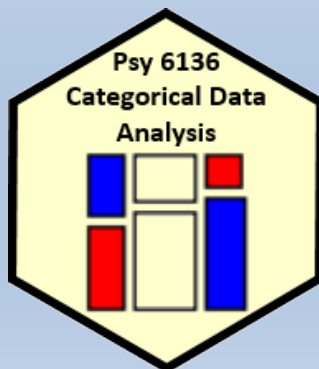


Categorical Data Analysis

Course overview

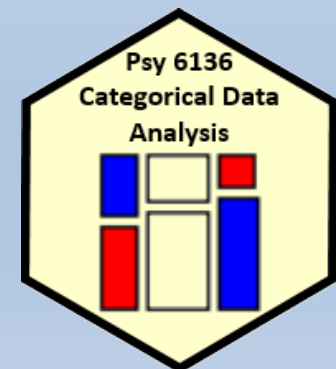


Michael Friendly
Psych 6136

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



@datavisFriendly || #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

Course schedule

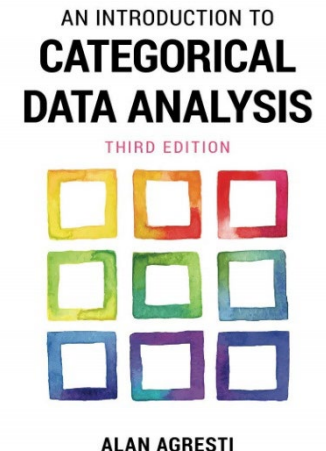
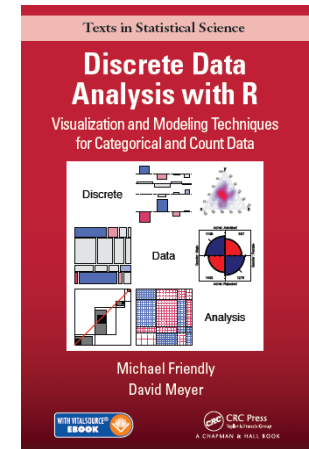
The [schedule](#) page provides links to slides, tutorials, readings & R scripts

Week	Topic	Readings	 R files	
1	Overview [slides] [4up] [Working with R Studio] [4up]	DDAR: Ch1 , Ch2 ; Agresti: Ch1	R-into.R [knitr]	
2	Discrete distributions [slides] [4up]	DDAR: Ch3	R-data.R [knitr] binomial.R [knitr]	
3	Two-Way Tables: Independence & Association [slides] [4up]	DDAR: Ch4 ; Agresti: Ch2	berk-4fold.R [knitr] vision-sieve.R [knitr]	
4	Two-Way Tables: Ordinal Data and Dependent Samples [Tutorial] on two-way tables	DDAR: Ch4 ; Agresti: Ch2	msdiag-agree.R [knitr] haireye-spineplot.R [knitr]	
5	Loglinear Models and Mosaic Displays [slides] [4up] [Tutorial] on loglin models; [Mosaic display animation]	DDAR: Ch5 ; Agresti: 2.7, Ch. 7	berkeley-glm.R [knitr] titanic-loglin.R [knitr]	
6	Correspondence Analysis [slides] [4up] [Tutorial] on CA;	DDAR: Ch6	mental-ca.R [knitr] mca-presex3.R [knitr]	
7	Logistic Regression I [slides] [4up] [Logistic regression tutorial]	DDAR: 7.1-7.3 ; Agresti: 3.1-3.2; Ch 4	arthritis-logistic.R [knitr] cowles-logistic.R [knitr] Arrests-logistic.R [knitr]	
8	Logistic Regression II [slides] [4up]	DDAR: 7.3-7.4 ; Agresti: Ch 4-5	cowles-effect.R [knitr] Arrests-effects.R [knitr] berkeley-diag.R [knitr]	
...				

Textbooks

Main texts

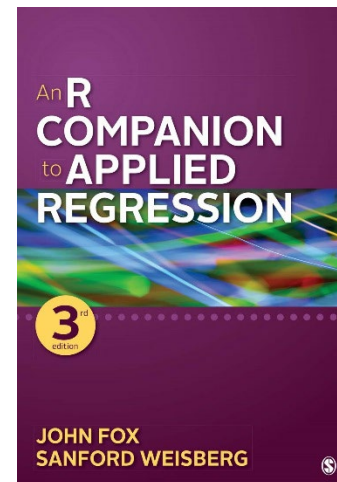
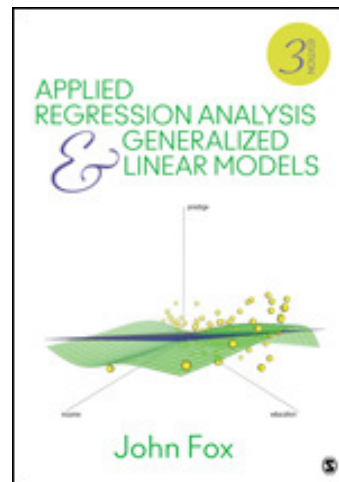
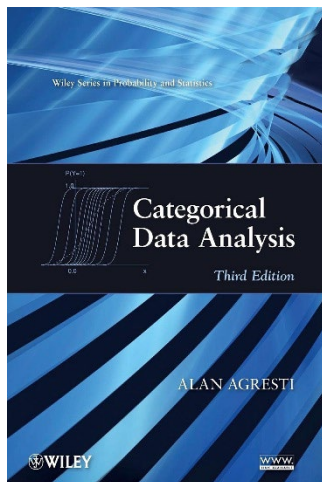
- Friendly & Meyer (2016). *Discrete Data Analysis with R: Visualizing & Modeling Techniques for Categorical & Count Data*
 - 30% discount on [Routledge web site](https://www.routledge.com/9781119440000) (code: ADC22)
 - Draft chapters linked in [Schedule](#)
 - DDAR web site: <https://ddar.datavis.ca>
- Agresti (2007). *An Introduction to Categorical Data Analysis*, 3rd E. Wiley & Sons.
 - eBook available: <https://bit.ly/3Wzqv0n>
 - Or, via [York Bookstore](#)



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. Particularly: Part IV on Generalized Linear Models
- Fox & Weisberg (2018). *An R Companion to Applied Regression*. Also, [web site for the book](#).

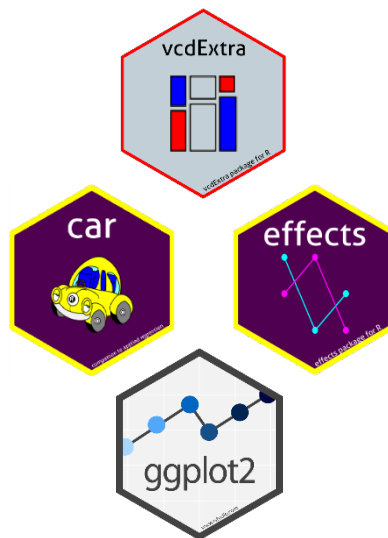


Expectations & grading

- I expect you will read chapters in *DDAR* & Agresti *Intro* each week
 - See [Topic Schedule](#) on course web site
 - R exercises & tutorials: Please work on these
 - R [Assignments](#) : Ungraded, but please submit them when assigned
 - Class discussion: Help make classes participatory
- [Evaluation](#): (tentative: subject to change)
 - (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

The R 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:
 - vcd
 - vcdExtra
 - car
 - effects
 - ggplot2
 - ...



R script to install packages:
<https://friendly.github.io/psy6136/R/install-vcd-pkgs.R>

Categorical data analysis: History

- Categorical data analysis is a relatively recent arrival
 - 1888 – Galton introduces the concept of correlation
 - 1908 – Student's t-distribution for the mean of small samples
 - 1931 – L. L. Thrustone: Multiple factor analysis
 - 1935 – R. A. Fisher's Design of Experiments – ANOVA
 - . . . (time passes)
 - 1972 – Nelder & Wedderburn develop the central ideas of [generalized linear models](#) (logistic & poisson regression)
 - 1973 – J-P. Benzecri: Correspondence analysis (analysis des données)
 - 1974 – Bishop , Fienberg, Holland introduce the [loglinear model](#) for discrete data, ANOVA for $\log(\text{Freq})$
 - 1984 – Leo Goodman enhanced loglinear models for complex data: RC models, mobility tables, panel data, ...

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()`

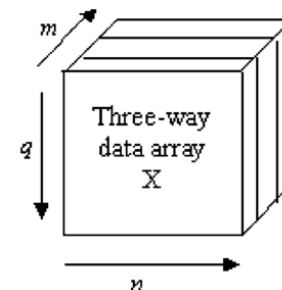
- 1-way
- 2-way – 2×2 , $r \times c$
- 3-way

Gender compared to handedness

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

← margins

- **Matrices**: `matrix()`, with row & col names
- **Arrays**: `array()`, with `dimnames()`
- **Data frames**



- Case form (individual observations)
- Frequency form

	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

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

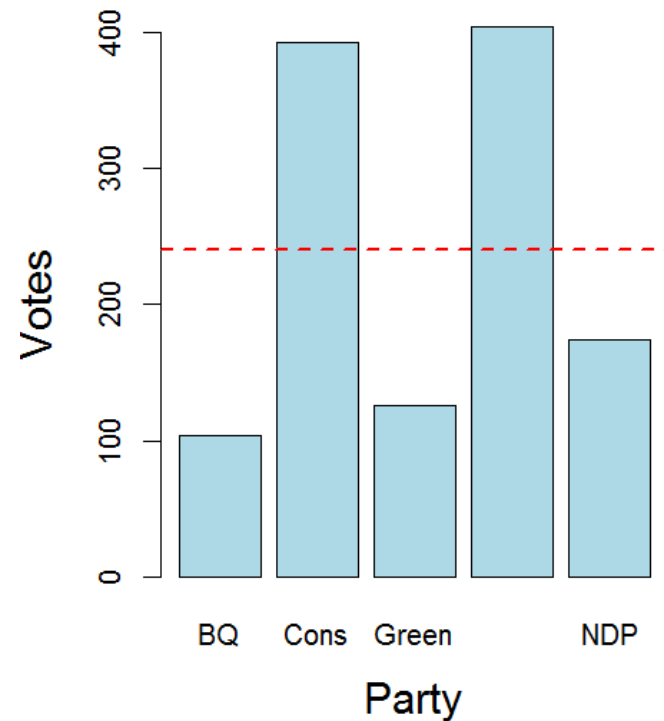
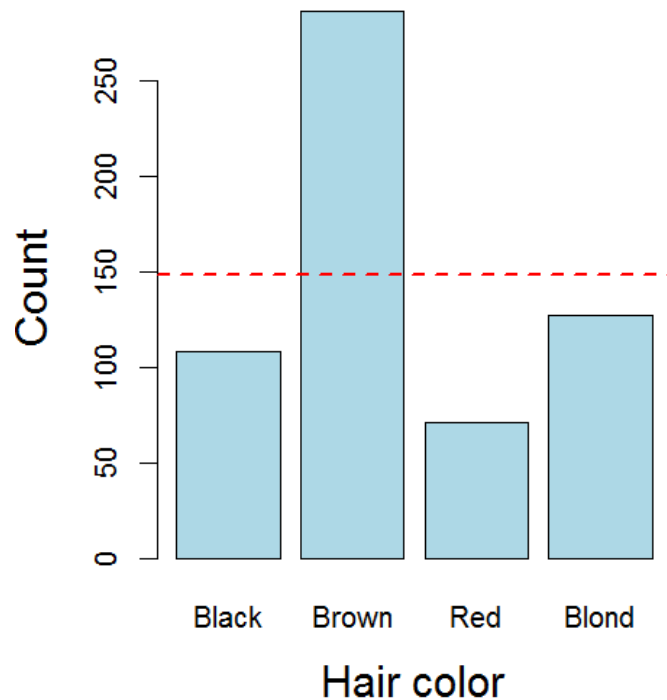
Voting intentions in Harris-Decima poll, 8/21/08

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

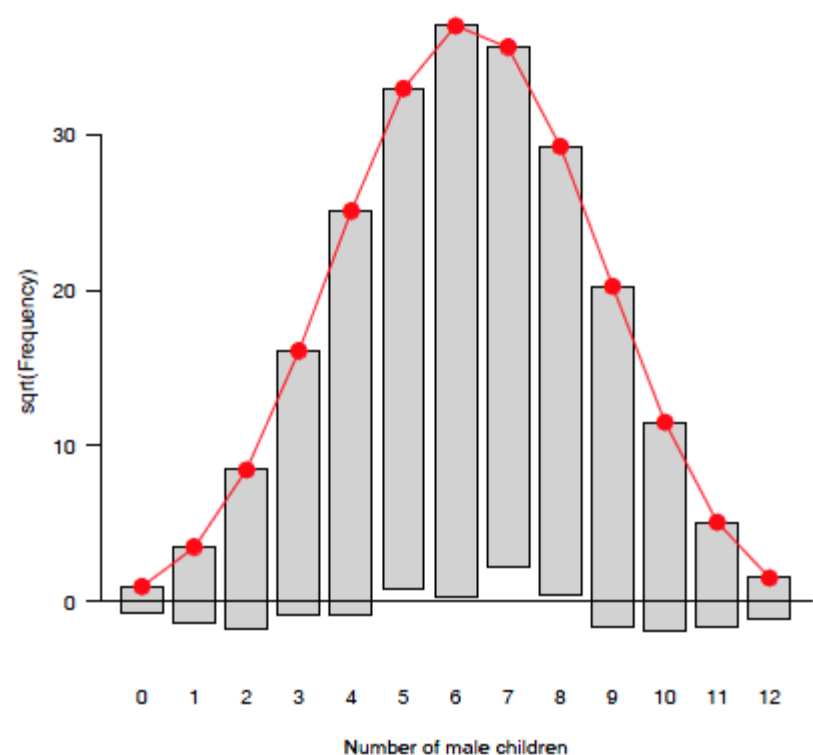
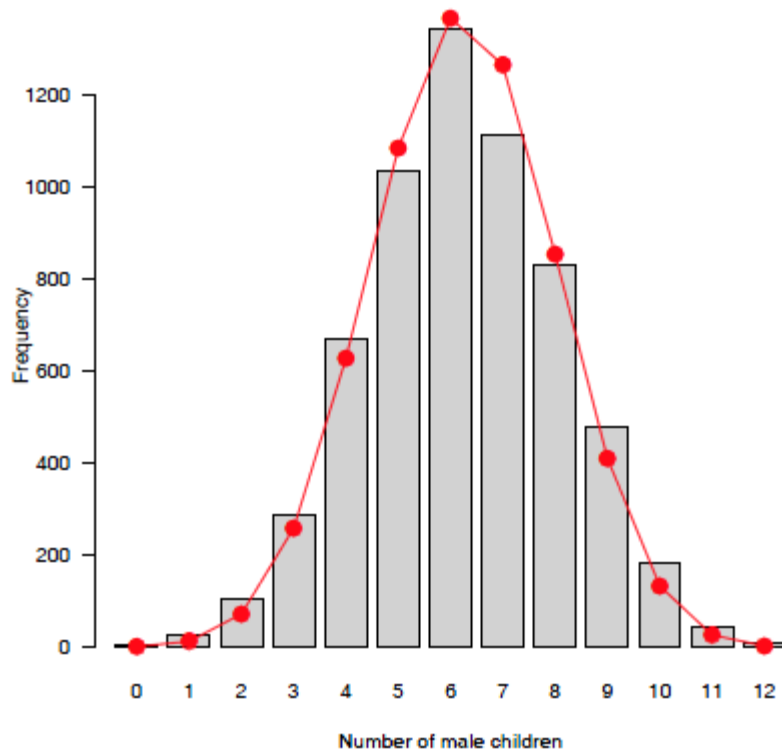
Questions:

- 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: $2 \times 2 \times \dots$

- Two-way

	Gender	Male	Female
Admi t			
Admitted		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							
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?

Larger tables: $r \times c \times \dots$

```
> 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 2-way
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

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

Questions:

- What are the dimensions of the table?
- What is the total frequency?

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 to frequency form
- case form is required if there are **continuous** variables
- case form is **tidy**
- not all CDA functions play well with tibbles

Converting data forms

R functions for CDA sometimes accept only tables (matrices), or data frames, in either case for frequency form.

You may have to convert your data from one form to another

From this ↓	To this ↓	To this ↓	To this ↓
	Case form	Freq form	Table form
Case form	---	<code>Z <- xtabs(~ A+ B)</code> <code>as.data.frame(Z)</code>	<code>table(A, B)</code>
Freq form	<code>expand.dft(X)</code>	---	<code>xtabs(Freq ~ A + B)</code>
Table form	<code>expand.dft(X)</code>	<code>as.data.frame(X)</code>	---

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: $\text{lm}(y \sim x1 + x2)$	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:

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

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

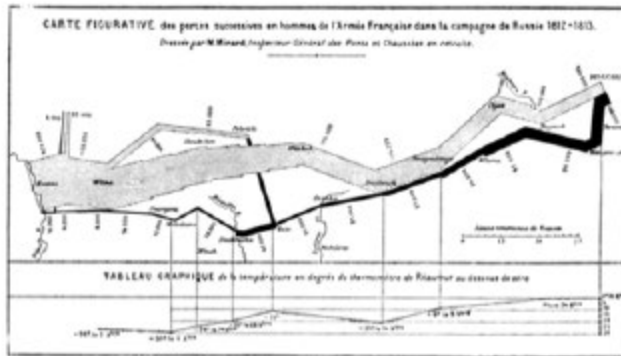
Tables vs. Graphs

- 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'

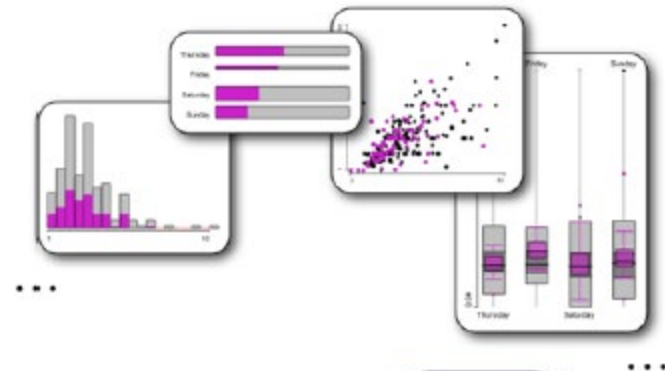
Graphical methods: Communication goals

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!)



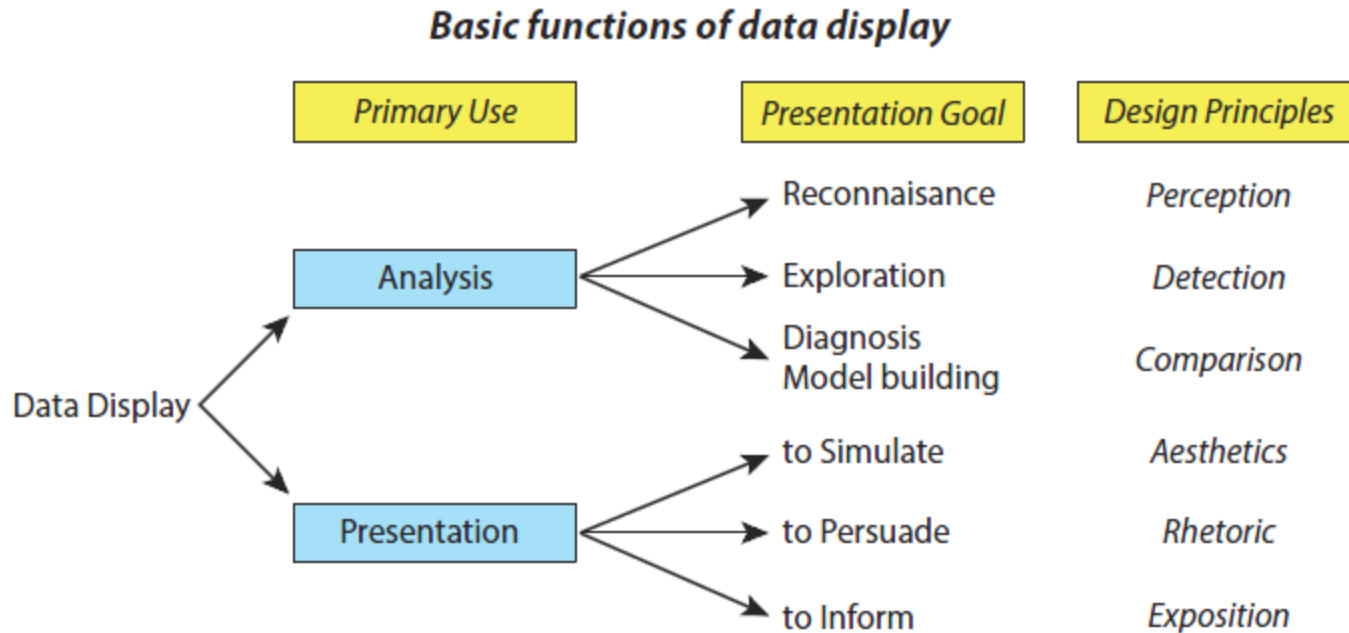
Presentation



Exploration

Graphical methods: Presentation goals

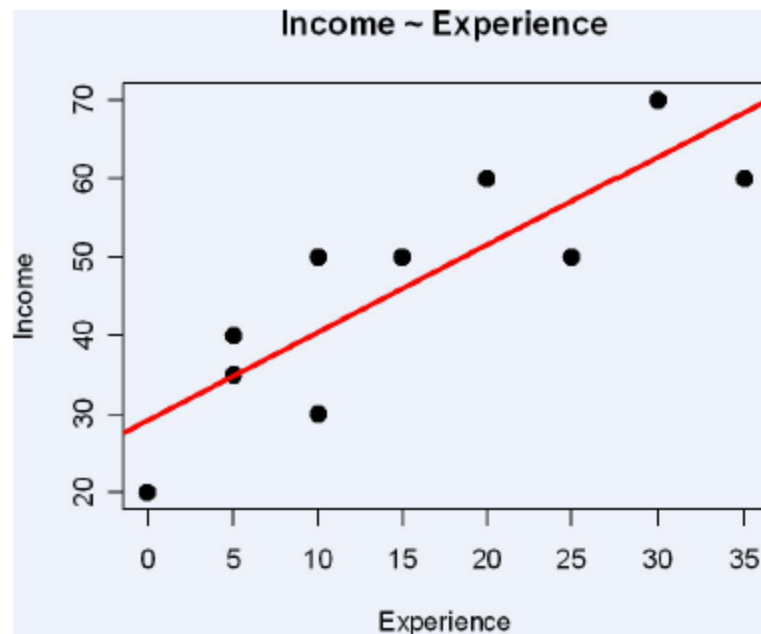
- Different presentation goals appeal to different design principles



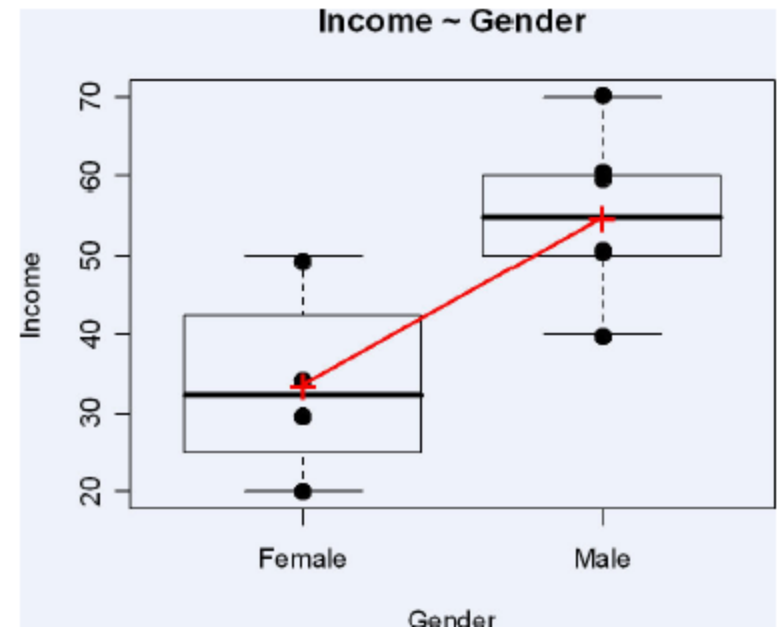
Think: What do I want to communicate? For what purpose?

Graphical methods: Quantitative data

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



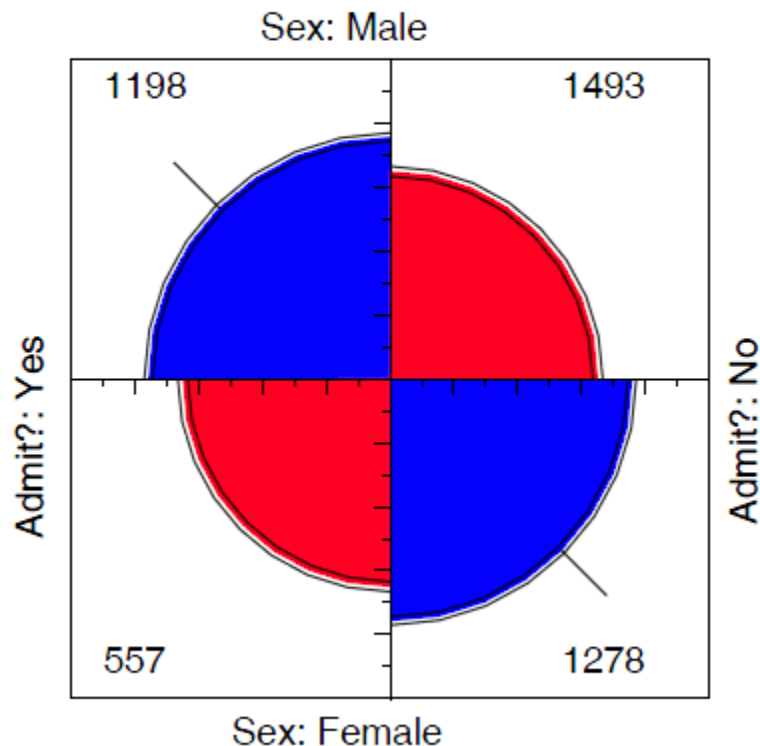
Scatterplot of Income vs.
Experience



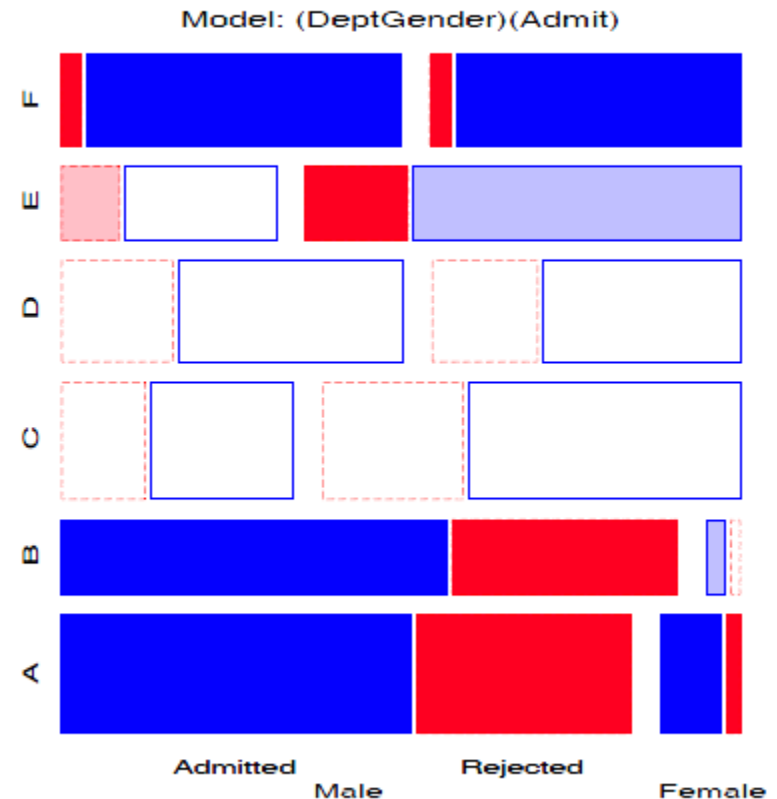
Boxplot of Income by Gender

Graphical methods: Categorical data

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



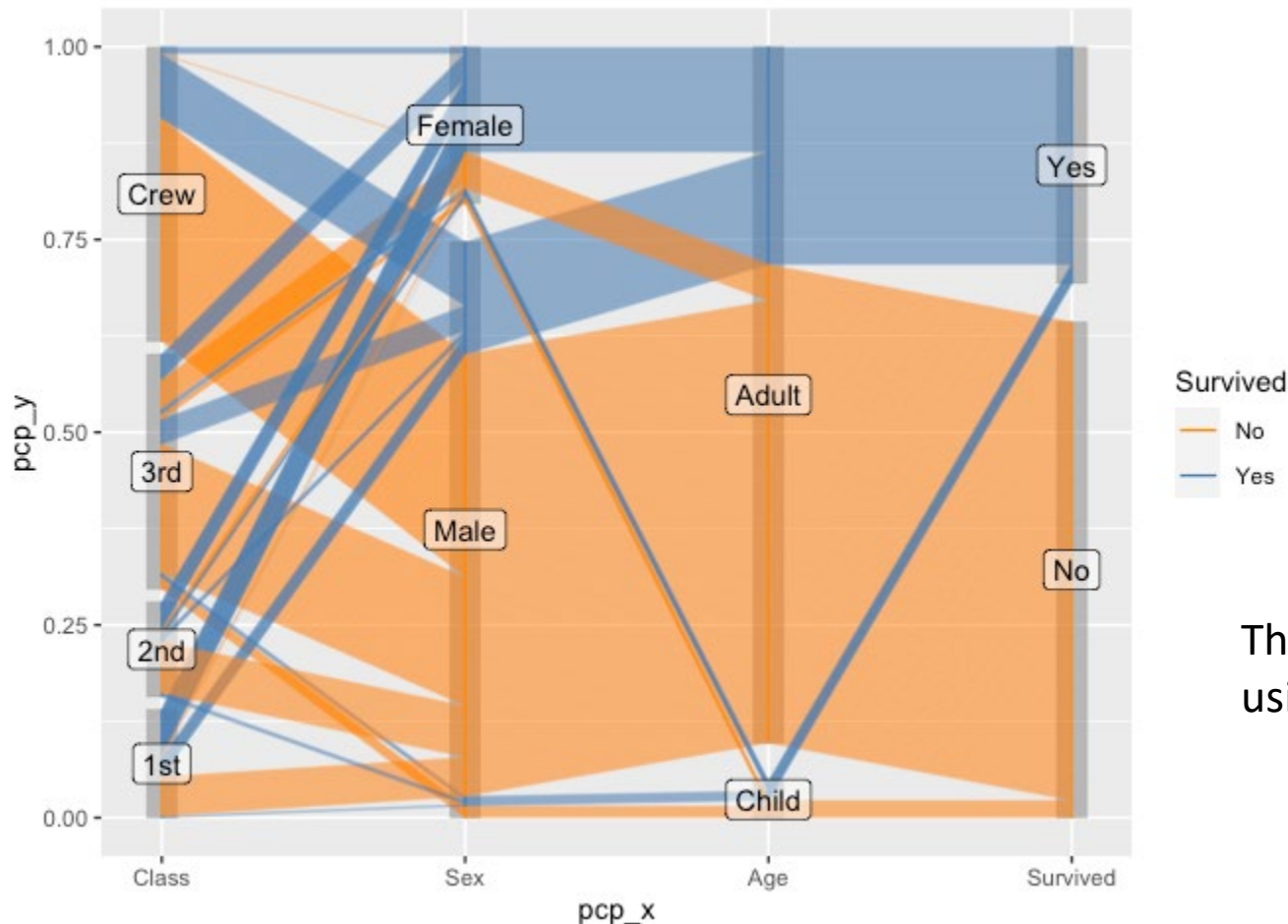
Fourfold display for 2x2 table



Mosaic plot for 3-way table

Categorical data: Parallel coordinates plot

Parallel coordinates plots show multiple variables, each along its' own || axis
The categorical version uses the width of the band to show frequency



Survival on the Titanic, classified by:

- Class,
- Sex,
- Age,
- Survived

Band width ~ Freq

This plot was produced using the **ggpcp** package

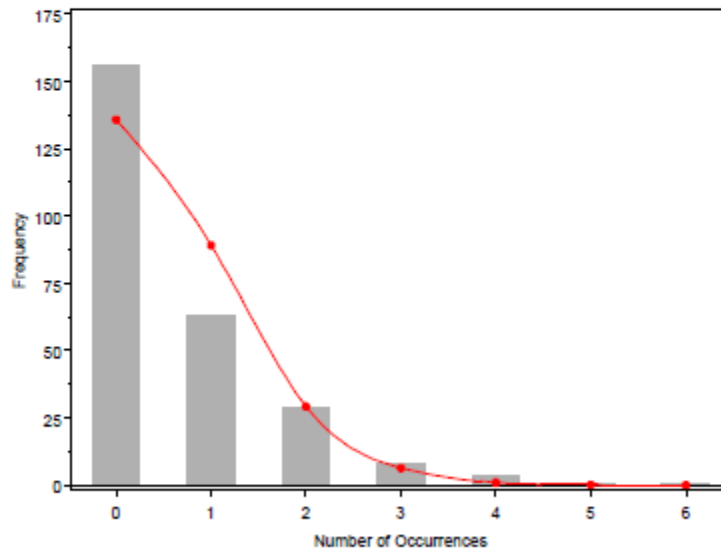
Effective data display

- Make the data stand out
 - Fill the data region (axes, ranges)
 - Use **visually distinct** symbols (shape, color) for different groups
 - Avoid **chart junk**, heavy grid lines that detract from the data
- Facilitate comparison
 - Emphasize the important comparisons **visually**
 - **Side-by-side** easier than in separate panels
 - “data” vs. a “standard” easier against a **horizontal** line
 - Show **uncertainty** where possible
- Effect ordering
 - For **variables** and **unordered** factors, arrange them according to the **effects** to be seen

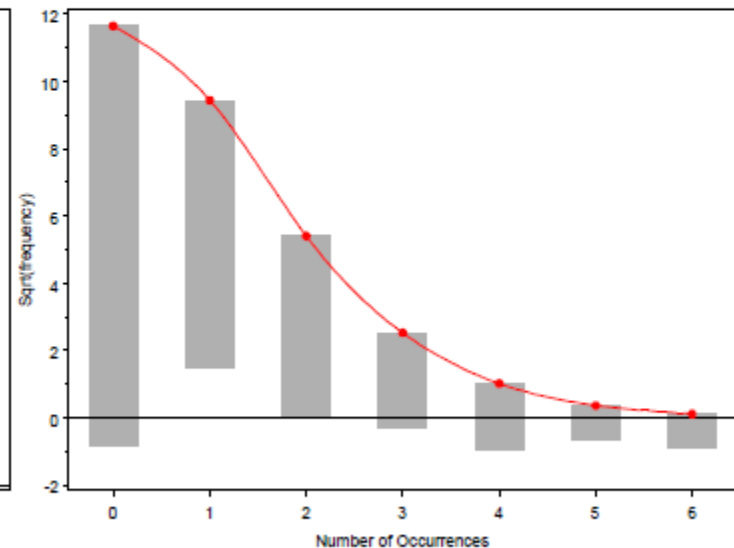
Facilitate comparison

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

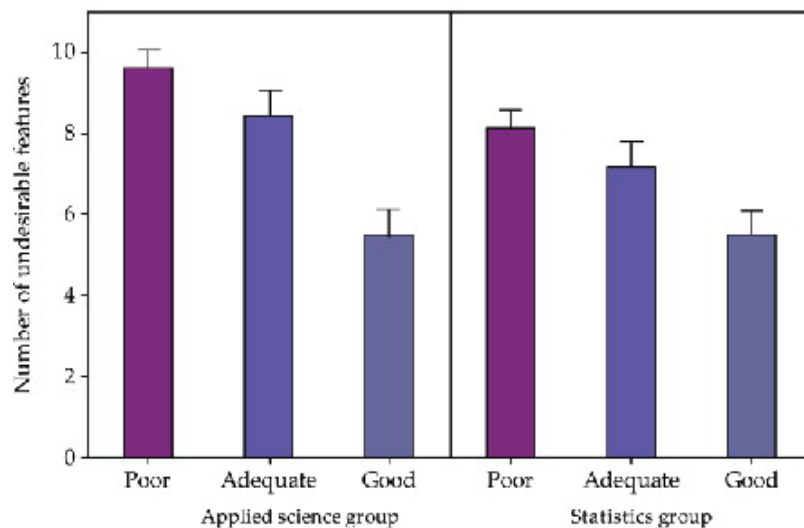


Suspended rootogram

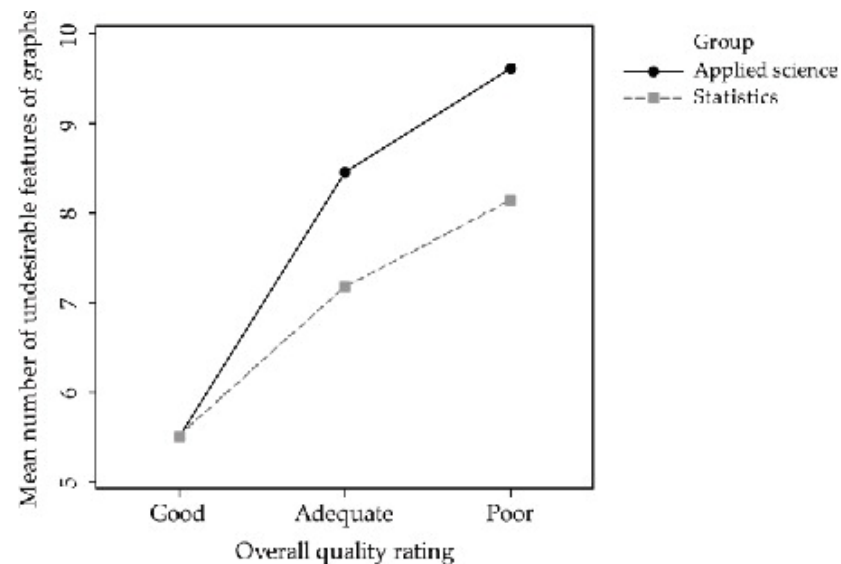
Make comparisons *direct*

- Use **points** not bars (and don't dynamite them with ineffective error bars!)
- Connect similar circumstances to be compared by **lines**
- **Same panel** comparisons easier than different panels

Is there evidence of an interaction here?



???

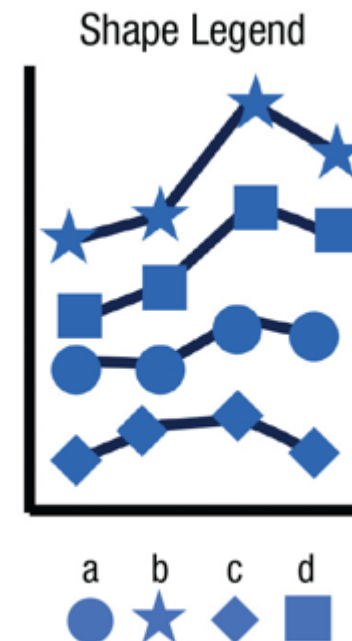
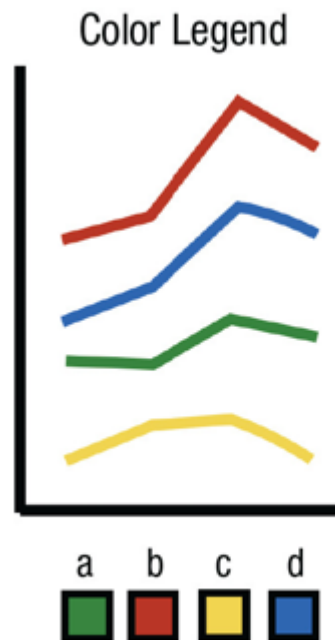
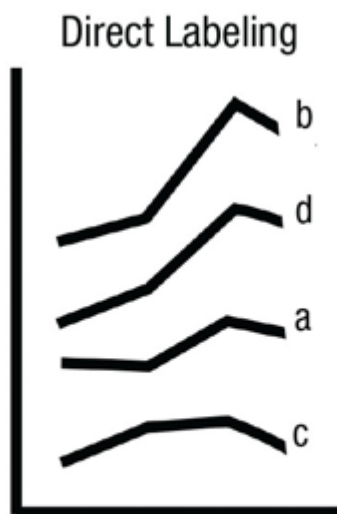


YES

Direct labels vs. legends

Direct labels for points, lines and regions are usually easier and faster than legends

- Give the names of the four groups shown in the line graph at left in top-to-bottom order. (Answer: b, d, a, c.)
- Now do so for the graphs using color or shape legends
- You need to look back and forth between the graph and legend

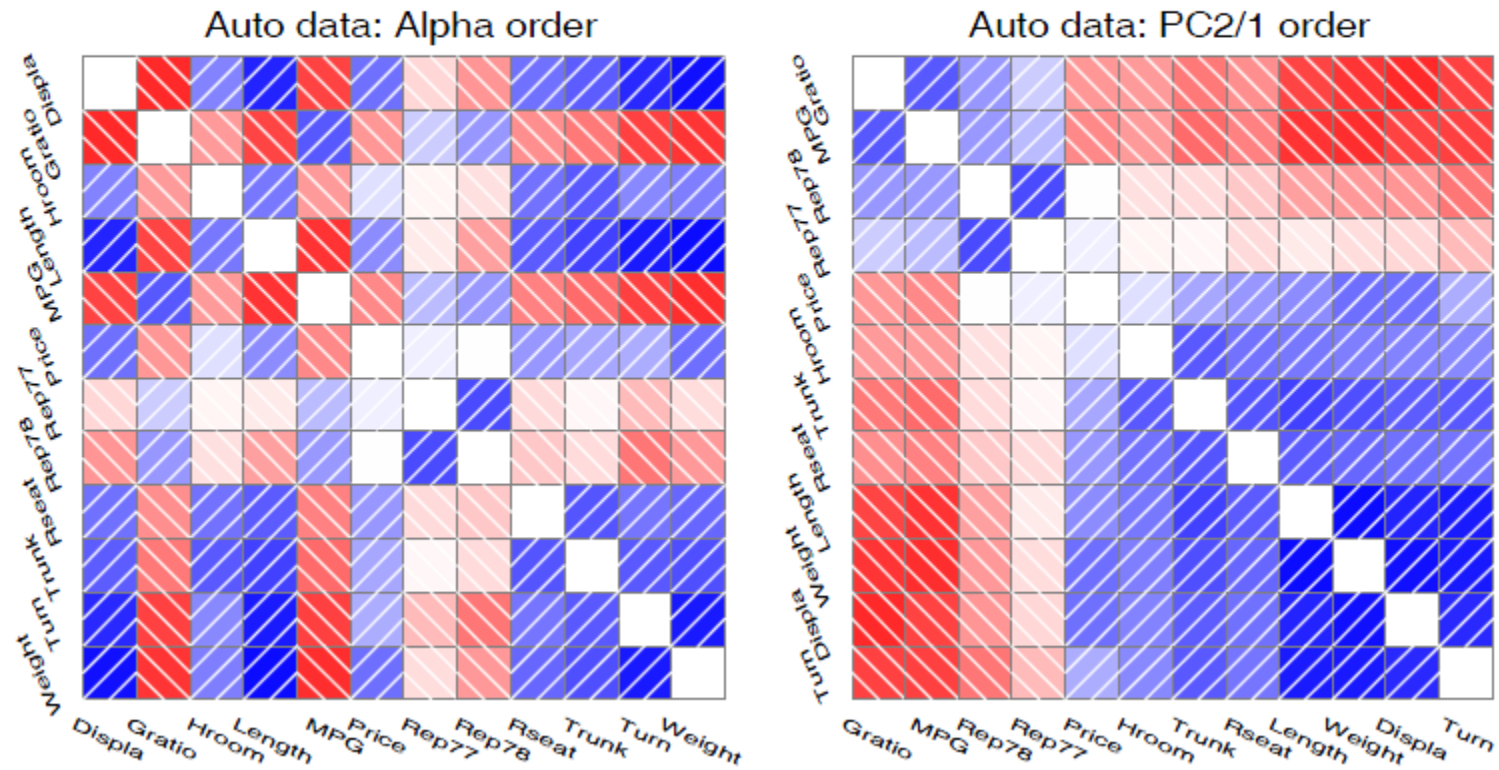


Effect ordering

- Information presentation is always **ordered**
 - in **time** or sequence (a talk or written paper)
 - in **space** (table or graph)
 - Constraints of time & space are dominant– can conceal or reveal the important message
- Effect ordering for data display
 - Sort the data by the **effects to be seen**
 - Order the data to **facilitate the task** at hand
 - **lookup** – find a value
 - **comparison** – which is greater?
 - **detection** – find patterns, trends, anomalies

Effect Ordering: Correlations

- **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: Main effect ordering

- Tables are often presented with rows/cols ordered **alphabetically**
 - good for lookup
 - bad for seeing patterns, trends, anomalies

Table 1: Average Barley Yields (rounded), Means by Site and Variety

Variety	Site 						Mean
	Crookston	Duluth	Grand Rapids	Morris	University Farm	Waseca	
Glabron	32	28	22	32	40	46	33.3
Manchuria	36	26	28	31	27	41	31.5
No. 457	40	28	26	36	35	50	35.8
No. 462	40	25	22	39	31	55	35.4
No. 475	38	30	17	33	27	44	31.8
Peatland	33	32	31	37	30	42	34.2
Svansota	31	24	23	30	31	43	30.4
Trebi	44	32	25	45	33	57	39.4
Velvet	37	24	28	32	33	44	33.1
Wisconsin No. 38	43	30	28	38	39	58	39.4
Mean	37.4	28.0	24.9	35.4	32.7	48.1	34.4

Tabular displays: Main effect ordering

- Better: sort rows/cols by **means/medians**
- Shade cells according to **residual** from additive model

Table 2: Average Barley Yields, sorted by Mean, shaded by residual from the model $\text{Yield} = \text{Variety} + \text{Site}$

Variety	Site						Mean
	Grand Rapids	Duluth	University Farm	Morris	Crookston	Waseca	
Svansota	23	24	31	30	31	43	30.4
Manchuria	28	26	27	31	36	41	31.5
No. 475	17	30	27	33	38	44	31.8
Velvet	28	24	33	32	37	44	33.1
Glabron	22	28	40	32	32	46	33.3
Peatland	31	32	30	37	33	42	34.2
No. 462	22	25	31	39	40	55	35.4
No. 457	26	28	35	36	40	50	35.8
Wisconsin No. 38	28	30	39	38	43	58	39.4
Trebi	25	32	33	45	44	57	39.4
Mean	24.9	28.0	32.7	35.4	37.4	48.1	34.4

Effect ordering: Frequency tables

- 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

Effect ordering: Frequency tables

- 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 **light** 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%

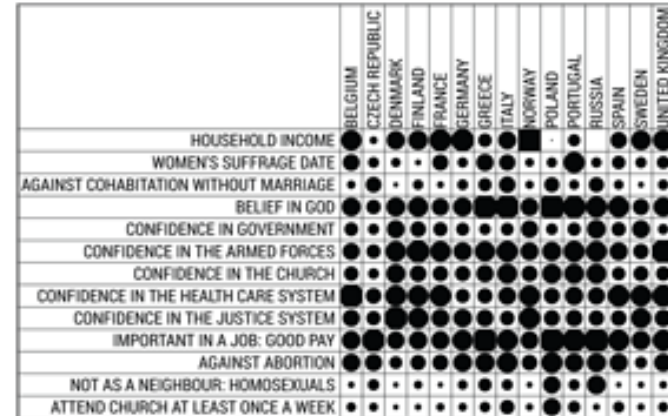
Source: [OECD](#) • [Get the data](#) • Created with [Datawrapper](#)

Bertifier: Turning tables into graphs

a attitudes & attributes

	Belgi	Czech	Dene	Finla	Fran	Gerr	Gredita	Nons	Poln	Porti	Russ	Sc	Swel	United	
Household income	2687	16957	2468	2573	2831	2879	2044	2431	145	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	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 Govern	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 just	50	35	87	73	56	58	50	36	78	44	48	41	42	69	51
Important in a job: god	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

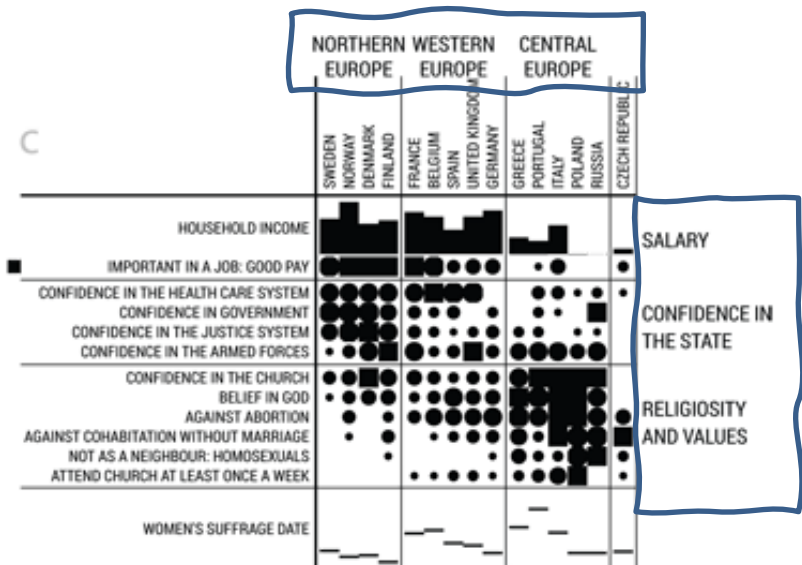
b encode values by size & shape



- (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>



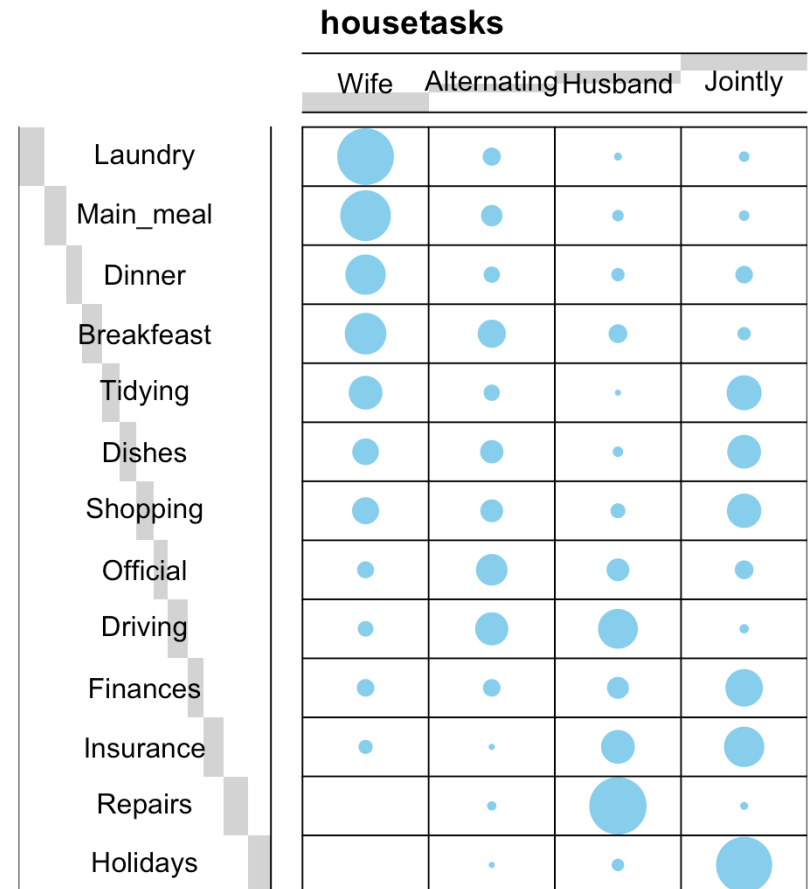
Example: Household tasks

Who does what in households?

Size of symbols in a **balloon plot** shows the frequencies

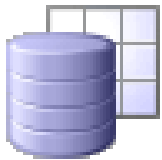
	Who_does_it?			
	Alternating	Husband	Jointly	Wife
Breakfast	36	15	7	82
Dinner	11	7	13	77
Dishes	24	4	53	32
Driving	51	75	3	10
Finances	13	21	66	13
Holidays	1	6	153	0
Insurance	1	53	77	8
Laundry	14	2	4	156
Main_meal	20	5	4	124
Official	46	23	15	12
Repairs	3	160	2	0
Shopping	23	9	55	33
Tidying	11	1	57	53

Rows and columns were permuted to show the relationship more clearly



Data, pictures, models & stories

Goal: Tell a credible story about
some real data problem



data

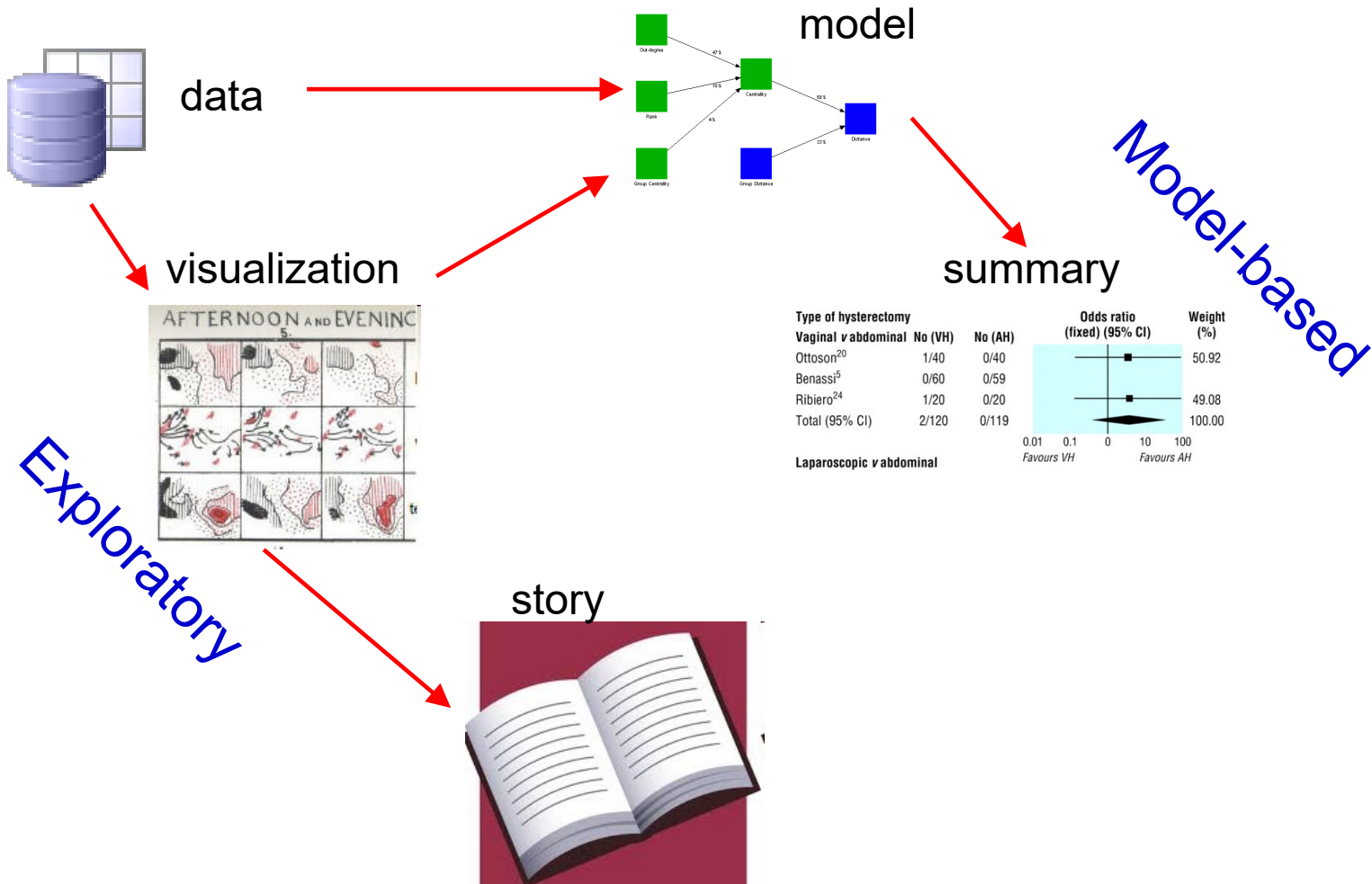
Gender bias
Measles vaccination
Global warming
...

story



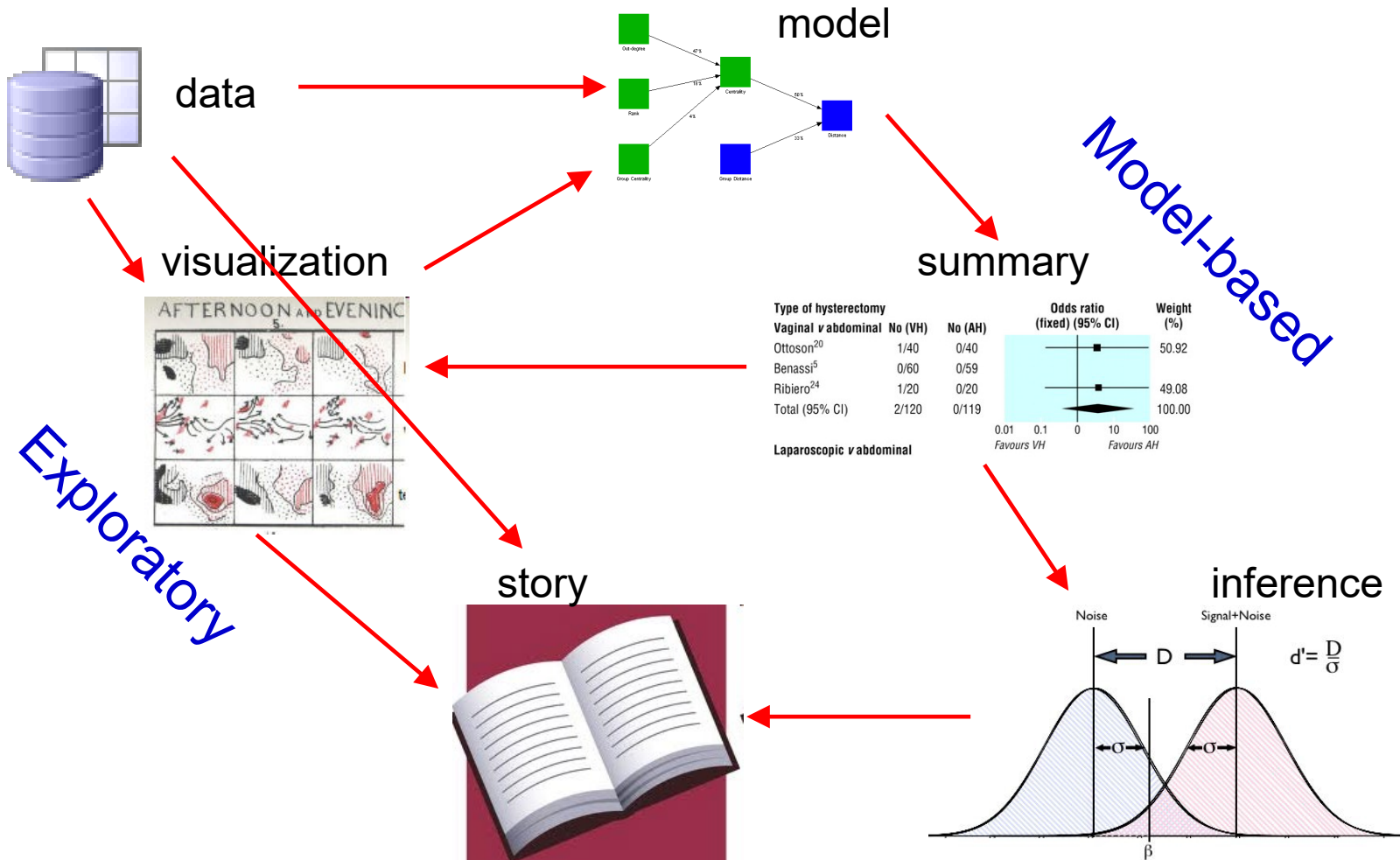
Data, pictures, models & stories

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

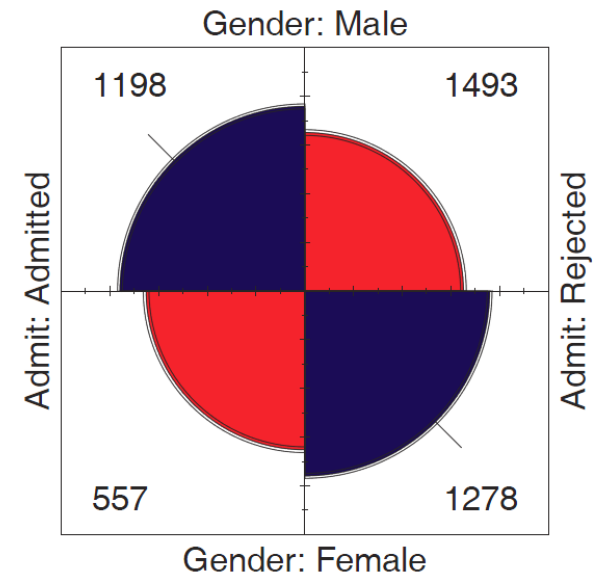
Table: Admissions to Berkeley graduate programs

	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!

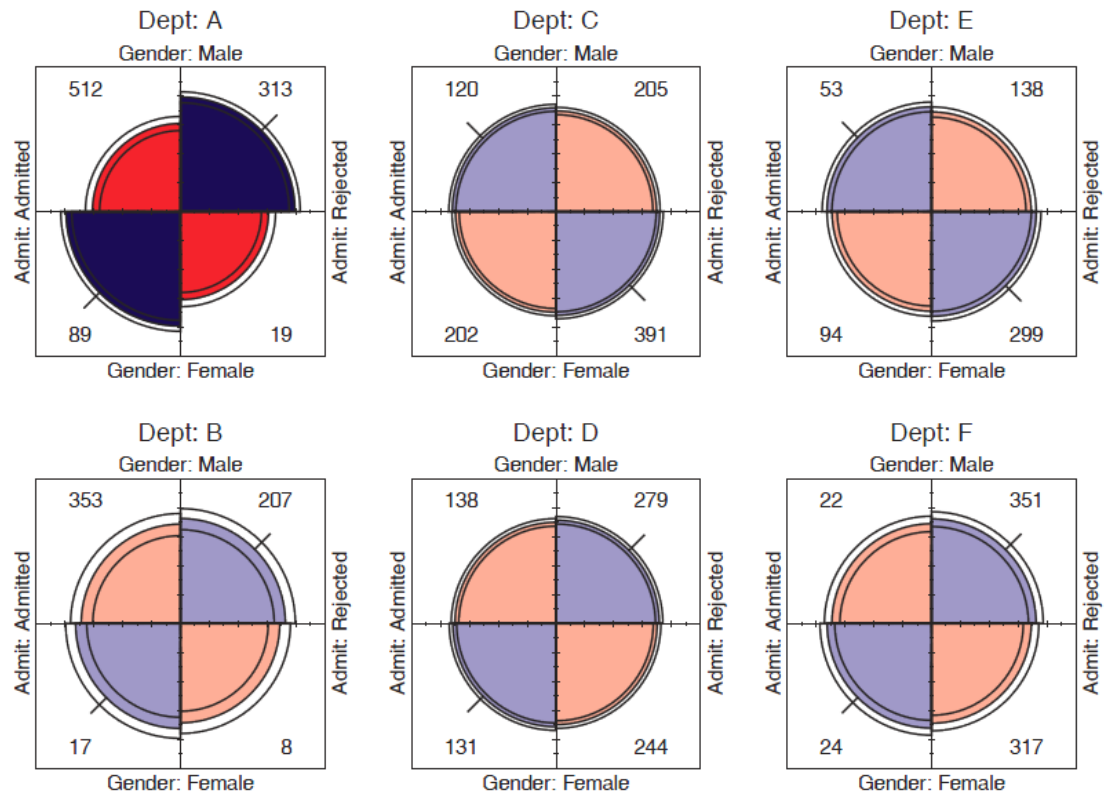
$2 \times 2 \times k$ Stratified tables

The data arose from 6 graduate departments

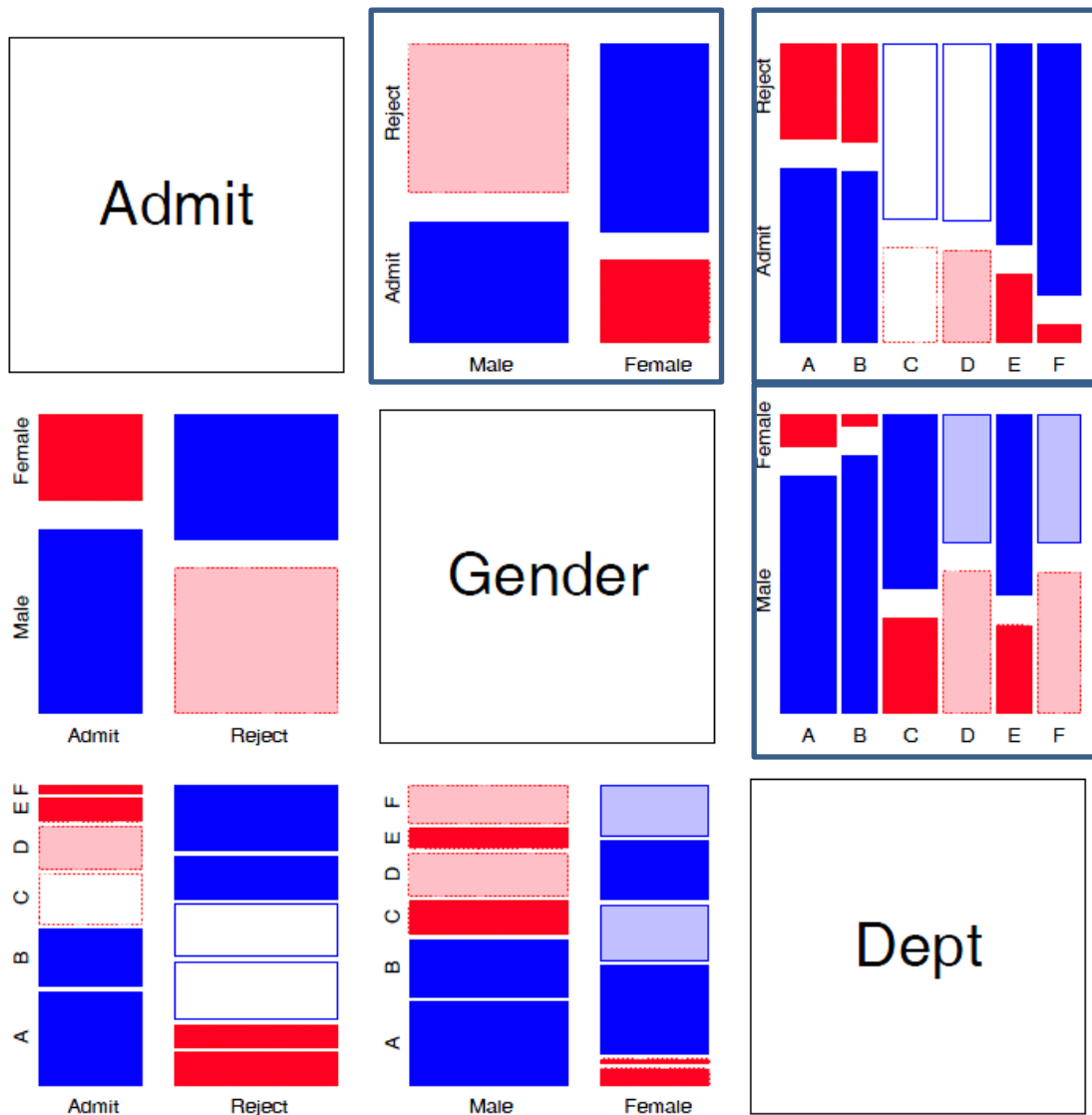
No difference between males & females, except in Dept A where **women more** likely to be admitted!

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**

Measures & models

If the focus is on the association between **gender** and **admission** for each department the **odds ratio**: $\text{odds}(\text{Admit} | \text{Male}) / \text{odds}(\text{Admit} | \text{Female})$ is a good summary



$\text{odds} = \text{Pr}(\text{Admit}) / \text{Pr}(\text{Reject})$

$\text{OR} = \text{odds}(\text{Admit} | \text{M}) / \text{odds}(\text{Admit} | \text{F})$

OR = 1 → M/F equally likely admitted

LOR = log(OR): 0 → M/F equally likely admitted

Std errors & CIs provide individual signif tests

Models provide a comprehensive summary

Now, we can tell the story !

Graphical methods for categorical data

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

Plots: Data, Model, Data+Model

- **Data plots:** well-known. Help to answer:
 - What do the data look like?
 - Are there unusual features? (outliers, non-linear relations)
 - What kinds of summaries would be useful?
- **Model plots**
 - What does the model look like? (plot predicted values)
 - How does the model change when parameters change? (plot competing models)
 - How does the model change when the data is changed? (influence plots)
- **Data+Model plots**
 - How well does model fit the data? (focus on residuals)
 - Does model fit uniformly good/bad, or just in some regions?
 - Model uncertainty: show confidence regions
 - Data support: where is data too thin to make a difference?

Summary

- Categorical data involves some new ideas
 - Discrete variables: `unordered` or `ordered`
 - Counts, frequencies as outcomes
- 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
 - Consider: graphical comparisons, effect order
- Models: Most are \cong natural extensions of `lm()`