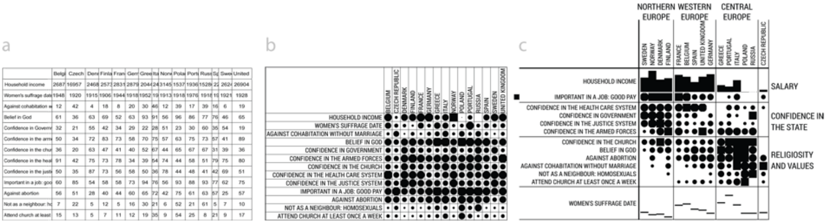
Categorical Data Analysis: Course Overview

Michael Friendly

Psych 6136

January 12, 2015



Course goals

This course is designed as a broad, applied introduction to the statistical analysis of categorical (or discrete) data, with an emphasis on:

Emphasis: visualization methods

- exploratory graphics: see patterns, trends, anomalies in your data
- model diagnostic methods: assess violations of assumptions
- model summary methods: provide an interpretable summary of your data

Emphasis: theory $\Rightarrow$ practice

- Understand how to translate research questions into statistical hypotheses and models
- Understand the difference between simple, non-parametric approaches (e.g., χ^2 test for indpendence) and model-based methods (logistic regression, GLM)
- Framework for thinking about categorical data analysis in *visual* terms

Course outline

1. Exploratory and hypothesis testing methods

- Week 1: Overview; Introduction to R
- Week 2: One-way tables and goodness-of-fit test
- Week 3: Two-way tables: independence and association
- Week 4: Two-way tables: ordinal data and dependent samples
- Week 5: Three-way tables: different types of independence
- Week 6: Correspondence analysis

2. Model-based methods

- Week 7: Logistic regression I
- Week 8: Logistic regression II
- Week 9: Multinomial logistic regression models
- Week 10: Log-linear models
- Week 11: Loglinear models: Advanced topics
- Week 12: Generalized Linear Models: Poisson regression
- Week 13: Course summary & additional topics

Textbooks

Main texts:

- Friendly, M. and Meyer, D. (2015). *Visualizing Categorical Data with R*. To be published by Chapman & Hall. Chapters will be made available on the web (password protected).
  http://euclid.psych.yorku.ca/www/psy6136/
- Agresti, Alan (2007). *An Introduction to Categorical Data Analysis*. 2nd ed. John Wiley & Sons, Inc.: New York. ISBN: 978-0-471-22618-5. Available in the bookstore.

Supplementary readings:
For those who desire a more in-depth treatment of categorical data analysis:

- Agresti, Alan (2013). *Categorical Data Analysis*. 3rd ed. New York: John Wiley & Sons, Inc. New York. ISBN: 978-0-470-46363-5

What is categorical data?

A ***categorical variable*** is one for which the possible measured or assigned values consist of a discrete set of categories, which may be *ordered* or *unordered*.

Some typical examples are:

- *Gender*, with categories "Male", "Female".
- *Marital status*, with categories "Never married", "Married", "Separated", "Divorced", "Widowed".
- *Party preference*, with categories "NDP", "Liberal", "Conservative", "Green".
- *Treatment outcome*, with categories "no improvement", "some improvement", or "marked improvement".
- *Age*, with categories "0-9", "10-19", "20-29", "30-39", $\dots$.
- *Number of children*, with categories $0, 1, 2, \dots$.

Categorical data structures: 1-way tables

Simplest case: 1-way frequency distribution

- Unordered factor

Hair	Black	Brown	Red	Blond
	108	286	71	127

Hair color among 592 students

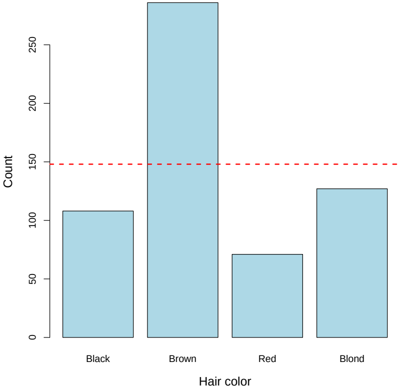
Party	BQ	Cons	Green	Liberal	NDP	Total
N	104	392	126	404	174	1200
%	8.7	32.6	10.5	33.7	14.5	100

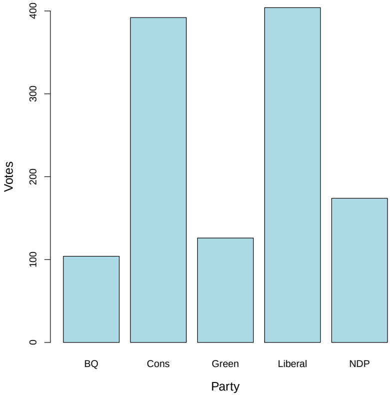
Voting intention in Harris-Decima poll, 8/21/08

- Questions:
  - Are all hair colors equally likely?
  - Do blondes have more fun?
  - Is there a difference in voting intentions between Liberal and Conservative?

Categorical data structures: 1-way tables

Even here, simple graphs are better than tables





But these don't really provide answers to the questions. Why?

Categorical data structures

Simplest case: 1-way frequency distribution

- 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

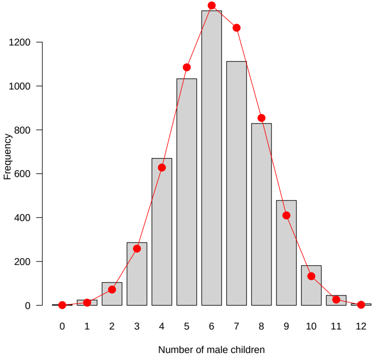
# of sons in Saxony families with 12 children

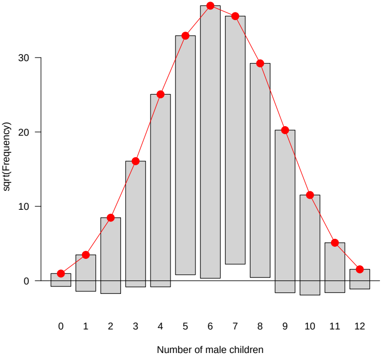
- Questions:
  - What is the *form* of this distribution?
  - Is it useful to think of this as a ***binomial distribution***?
  - If so, is $\Pr(male) = .5$ reasonable?
  - How could so many families have 12 children?

Categorical data structures: 1-way tables

When a particular distribution is in mind,

- better to plot the data together with the fitted frequencies
- better still: a ***hanging rootogram***– plot frequencies on sqrt scale, and hang the bars from the fitted values.





Categorical data structures: 2x2 tables

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, stratified by another factor

... 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

Questions:

- Is admission associated with gender?
- Does admission rate vary with department?

Categorical data structures: Larger tables

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 men more likely to be admitted than women?
  - Are hair color and eye color associated?
  - Does the binomial distribution fit these data?
- Mostly for two-way tables (possibly stratified)
- R:
  - Pearson Chi-square: `chisq.test()`
  - Fisher's exact test (for small expected frequencies): `fisher.test()`
  - Mantel-Haenszel tests (ordered categories: test for *linear* association): `CMHtest()`
- SAS: PROC `FREQ` — can do all the above
- 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
- R: `glm()` family, Packages: car, gnm, vcd, ...
  - estimate standard errors, covariances for model parameters
  - confidence intervals for parameters, predicted Pr{response}
- SAS: PROC `LOGISTIC`, `CATMOD`, `GENMOD` , `INSIGHT` (Fit YX), ...
- 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 among variables
- 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

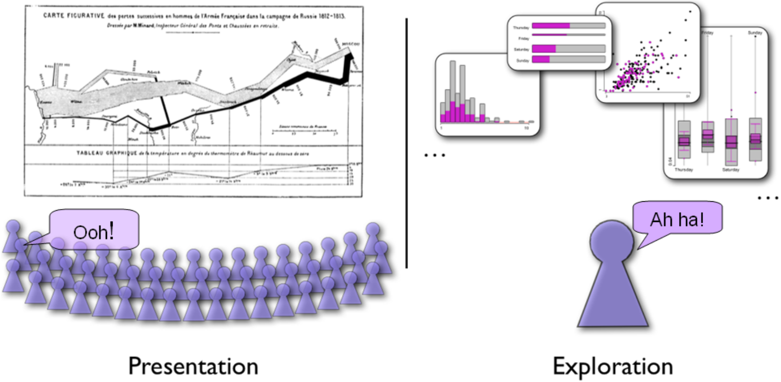
Tables vs. Graphs

- Tables are best suited for *look-up* and calculation—
  - read off exact numbers
  - show 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?
  - $\Longrightarrow$ design graphs and tables to 'speak to the eyes'

Graphical methods: Communication goals

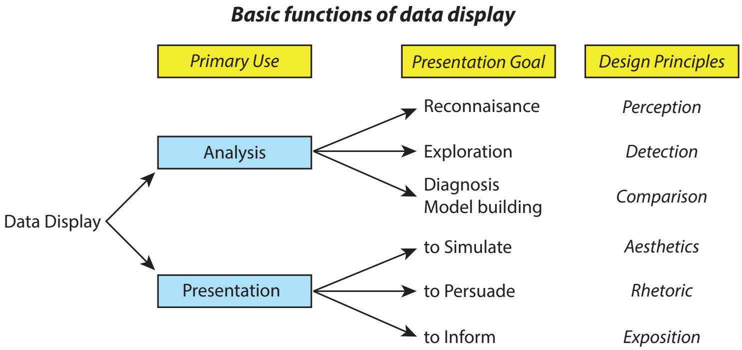
Different audiences require different graphs:

- **Presentation**: A single, carefully crafted graph to appeal to a wide audience
- **Exploration, analysis**: Many related graphics from different perspectives, for a narrow audience (often: you!)



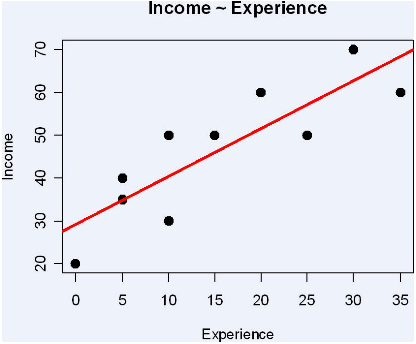
Graphical methods: Presentation goals

Different presentation goals appeal to different design principles

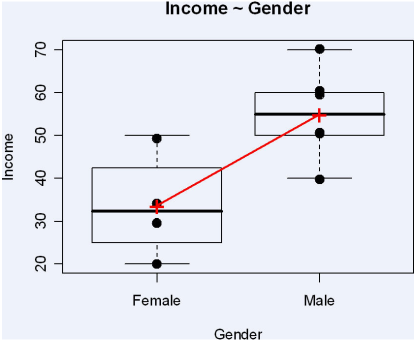


Graphical methods: Quantitative data

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



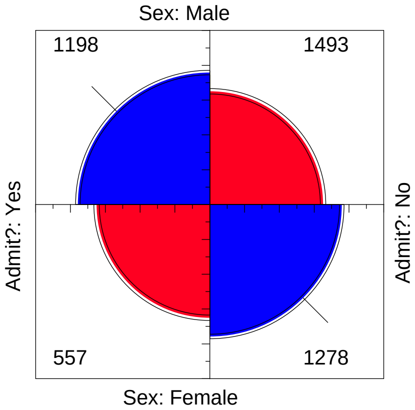
Scatterplot of Income vs. Experience



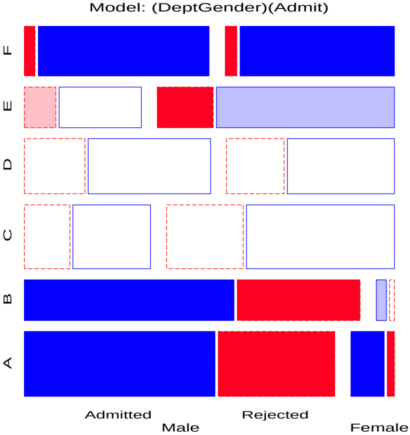
Boxplot of Income by Gender

Graphical methods: Categorical data

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

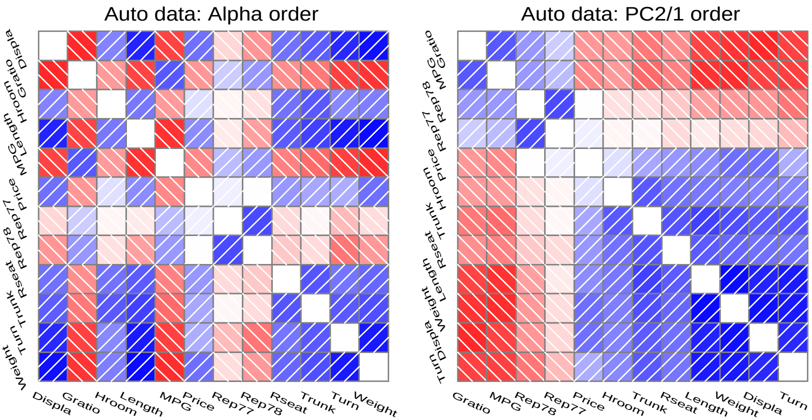


Fourfold display for 2×2 table



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)

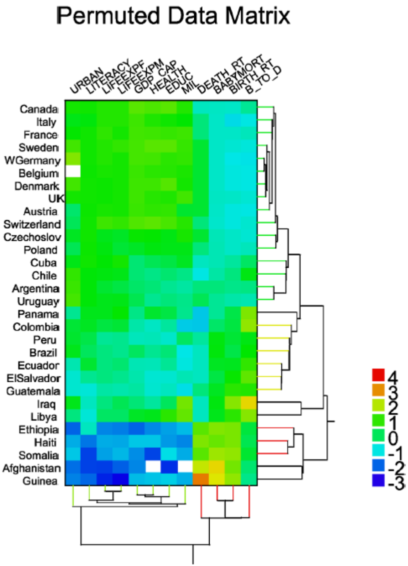
- Effect ordering and high-lighting for tables

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		

Clustered heat map: Showing patterns in tables



The clustered heat map is one method for making large tables more visually understandable.

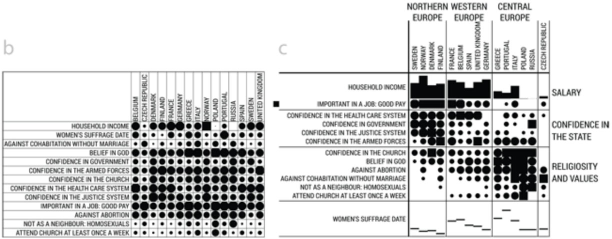
- Social statistics from UN survey
- Rows and columns are sorted, using cluster analysis
- Standardized data values are encoded using color

Bertifier: Turning tables into graphics

Bertifier: A web app implementing Bertin's idea of the *reorderable matrix*.
See: http://www.aviz.fr/bertifier

a

	Belgi	Czech	Denr	Finla	Fran	Gern	Gree	Ita	Norw	Polar	Portu	Russ	Sp	Swe	United
Household income	2687	16957	2468	2573	2831	2879	2044	24	3145	1537	1936	1528	22	2624	26904
Women's suffrage date	1948	1920	1915	1906	1944	1918	1952	19	1913	1918	1976	1918	19	1921	1928
Against cohabitation w	12	42	4	18	8	20	30	46	12	39	17	39	16	6	19
Belief in God	61	36	63	69	52	63	93	91	56	96	86	77	76	46	65
Confidence in Governm	32	21	55	42	34	29	22	28	51	23	30	60	35	54	19
Confidence in the arm	50	34	72	83	73	58	70	75	57	63	75	73	57	41	89
Confidence in the chur	36	20	63	47	41	40	52	67	44	65	67	67	31	39	36
Confidence in the heal	91	42	75	73	78	34	39	54	74	44	58	51	79	75	80
Confidence in the justi	50	35	87	73	56	58	50	36	78	44	48	41	42	69	51
Important in a job: goo	60	85	54	58	58	73	94	76	56	93	88	93	77	62	75
Against abortion	56	51	28	40	44	60	65	72	42	75	61	63	57	25	57
Not as a neighbour: ho	7	22	5	12	5	16	30	21	6	52	21	61	5	7	10
Attend church at least	15	13	5	7	11	12	19	35	9	54	25	8	21	9	17



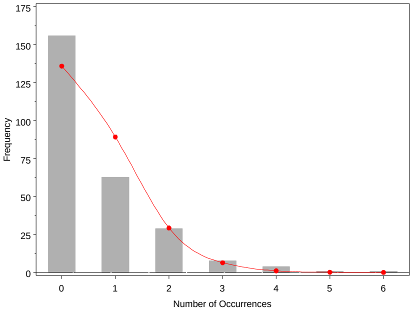
1. A table: Attitudes and attributes by country
2. Values encoded by size and shape
3. Sorted and grouped by themes and country regions

Watch: Youtube video of Bertifier

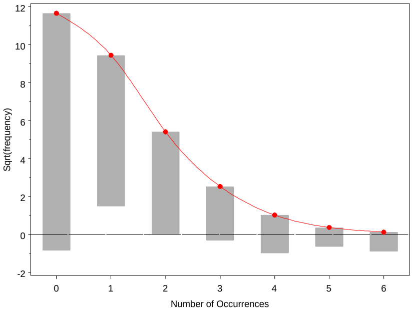
Visual comparisons

Comparisons— Make visual comparisons easy

- Visual grouping— connect with lines, make key comparisons contiguous
- Baselines— compare *data* to *model* against a line, preferably horizontal
- Frequencies often better plotted on a square-root scale

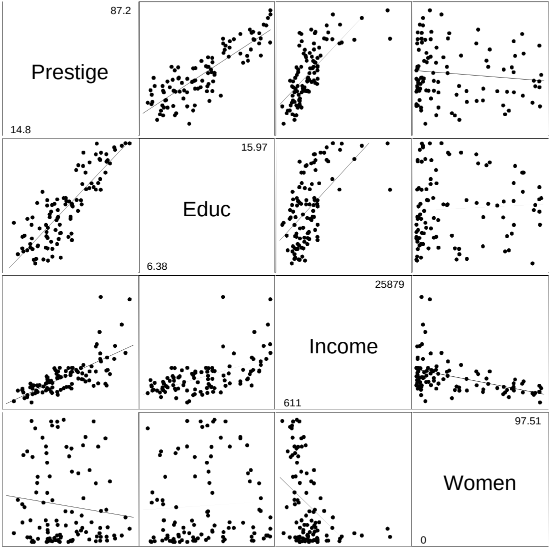


Standard histogram with fit

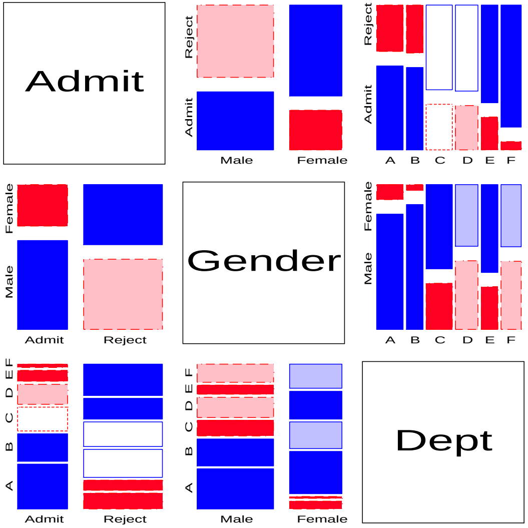


Suspended rootogram

- **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*
- But can add summaries: smoothed curve(s), trend lines, ...
- 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

References I

Friendly, M. Conceptual and visual models for categorical data. *The American Statistician*, 49:153–160, 1995.

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.

Tufte, E. R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.