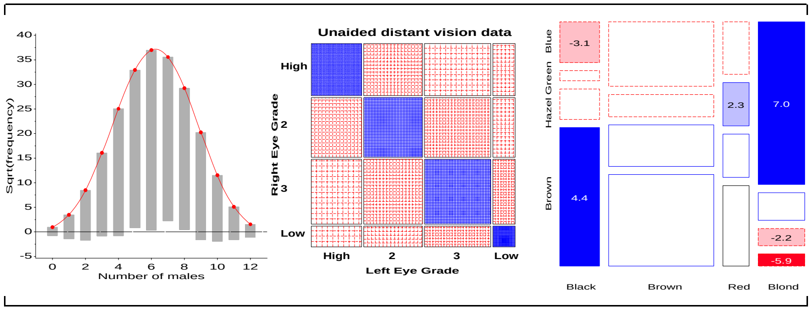
Categorical Data Analysis with Graphics



Michael Friendly

York University, <friendly@yorku.ca>

SCS Short Course

Feb-Mar., 2007

Outline

- Overview: Categorical Data *and* Graphics
- Methods for discrete distributions – testing goodness of fit
  - Hanging rootograms
  - Robust distribution plots
- Methods for two-way frequency tables – understanding association
  - Fourfold displays
  - Sieve diagrams
- Mosaic displays and loglinear models for *n*-way tables
  - Mosaic displays
  - Mosaic matrices
  - Correspondence analysis and MCA
- Logistic and logit regression
  - Logit plots, effect plots
  - Diagnostic plots

Color version of these slides:
`http://www.math.yorku.ca/SCS/Courses/grcat/`

Categorical Data Analysis

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

- **Non-parametric, randomization-based methods**
  - make minimal assumptions
  - useful for hypothesis-testing
  - SAS: `PROC FREQ`; SPSS: Crosstabs
    - Pearson Chi-square
    - Fisher's exact test (for small expected frequencies)
    - Mantel-Haenszel tests (ordered categories: test for *linear* association)
- **Model-based methods**
  - Must assume random sample (possibly stratified)
  - Useful for estimation purposes
  - Greater flexibility; fitting specialized models (e.g., symmetry)
  - More suitable for multi-way tables
  - SAS: `PROC LOGISTIC`, `CATMOD`, `GENMOD` , `INSIGHT` (Fit YX)
    - estimate standard errors, covariances for model parameters
    - confidence intervals for parameters, predicted Pr{response}
  - SPSS: Hiloglinear, Loglinear, Generalized linear models

Graphical Methods for Categorical Data

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

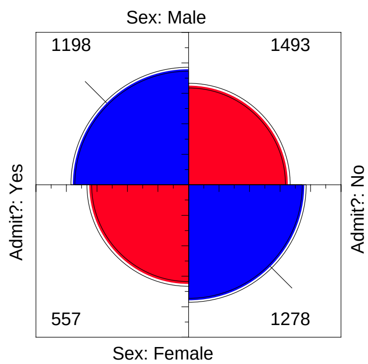
Getting information from a table is like extracting sunlight from a cucumber.

Farquhar & Farquhar, 1891

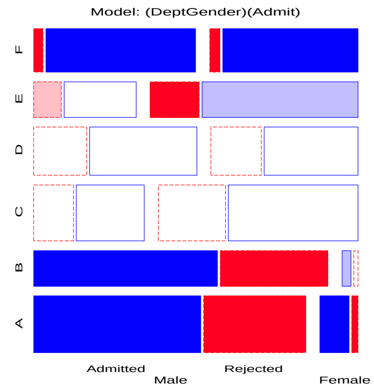
- **Tables vs. Graphs**
  - Tables are best suited for *look-up*— read off exact numbers
  - Graphs are better for showing *patterns, trends, anomalies*, making *comparisons*
  - Visual presentation as *communication*: what do you want to say?

Visual metaphors

- Quantitative data: **magnitude** $\sim$ **position along an axis**
- Frequency data: **count** $\sim$ **area** (Friendly, 1995)



Fourfold display for 2$\times$2 table

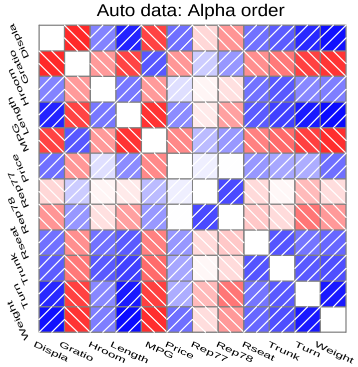


Mosaic plot for 3-way table

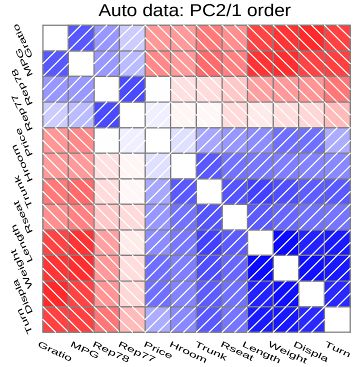
Principles of Graphical Displays

- **Effect ordering** (Friendly and Kwan, 2002)— In tables and graphs, sort unordered factors according to the effects you want to see/show.

Auto data: Alpha order



Auto data: PC2/1 order



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

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

Table 1: Hair color - Eye color data: Alpha ordered

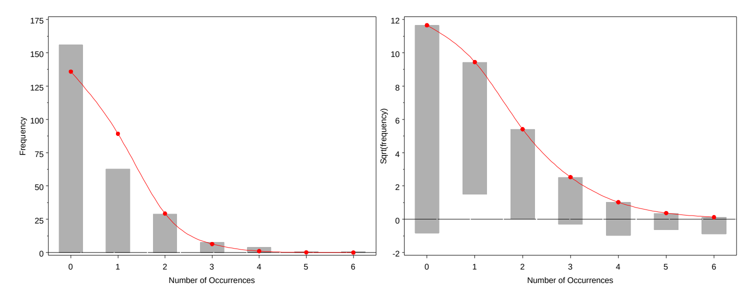
Eye color	Hair color: Blond	Black	Brown	Red
Blue	94	20	17	84
Brown	7	68	26	119
Green	10	15	14	54
Hazel	16	5	14	29

Table 2: Hair color - Eye color data: Effect ordered

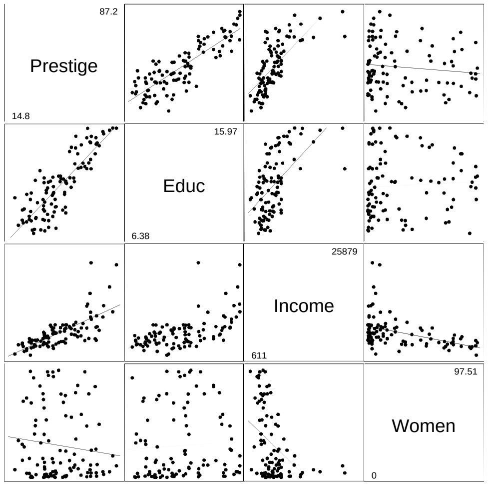
Eye color	Hair color: Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

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		

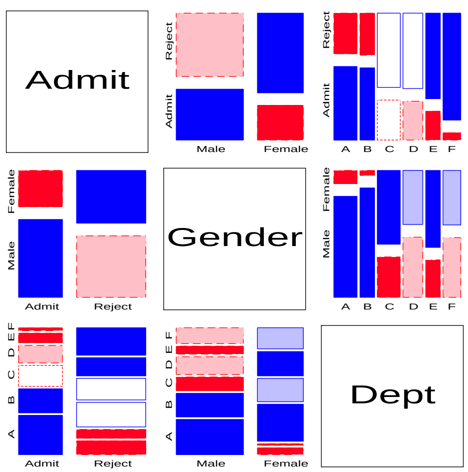
- **Comparisons**— Make visual comparisons easy
  - Visual grouping— connect with lines, make key comparisons contiguous
  - Baselines— compare *data* to *model* against a line, preferably horizontal



- **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



- **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 *SSSG* - Make these methods *available* and *accessible* in SAS
  - **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 `www.math.yorku.ca/SCS/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

`AGREE`	Observer agreement chart
`CORRESP`	Plot `PROC CORRESP` results
`FFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (macro)
`FOURFOLD`	Fourfold displays for $2 \times 2 \times k$ tables (SAS/IML)
`SIEVEPLOT`	Sieve diagrams
`MOSAIC`	Mosaic displays (macro)
`MOSAICS`	SAS/IML modules for mosaic displays
`MOSMAT`	Mosaic matrices (macro)
`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
POWERRxC	Power calculations for two-way frequency table
POWER2x2	Power calculations for a 2×2 table
ROBUST	Robust fitting for linear models

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 to purchasers at:
`support.sas.com/publishing/bbu/56571_sample.html`

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
sieveplot	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 and the `vcd` package

Model-based methods

glm	Fitting generalized linear models
loglm	MASS package: Fitting loglinear models
R Commander	Menu-driven package for statistical analysis and graphics
car	Package for graphics and extensions of generalized linear models
effects	Effects plots for generalized linear 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.

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://www.math/yorku.ca/SCS/vcd/`
- Use as:

  ```
  %macname(data=dataset, var=variables, ...);
  ```

- 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,
    var=count,
    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, $\hat{\lambda} = \sum k n_k / \sum n_k$ = mean
  - Binomial, $\hat{p} = \sum k n_k / (n \sum n_k)$ = 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:

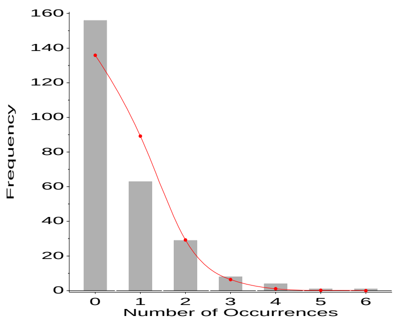
```
 Goodness-of-fit test for data set MADISON

 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:

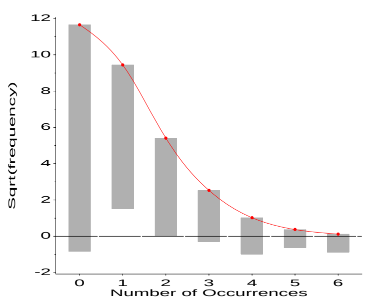
- largest frequencies dominate display
- must assess deviations vs. a curve

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.

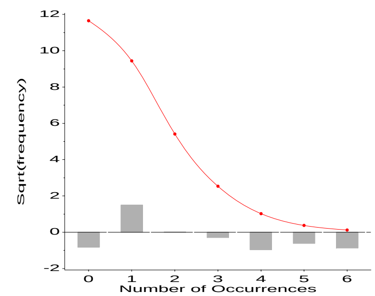
```
%goodfit(data=madison, var=count, freq=blocks,
   dist=poisson, out=fit);
%rootgram(data=fit, var=count, obs=blocks);
```



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)	$\theta = b$
			$\theta = -a$

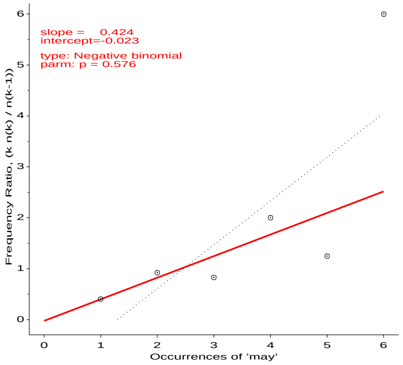
- Fit line by WLS, using $\sqrt{n_k}$ as weights

Ord plots

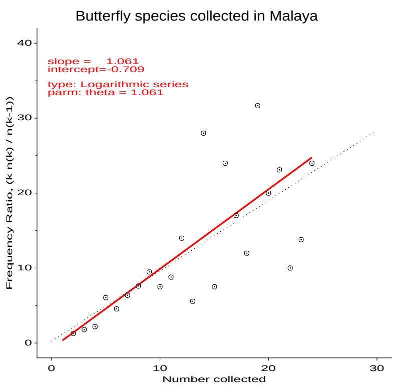
- ORDPLOT macro

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

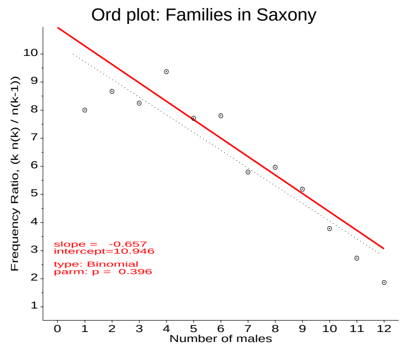
  - Diagnoses distribution as NegBin
  - Estimates $\hat{p} = 0.576$



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

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

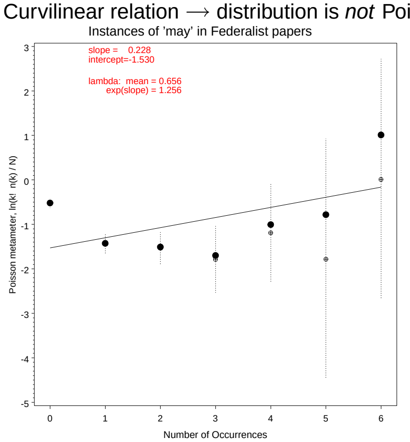
- $\Rightarrow$ if the distribution is Poisson, plotting $\phi(n_k)$ vs. k should give a line with
  - intercept = $-\lambda$
  - slope = $\log\, \lambda$
- Nonlinear relation $\longrightarrow$ distribution is *not* Poisson
- Hoaglin and Tukey (1985) give details on calculation of confidence intervals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

`POISPLOT` macro: example

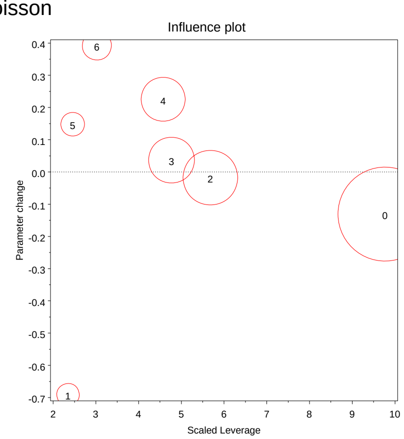
```
title "Instances of 'may' in Federalist papers";
data madison;
   input count blocks;
   label count='Number of Occurrences'
         blocks='Blocks of Text';
datalines;
  0    156
  1     63
  2     29
  3      8
  4      4
  5      1
  6      1
;
%poisplot(data=madison,count=count, freq=blocks);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


`POISPLOT` macro: output

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



Poissonness plot



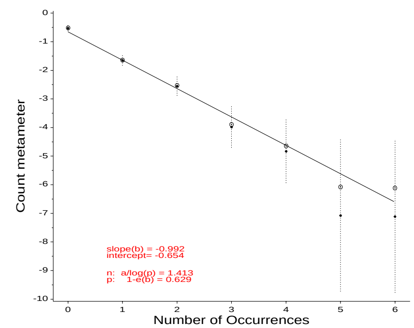
Influence plot for change in λ

Generalized robust distribution plots

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

- Linear relation $\Rightarrow$ 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);
```



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 already a frequency table */
   table factor * response / chisq;
```

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

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

Example: Cholesterol 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
```


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 anlysis 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

Treatment	Sex	None	Some	Marked	Total
		Outcome			
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.
- The `chisq` option gives the usual χ^2 tests (Pearson, Fisher's, etc.). The `cmh` option requests the Cochran-Mantel-Haenszel tests for ordinal variables.

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

- **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 should 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	

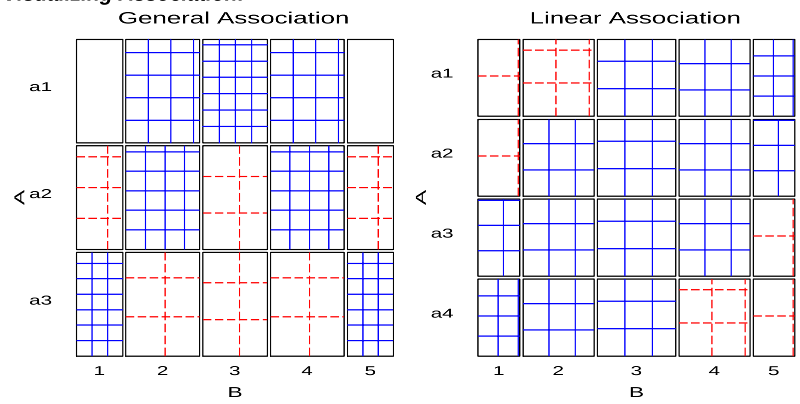
Output

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:



Stratified analysis

- **Overall analysis**
  - ignores other variables (like sex), by collapsing over them
  - risks losing important interactions (e.g., different associations for M and F)
- **Stratified analysis**:
  - controls for the effects of one or more background variables
  - list stratification variable(s) *first* on the `TABLES` statement

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

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 table (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 is the same over all strata?
- 2 × 2 tables: ⟶ Equal odds ratios across all strata?
  - `PROC FREQ`: `MEASURES` option on `TABLES` statement ⟶ Breslow-Day test
- Larger tables: Use `PROC CATMOD` to test for *no three-way association* = same association for the primary factor & response variables.
- 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

- the likelihood ratio χ^2 (the badness-of-fit for the No 3-Way model) is the test for homogeneity
- clearly non-significant $\longrightarrow$ 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: