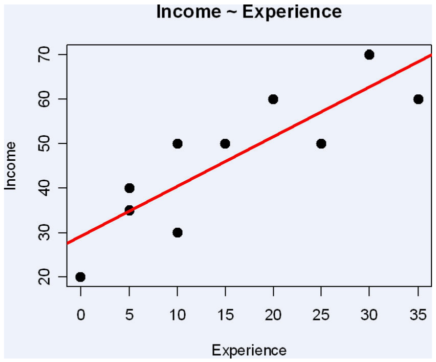
Getting information from a table is like extracting sunlight from a cucumber. Farquhar & Farquhar, 1891

Tables vs. Graphs

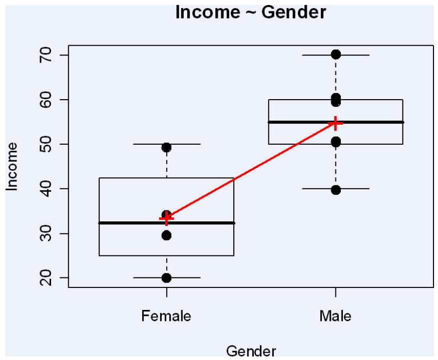
- Tables are best suited for *look-up* and calculation—
  - read off exact numbers
  - additional calculations (e.g., % change)
- Graphs are better for:
  - showing *patterns, trends, anomalies*,
  - making *comparisons*
  - seeing the unexpected!
- Visual presentation as *communication*:
  - what do you want to say or show?
  - design graphs and tables to 'speak to the eyes'

Graphical methods: Quantitative data

Quantitative data (amounts) are naturally displayed in terms of **magnitude ∼ position along a scale**



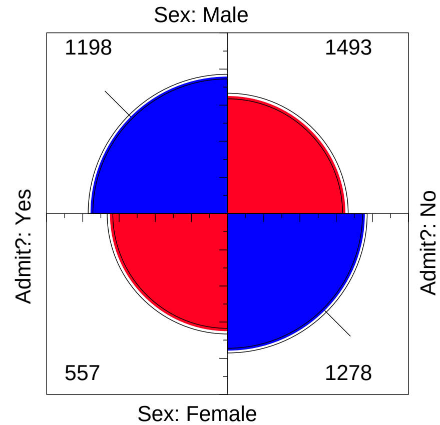
Scatterplot of Income vs. Experience



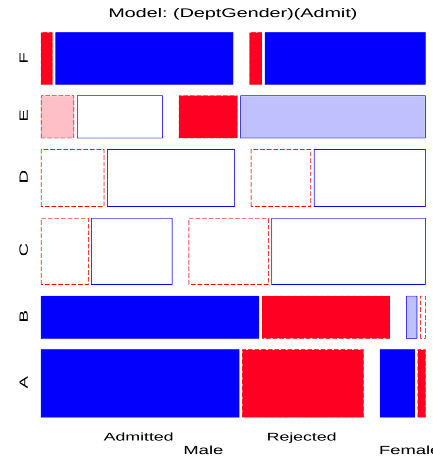
Boxplot of Income by Gender

Graphical methods: Categorical data

Frequency data (counts) are more naturally displayed in terms of **count** ∼ **area** (Friendly, 1995)

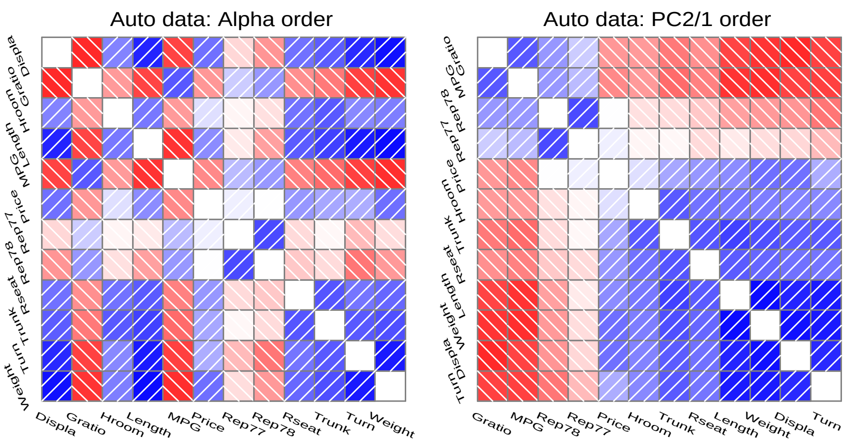


Fourfold display for 2×2 table



Mosaic plot for 3-way table

- **Principles of Graphical Displays**
  - **Effect ordering** (Friendly and Kwan, 2003)— In tables and graphs, sort unordered factors according to the effects you want to see/show.



"Corrgrams: Exploratory displays for correlation matrices" (Friendly, 2002)

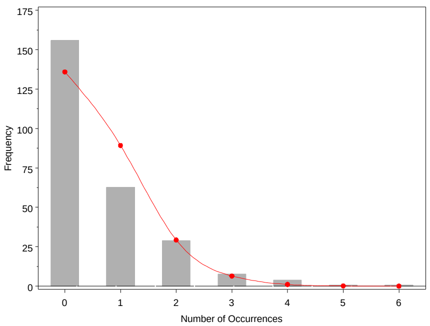
- Effect ordering and high-lighting for tables (Friendly, 2000)

Table: Hair color - Eye color data: Effect ordered

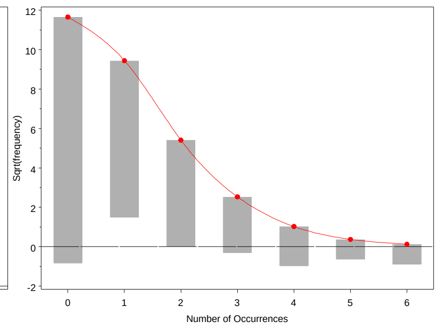
Eye color	Hair color: Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Model:	*Independence*: [Hair][Eye] χ^2 (9)= 138.29						
Color coding:	<-4	<-2	<-1	0	>1	>2	>4
n in each cell:	$n <$ expected				$n >$ expected		

- **Comparisons**— Make visual comparisons easy
  - Visual grouping— connect with lines, make key comparisons contiguous
  - Baselines— compare *data* to *model* against a line, preferably horizontal

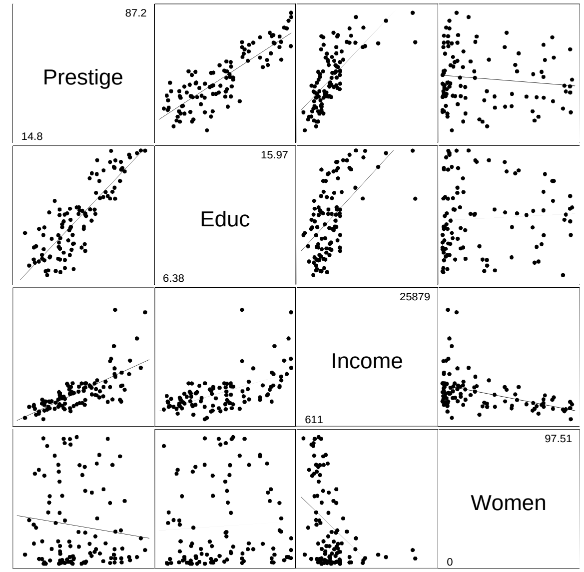


Standard histogram with fit

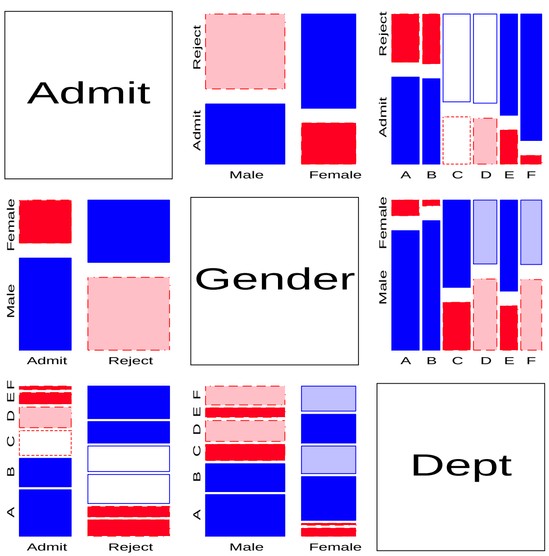


Suspended rootogram

- **Small multiples**— combine stratified graphs into coherent displays (Tufte, 1983)
  - e.g., scatterplot matrix for quantitative data: all pairwise scatterplots



- e.g., mosaic matrix for quantitative data: all pairwise mosaic plots



Graphical methods: Categorical data

Exploratory methods

- Minimal assumptions (like non-parametric methods)
- Show the *data*, not just *summaries*
- Help detect *patterns, trends, anomalies*, suggest hypotheses

Plots for model-based methods

- Residual plots - departures from model, omitted terms, ...
- Effect plots - estimated probabilities of response or log odds
- Diagnostic plots - influence, violation of assumptions

Goals

- *VCD* and R vcd package - Make these methods *available* and *accessible* in SAS & R
- **Practical power = Statistical power** $\times$ **Probability of Use**
- Today's goal: take-home knowledge
- Tomorrow's goal: dynamic, interactive graphics for categorical data

VCD Macros & SAS/IML programs

- Macros, datasets available at datavis.ca/vcd/

Discrete distributions

DISTPLOT	Plots for discrete distributions
GOODFIT	Goodness-of-fit for discrete distributions
ORDPLOT	Ord plot for discrete distributions
POISPLOT	Poissonness plot
ROOTGRAM	Hanging rootograms

Two-way and n-way tables

AGREEPLOT	Observer agreement chart
CORRESP	Plot PROC CORRESP results
FFOLD	Fourfold displays for $2 \times 2 \times k$ tables
SIEVEPLOT	Sieve diagrams
MOSAIC	Mosaic displays
MOSMAT	Mosaic matrices
TABLE	Construct a grouped frequency table, with recoding
TRIPLOT	Trilinear plots for $n \times 3$ tables

Model-based methods

ADDVAR	Added variable plots for logistic regression
CATPLOT	Plot results from PROC CATMOD
HALFNORM	Half-normal plots for generalized linear models
INFLGLIM	Influence plots for generalized linear models
INFLOGIS	Influence plots for logistic regression
LOGODDS	Plot empirical logits and probabilities for binary data
POWERLOG	Power calculations for logistic regression

Utility macros

DUMMY	Create dummy variables
LAGS	Calculate lagged frequencies for sequential analysis
PANELS	Arrange multiple plots in a panelled display
SORT	Sort a dataset by the value of a statistic or formatted value
Utility	Graphics utility macros: BARS, EQUATE, GDISPLA, GENSYM, GSKIP, LABEL, POINTS, PSCALE

VCD Archive (vcdprog.zip) available at:
http://datavis.ca/courses/VCD/vcdprog.zip

R software and the vcd package

- R software and the vcd package, available at www.r-project.org

Discrete distributions

distplot	Plots for discrete distributions
goodfit	Goodness-of-fit for discrete distributions
ordplot	Ord plot for discrete distributions
poisplot	Poissonness plot
rootgram	Hanging rootograms

Two-way and n-way tables

agreementplot	Observer agreement chart
fourfold	Fourfold displays for $2 \times 2 \times k$ tables
sieve	Sieve diagrams
mosaic	Mosaic displays
pairs.table	Matrix of pairwise association displays
structable	Manipulate high-dimensional contingency tables
triplot	Trilinear plots for $n \times 3$ tables

R software: Other packages

Model-based methods

glm	Fitting generalized linear models
gnm	Fitting generalized *non-linear* models, e.g., RC(1) model
loglm	MASS package: Fitting loglinear models
Rcmdr	Menu-driven package for statistical analysis and graphics
car	Graphics and extensions of generalized linear models
effects	Effects plots for generalized linear models

vcdExtra package

vcd-tutorial	Vignette on working with categorical data and the vcd package
mosaic.glm	mosaic displays for GLMs and GNMs
mosaic3d	3D mosaic displays
glmlist	Methods for working with lists of models

Discrete distributions

- **Counts of occurrences:** accidents, words in text, blood cells with some characteristic.
- **Data:** Basic outcome value, k , $k = 0, 1, \ldots$, and number of observations, n_k, with that value.
- **Example:** distributions of key "marker" words: *from*, *may*, *whilst*, ... in *Federalist Papers* by James Madison, e.g., blocks of 200 words with *may*:

Occurrences (k)	0	1	2	3	4	5	6
Blocks (n_k)	156	63	29	8	4	1	1

- **Example:** Saxony families with 12 children having $k = 0, 1, \ldots 12$ sons.

k	0	1	2	3	4	5	6	7	8	9	10	11	12
n_k	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Discrete distributions

Questions:

- What process gave rise to the distribution?
- Form of distribution: uniform, binomial, Poisson, negative binomial, geometric, etc.?
- Estimate parameters
- Visualize goodness of fit

For example:

- *Federalist Papers:* might expect a Poisson(λ) distribution.
- *Families in Saxony:* might expect a Bin(n, p) distribution with $n = 12$. Perhaps $p = 0.5$ as well.

Discrete distributions

Lack of fit:

- Lack of fit tells us something about the process giving rise to the data
- Poisson: assumes constant small probability of the basic event
- Binomial: assumes constant probability and independent trials

Motivation:

- Models for more complex categorical data often use these basic discrete distributions
- Binomial (with predictors) $\rightarrow$ logistic regression
- Poisson (with predictors) $\rightarrow$ poisson regression, loglinear models
- $\Rightarrow$ many of these are special cases of *generalized linear models*

Fitting and graphing discrete distributions

VCD

methods to fit, visualize, and diagnose discrete distributions:

- **Fitting:** `GOODFIT` macro fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
- **Hanging rootograms:** Sensitively assess departure between Observed, Fitted counts (`ROOTGRAM` macro)
- **Ord plots:** Diagnose form of a discrete distribution (`ORDPLOT` macro)
- **Poissonness plots:** Robust fitting and diagnostic plots for Poisson (`POISPLOT` macro)
- **Robust distribution plots** (`DISTPLOT` macro)

Sidebar: Using SAS macros

- SAS macros are high-level, general programs consisting of a series of `DATA` steps and `PROC` steps.
- Keyword arguments substitute your data names, variable names, and options for the named macro parameters.
- Use as:

  `%macname(data=dataset, var=variables, ...);`
- Most arguments have default values (e.g., `data=_last_`)
- All *VCD* macros have internal and online documentation, http://datavis.ca/sasmac/
- Macros can be installed in directories automatically searched by SAS. Put the following `options` statement in your `AUTOEXEC.SAS` file:

  `options sasautos=('c:\sasuser\macros' sasautos);`

Sidebar: Using SAS macros

E.g., the `GOODFIT` macro is defined with the following arguments:

··· goodfit.sas ···

```
%macro goodfit(
  data=_last_,    /* name of the input data set             */
  var=,           /* analysis variable (basic count)        */
  freq=,          /* frequency variable                     */
  dist=,          /* name of distribution to be fit         */
  parm=,          /* required distribution parameters?      */
  sumat=100000,   /* sum probs. and fitted values here      */
  format=,        /* format for ungrouped analysis variable */
  out=fit,        /* output fit data set                    */
  outstat=stats); /* output statistics data set             */
```

Typical use:

```
%goodfit(data=madison, /* data set       */
    var=count,         /* count variable */
    freq=blocks,
    dist=poisson);
```


Fitting discrete distributions

- **Distributions:**
  - Poisson, $p(k) = e^{-\lambda}\lambda^k / k!$
  - Binomial, $p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
  - Negative binomial, $p(k) = \binom{n+k-1}{k} p^n (1-p)^k$
  - Geometric, $p(k) = p(1-p)^k$
  - Logarithmic series, $p(k) = \theta^k / [-k \log(1-\theta)]$
- **Estimate parameter(s):**
  - Poisson, $\hat{\lambda} = \sum k n_k / \sum n_k = \text{mean}$
  - Binomial, $\hat{p} = \sum k n_k / (n \sum n_k) = \text{mean / n}$
- **Goodness of fit:**

$$\chi^2 = \sum_{k=1}^{K} \frac{(n_k - N\hat{p}_k)^2}{N\hat{p}_k} \sim \chi^2_{(K-1)}$$

where $\hat{p}_k$ is the estimated probability of each basic count, under the hypothesis that the data follows the chosen distribution.

GOODFIT macro: Fitting discrete distributions

- `GOODFIT` macro fits uniform, binomial, Poisson, negative binomial, geometric, logarithmic series distributions (or any specified multinomial)
- E.g., Try fitting Poisson model

madfit.sas

```
title "Instances of 'may' in Federalist papers";
data madison;
   input count blocks;
   label count='Number of Occurrences'
         blocks='Blocks of Text';
datalines;
  0    156
  1     63
  2     29
  3      8
  4      4
  5      1
  6      1
;
 %goodfit(data=madison, var=count, freq=blocks,
      dist=poisson);
```

Fitting discrete distributions

The GOODFIT macro gives a table of observed and fitted frequencies, Pearson χ^2 residuals (CHI) and likelihood-ratio deviance residuals (DEV).

```
           Instances of 'may' in Federalist papers

COUNT    BLOCKS      PHAT        EXP        CHI         DEV

  0        156     0.51867    135.891     1.72499     6.56171
  1         63     0.34050     89.211    -2.77509    -6.62056
  2         29     0.11177     29.283    -0.05231    -0.75056
  3          8     0.02446      6.408     0.62890     1.88423
  4          4     0.00401      1.052     2.87493     3.26912
  5          1     0.00053      0.138     2.31948     1.98992
  6          1     0.00006      0.015     8.01267     2.89568
         ======    =======    =======
           262     0.99999    261.998
```


Fitting discrete distributions

In addition, it provides the overall goodness-of-fit tests:

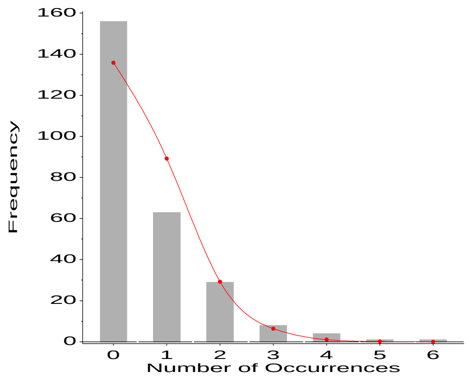
```
Goodness-of-fit test for data set MADISON

Analysis variable:        COUNT Number of Occurrences
Distribution:             POISSON
Estimated Parameters:     lambda = 0.6565

Pearson chi-square     = 88.92304707
Prob > chi-square      = 0

Likelihood ratio G2    = 25.243121314
Prob > chi-square      = 0.0001250511

Degrees of freedom     = 5
```

The poisson model does not fit! Why?

What's wrong with histograms?

- Discrete distributions often graphed as histograms, with a theoretical fitted distribution superimposed.

```
%goodfit(data=madison, var=count, freq=blocks,
   dist=poisson);
```



Problems:

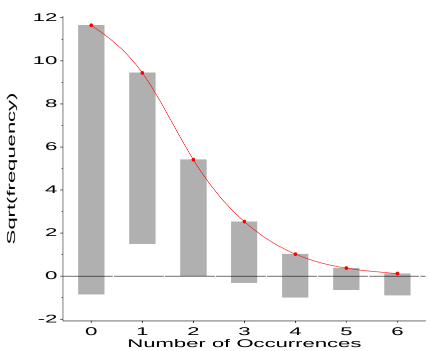
- largest frequencies dominate display
- must assess deviations vs. a curve

Hang & root them → Hanging rootograms

Tukey (1972, 1977):

- shift histogram bars to the fitted curve → judge deviations vs. horizontal line.
- plot $\sqrt{\text{freq}}$ → smaller frequencies are emphasized.

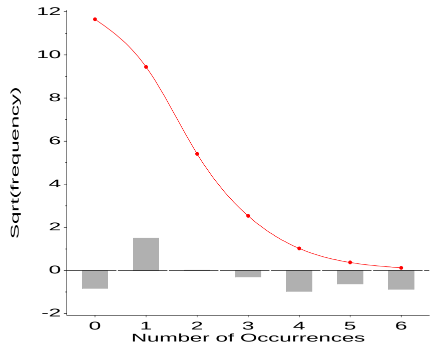
```
%goodfit(data=madison, var=count, freq=blocks,
  dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks);
```



Highlight differences → Deviation rootograms

- Emphasize differences between observed and fitted frequencies
- Draw bars to show the gaps (btype=dev)

```
%goodfit(data=madison, var=count, freq=blocks,
  dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks, btype=dev);
```



Ord plots: Diagnose form of discrete distribution

- How to tell which discrete distributions are likely candidates?
- Ord (1967): for each of Poisson, Binomial, Negative Binomial, and Logarithmic Series distributions,
  - plot of $k p_k / p_{k-1}$ against k is linear
  - signs of intercept and slope → determine the form, give rough estimates of parameters

Slope (b)	Intercept (a)	Distribution (parameter)	Parameter estimate
0	+	Poisson (λ)	$\lambda = a$
−	+	Binomial (n, p)	$p = b/(b-1)$
+	+	Neg. binomial (n,p)	$p = 1 - b$
+	−	Log. series (θ)	$\theta = b$
			$\theta = -a$

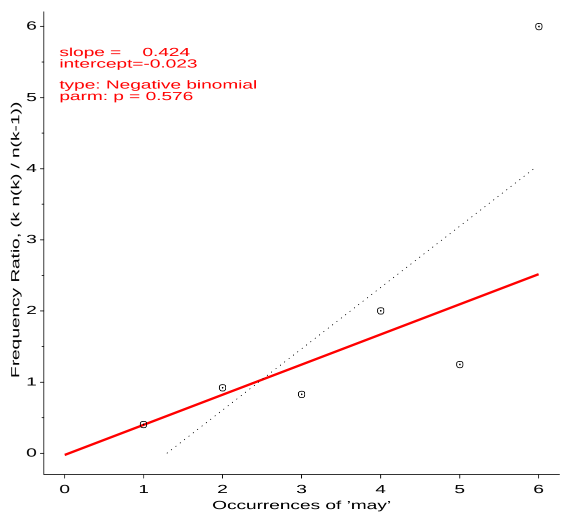
- Fit line by WLS, using $\sqrt{n_k}$ as weights

Ord plots

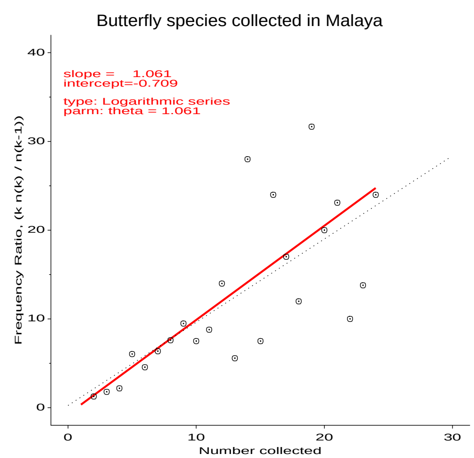
- ORDPLOT macro

```
%ordplot(data=madison, count=Count, freq=blocks);
```

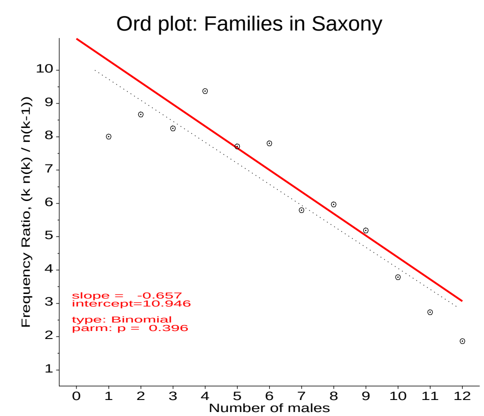
- Diagnoses distribution as NegBin
- Estimates $\hat{p} = 0.576$



Ord plots: Other distributions



Logarithmic series



Binomial

Robust distribution plots: Poisson

- Ord plots lack robustness
  - one discrepant freqency, n_k affects points for both k and $k+1$
- Robust plots for Poisson distribution (Hoaglin and Tukey, 1985)
  - For Poisson, plot **count metameter** $= \phi(n_k) = \log_e(k!\, n_k / N)$ vs. k
  - Linear relation $\Rightarrow$ Poisson, slope gives $\hat{\lambda}$
  - CI for points, diagnostic (influence) plot
  - POISPLOT macro

Poissonness plots: Details

- If the distribution of n_k is Poisson(λ) for some fixed λ, then each observed frequency, $n_k \approx m_k = N p_k$.
- Then, setting $n_k = N p_k = e^{-\lambda}\, \lambda^k / k!$, and taking logs of both sides gives

$$\log(n_k) = \log N - \lambda + k \log \lambda - \log k!$$

which can be rearranged to

$$\phi(n_k) \equiv \log\left(\frac{k!\, n_k}{N}\right) = -\lambda + (\log \lambda)\, k$$

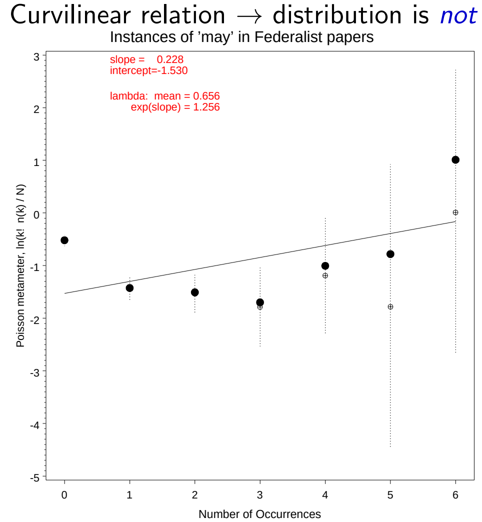
- $\Rightarrow$ if the distribution is Poisson, plotting $\phi(n_k)$ vs. k should give a line with
  - intercept $= -\lambda$
  - slope $= \log \lambda$
- Nonlinear relation $\rightarrow$ distribution is *not* Poisson
- Hoaglin and Tukey (1985) give details on calculation of confidence intervals and influence measures.

POISPLOT macro: example

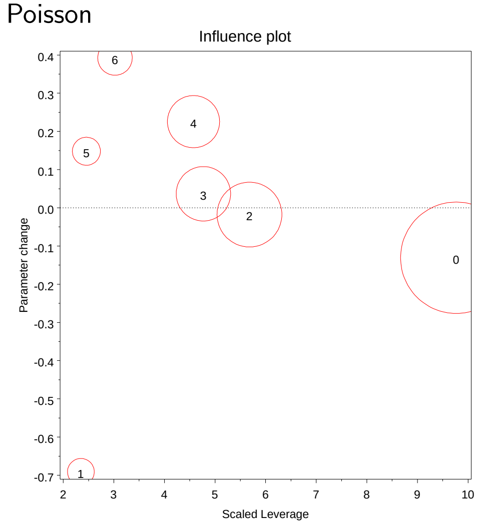
```
title "Instances of 'may' in Federalist papers";
data madison;
   input count blocks;
   label count='Number of Occurrences'
         blocks='Blocks of Text';
datalines;
  0    156
  1     63
  2     29
  3      8
  4      4
  5      1
  6      1
;
%poisplot(data=madison,count=count, freq=blocks);
```


POISPLOT macro: output

Curvilinear relation $\rightarrow$ distribution is *not* Poisson



Poissonness plot



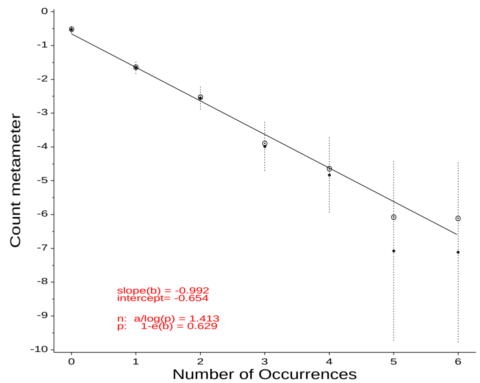
Influence plot for change in λ

Generalized robust distribution plots

Other distributions: Analogous plots, for suitable count metameter, $\phi(n_k)$ vs. k.

- Linear relation ⇒ correct distribution, slope gives parameter estimates
- CI reflect variability of the individual counts, n_k
- DISTPLOT macro

```
%distplot(data=madison, count=count, freq=blocks,
      dist=negbin);
```



Discrete distributions with R and the vcd package

In R, discrete distributions are conveniently represented as one-way frequency tables,

```
> library(vcd)
> data(Federalist)
> Federalist
```

```
nMay
  0   1   2   3   4   5   6
156  63  29   8   4   1   1
```

The goodfit() function in vcd fits a variety of discrete distributions:

```
> # fit the poisson model
> gf1 <- goodfit(Federalist, type="poisson")
> gf1
```

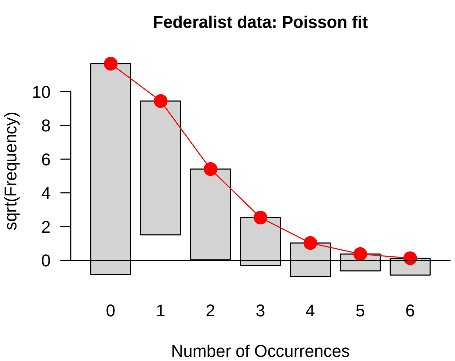
```
Observed and fitted values for poisson distribution
with parameters estimated by `ML'

 count observed       fitted
     0      156 135.89138870
     1       63  89.21114067
     2       29  29.28304617
     3        8   6.40799484
     4        4   1.05169381
     5        1   0.13808499
     6        1   0.01510854
```

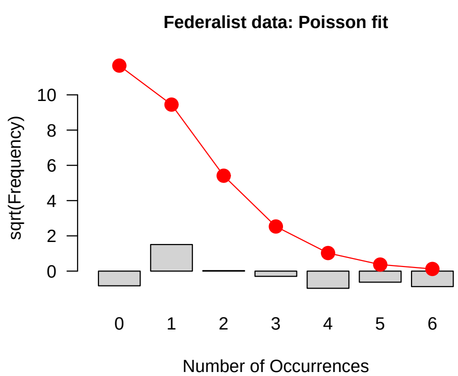

R is object-oriented. A goodfit object has print(), summary() and plot() methods:

```
> summary(gf1)
```

```
        Goodness-of-fit test for poisson distribution

                      X^2 df     P(> X^2)
Likelihood Ratio 25.24312  5 0.0001250511
```

```
> plot(gf1, main="Federalist data: Poisson fit")
> plot(gf1, main="Federalist data: Poisson fit", type="dev")
```





Discrete distributions with R and the vcd package

The Poisson distribution

```
> # In a poisson, mean = var; this is 'over-dispersed'
>  mean(rep(0:6, times=Federalist))
```

```
[1] 0.6564885
```

```
>  var(rep(0:6, times=Federalist))
```

```
[1] 1.007985
```

The negative binomial distribution, Nbin(r, p) allows the data to deviate from a true Poisson according to a parameter $r > 0$.

```
> ## try negative binomial distribution (r, p)
> gf2 <- goodfit(Federalist, type = "nbinomial")
> summary(gf2)
```

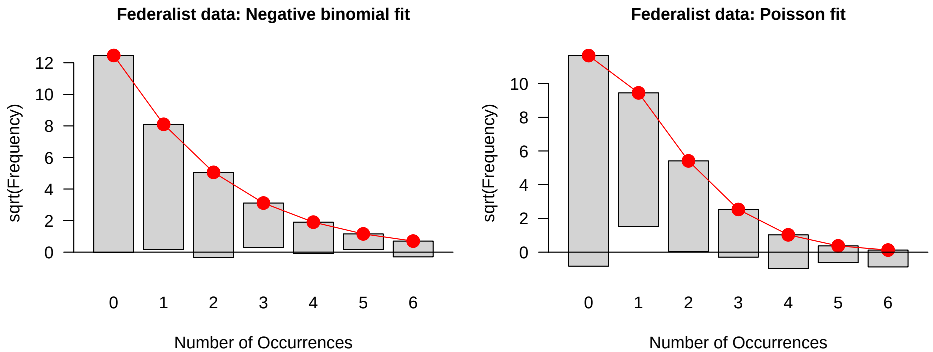
```
        Goodness-of-fit test for nbinomial distribution

                      X^2 df  P(> X^2)
Likelihood Ratio 1.964028  4 0.7423751
```

This has an acceptable fit to the Federalist data

Discrete distributions with R and the vcd package

Compare the fits side-by-side:

```
> plot(gf2, main="Federalist data: Negative binomial fit")
> plot(gf1, main="Federalist data: Poisson fit")
```

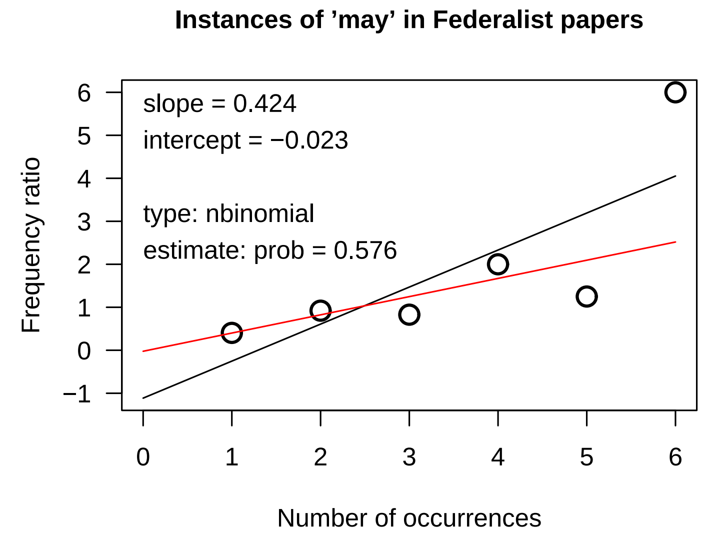


Conclusions:

- Perhaps marker words like 'may' do not occur with constant probability in all blocks of text
- Perhaps the blocks of text were written under different circumstances

vcd includes Ord_plot() and distplot() functions. E.g.,

```
> Ord_plot(Federalist,
      main = "Instances of 'may' in Federalist papers")
```



Testing Association in Two-Way Tables

Typical analysis: Nominal factors

- Pearson χ^2 (or LR χ^2)— when most expected frequencies ≥ 5.

```
proc freq;
   weight count;        /* if in frequency form */
   table factor * response / chisq;
```

- Exact tests— small tables, small sample sizes (e.g., Fisher's)

```
proc freq;
   weight count;        /* if in frequency form */
   table factor * response / chisq;
   exact pchi;
```


Example: Cholesterol diet and heart disease

Is there a relation between Hi/Lo cholesterol diet and heart disease?

fat.sas

```
title 'Cholesterol diet and heart disease';
data fat;
   input diet $ disease $ count;
datalines;
LoChol  No   6
LoChol  Yes  2
HiChol  No   4
HiChol  Yes 11
;

proc freq data=fat;
  weight count;
  tables diet * disease / chisq nopercent nocol;
  exact pchi;
```

Standard output:

```
                    Table of diet by disease

              diet      disease

              Frequency|
              Row Pct  |No      |Yes     |  Total
              ---------+--------+--------+
              HiChol   |      4 |     11 |     15
                       |  26.67 |  73.33 |
              ---------+--------+--------+
              LoChol   |      6 |      2 |      8
                       |  75.00 |  25.00 |
              ---------+--------+--------+
              Total          10       13       23


            Statistics for Table of diet by disease

     Statistic                     DF       Value      Prob
     ------------------------------------------------------
     Chi-Square                     1      4.9597    0.0259
     Likelihood Ratio Chi-Square    1      5.0975    0.0240
     Continuity Adj. Chi-Square     1      3.1879    0.0742

  WARNING: 50% of the cells have expected counts less than 5.
           (Asymptotic) Chi-Square may not be a valid test.
```

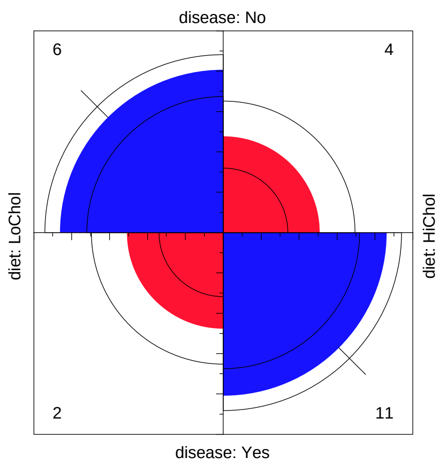
- The Pearson and LR χ^2 tests are not valid
- The conservative continuity-adjusted test fails significance

- Exact tests are valid and significant.

Exact test output:

```
         Pearson Chi-Square Test
    ----------------------------------
    Chi-Square                  4.9597
    DF                               1
    Asymptotic Pr >  ChiSq      0.0259
    Exact      Pr >= ChiSq      0.0393


           Fisher's Exact Test
    ----------------------------------
    Cell (1,1) Frequency (F)         4
    Left-sided Pr <= F          0.0367
    Right-sided Pr >= F         0.9967

    Table Probability (P)       0.0334
    Two-sided Pr <= P           0.0393
```


Preview: Visualizing association in 2 $\times$ 2 tables



- Fourfold display: area $\sim$ frequency
- Color: blue (+), red(−)
- Confidence bands: significance of odds ratio
- Interp: Hi cholesterol $\rightarrow$ Heart disease

```
%ffold(data=fat, var=diet disease);
```


Ordinal factors and Stratified analyses

More powerful CMH tests

- When either the row (factor) or column (response) levels are ordered, more specific (CMH = Cochran - Mantel - Haentzel) tests which take order into account have greater power to detect ordered relations.

```
proc freq;
   weight count;
   table factor * response / chisq cmh;
```

Control for other background variables

- Stratified analysis tests the association between a main factor and response *within* levels of the control variable(s)
- Can also test for homogeneous association across strata

```
proc freq;
   weight count;
   table strata * factor * response / chisq cmh;
```

Example: Arthritis treatment

Data on treatment for rheumatoid arthritis (Koch and Edwards, 1988)

- **Ordinal response**: none, some, or marked improvement
- **Factor**: active treatment vs. placebo
- **Strata**: Sex

		Outcome			
Treatment	Sex	None	Some	Marked	Total
Active	Female	6	5	16	27
	Male	7	2	5	14
Placebo	Female	19	7	6	32
	Male	10	0	1	11
Total		42	14	28	84

Overall analysis, ignoring sex

:

arthfreq.sas ···

```
title 'Arthritis Treatment: PROC FREQ Analysis';
data arth;
   input sex$ treat$ @;
   do improve = 'None  ', 'Some', 'Marked';
      input count @;
      output;
      end;
datalines;
Female  Active    6  5  16
Female  Placebo  19  7   6
Male    Active    7  2   5
Male    Placebo  10  0   1
;
*-- Ignoring sex;
proc freq order=data;
   weight count;
   tables treat * improve / cmh chisq nocol nopercent;
   run;
```

Notes:

- PROC FREQ orders character variables alphabetically (i.e., 'Marked', 'None', 'Some') by default.
- To treat the IMPROVE variable as ordinal, use `order=data` on the PROC FREQ statement.

Overall analysis, ignoring sex: Results (`chisq` option)

STATISTICS FOR TABLE OF TREAT BY IMPROVE

Statistic	DF	Value	Prob
Chi-Square	2	13.055	0.001
Likelihood Ratio Chi-Square	2	13.530	0.001
Mantel-Haenszel Chi-Square	1	12.859	0.000
Phi Coefficient		0.394	
Contingency Coefficient		0.367	
Cramer's V		0.394	

Cochran-Mantel-Haenszel tests: (`cmh` option)

SUMMARY STATISTICS FOR TREAT BY IMPROVE
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	12.859	0.000
2	Row Mean Scores Differ	1	12.859	0.000
3	General Association	2	12.900	0.002

CMH tests for ordinal variables

Three types of test:

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 Mean Scores Differ

- Use when only *column* variable is ordinal
- Analogous to the Kruskal-Wallis non-parametric test (ANOVA on rank scores)
- Ordinal variable must be listed last in the TABLES statement

General Association

- Use when *both* row and column variables are nominal.
- Similar to overall Pearson χ^2 and Likelihood Ratio χ^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	0.000

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	

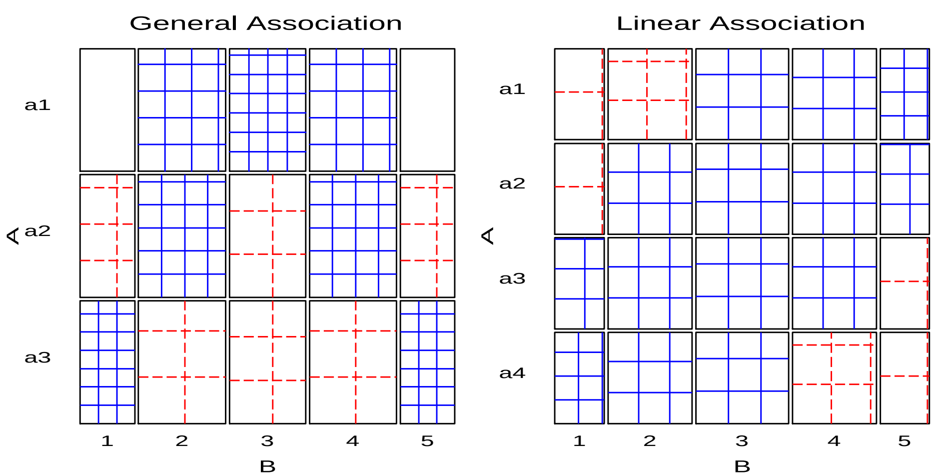
Output:

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	10.639	0.001
2	Row Mean Scores Differ	3	10.676	0.014
3	General Association	12	13.400	0.341

Sample CMH Profiles

Visualizing Association: Sieve diagrams



Stratified analysis

Overall analysis

- ignores other variables (like sex), by collapsing over them
- risks losing important interactions (e.g., different associations for M & F)

Stratified analysis

- controls for the effects of one or more background variables
- list stratification variable(s) *first* on the TABLES statement

```
proc freq;
   tables age * sex * treat * improve;
```

Looking forward: Loglinear models

- allow more general hypotheses to be stated and tested
- closer connection between testing and visualization (how are variables associated)

Stratified analysis

The statements below request a stratified analysis with CMH tests, controlling for sex.

··· arthfreq.sas ···

```
*-- Stratified analysis, controlling for sex;
proc freq order=data;
   weight count;
   tables sex * treat * improve / cmh chisq nocol nopercent;
   run;
```

→ separate tables (partial tests) for Females and Males

```
       STATISTICS FOR TABLE 1 OF TREAT BY IMPROVE
               CONTROLLING FOR SEX=Female

Statistic                     DF      Value        Prob
------------------------------------------------------
Chi-Square                     2     11.296       0.004
Likelihood Ratio Chi-Square    2     11.731       0.003
Mantel-Haenszel Chi-Square     1     10.935       0.001
 ...
```

- Strong association between TREAT and IMPROVE for females

Males:

```
       STATISTICS FOR TABLE 2 OF TREAT BY IMPROVE
                CONTROLLING FOR SEX=Male

Statistic                     DF      Value        Prob
------------------------------------------------------
Chi-Square                     2      4.907       0.086
Likelihood Ratio Chi-Square    2      5.855       0.054
Mantel-Haenszel Chi-Square     1      3.713       0.054
 ...

WARNING:  67% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.
```

- Weak association between TREAT and IMPROVE for males
- Sample size $N = 29$ for males is small

Stratified tests

- Individual (*partial*) tests are followed by a *conditional* test, controlling for strata (SEX)
- These tests **do not** require large sample size in the individual strata— just a large total sample size.
- They *assume*, but do not *test* that the association is the same for all strata.

```
           SUMMARY STATISTICS FOR TREAT BY IMPROVE
                     CONTROLLING FOR SEX

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

Statistic    Alternative Hypothesis    DF       Value      Prob
--------------------------------------------------------------
   1        Nonzero Correlation        1      14.632     0.000
   2        Row Mean Scores Differ     1      14.632     0.000
   3        General Association        2      14.632     0.001
```


Homogeneity of association

- Is the association between the primary table variables the same over all strata?
- 2×2 tables: → Equal odds ratios across all strata?
  - PROC FREQ: MEASURES option on TABLES statement → Breslow-Day test

```
proc freq;
   tables strata * factor * response / measures cmh ;
```

- Larger tables: Use PROC CATMOD to test for *no three-way association*
  - ≡ same association for the primary factor & response variables ∀ strata
  - ≡loglinear model: [Strata Factor] [Strata Response] [Factor Response]

```
proc catmod;
   ...
   loglin strata | factor | response @2;
```

Homogeneity of association: Example

- Arthritis data: homogeneity ↔ no 3-way sex * treatment * outcome association
  - ≡ loglinear model: [SexTreat] [SexOutcome] [TreatOutcome]
  - ≡ `loglin sex|treat|improve@2` for `PROC CATMOD`
  - Zero frequencies: `PROC CATMOD` treats as “structural zeros” by default; recode if necessary.

··· arthfreq.sas

```
title2 'Test homogeneity of treat*improve association';
data arth;
   set arth;
   if count=0 then count=1E-20;    *-- sampling zeros;
proc catmod order=data;
   weight count;
   model sex * treat * improve = _response_ / ml ;
   loglin sex|treat|improve @2 / title='No 3-way association';
run;
   loglin sex treat|improve   / title='No Sex Associations';
```


Homogeneity of association: Example

- the likelihood ratio χ^2 (the badness-of-fit for the No 3-Way model) is the test for homogeneity
- clearly non-significant → treatment-outcome association can be considered to be the same for men and women.

```
   Test homogeneity of treat*improve association
                No 3-way association
   MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

 Source                    DF   Chi-Square      Prob
 --------------------------------------------------
 SEX                        1        14.13    0.0002
 TREAT                      1         1.32    0.2512
 SEX*TREAT                  1         2.93    0.0871
 IMPROVE                    2        13.61    0.0011
 SEX*IMPROVE                2         6.51    0.0386
 TREAT*IMPROVE              2        13.36    0.0013

 LIKELIHOOD RATIO           2         1.70    0.4267
```

- But, associations of `SEX*TREAT` and `SEX*IMPROVE` are both small.
- Suggests stronger model of homogeneity, [Sex] [TreatOutcome], tested by `loglin sex treat|improve;` statement.

Homogeneity of association: Reduced model

··· arthfreq.sas

```
proc catmod order=data;
   weight count;
   model sex * treat * improve = _response_ / ml ;
   loglin sex|treat|improve@2 / title='No 3-way association';
run;
   loglin sex treat|improve   / title='No Sex Associations';
```

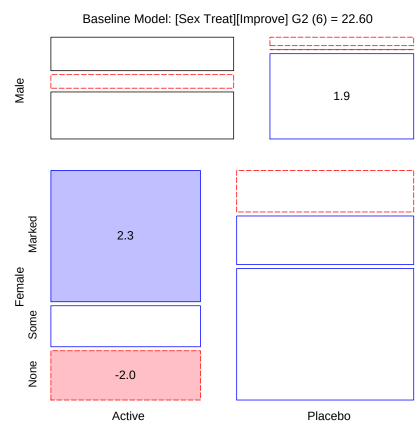
Output:

```
                  No Sex Associations
     MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

   Source                    DF   Chi-Square      Prob
   --------------------------------------------------
   SEX                        1        12.95    0.0003
   TREAT                      1         0.15    0.6991
   IMPROVE                    2        10.99    0.0041
   TREAT*IMPROVE              2        12.00    0.0025

   LIKELIHOOD RATIO           5         9.81    0.0809
```

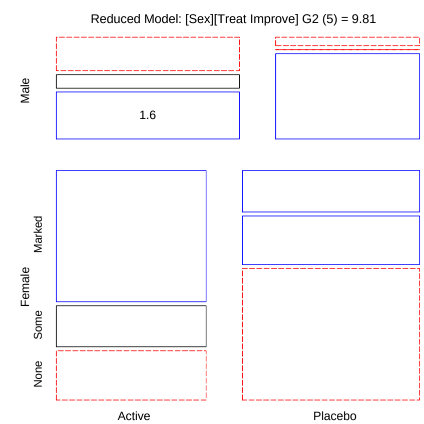
- Fits reasonably well
- How to interpret?

Homogeneity of association

Visualizing Association: Mosaic displays



Baseline model



Reduced model

Summary: Part 1

- **Categorical data**
  - Table form vs. case form
  - Non-parametric methods vs. model-based methods
  - Response models vs. association models
- **Graphical methods for categorical data**
  - Frequency data more naturally displayed as **count** $\sim$ area
  - Sieve diagram, fourfold & mosaic display: compare observed vs. expected frequency
  - Graphical principles: Visual comparison, effect-ordering, small multiples
- **Discrete distributions**
  - Fit: `GOODFIT`; Graph: hanging rootograms to show departures
  - Ord plot: diagnose form of distribution
  - `POISPLOT`, `DISTPLOT` for robust distribution plots
- **Testing association**
  - Pearson χ^2, L.R. χ^2 (largish samples) vs. Fisher exact test (small samples)
  - CMH tests more powerful for ordinal factors
  - Three-way+ tables: Stratified analysis, homogeneity of association
  - Visualize with Sieve diagram, fourfold & mosaic display