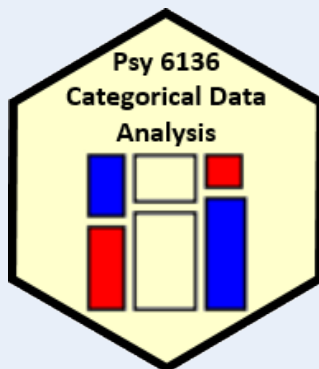


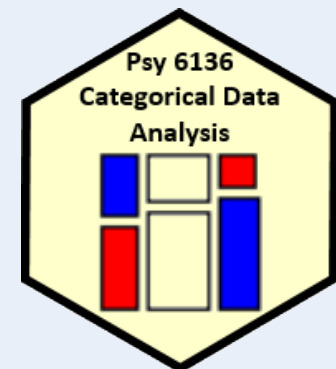
Categorical Data Analysis

Course overview



Michael Friendly
Psych 6136

<http://friendly.github.io/psy6136>



Course goals

This course is designed as a broad, applied introduction to the statistical analysis of categorical 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 independence) 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

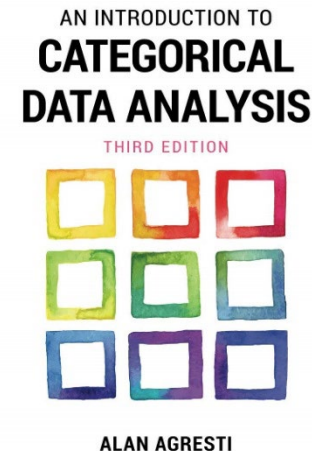
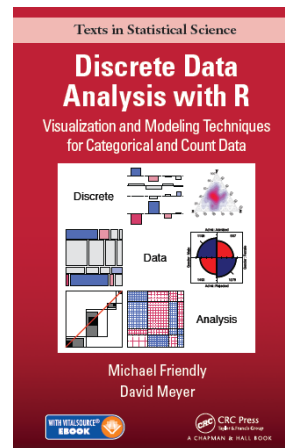
Textbooks

Main texts

- Friendly & Meyer (2016). *Discrete Data Analysis with R: Visualizing & Modeling Techniques for Categorical & Count Data*
 - 30% discount on Routledge web site (code: ADC22)
 - Draft chapters on <http://euclid.psych.yorku.ca/www/psy6136>
 - DDAR web site: <https://ddar.datavis.ca>
- Agresti (2007). *An Introduction to Categorical Data Analysis*, 3rd E. Wiley & Sons.

eBook available

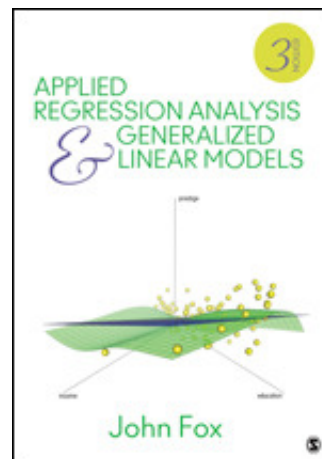
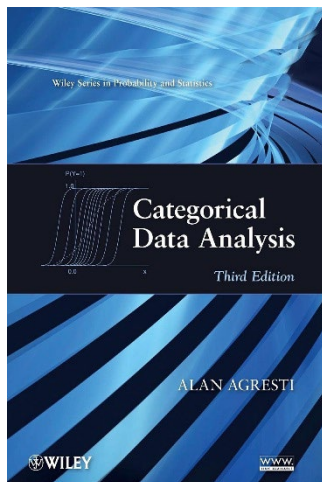
PDF on course web site



Textbooks

Supplementary readings

- Agresti (2013). *Categorical Data Analysis*, 3rd ed. [More mathematical, but the current Bible of CDA]
 - PDF available: <https://bityl.co/FG9c>
- Fox (2016). *Applied Regression Analysis and Generalized Linear Models*, 3rd ed.



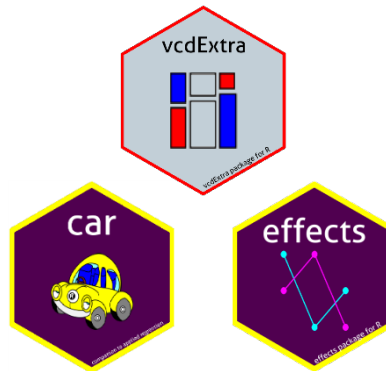
Expectations & grading

- I expect you will read chapters in *DDAR* & Agresti *Intro* each week
 - See **Topic Schedule** on course web site
 - R exercises: A few are listed as (ungraded) **Assignments**
 - Class discussion: Help make classes participatory
- **Evaluation:**
 - (2 x 40%) Two take-home projects: Analysis & research report, based on assignment problems or your own data
 - (20%)
 - Assignment portfolio: best work, enhanced
 - Research report on journal article(s) of theory / application of CDA
 - In-class presentation (~15 min) on application of general interest

What you need

- R, version ≥ 3.6 [R 4.2 is current]
 - Download from <https://cran.r-project.org/>
- RStudio IDE, highly recommended
 - <https://www.rstudio.com/products/rstudio/>
- R packages: see course web page

- vcd
- vcdExtra
- car
- effects
- ...



R script to install packages:
<https://friendly.github.io/6136/R/install-vcd-pkgs.R>

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", "trans"}
- Marital status: {"Never married", "Married", "Separated", "Divorced", "Widowed" }
- Party preference: {"NDP", "Liberal", "Conservative", "Green"}
- Treatment improvement: {"none", "some", "marked"}
- Age: {"0-9", "10-19", "20-29", "30-39", ... }.
- Number of children: 0, 1, 2, 3,

Questions:

- Which of these are ordered (ordinal)?
- Which could be treated as numeric? How?
- Which have missing categories, sometimes ignored, or treated as "Other"

Categorical data: Structures

Categorical (frequency) data appears in various forms

- Tables: often the result of `table()` or `xtabs()`

- 1-way

- 2-way – 2×2 , $r \times c$

- 3-way

Gender compared to handedness

	Handed		
	Left	Right	
Female	7	46	53
Male	5	63	68
	12	109	121

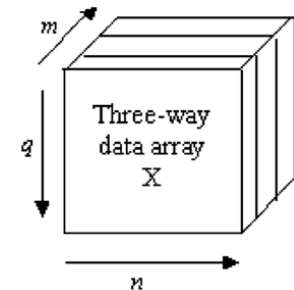
- Matrices: `matrix()`, with row & col names

- Arrays: `array()`, with `dimnames()`

- Data frames

- Case form (individual observations)

- Frequency form



1-way tables

- Unordered factors

	Black	Brown	Red	Blond
n	108	286	71	127
%	0.18	0.48	0.12	0.21

Hair color of 592 students

	BQ	Cons	Green	Liberal	NDP
n	104	392	126	404	174
%	0.087	0.33	0.1	0.34	0.14

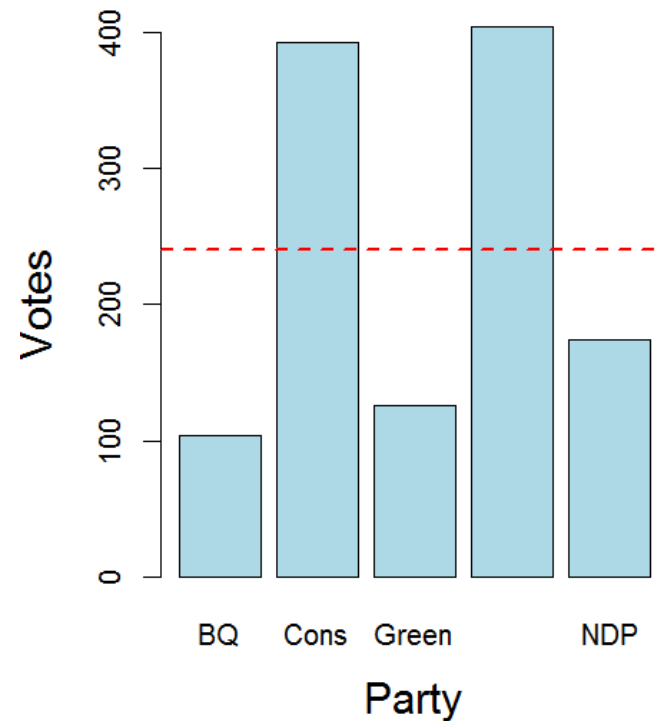
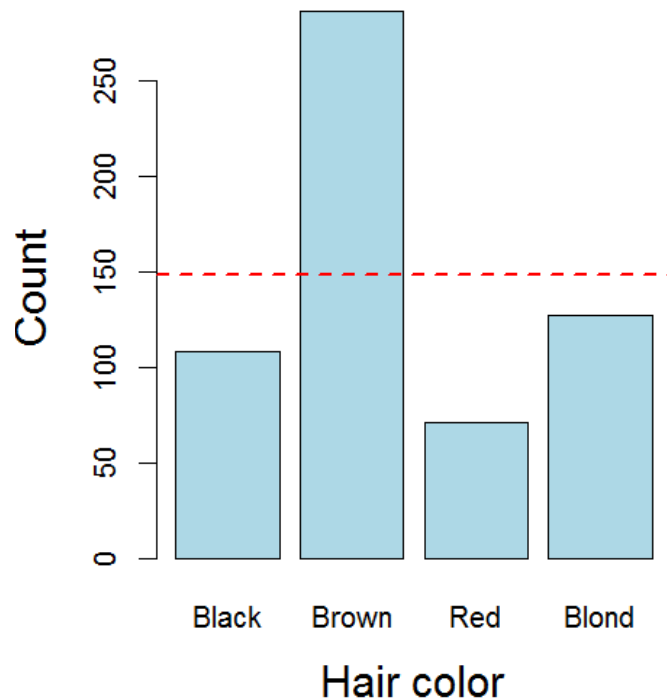
Voting intentions in Harris-Decima poll, 8/21/08

Questions:

- Are all hair colors equally likely?
- Aside from Brown hair, are others equally likely?
- Is there a diff in voting intentions for Liberal vs. Conservative

1-way tables

- Even here, simple graphs are more informative than tables



But these don't really answer the questions. Why?

1-way tables

- Ordered, quantitative factors
 - Number of sons in Saxony families with 12 children

```
> data(Saxony, package="vcd")
```

```
> Saxony
```

```
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

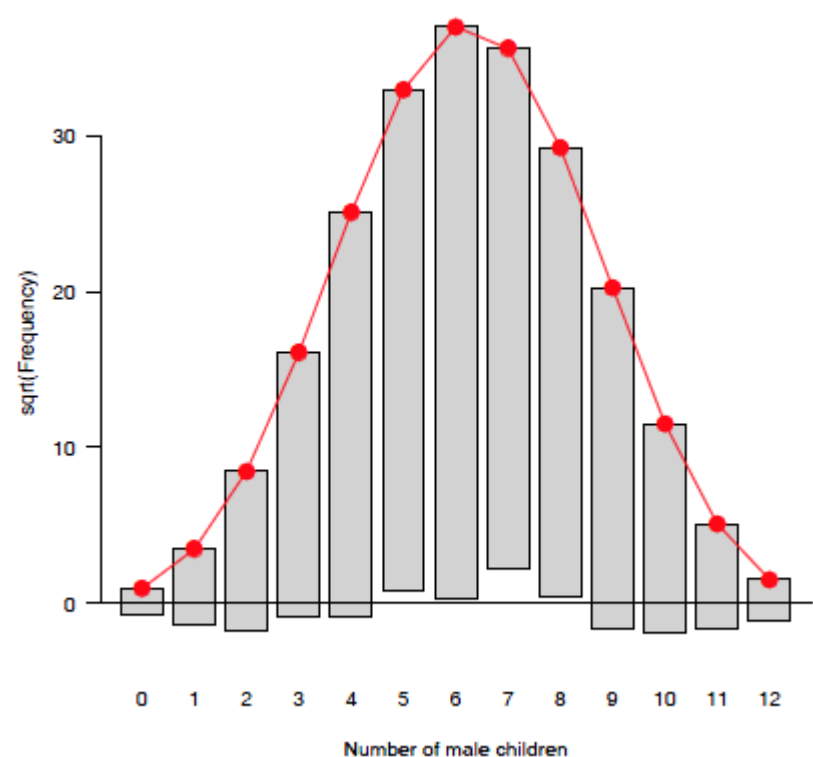
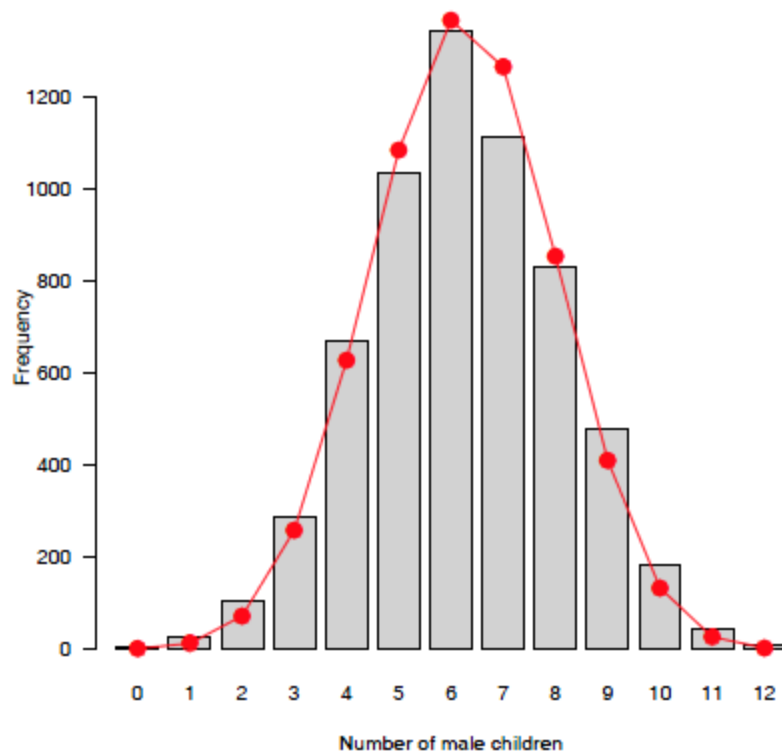
Questions:

- What is the **form** of this distribution?
- Is it useful to think of this as a **binomial distribution**?
- If so, is $\text{Pr}(\text{male}) = 0.5$ reasonable to describe the data?
- How could families have > 10 children?

1-way tables: graphs

For a particular distribution in mind:

- Plot the data together with the fitted frequencies
- Better still: **hanging rootogram**: freq on sqrt scale; hang bars from fitted values



2-way tables: 2 x 2 x ...

- Two-way

	Gender	Male	Female
Admi t			
Admi tted		1198	557
Rejected		1493	1278

Admission to
graduate programs
at UC Berkeley

- Three-way, stratified by another factor

... by Department

		Dept	A	B	C	D	E	F
Admi t	Gender							
Admi tted	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?

Larger tables

```
> margin.table(HairEyeColor, 1:2)
```

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

2-way

Actually, this is a 2D
margin of a 3-way table

```
> ftable(Eye ~ Sex + Hair, data=HairEyeColor)
```

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

3-way (& higher) can
be “flattened” for a
more convenient
display

formula notation:
row vars ~ col vars

Table form

- Table form is convenient for display, but information is **implicit**
 - a table has dimensions, `dim()` and `dimnames()`
 - the “observations” are the cells in the tables
 - the “variables” are the dimensions of the table (factors)
 - the cell value is the count or frequency

```
> dim(haireye)
[1] 4 4
> dimnames(haireye)
$Hair
[1] "Black" "Brown" "Red"   "Blond"

$Eye
[1] "Brown" "Blue"  "Hazel" "Green"
```

```
> names(dimnames(haireye)) # factor names
[1] "Hair" "Eye"
> prod(dim(haireye))       # of cells
[1] 16
> sum(haireye)             # total count
[1] 592
```


Datasets: frequency form

- Another common format is a dataset in **frequency form**

```
> as.data.frame(haireye)
  Hair   Eye Freq
1 Black Brown  68
2 Brown Brown 119
3  Red Brown  26
4 Blond Brown   7
5 Black  Blue  20
6 Brown  Blue  84
7  Red  Blue  17
8 Blond  Blue  94
9 Black Hazel  15
10 Brown Hazel  54
11  Red Hazel  14
12 Blond Hazel  10
13 Black Green   5
14 Brown Green  29
15  Red Green  14
16 Blond Green  16
```

- Use `as.data.frame(table)`
- One row for each cell
- Columns: factors + Freq or count

Datasets: case form

- Raw data often arrives in **case form**

```
> expand.dft(as.data.frame(haireye)) |>
+   as_tibble() |>
+   mutate(age = round( runif( n =
+     sum(haireye), min=17, max=29)))
# A tibble: 592 x 3
   Hair Eye   age
  <chr> <chr> <dbl>
1 Black Brown   19
2 Black Brown   19
3 Black Brown   27
4 Black Brown   23
5 Black Brown   19
6 Black Brown   29
7 Black Brown   25
8 Black Brown   29
9 Black Brown   17
10 Black Brown   23
# ... with 582 more rows
```

- One obs. per case
- # rows = sum of counts
- `vcdExtra::expand.dft()` expands frequency form
- case form is required if there are continuous variables
- case form is **tidy**
- not all CDA functions play well with tibbles

Categorical data analysis: Methods

Methods for categorical data analysis 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 $\Pr\{\text{response}\}$
- SAS: PROC LOGISTIC, CATMOD, GENMOD , INSIGHT (Fit YX), ...
- SPSS: Hiloglinear, Loglinear, Generalized linear models

Models: Response vs. Association

Response models

- Sometimes, one variable is a natural discrete response.
 - Q: How does the response relate to explanatory variables?
 - $\text{Admit} \sim \text{Gender} + \text{Dept}$
 - $\text{Party} \sim \text{Age} + \text{Education} + \text{Urban}$
- ⇒ Logit models, logistic 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]
- ⇒ Loglinear models

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

Models: Response vs. Association

Response models

- Sometimes, one variable is a natural discrete response.
 - Q: How does the response relate to explanatory variables?
 - $\text{Admit} \sim \text{Gender} + \text{Dept}$
 - $\text{Party} \sim \text{Age} + \text{Education} + \text{Urban}$
- ⇒ Logit models, logistic 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]
- ⇒ Loglinear models

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

Response models

Analysis methods for categorical outcome (response) variables have close parallels with those for quantitative outcomes

	Quantitative outcome	Categorical outcome
Continuous predictor	Regression: $\text{lm}(y \sim x_1 + x_2)$	Logistic regression: <code>glm()</code> Loglinear model: <code>loglm()</code> Ordered: prop. odds model: <code>polr()</code>
Categorical predictor	ANOVA: $\text{lm}(y \sim A + B)$ Ordered: polynomial contrasts	χ^2 tests: <code>chisq.test()</code> Ordered: CMH tests, <code>CMHtest()</code> Loglinear model: <code>loglm()</code>
Both	ANCOVA: $\text{lm}(y \sim A + B + x)$	Logistic regression: <code>glm()</code> Loglinear model: <code>loglm()</code>

All use similar model formulas:

```
lm(y ~ A)
```

```
lm(y ~ A*B)
```

```
lm(y ~ X + A)
```

```
lm(y ~ (A+B+C)^2)
```

```
# one way ANOVA
```

```
# two way: A + B + A:B
```

```
# one-way ANCOVA
```

```
# 3-way ANOVA: A, B, C, A:B, A:C, B:C
```

Response models

For quantitative outcomes, `lm()` for everything, formula notation

```
lm(y ~ A)                # one way ANOVA
lm(y ~ A*B)              # two way: A + B + A:B
lm(y ~ X + A)            # one-way ANCOVA
lm(y ~ (A+B+C)^2)        # 3-way ANOVA: A, B, C, A:B, A:C, B:C
```

For categorical outcomes, different modeling functions for different outcome types

```
glm(binary ~ X + A, family="binomial")  # logistic regression
glm(Freq ~ X + A, family="poisson")      # poisson regression
MASS::polr(multicat ~ X + A)             # ordinal regression
nnet::multinom(multicat ~ X + A)         # multinomial regression
loglin(table, margins)                  # loglinear model
MASS::loglm(Freq ~ .)                   # loglinear model, . = A+B+C+ ...
MASS::loglm(Freq ~ .^2)                  # + all two-way associations
```


Data display: Tables vs. 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

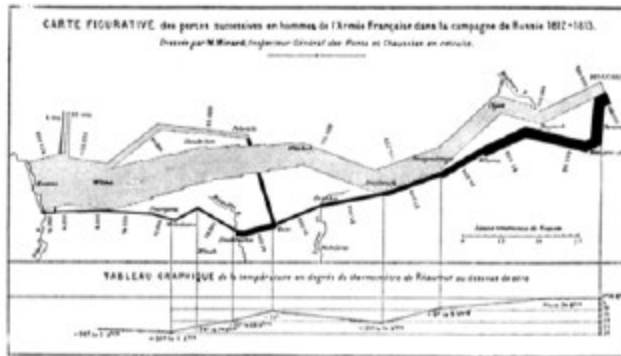
Tables vs. Graphs

- 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?
 - \implies design graphs and tables to 'speak to the eyes'

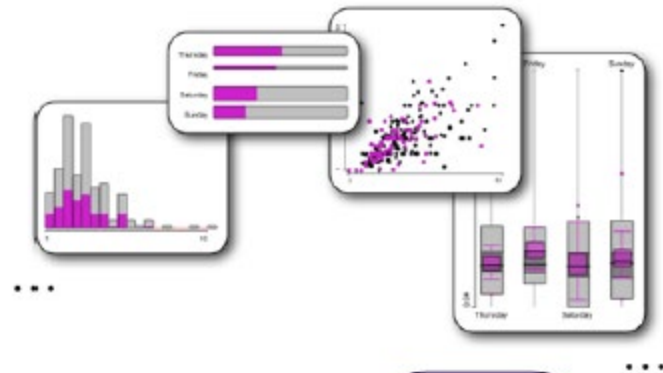
Graphical methods: Communication goals

Different graphs for different audiences

- **Presentation:** A carefully crafted graph to appeal to a wide audience
- **Exploration, analysis:** Possibly many related graphs, different perspectives, narrow audience (often: just you!)



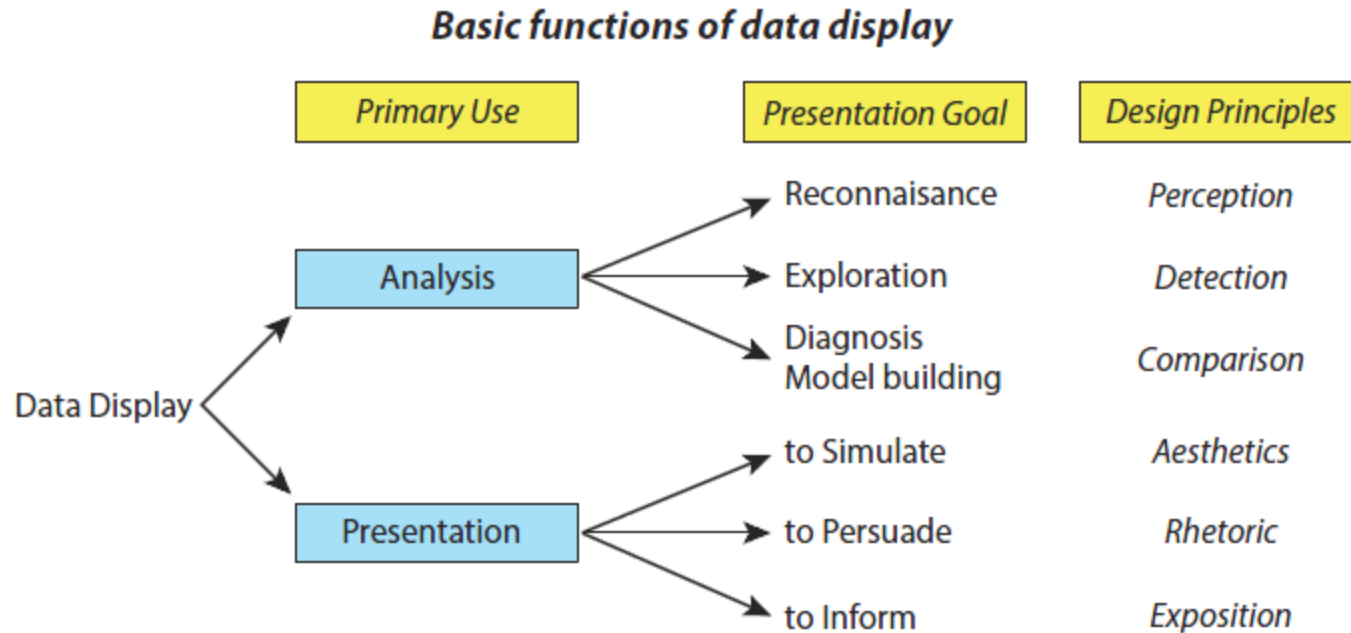
Presentation



Exploration

Graphical methods: Presentation goals

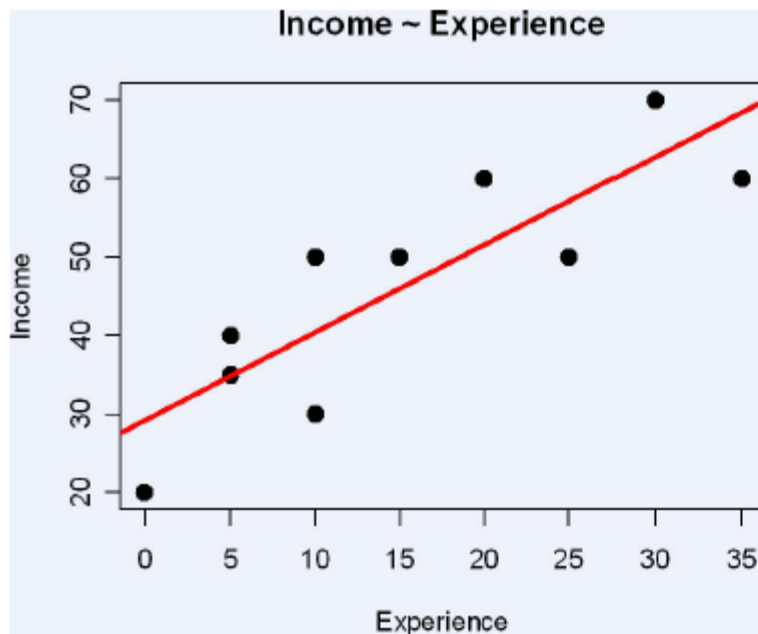
- Different presentation goals appeal to different design principles



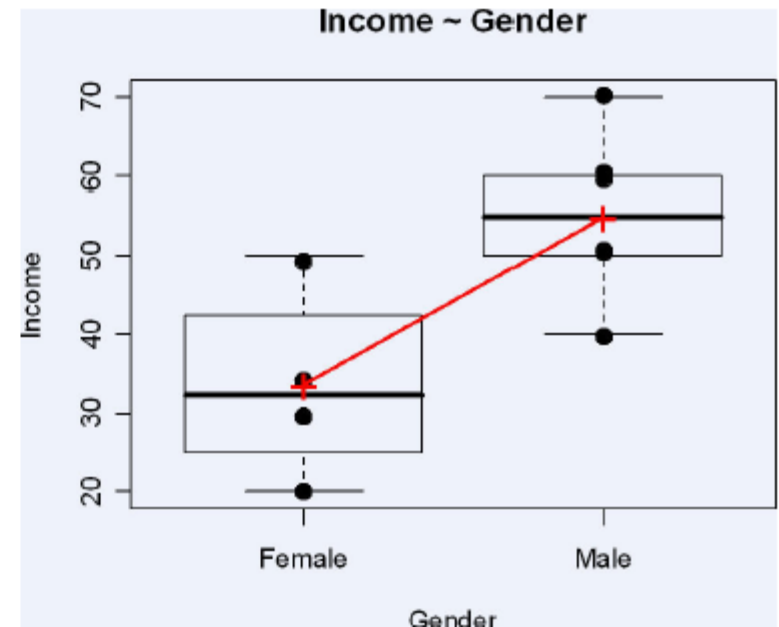
Think: What do I want to communicate? For what purpose?

Graphical methods: Quantitative data

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



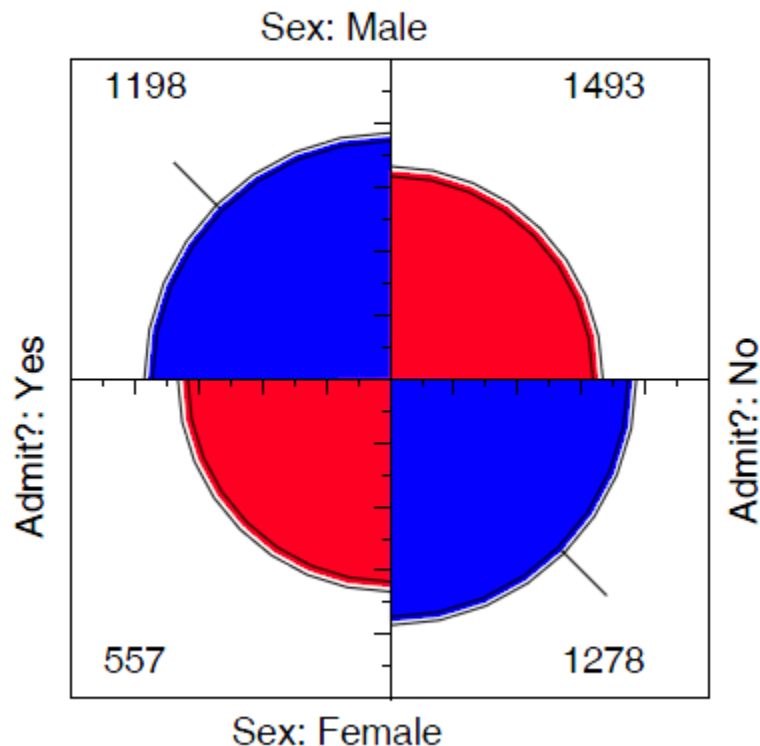
Scatterplot of Income vs.
Experience



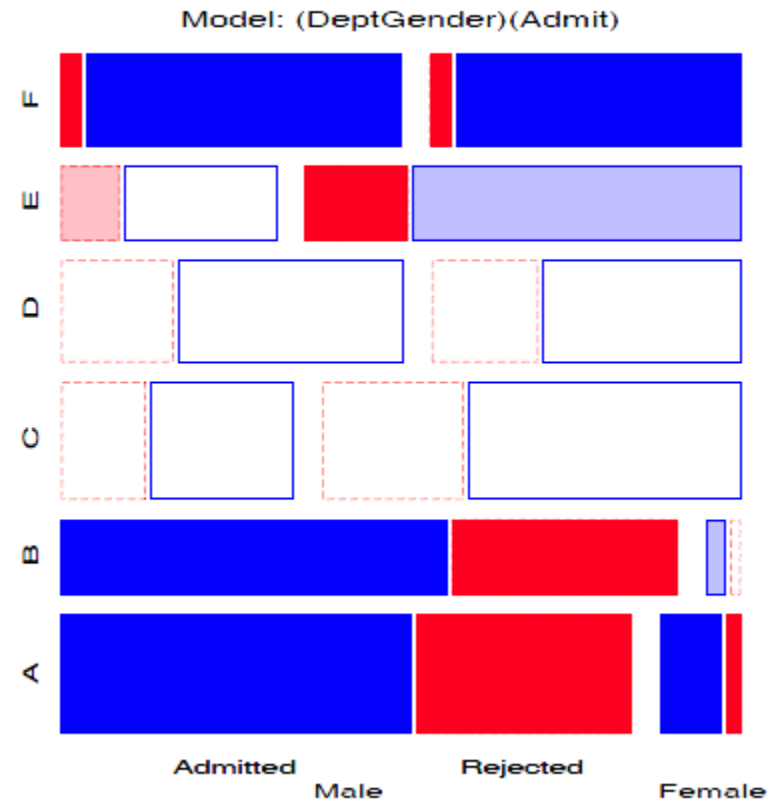
Boxplot of Income by Gender

Graphical methods: Categorical data

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



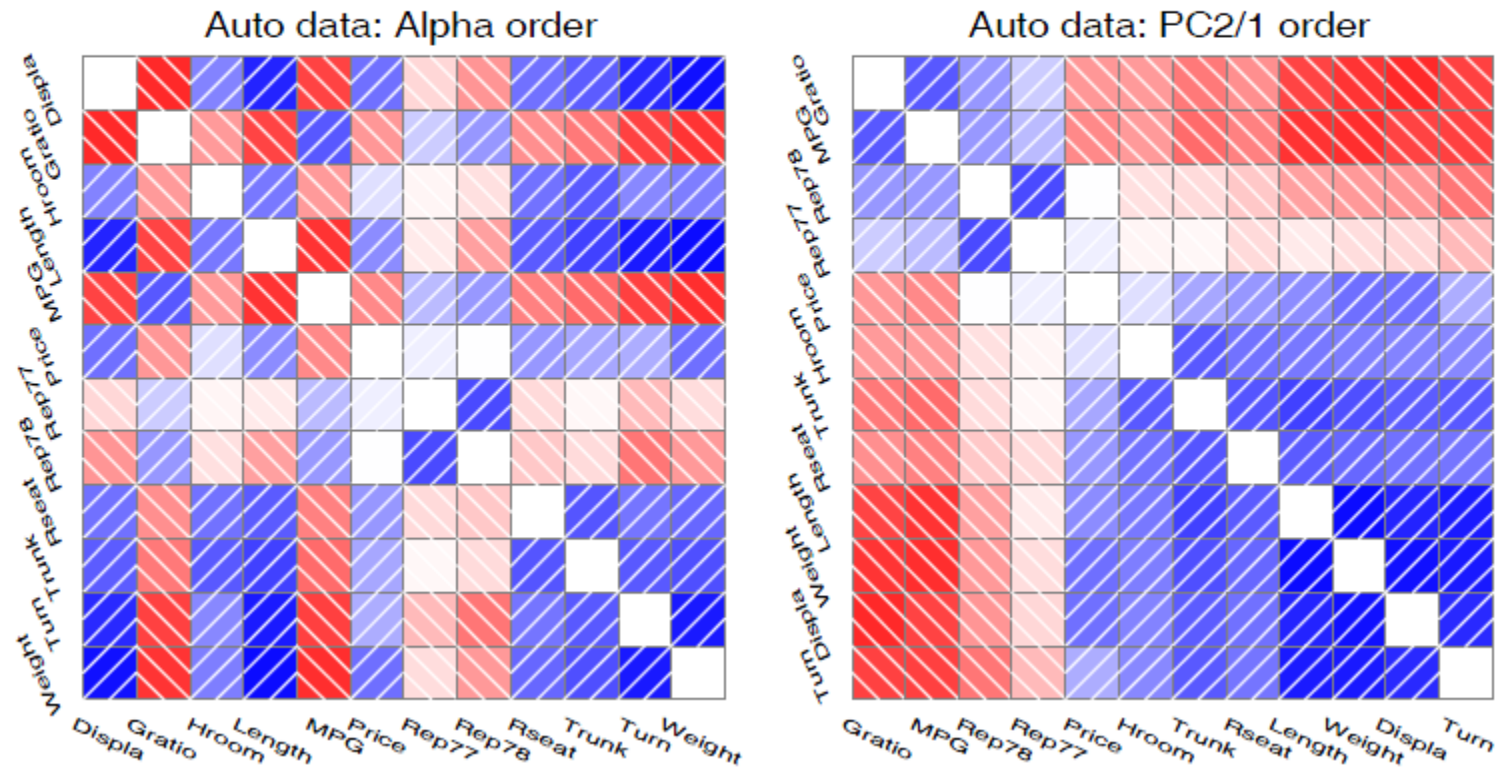
Fourfold display for 2x2 table



Mosaic plot for 3-way table

Principles of graphical display

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



Friendly & Kwan (2003). [Corrgrams: Exploratory displays for correlation matrices](#). *American Statistician*, **54**(4): 316-324.

Tabular displays

- Effect ordering and high-lighting for tables

Table: 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

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

There is an association, but it is hard to see the general pattern

Tabular displays

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

The pattern is clearer when the eye colors are permuted: light hair goes with light eyes & vice-versa

Sometimes, don't need numbers at all

COVID transmission risk ~ Occupancy * Ventilation * Activity * Mask? * Contact.time

A complex 5-way table,
whose message is clearly
shown w/o numbers

A semi-graphic table shows
the **patterns** in the data

There are 1+ unusual cells
here. Can you see them?

Type and level of group activity	Low occupancy			High occupancy		
	Outdoors and well ventilated	Indoors and well ventilated	Poorly ventilated	Outdoors and well ventilated	Indoors and well ventilated	Poorly ventilated
Wearing face coverings, contact for short time						
Silent	Low	Low	Low	Low	Low	Medium
Speaking	Low	Low	Low	Low	Low	Medium
Shouting, singing	Low	Low	Medium	Medium	Medium	High
Wearing face coverings, contact for prolonged time						
Silent	Low	Low	Medium	Low	Medium	High
Speaking	Low	Low*	Medium	Medium	Medium	High
Shouting, singing	Low	Medium	High	Medium	High	High
No face coverings, contact for short time						
Silent	Low	Low	Medium	Medium	Medium	High
Speaking	Low	Medium	Medium	Medium	High	High
Shouting, singing	Medium	Medium	High	High	High	High
No face coverings, contact for prolonged time						
Silent	Low	Medium	High	Medium	High	High
Speaking	Medium	Medium	High	High	High	High
Shouting, singing	Medium	High	High	High	High	High

Risk of transmission
Low ■ Medium ■ High ■

* Borderline case that is highly dependent on quantitative definitions of distancing, number of individuals, and time of exposure

From: N.R. Jones et-al (2020). Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ* 2020;370:m3223, doi: <https://doi.org/10.1136/bmj.m3223>

Visual table ideas: Heatmap shading

Heatmap shading: Shade the **background** of each cell according to some criterion

The trends in the US and Canada are made obvious

NB: Table rows are sorted by Jan. value, lending coherence

Background shading ~ value:
US & Canada are made to stand out.

Tech note: use white text on a darker background

Unemployment rate in selected countries

January-August 2020, sorted by the unemployment rate in January.

country	Jan ▲	Feb	Mar	Apr	May	Jun	Jul	Aug
Japan	2.4%	2.4%	2.5%	2.6%	2.9%	2.8%	2.9%	3.0%
Netherlands	3.0%	2.9%	2.9%	3.4%	3.6%	4.3%	4.5%	4.6%
Germany	3.4%	3.6%	3.8%	4.0%	4.2%	4.3%	4.4%	4.4%
Mexico	3.6%	3.6%	3.2%	4.8%	4.3%	5.4%	5.2%	5.0%
US	3.6%	3.5%	4.4%	14.7%	13.3%	11.1%	10.2%	8.4%
South Korea	4.0%	3.3%	3.8%	3.8%	4.5%	4.3%	4.2%	3.2%
Denmark	4.9%	4.9%	4.8%	4.9%	5.5%	6.0%	6.3%	6.1%
Belgium	5.1%	5.0%	5.0%	5.1%	5.0%	5.0%	5.0%	5.1%
Australia	5.3%	5.1%	5.2%	6.4%	7.1%	7.4%	7.5%	6.8%
Canada	5.5%	5.6%	7.8%	13.0%	13.7%	12.3%	10.9%	10.2%
Finland	6.8%	6.9%	7.0%	7.3%	7.5%	7.8%	8.0%	8.1%

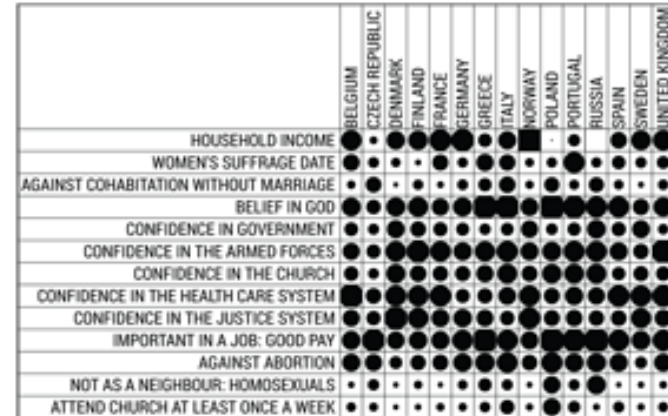
Source: [OECD](#) • [Get the data](#) • Created with [Datawrapper](#)

Bertifier: Turning tables into graphs

a attitudes & attributes

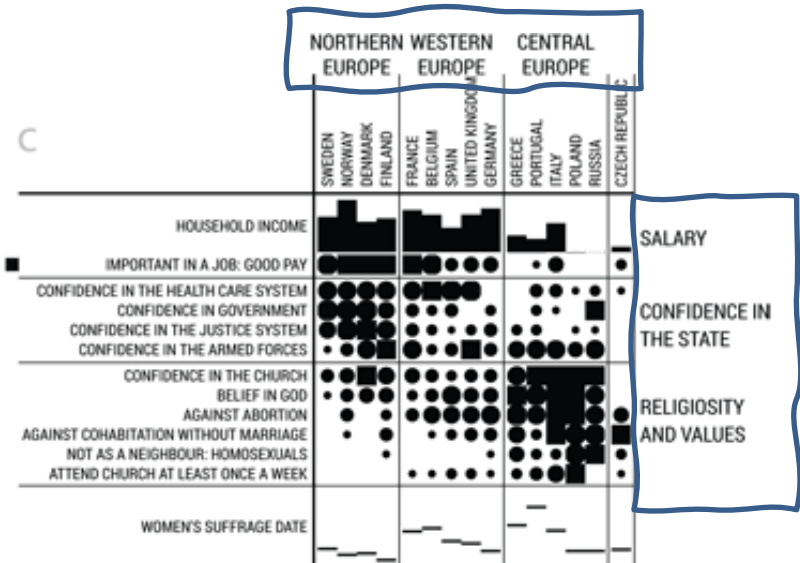
	Belgi	Czech	Dene	Finla	Frani	Gerr	Greeita	Noni	Polis	Porti	Rusi	Sg	Swei	United
Household income	2687	16957	2468	2573	2831	2879	2044	243145	1537	1936	1528	22	2624	26904
Women's suffrage date	1948	1920	1915	1906	1944	1918	1952	19	1913	1918	1976	1918	1921	1928
Against cohabitation	12	42	4	18	8	20	30	46	12	39	17	39	16	19
Belief in God	61	36	63	69	52	63	93	91	56	96	86	77	76	65
Confidence in Govern	32	21	55	42	34	29	22	28	51	23	30	60	35	19
Confidence in the arm	50	34	72	83	73	58	70	75	57	63	75	73	57	89
Confidence in the chur	36	20	63	47	41	40	52	67	44	65	67	67	31	36
Confidence in the heal	91	42	75	73	78	34	39	54	74	44	58	51	79	80
Confidence in the just	50	35	87	73	56	58	50	36	78	44	48	41	42	69
Important in a job: god	60	85	54	58	58	73	94	76	56	93	88	93	77	62
Against abortion	56	51	28	40	44	60	65	72	42	75	61	63	57	57
Not as a neighbour: ho	7	22	5	12	5	16	30	21	6	52	21	61	5	10
Attend church at least	15	13	5	7	11	12	19	35	9	54	25	8	21	17

b encode values by size & shape



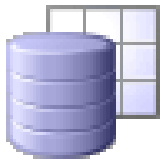
- (a) Table: attitudes and attributes by country
- (b) Visual: encode values by size, shape
- (c) Sort & group by themes, country regions

Bertifier: Bertin's reorderable matrix
See: <http://www.aviz.fr/bertifier>



Data, pictures, models & stories

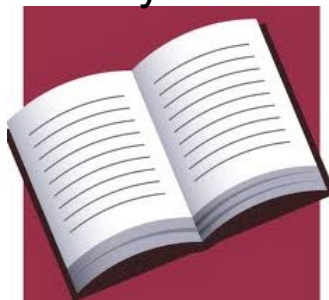
Goal: Tell a credible story about
some real data problem



data

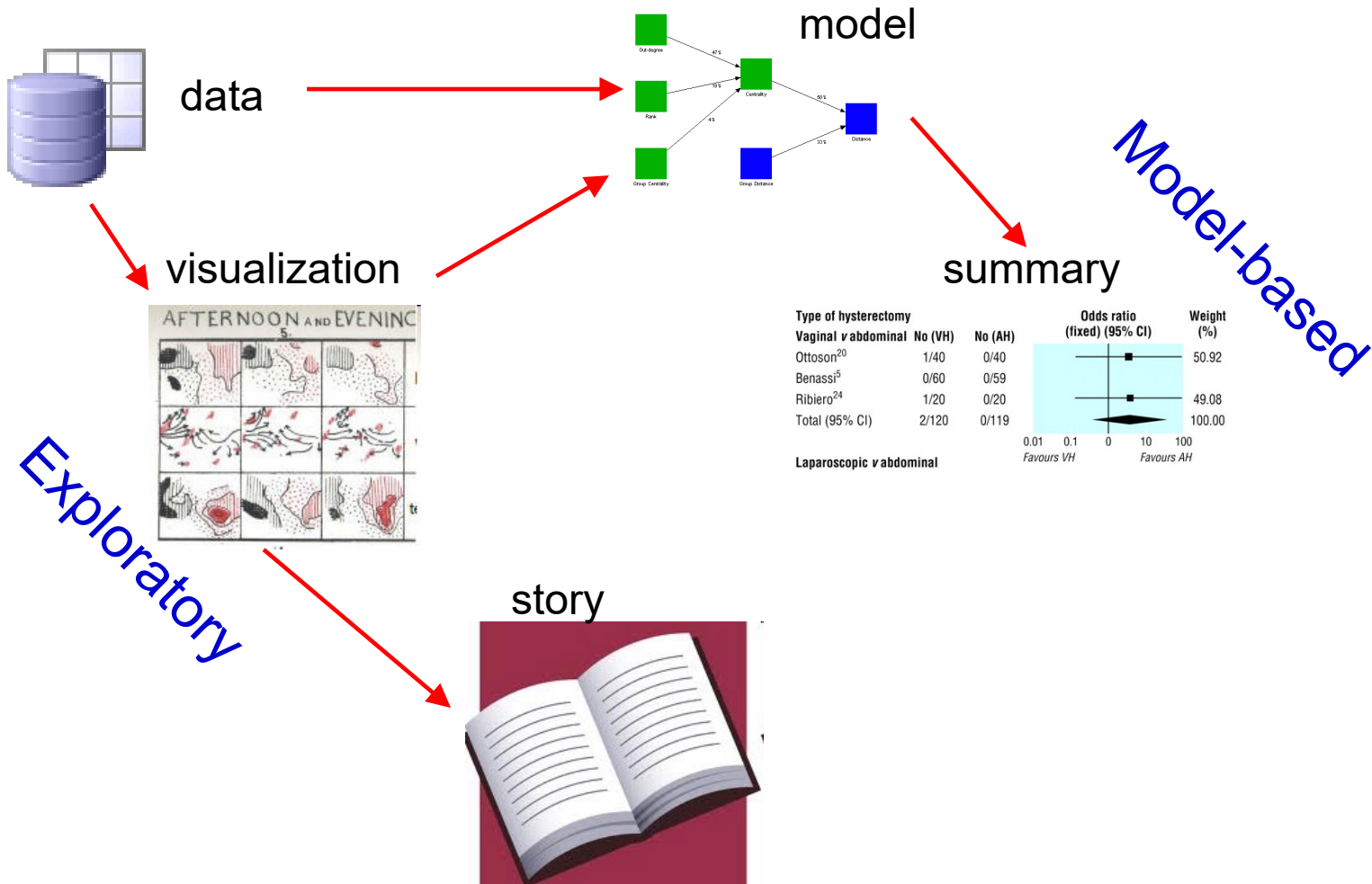
Gender bias
Measles vaccination
Global warming
...

story



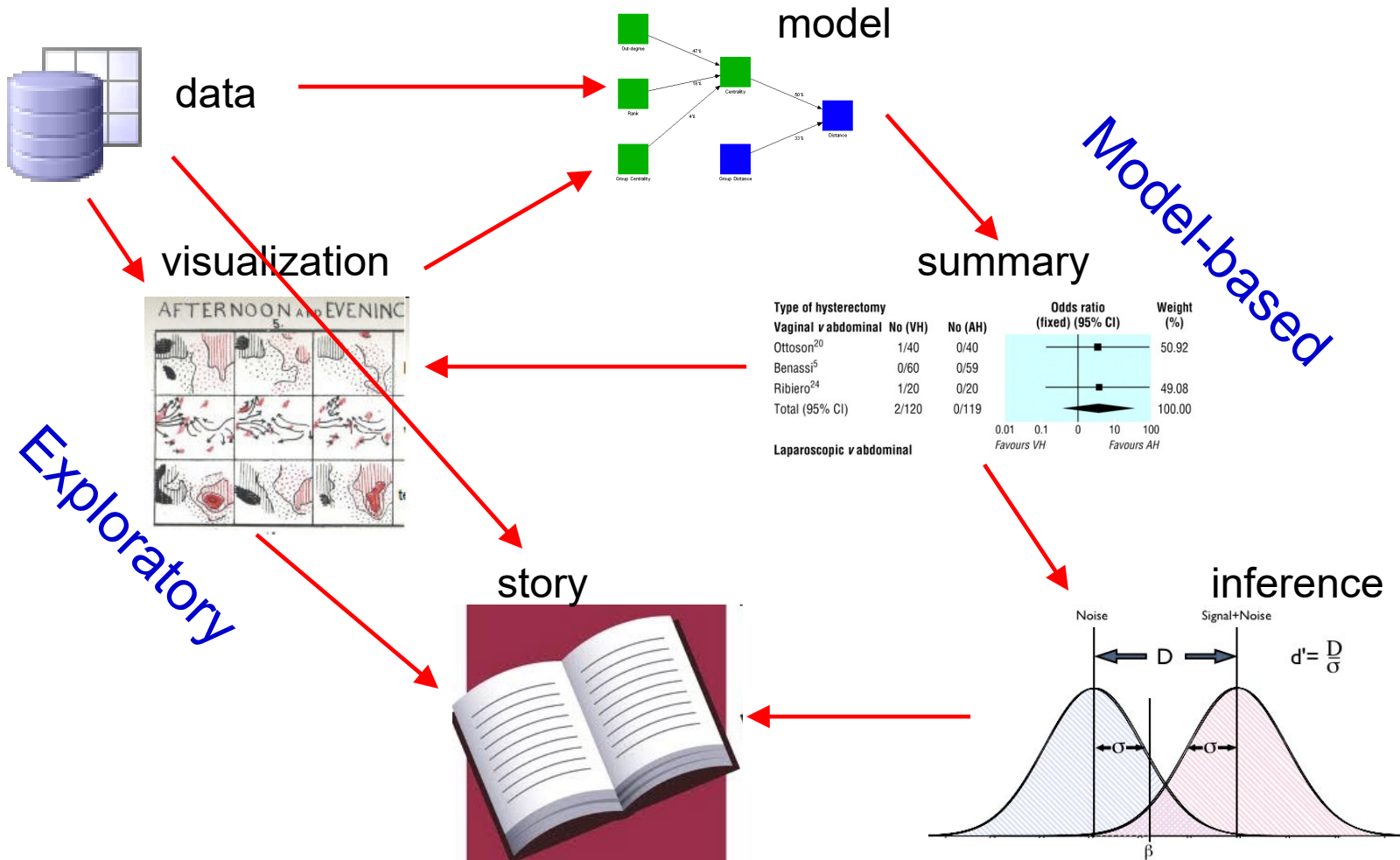
Data, pictures, models & stories

Two paths to enlightenment



Data, pictures, models & stories

Now, tell the story!



Gender Bias at UC Berkeley?

Science, 1975, **187**: 398--403

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed,
and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Determining whether discrimination because of sex or ethnic identity is being practiced against persons seeking passage from one social status or locus to another is an important problem in our society today. It is legally impor-

deceision to admit or to deny admission.

The question we wish to pursue is whether the decision to admit or to deny was influenced by the sex of the applicant.

We cannot know with any certainty the influences on the evaluators in the

by using a
As already
pitfalls ah
but we ir
one of the

We mu
sumptions
of the da
approach.
given disc
plicants de
intelligence
ise, or ot
mately per
students. I
that make
meaningfu
any differ
plicants by
differences
ise as scho
ly one co
example, b
biased est

2 × 2 Frequency Tables: Fourfold displays

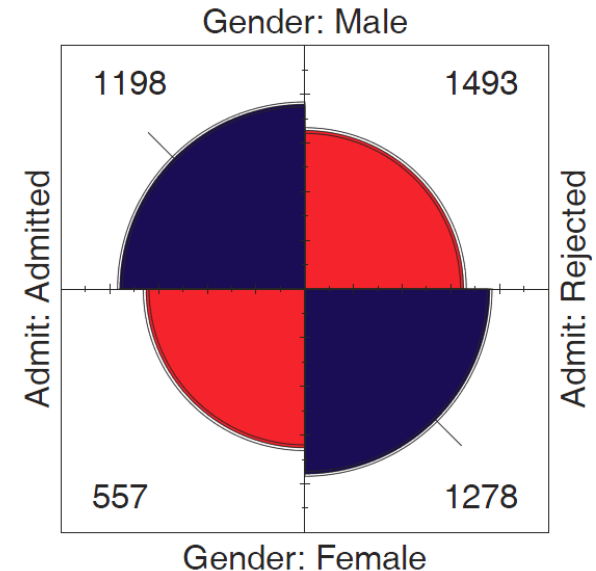
Table: Admissions to Berkeley graduate programs

	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

odds ratio (θ) = 1.84

Males nearly **twice** as likely to be admitted

- Is this a “significant” association?
- Is it evidence for gender bias?
- How to measure strength of association?
- How to visualize?



Fourfold display:

- quarter circles, area \sim frequency
- ratio of areas: odds ratio (θ)
- confidence bands: overlap iff $\theta \approx 1$
- visualize significance!

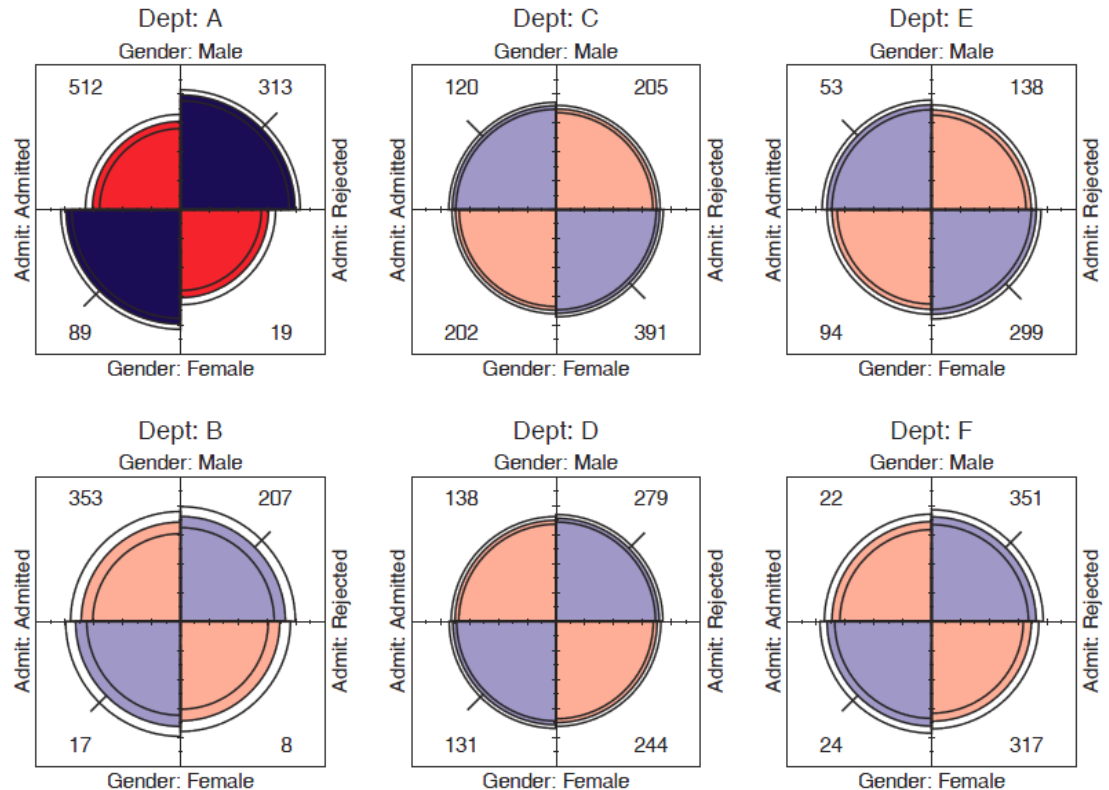
$2 \times 2 \times k$ Stratified tables

The data arose from 6 graduate departments

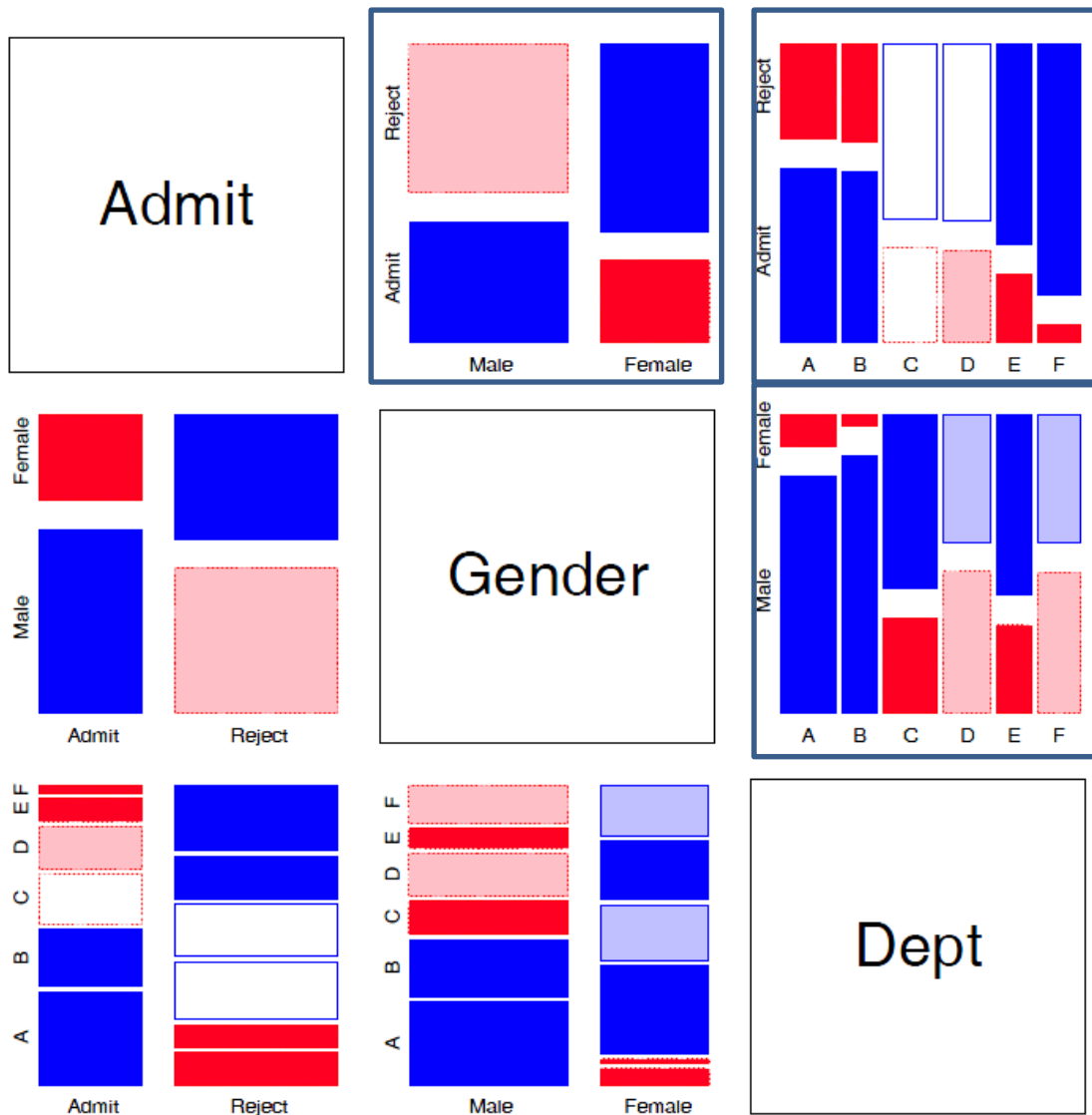
No difference between males & females, except in Dept A where **women more** likely to be admitted!

Design:

- small multiples
- encode direction by color
- encode signif. by shading



Mosaic matrices



Scatterplot matrix analog for categorical data

All pairwise views
Small multiples → comparison

The answer: **Simpson's Paradox**

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

Graphical methods for categorical data

In general, these share similar ideas & scope with methods for quantitative 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

Summary

- Categorical data involves some new ideas
 - Discrete variables: unordered or ordered
 - Counts, frequencies
- New / different data structures & functions
 - tables – 1-way, 2-way, 3-way, ... `table()`, `xtabs()`
 - similar in matrices or arrays `matrix()`, `array()`
 - datasets:
 - frequency form
 - case form
- Graphical methods: often use area \sim Freq
- Models: Most are \approx natural extensions of `lm()`