How To Communicate Graphically (and How To Not Communicate Graphically)

Ken Rice

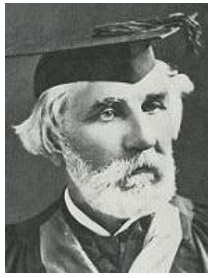
Summer BIOST RA projects

July 27, 2010

Obligatory quotations

One picture is worth 10,000 words

Fred Barnard (in a fake Chinese proverb)
Printer's Ink 1927

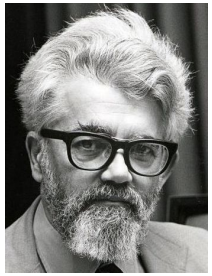


A picture shows me at a glance what it takes dozens of pages of a book to expound.

Ivan Turgenev (Russian Novelist), 1862

Un bon croquis vaut mieux qu'un long discours
(A good sketch is better than a long speech)

Napoleon Bonaparte



1001 words are worth more than a picture

John McCarthy, computer scientist

Why communicate graphically?

This is a poster session;

Your presentation of information must be;

- Comprehensible – easily/quickly
- Captivating (or near the bar)

Some Recommended Reading

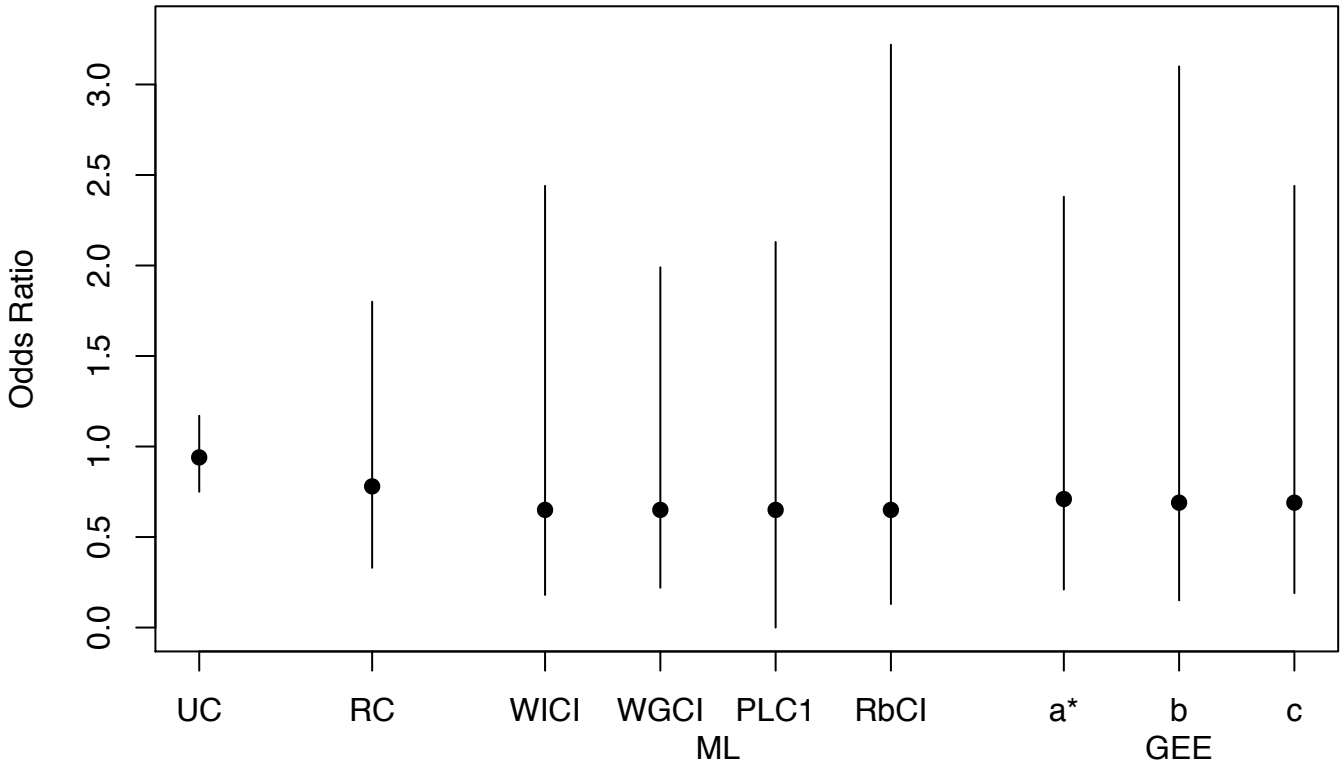
Why not tabulate everything?

Gelman *et al* (*Let's Practice What We Preach: Turning Tables Into Graphs*) compare tables for **lookup** ...

Method	$\widehat{OR}$	95% Interval
UC	0.94	0.75–1.17
RC	0.78	0.33–1.80
ML-WICI	0.65	0.18–2.44
ML-WGCI	0.65	0.22–1.99
ML-PLCI	0.65	0.00–2.13
ML-RbCI	0.65	0.13–3.22
GEEa*-RBCI	0.71	0.21–2.38
GEEb-RbCI	0.69	0.15–3.10
GEEc-RbCI	0.69	0.19–2.44

Some Recommended Reading

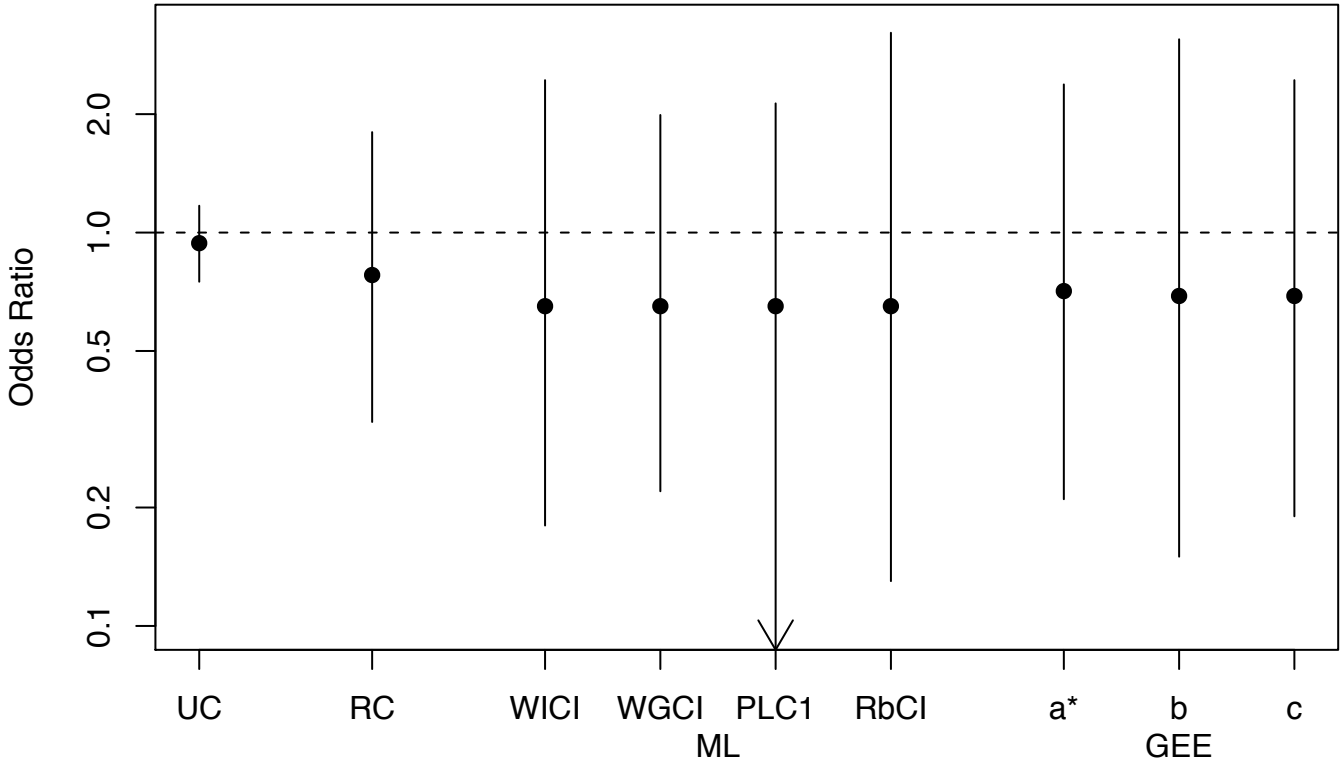
... to graphs, for **comparison**;



- Grouping helps (can also do in tables)
- Comparisons are far easier, faster

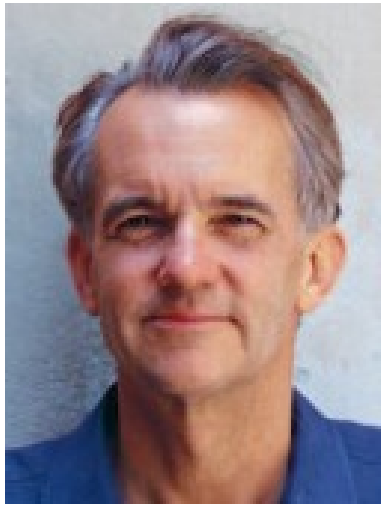
Some Recommended Reading

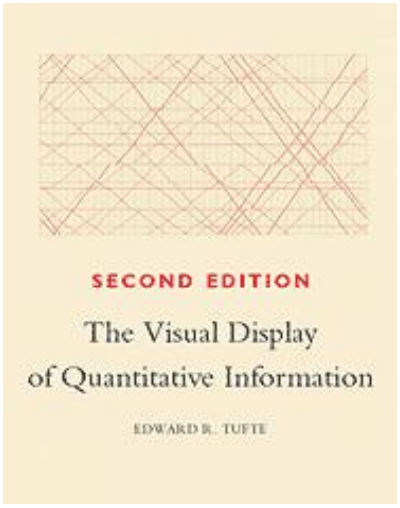
... to graphs, for **comparison**;

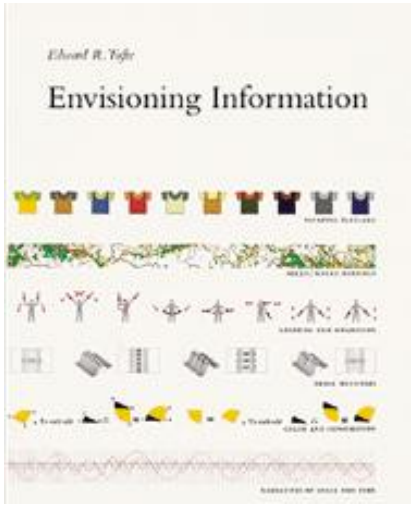


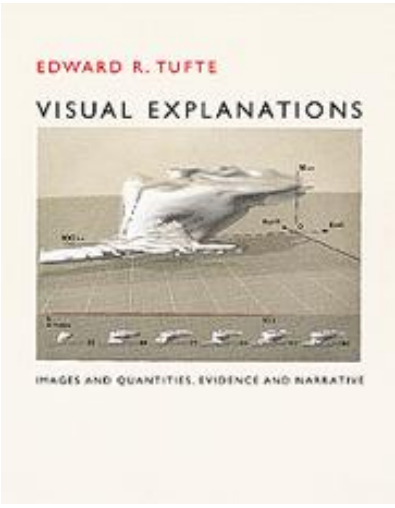
- Log-scale helps compare estimates *and* Std Errs, in this case
- ... but zeroes require extra work

More Recommended Reading







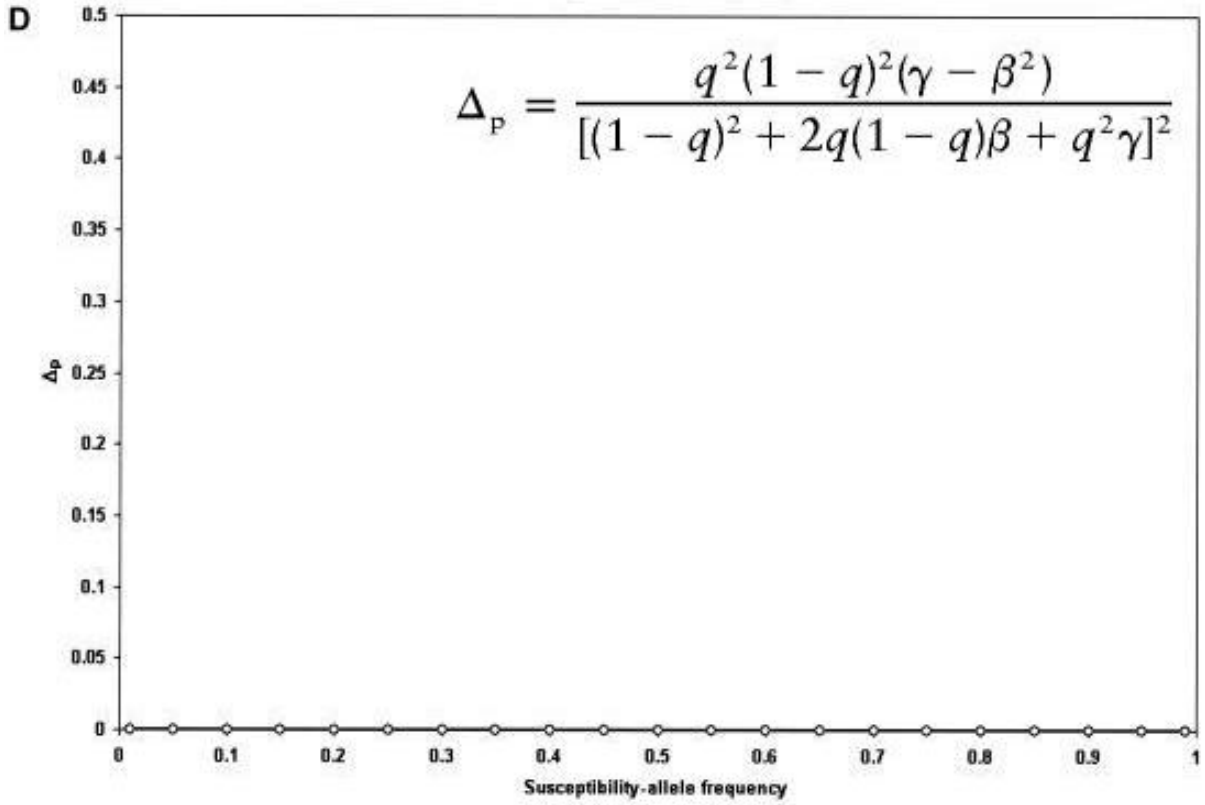


Some principles;

- Serve a reasonably clear purpose
- Show the data
- Avoid distorting what the data have to say
- Encourage the eye to compare different pieces of data

Good graphs: serving a purpose

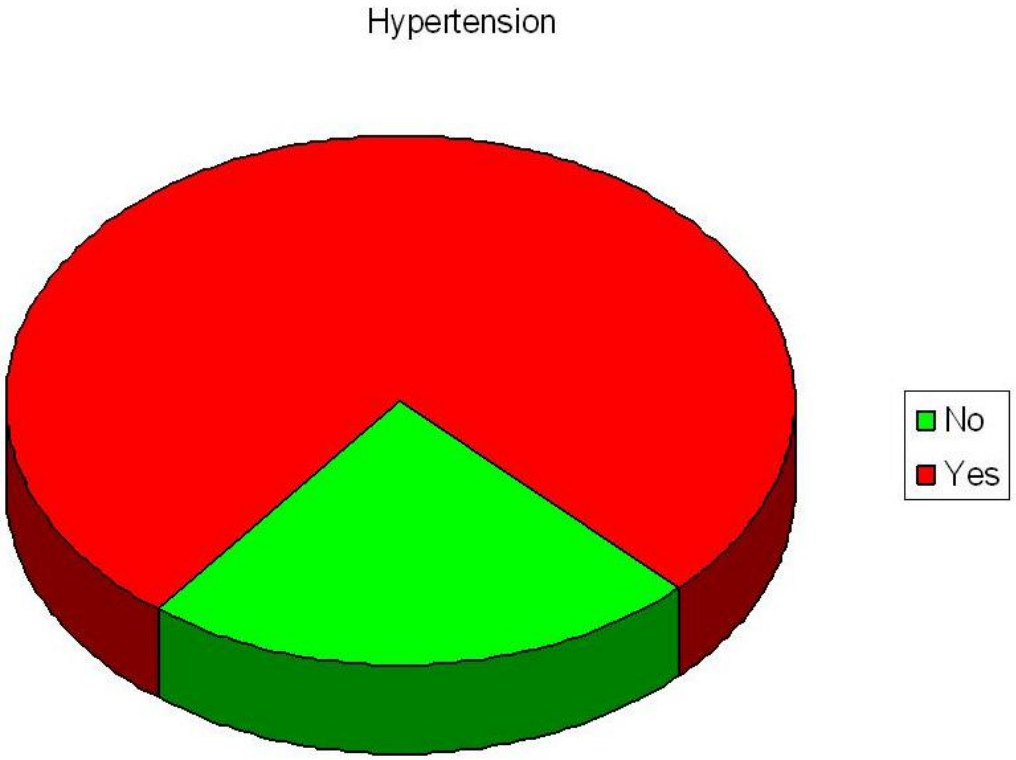
Serve a reasonably clear purpose ?



Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986

Good graphs: serving a purpose

Serve a reasonably clear purpose ?



From a real poster; (American Heart Association); three of these (percentages Yes, Female, Yes & Female) were worth a 2×2 table

Good graphs: serving a purpose

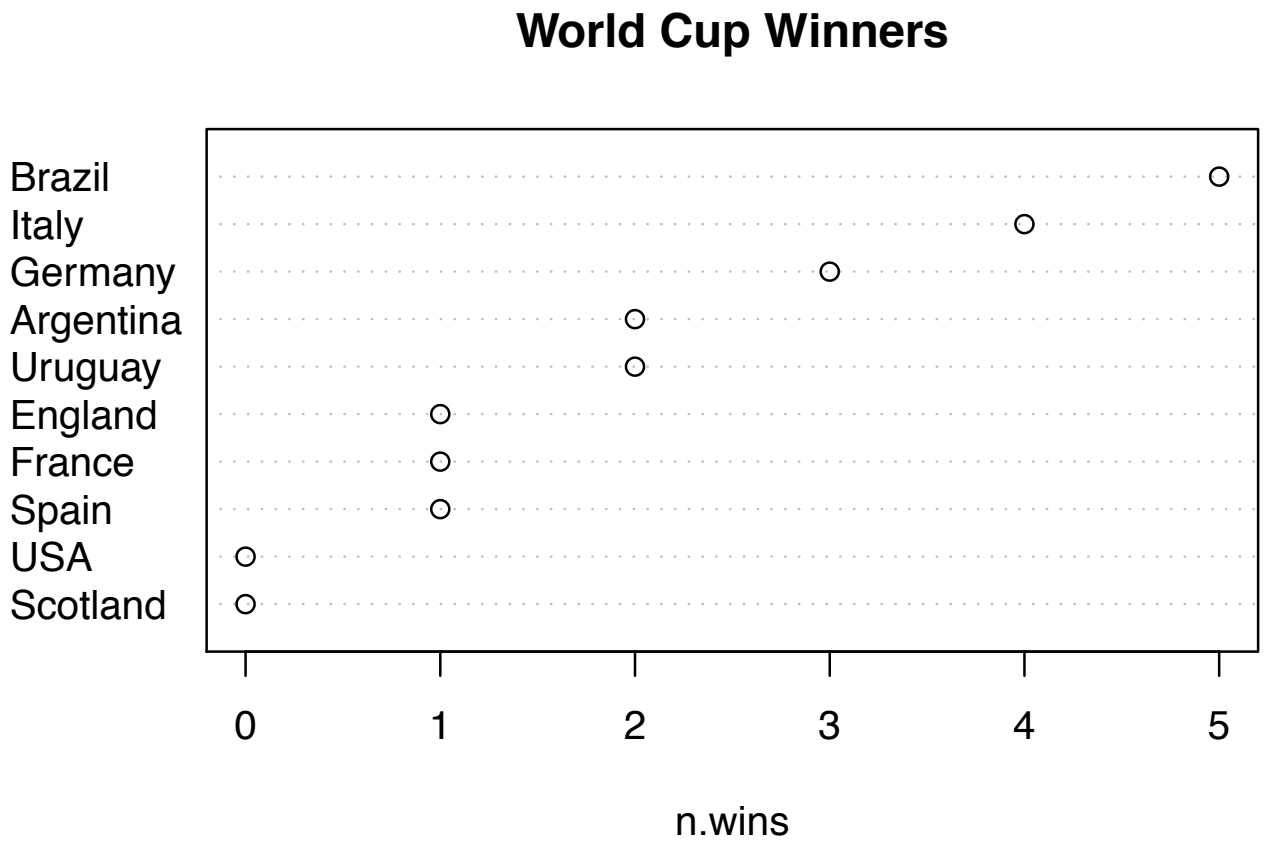
What *is* the graph's purpose?

- Histogram/Dotchart: summarize one-dimensional continuous data
- Barchart: compare one-dimensional categorical data
- Scatterplots: show association of continuous Y and X (or lack of association)
- Mosaic plots: show association of categorical Y and X (or lack of association)
- Boxplots: show association of continuous Y and categorical X (or lack of association)
- QQ plots: show two continuous distributions; talk about the shift, spread, heavy tails, light tails etc

Note the close connections to 'Table 1', t-tests, regression etc

Good graphs: serving a purpose

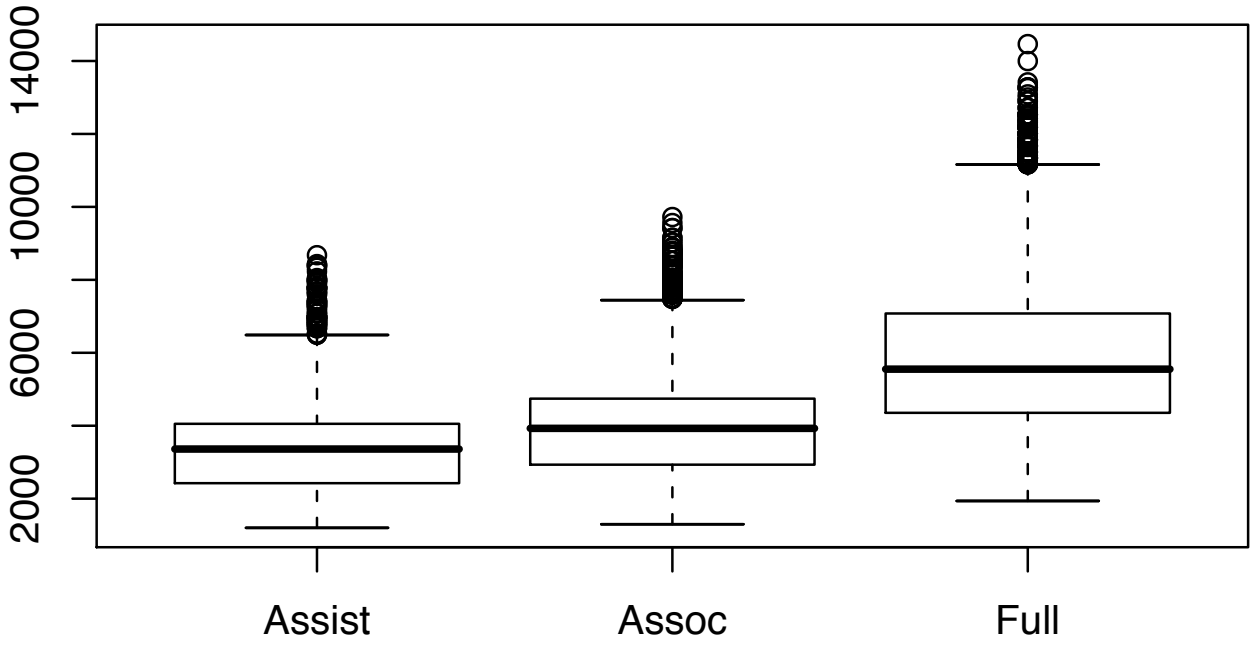
A dotchart – see `dotchart()`, `stripchart()`



– can be easily grouped; good for small samples

Good graphs: serving a purpose

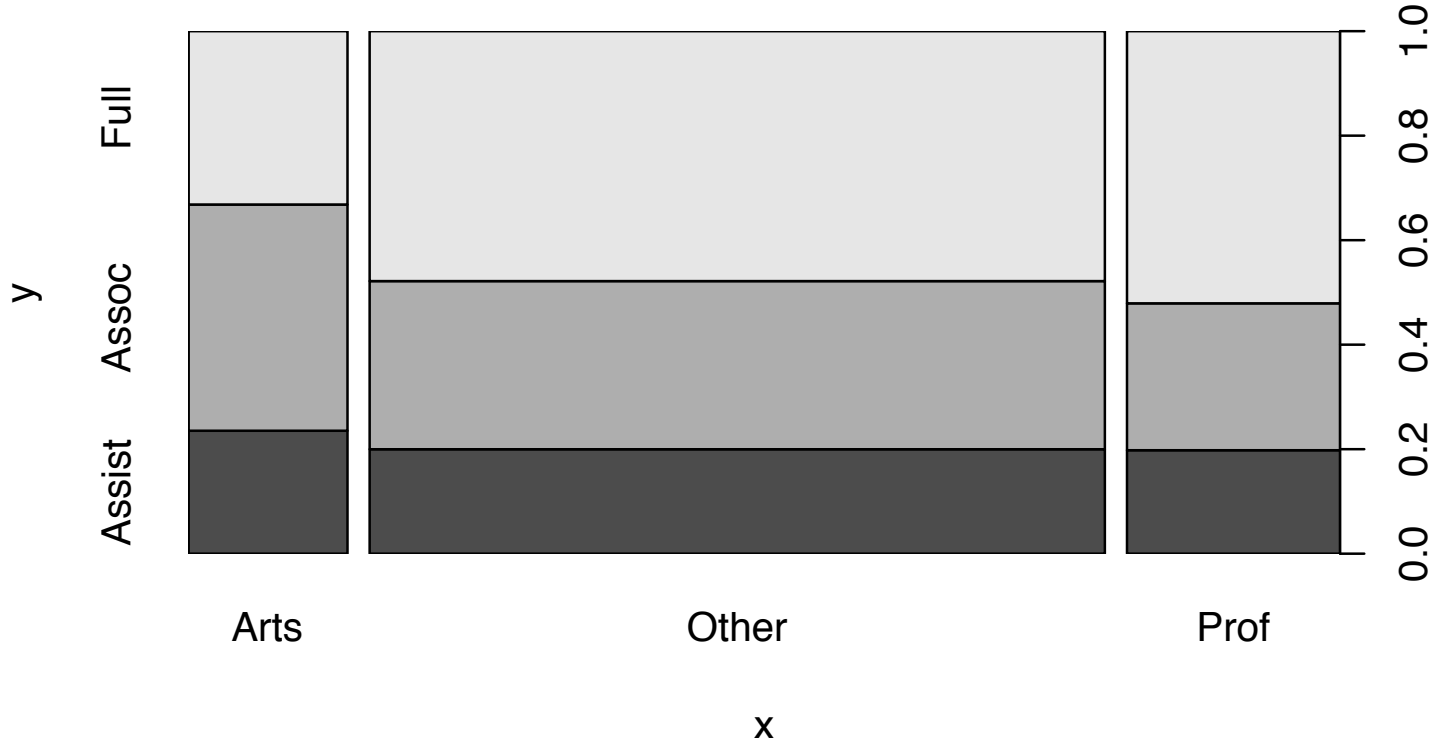
More examples; (all with `plot(y~x)`)



With only 2 groups, a QQ plot contains more information – but less familiar, and more complex

Good graphs: serving a purpose

A mosaic plot; (areas indicate counts)



... have to condition on one variable (just like regression)

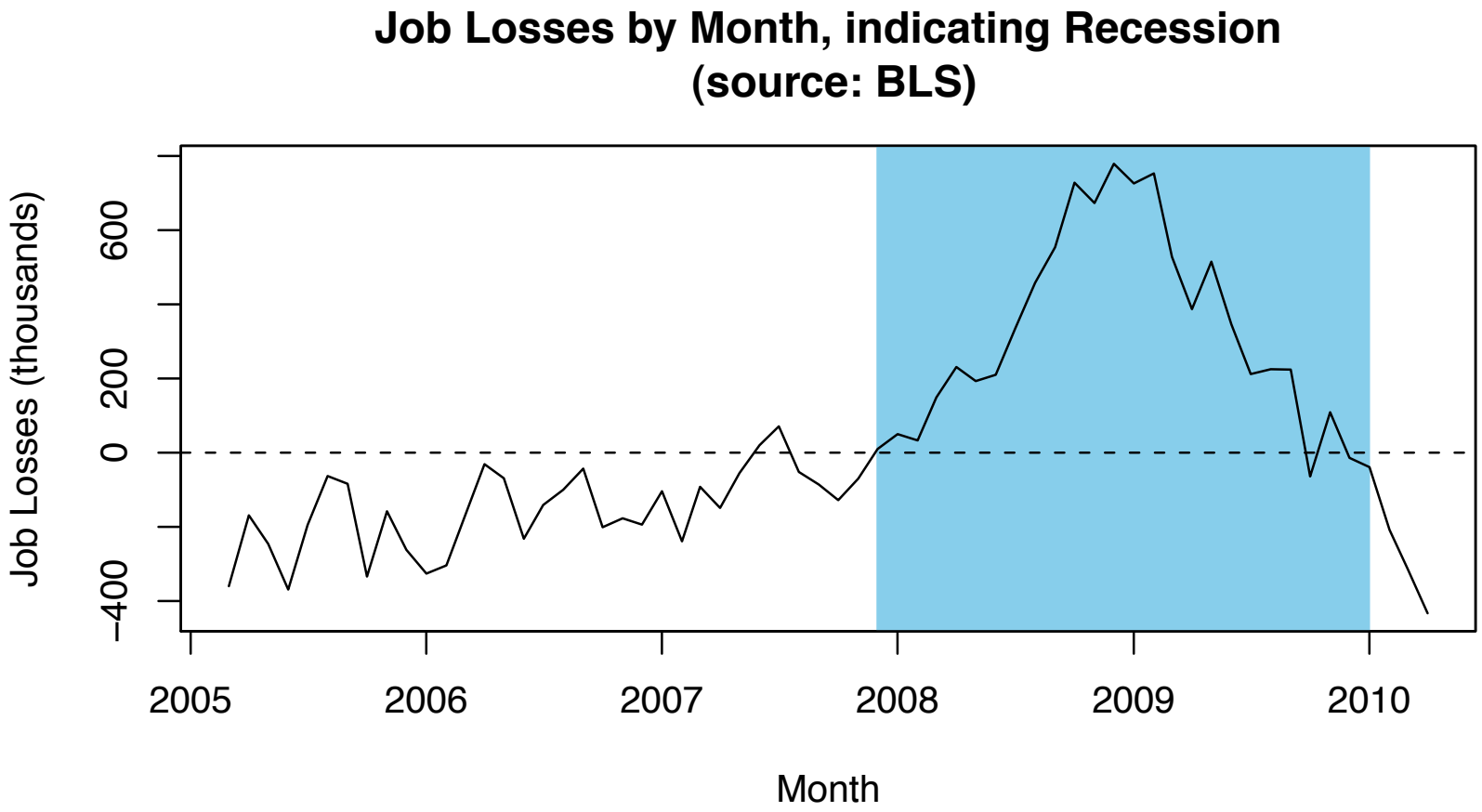
Good graphs: not distorting data

"Fair and Balanced" Fox News: "We Report. You Decide"

– presented on June 28, 2010. What's wrong with it?

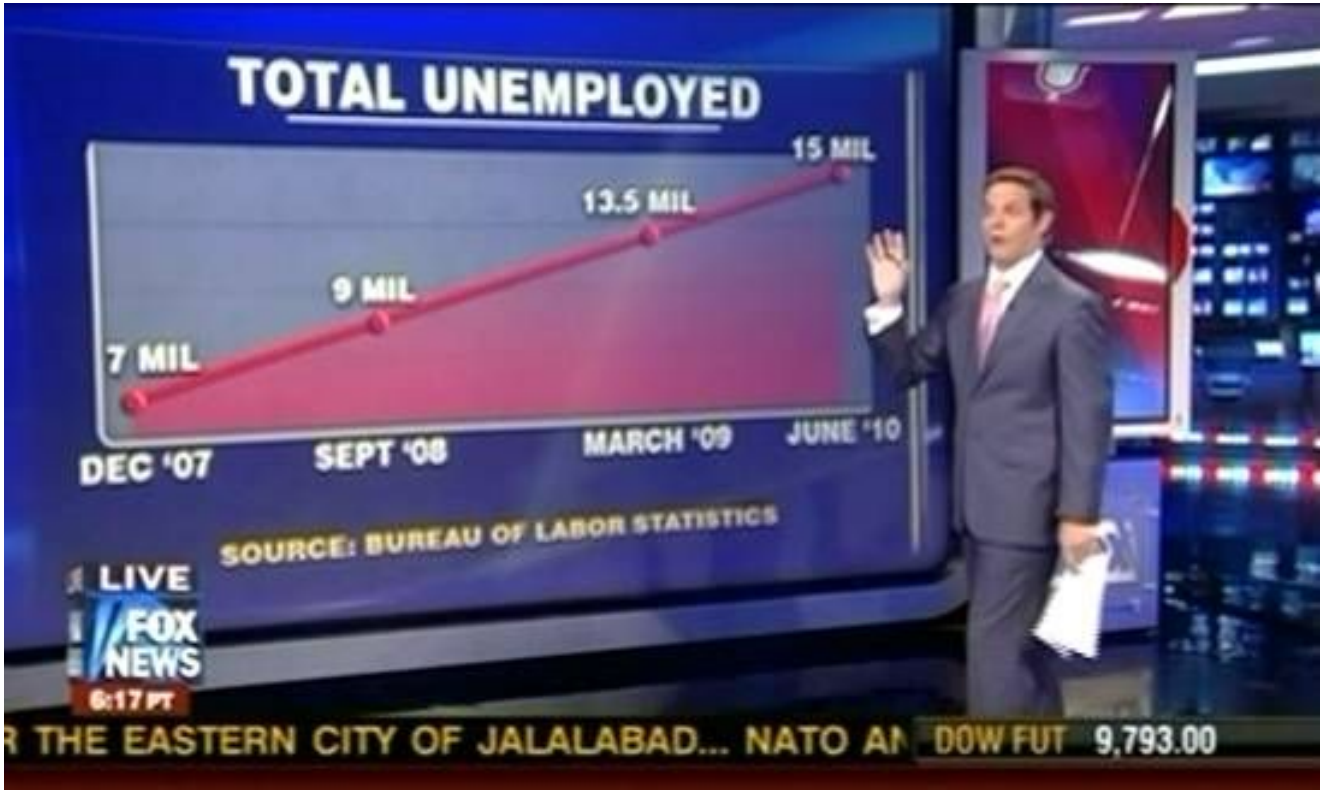
Good graphs: not distorting data

Here's the original data;



Good graphs: not distorting data

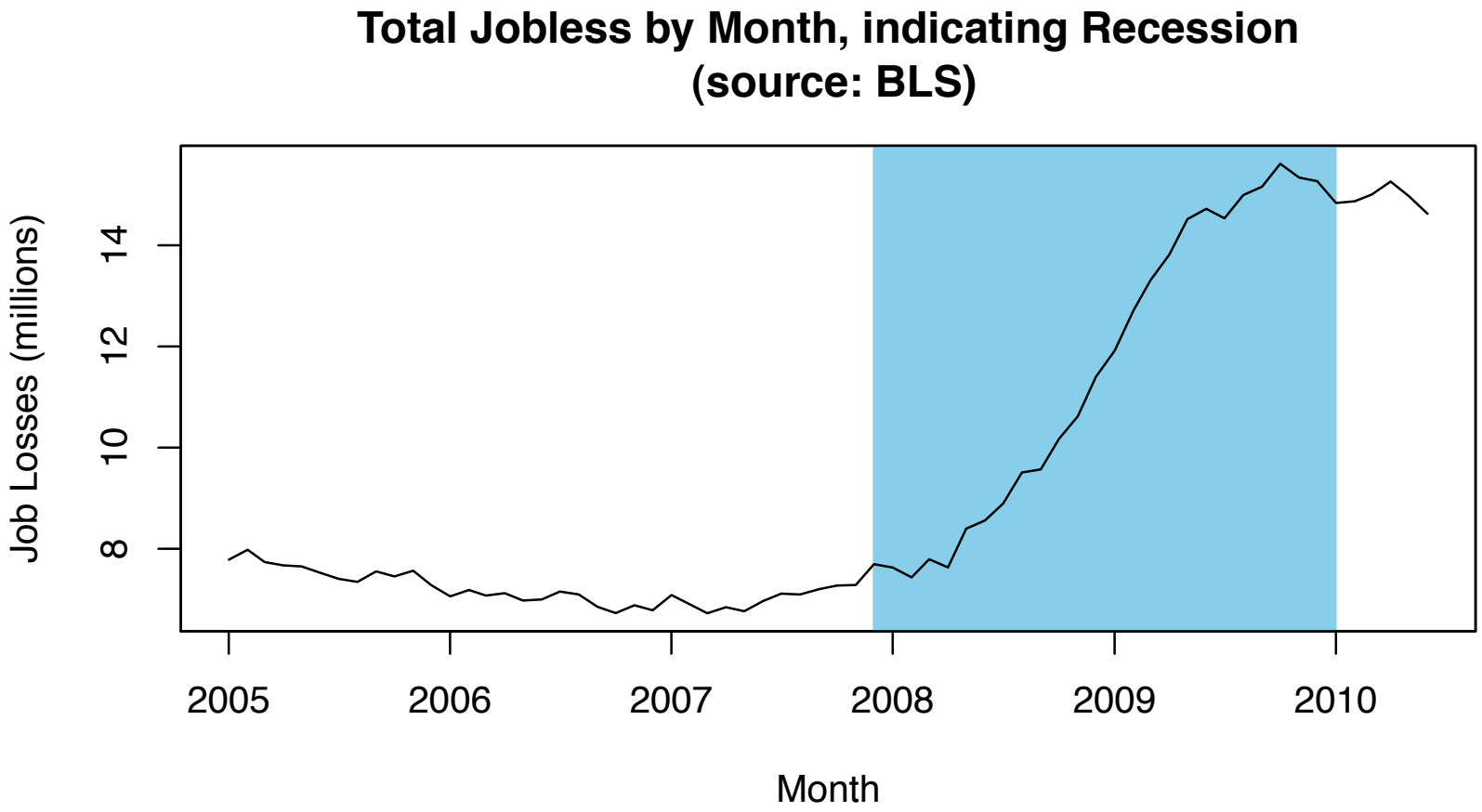
They did correct the title;



... but are *still* distorting the BLS's hard work;

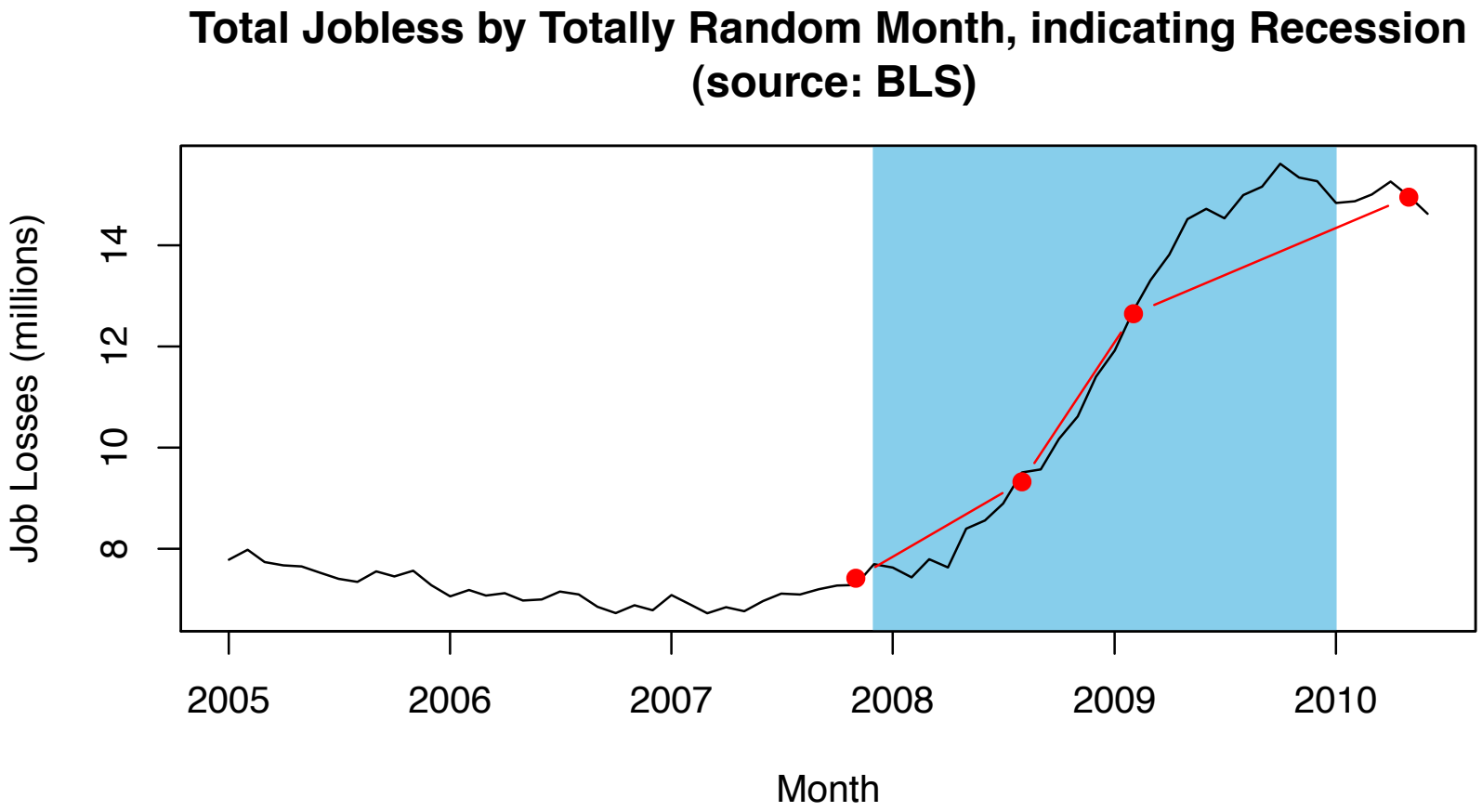
Good graphs: not distorting data

The full original data...



Good graphs: not distorting data

... and what Fox chose to show of it;



Good graphs: not distorting data

Graphics *reveal* data – or what the data tell us. Like good statistical analysis, good graphs help your reader accurately assess whether;

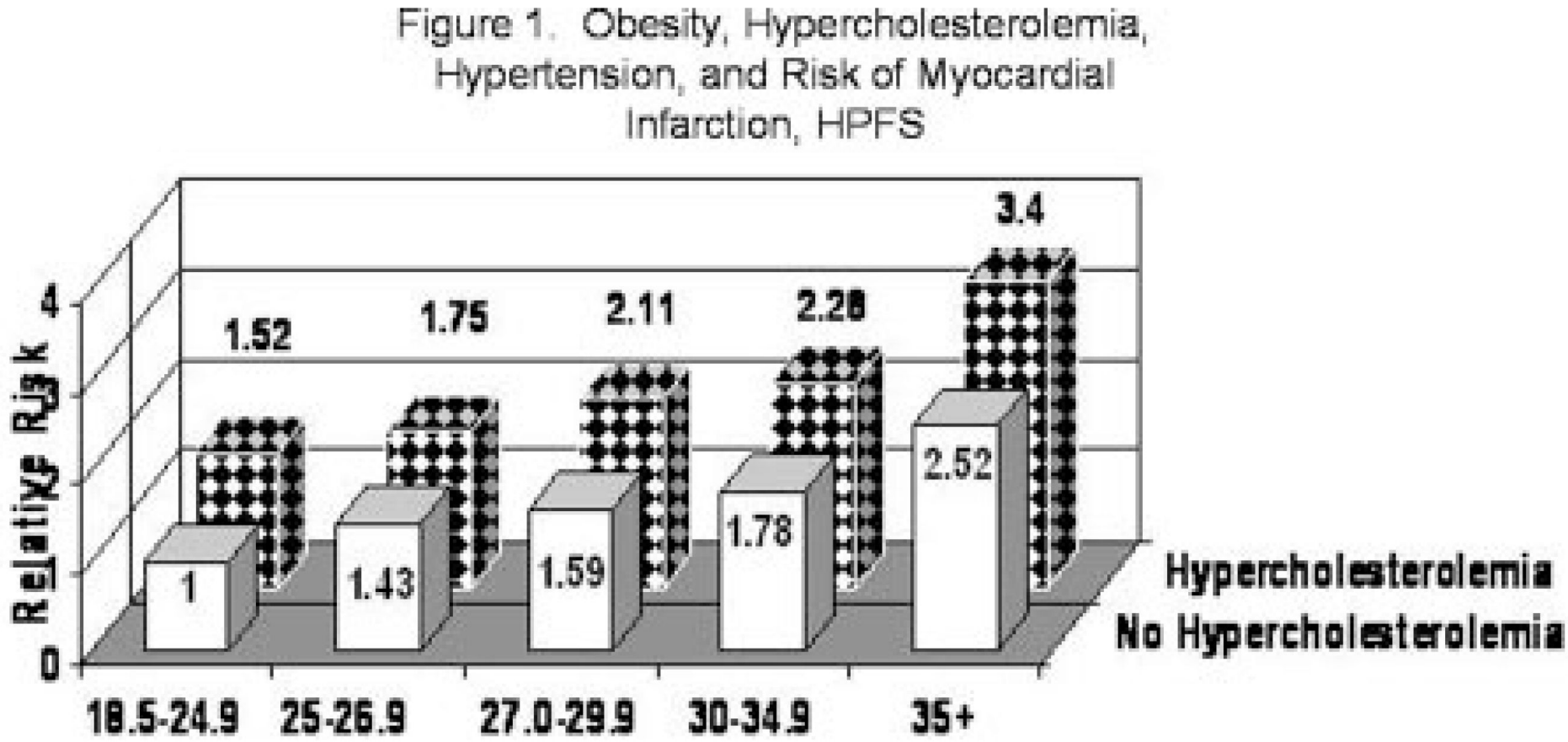
- The effect is there
- The effect is not there
- The data are so uninformative that no-one can tell

Fox were cherry-picking results to show a non-existent effect – see also Huff 1954, *How To Lie With Statistics*.

A more *honest* mistake is to use an uninformative graph, that does not 'show the data'.

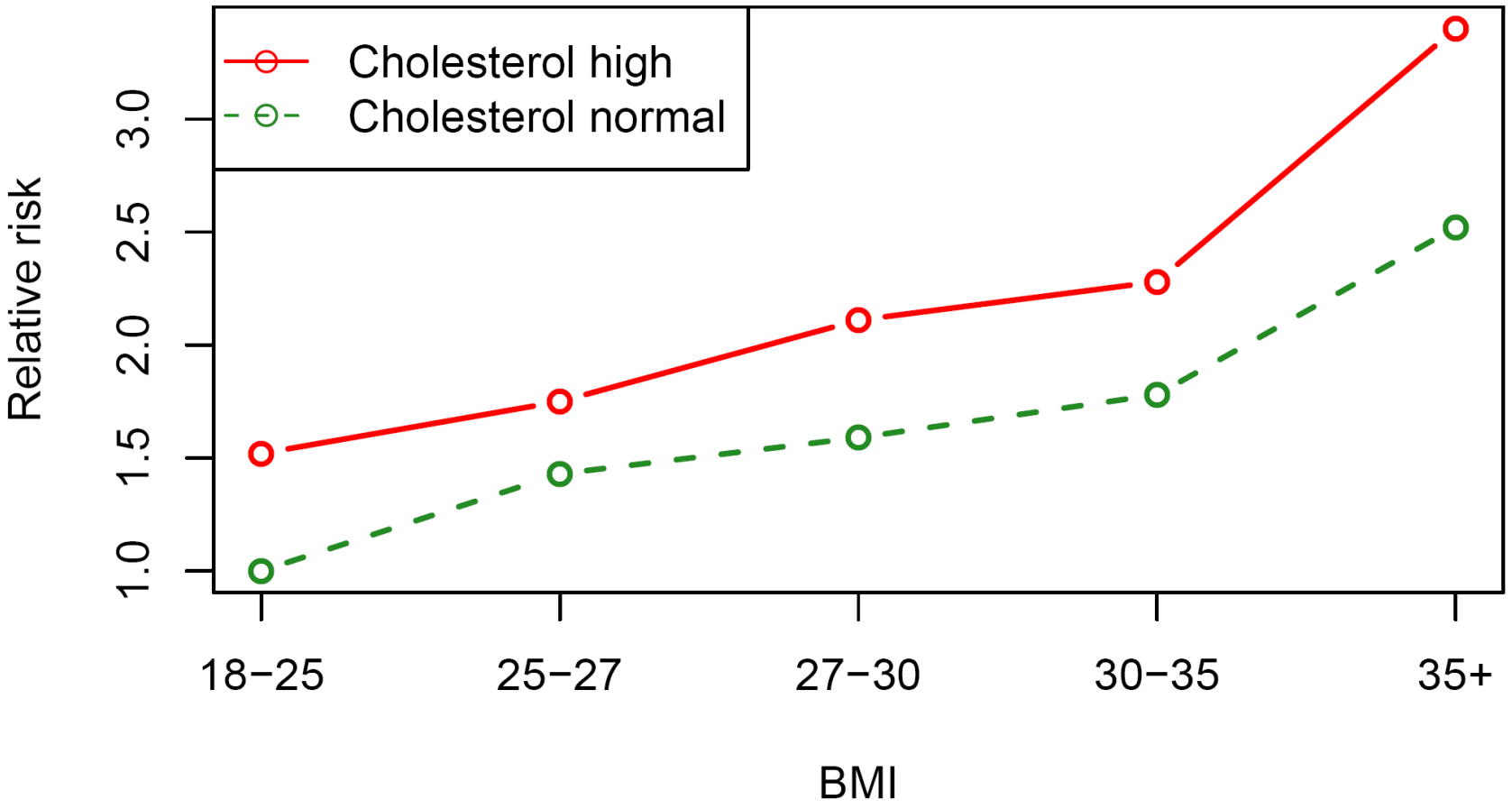
Good graphs: being informative

More from the AHA Epi conference;



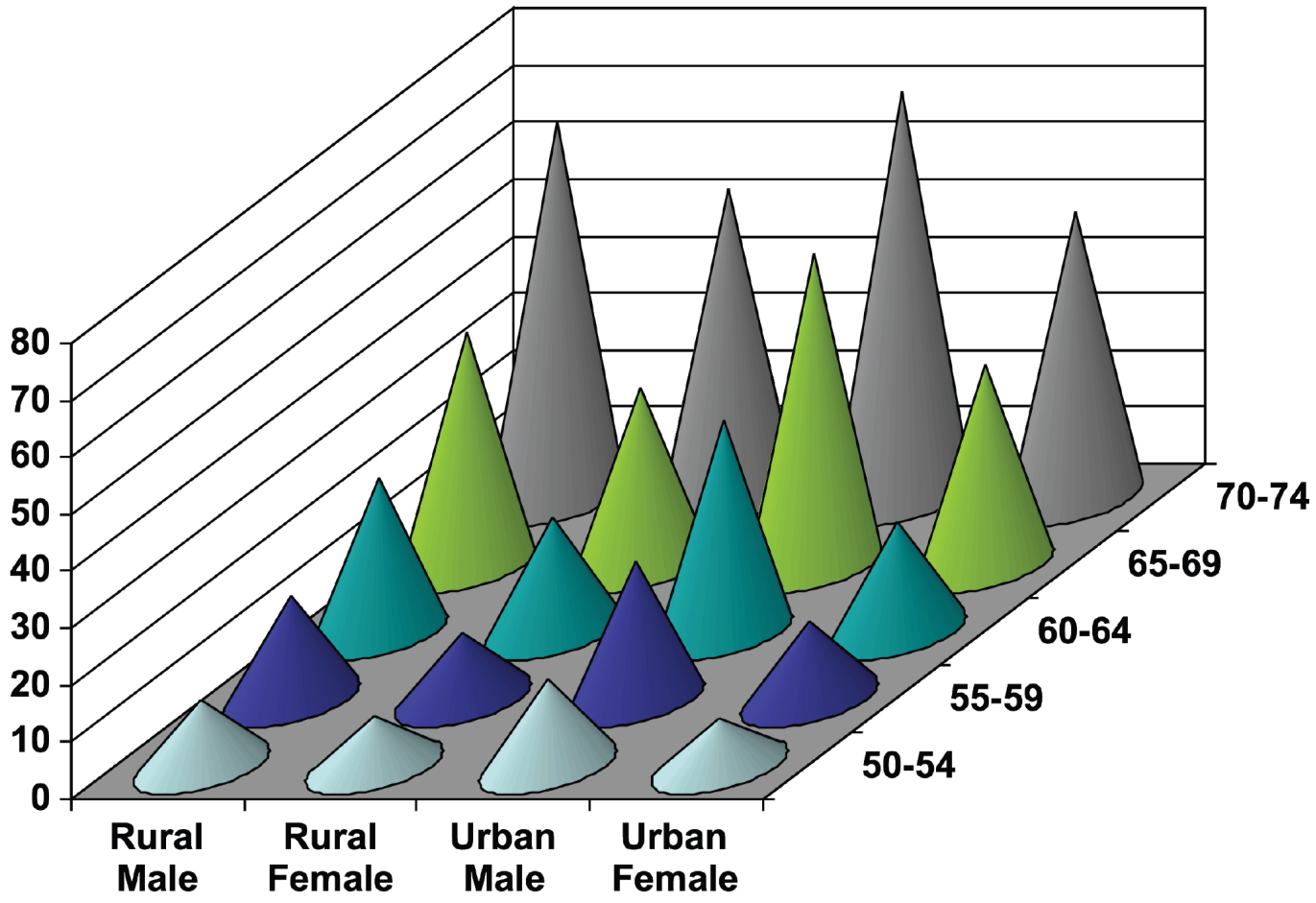
Good graphs: being informative

Re-imagined; (confidence intervals would help too)



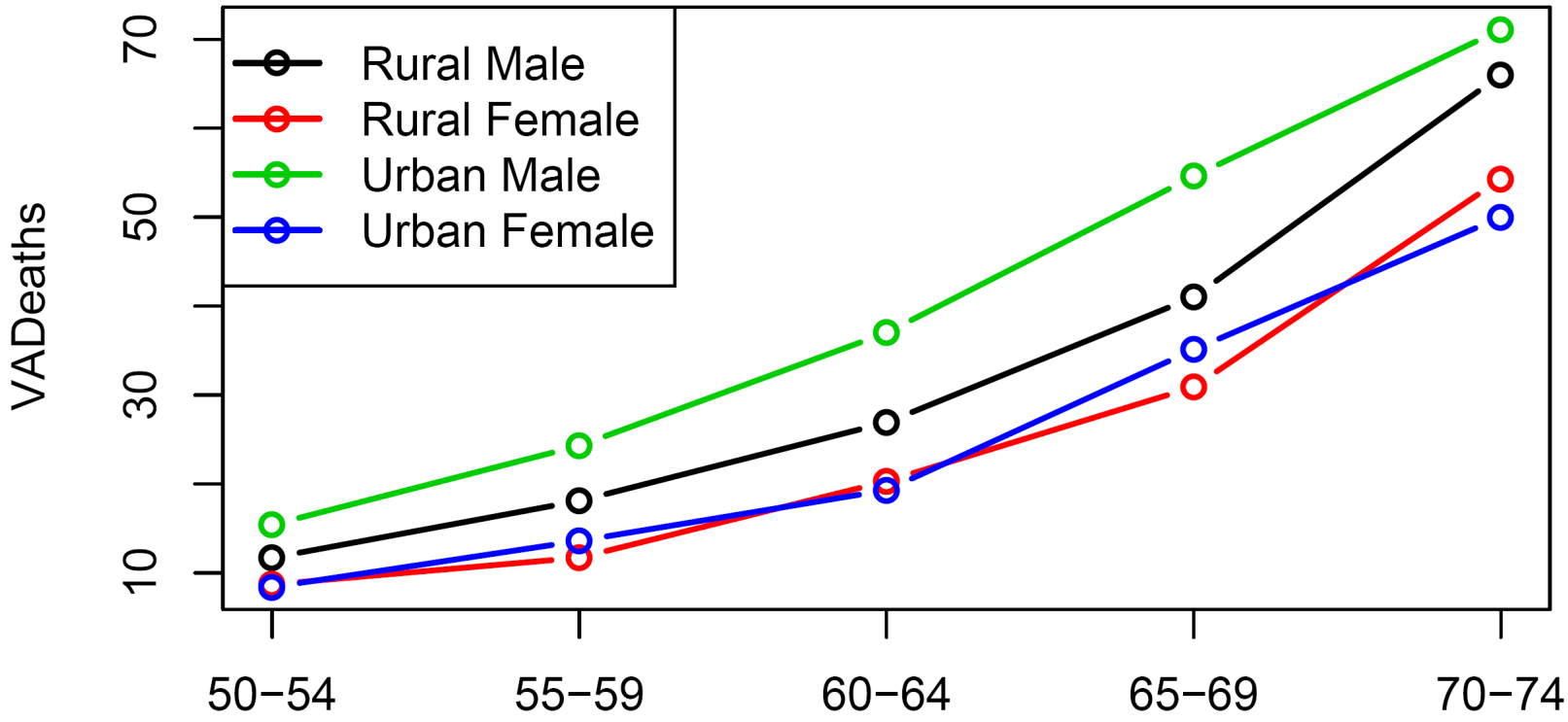
Good graphs: being informative

The ‘Bed of Nails’, from an AHA poster;



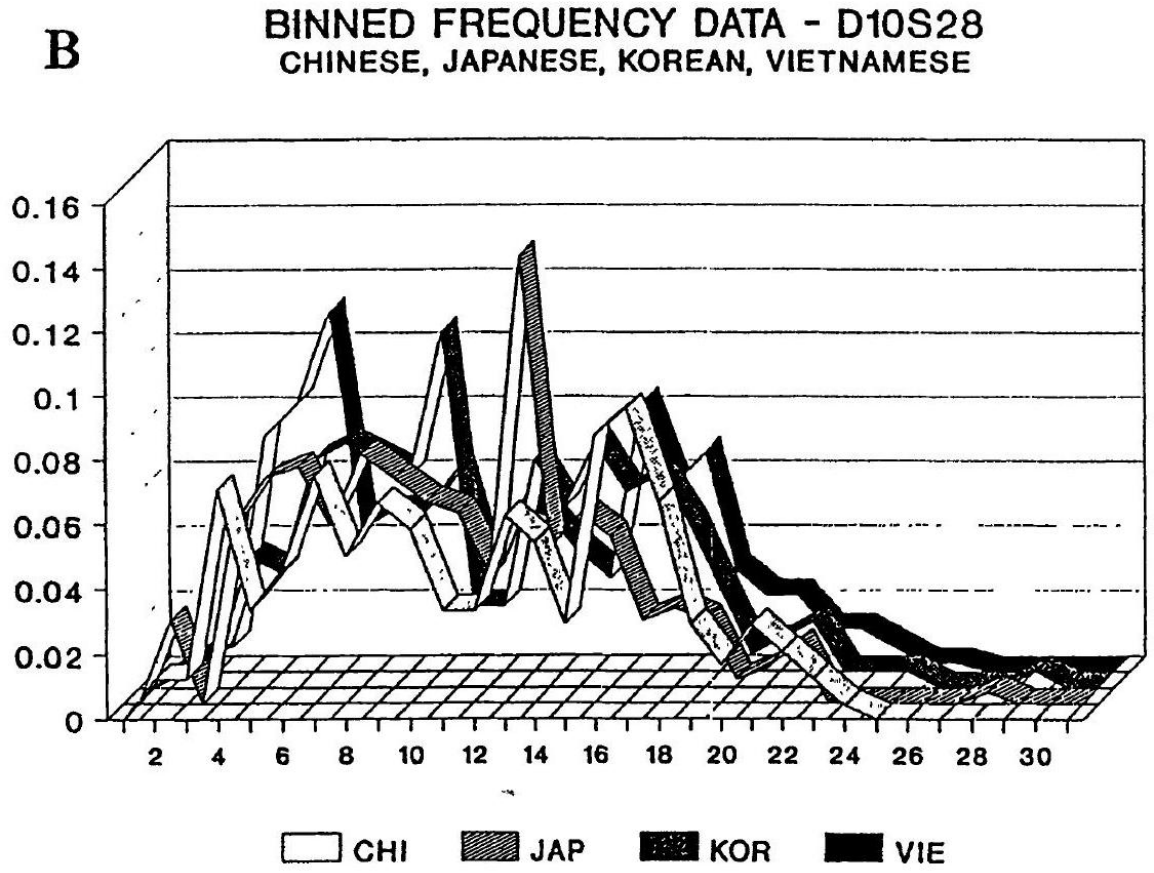
Good graphs: being informative

A comprehensible version;



Good graphs: being informative

A good statistician making a very bad graph;



Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4

Effective Comparisons

"Should I use this graph?" and "Does it look cool?" are not the same question.

For **data** on the utility of graphical measures for comparisions, see e.g. Cleveland and McGill (JASA 1984, JRSSA 1987);

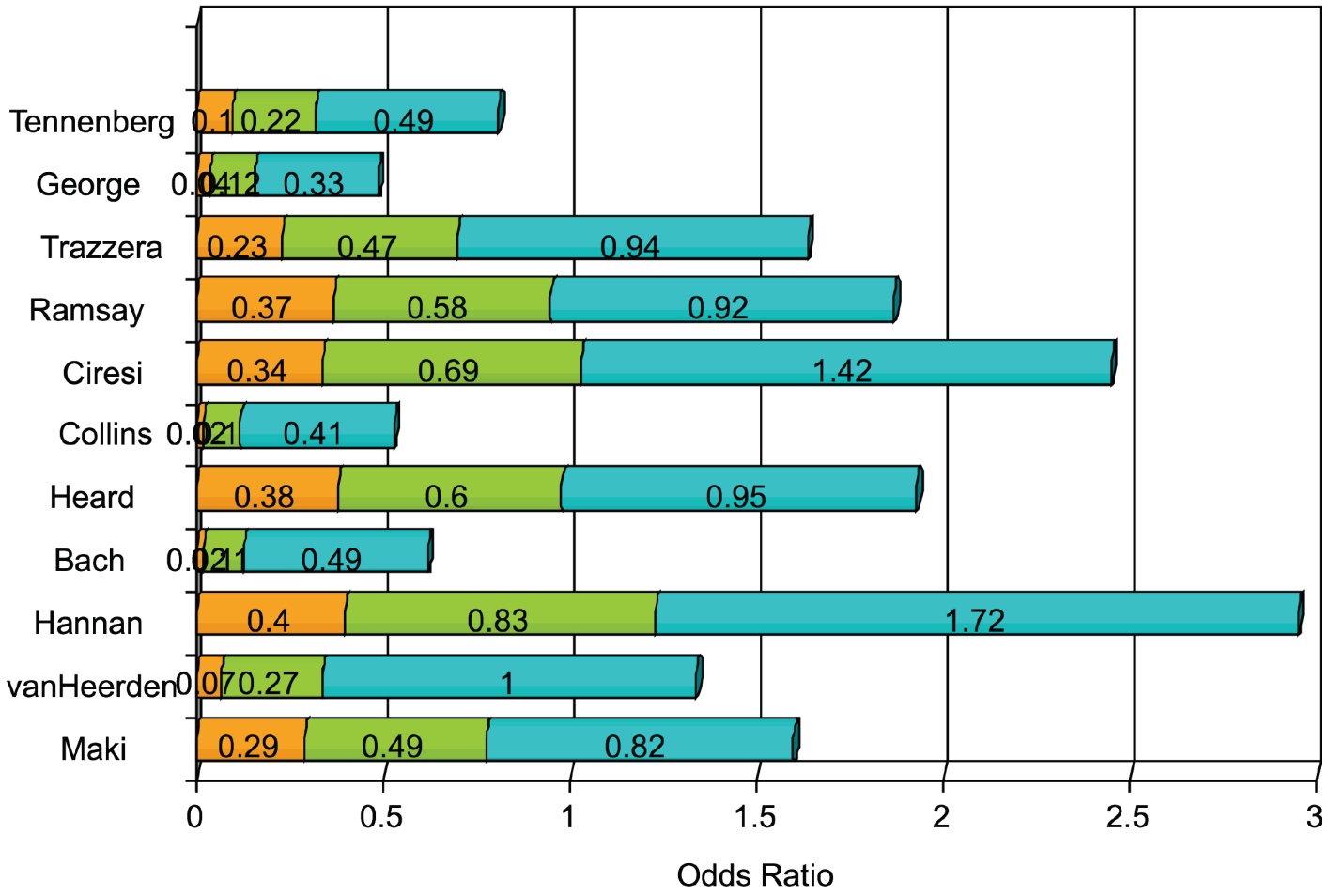
Metric	Usage	Accuracy
Position on common scale	Dot Plot	Best
Length	Bar chart	
Angle/Slope	Pie chart	
Area	Bubble Plot	
Volume/Curvature	Fake3D	
Color hue, density	e.g. Heat map	Worst

Note that 'Position on common scale'^{2} = scatterplot.

See also the `lattice` package, for 'small multiples' of these

Effective Comparisons

Possibly the *worst* way to compare intervals;



Effective Comparisons: Examples

Some (published!) data on favorite color;

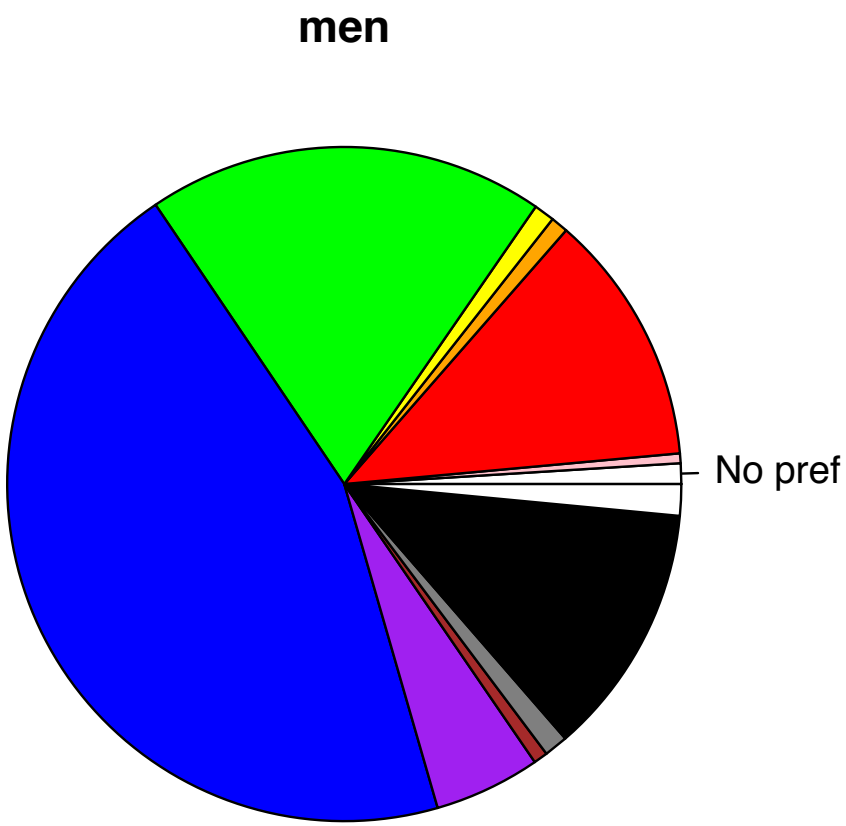
color	M	F	color	M	F
No pref	19	95	blue	866	938
pink	9	199	purple	98	459
red	233	447	brown	13	19
orange	16	66	grey	22	7
yellow	19	100	black	233	306
green	367	1051	white	29	79

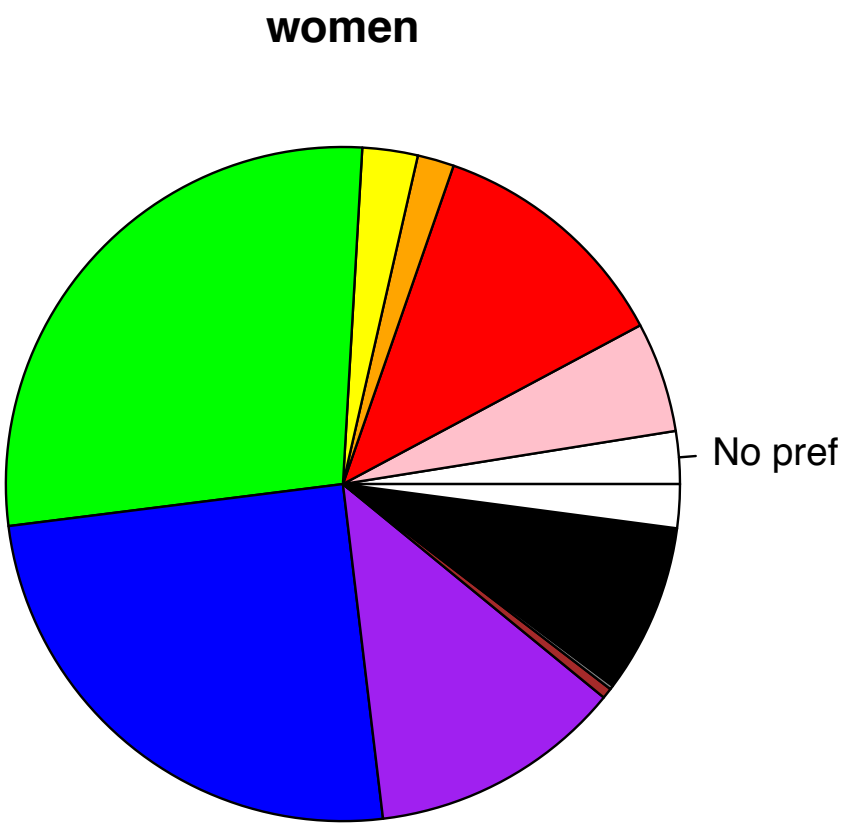
Source: Ellis and Ficek, Color preferences according to gender and sexual orientation *Personality and Individual Differences* (2001) 31:8

– who suggest differences are "inclined to suspect the involvement of neurohormonal factors" noting there are "sex differences in retinal biochemistry and in how the brain processes color information".

Effective Comparisons: Examples

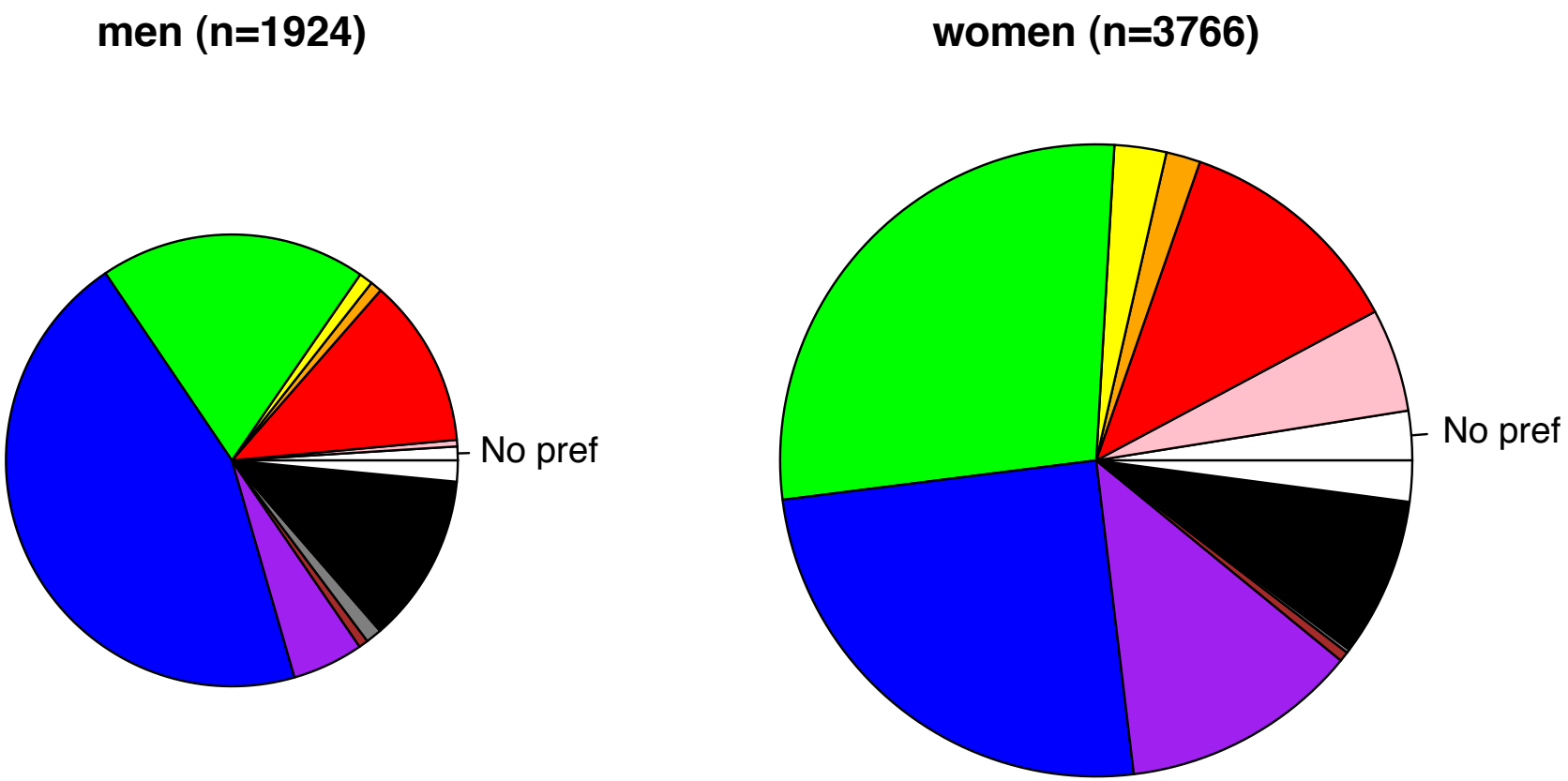
A first attempt; no intervals, comparisons hard





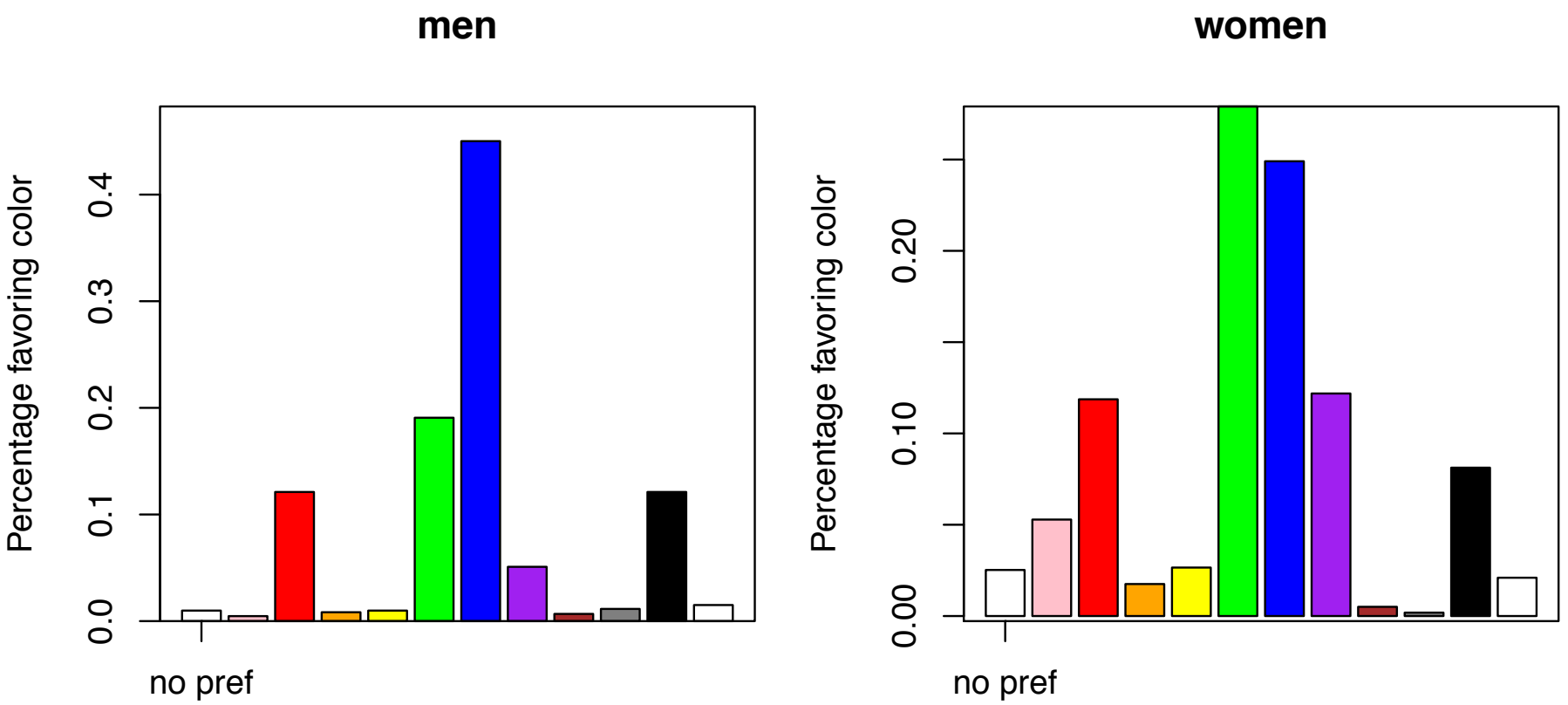
Effective Comparisons: Examples

With a *rough* attempt at intervals;



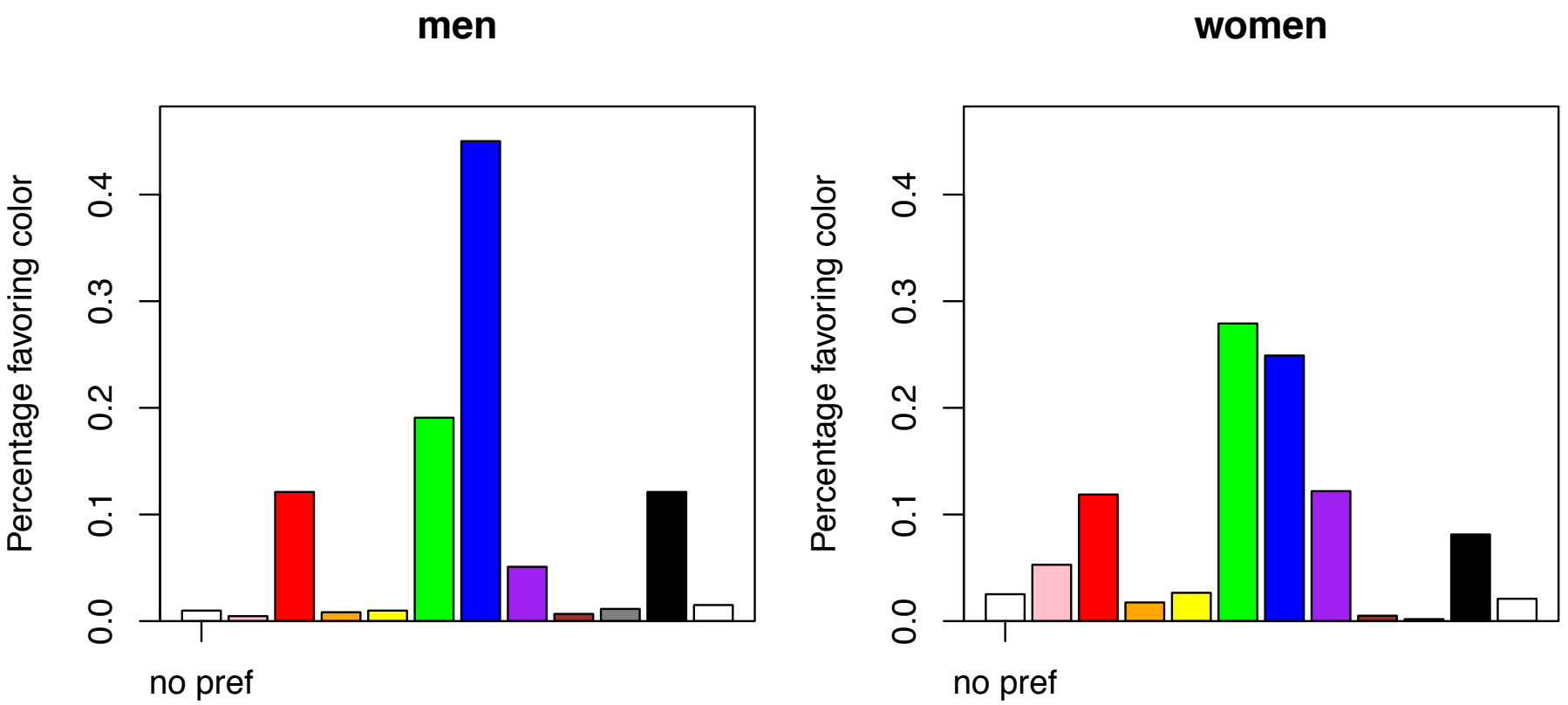
Effective Comparisons: Examples

Using `barplot()` ;



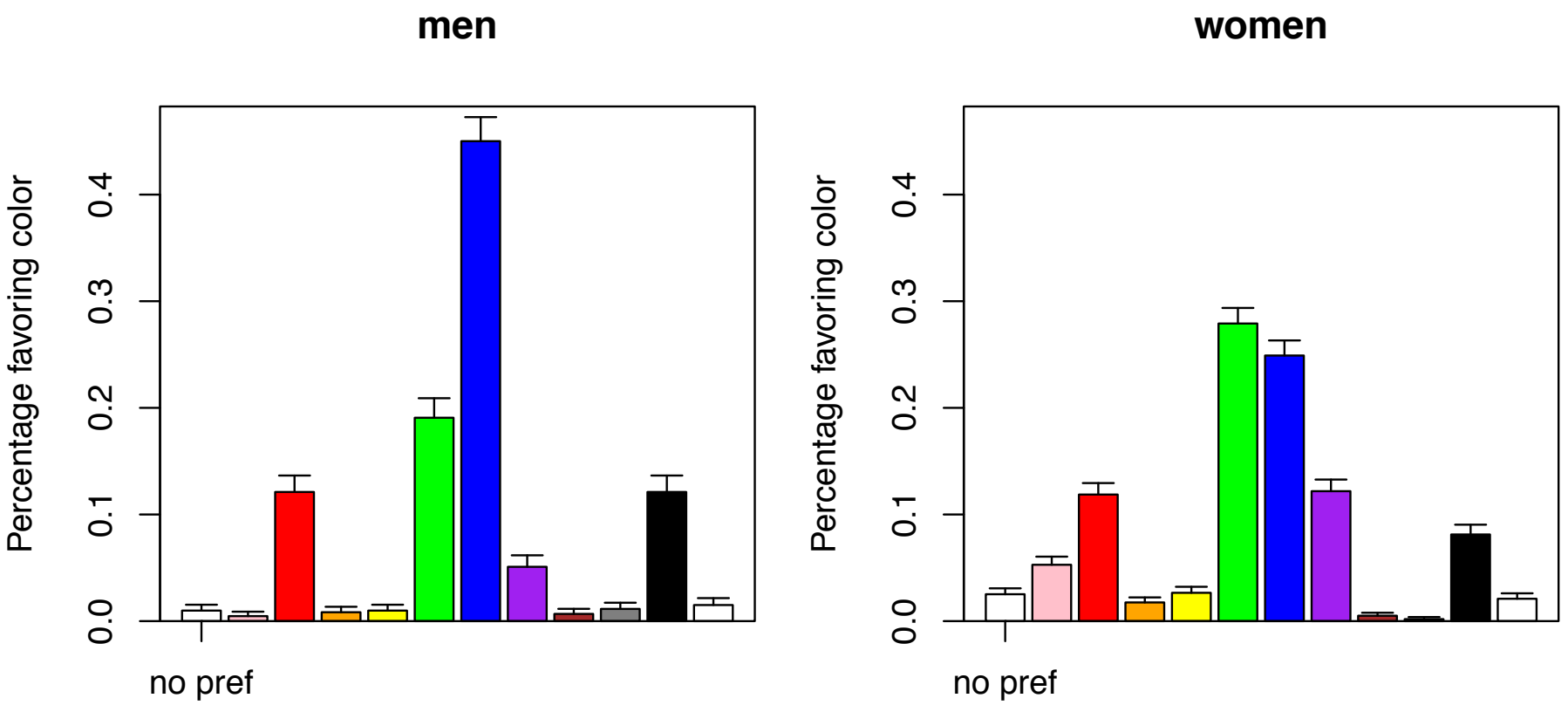
Effective Comparisons: Examples

Using `barplot()` and a common scale;



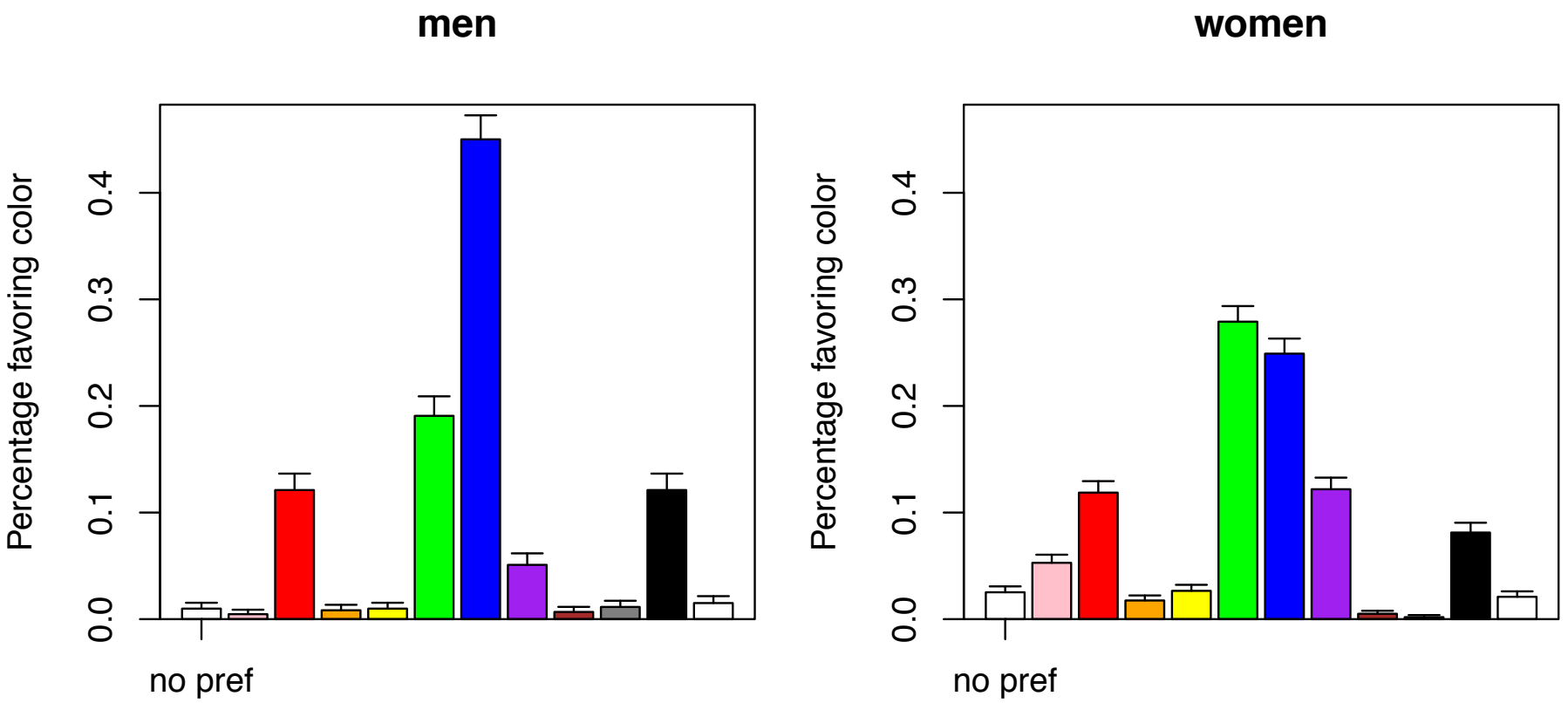
Effective Comparisons: Examples

Tower blocks with antennae – use `segments()`



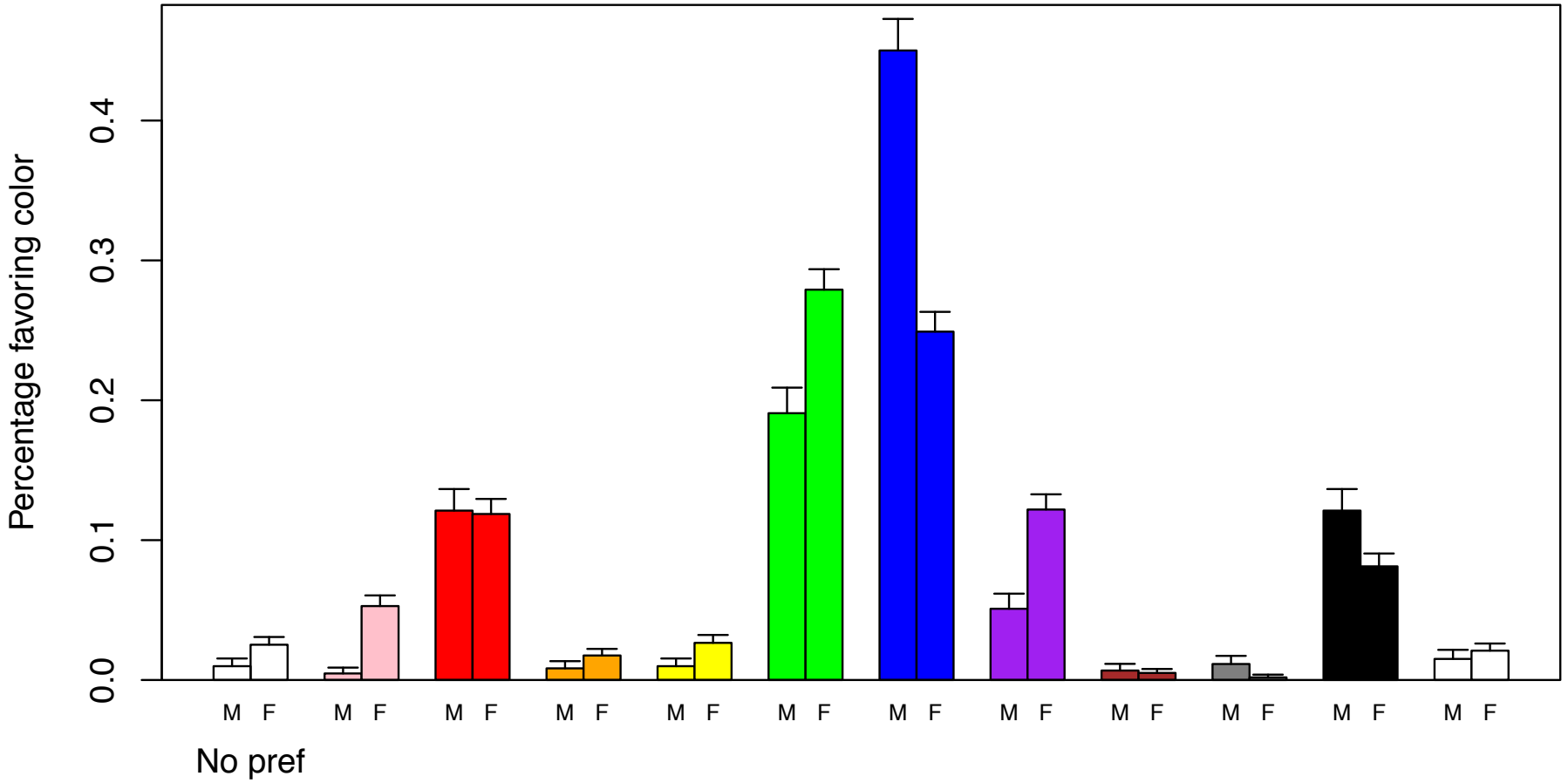
Effective Comparisons: Examples

Tower blocks with antennae; what do we compare?



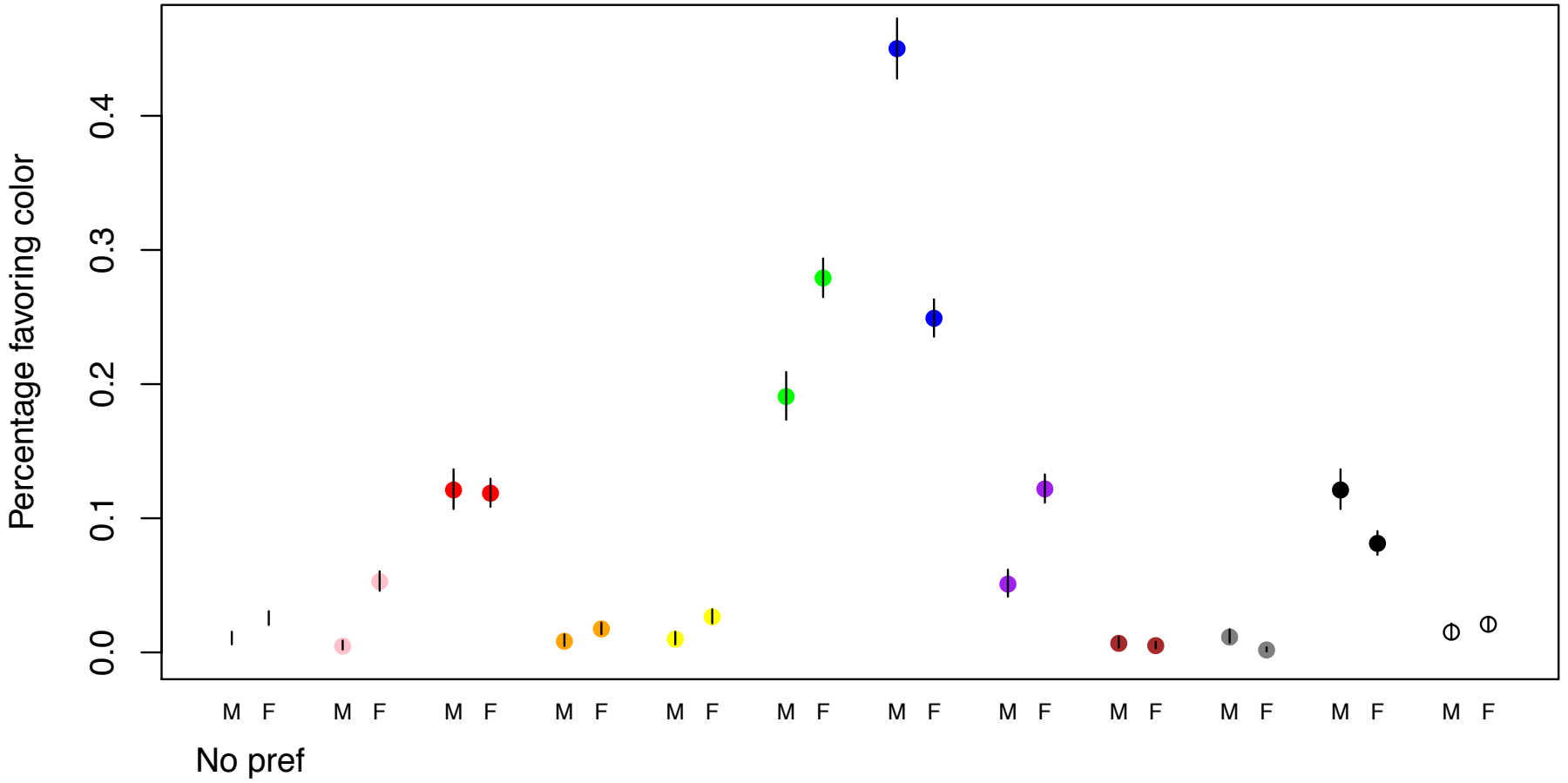
Effective Comparisons: Examples

Tower blocks with antennae; what do we compare?



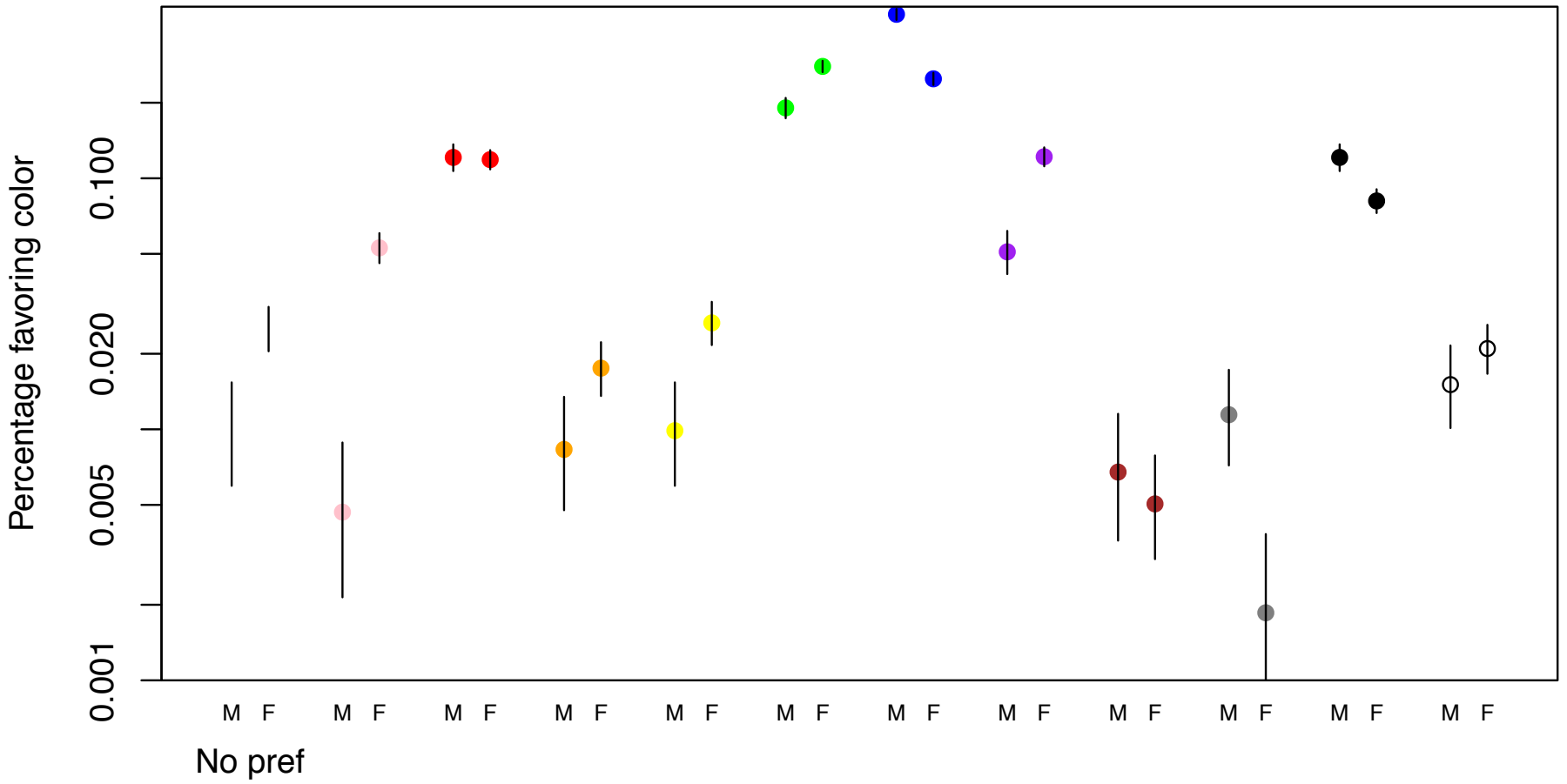
Effective Comparisons: Examples

Dump the blocks; 'position on common scale'



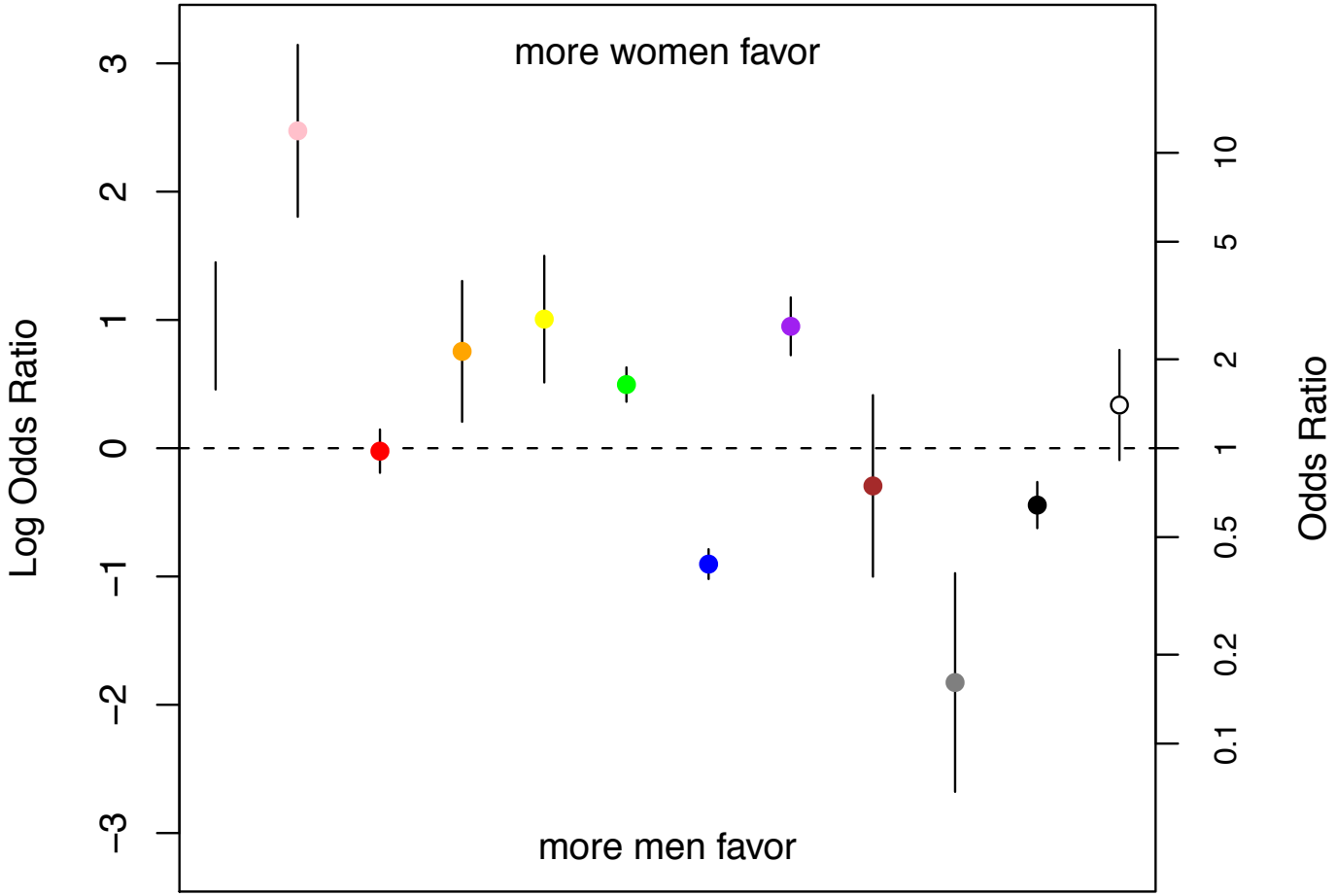
Effective Comparisons: Examples

Stresses unpopular colors;



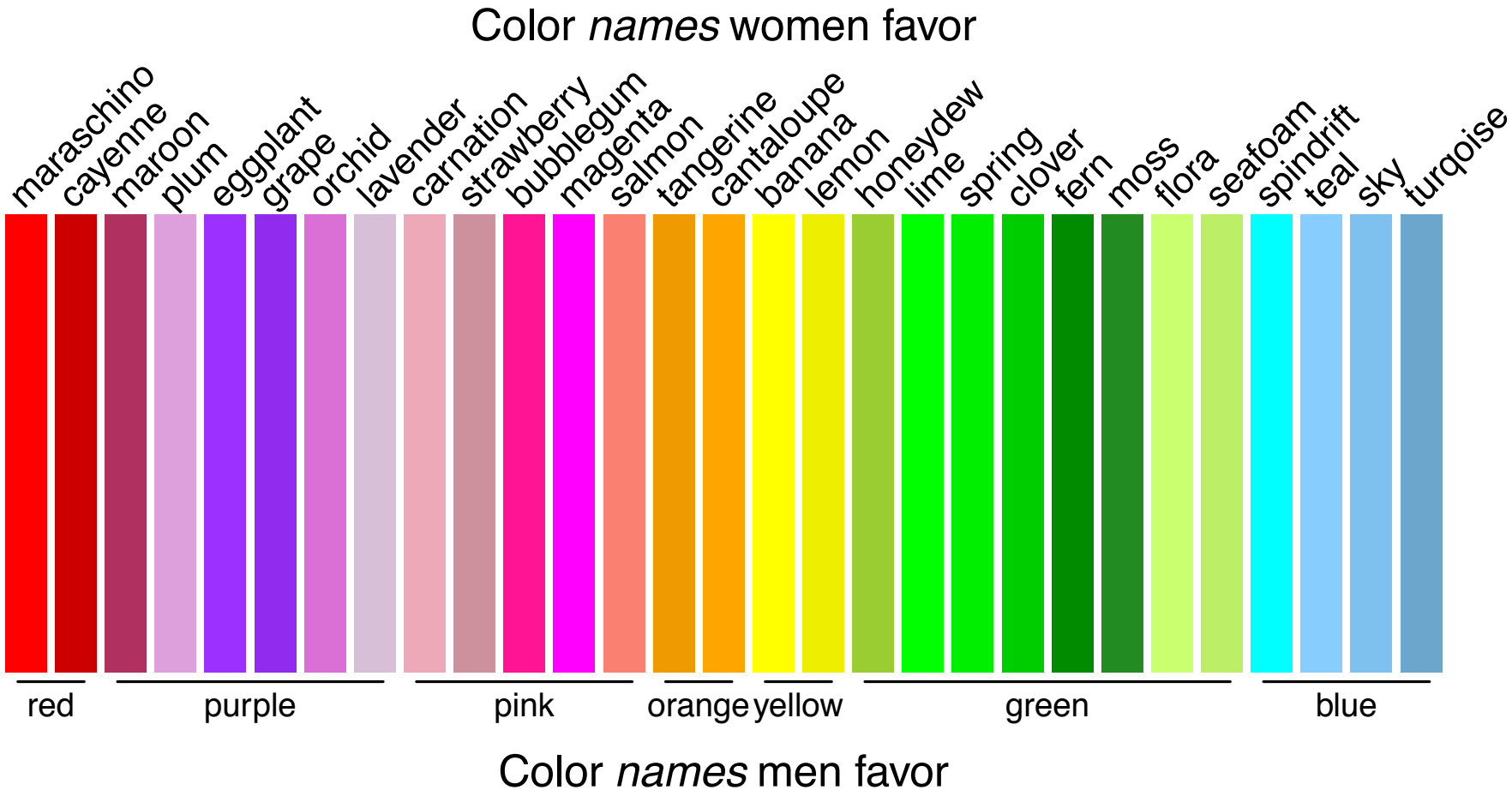
Effective Comparisons: Examples

Stresses only differences; (baseline group irrelevant)



Effective Comparisons: Examples

And finally;



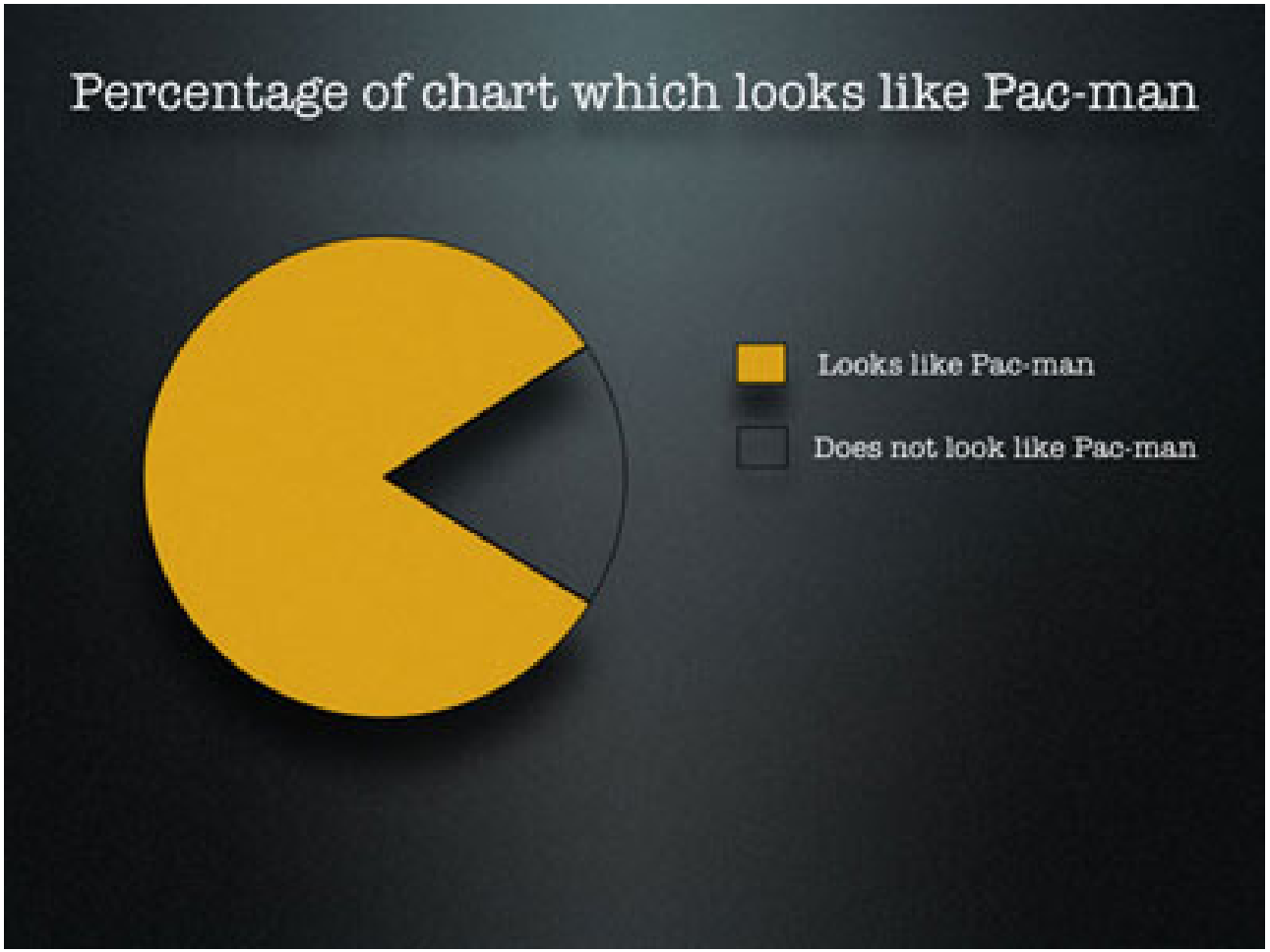
Effective Comparisons: Examples

Some lessons from all that;

- Decide what you want to compare; differences or absolute values?
- Often it will be differences – recall plotting residuals for model-checking, not data
- If you want to compare items, put them beside each other
- Minimalist representations (e.g. use of points not areas) are aesthetically 'clean' – and permit e.g. confidence intervals
- Plots will/should evolve, as you decide to stress different results
- Avoid pie charts...

Effective Comparisons: Examples

... one acceptable pie chart;

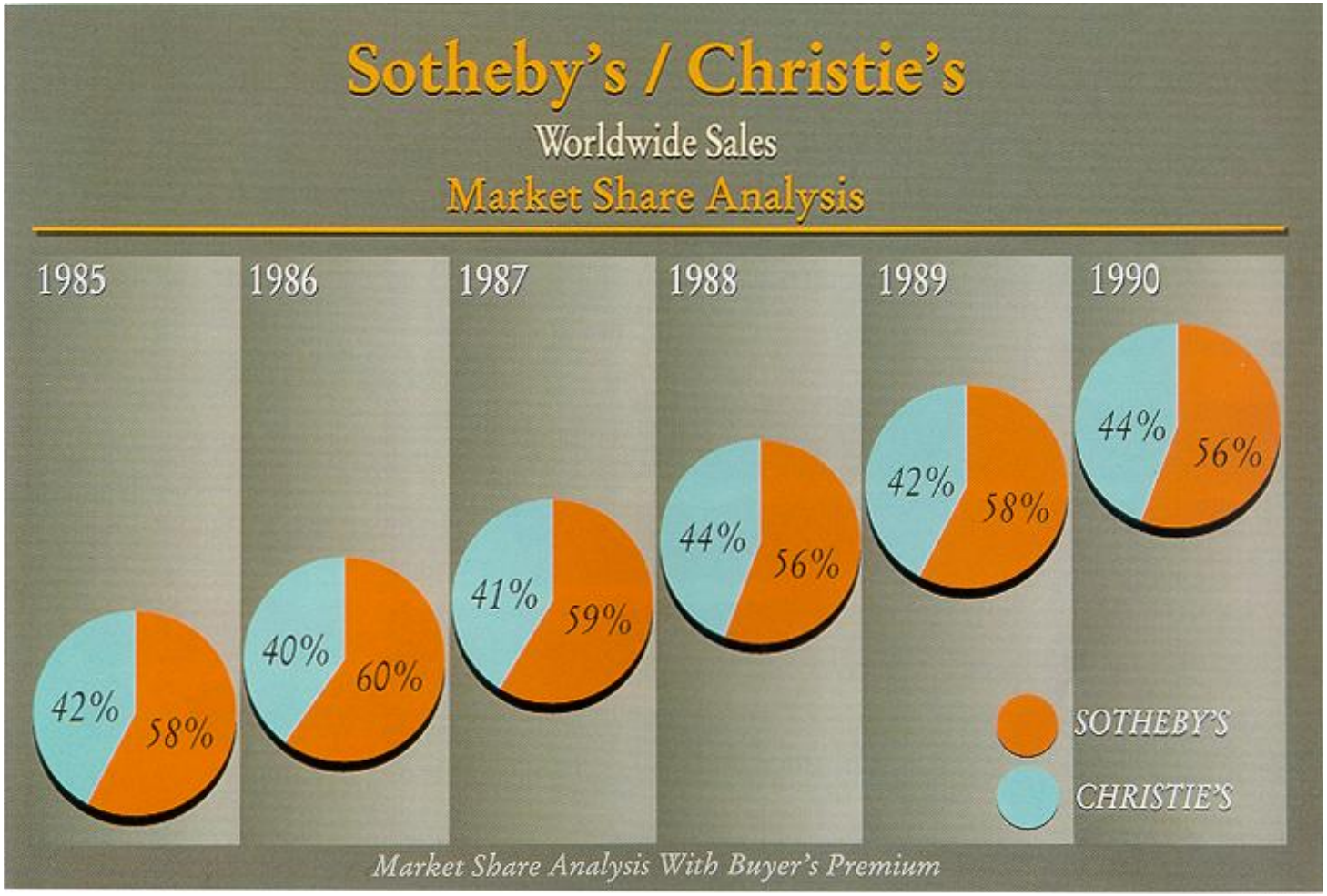


Effective Comparisons: Examples

... one acceptable pie;

Effective Comparisons: Examples

What visual comparison is bonkers, here?



Tufte's theory of data graphics

The 'minimalism' approach is formalized by Tufte, in his principles for better graphics;

- Above all else, show the data

- Maximize the data-ink ratio (i.e. data ink / total ink)
- Erase non-data-ink (*chartjunk*)
- Erase redundant data-ink
- Revise and edit

Let's apply these, for another small dataset;

Tufte theory: Berkeley 1973 data

In 1973, sex discrimination was suspected in admission to Berkeley;

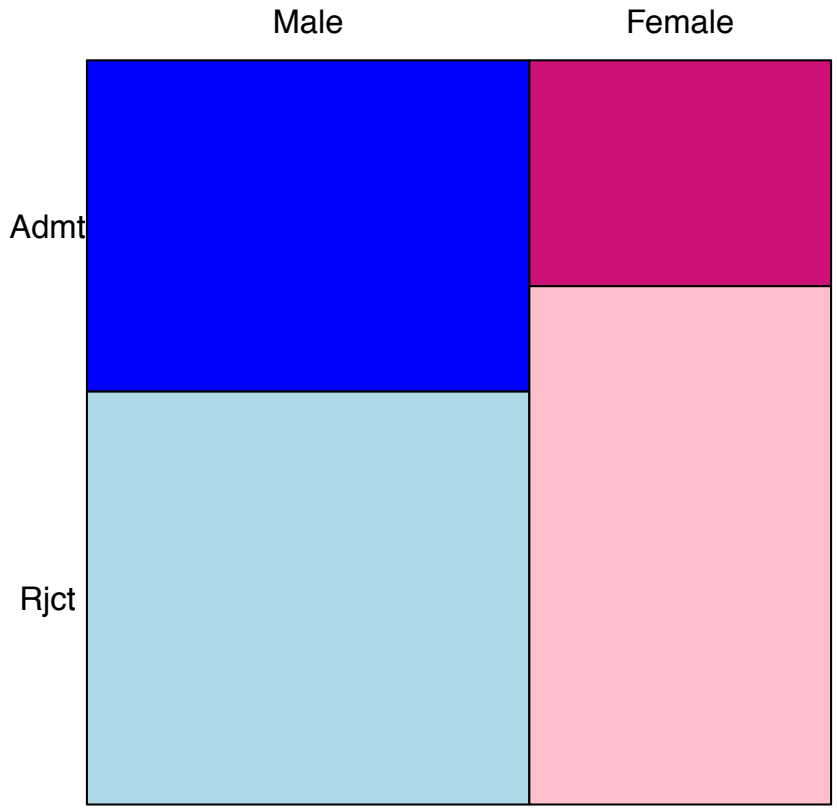
Dept	Men		Women	
	n	Admit	n	Admit
A	825	0.62	108	0.82
B	560	0.63	25	0.68
C	325	0.37	593	0.34
D	417	0.33	375	0.35
E	191	0.28	393	0.24
F	373	0.06	341	0.07
Total	2691	0.45	1835	0.30

– the 'headlines' compared 45% to 30%.

How can we turn this table into a graph?

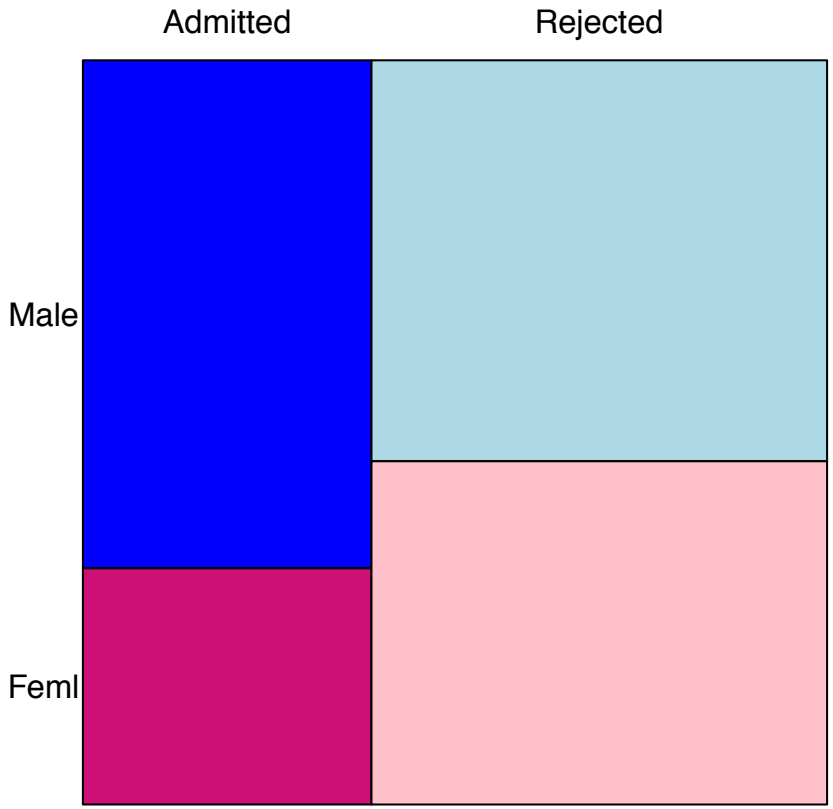
Tufte theory: Berkeley 1973 data

Mosaic plots are a fairly 'old school' method...



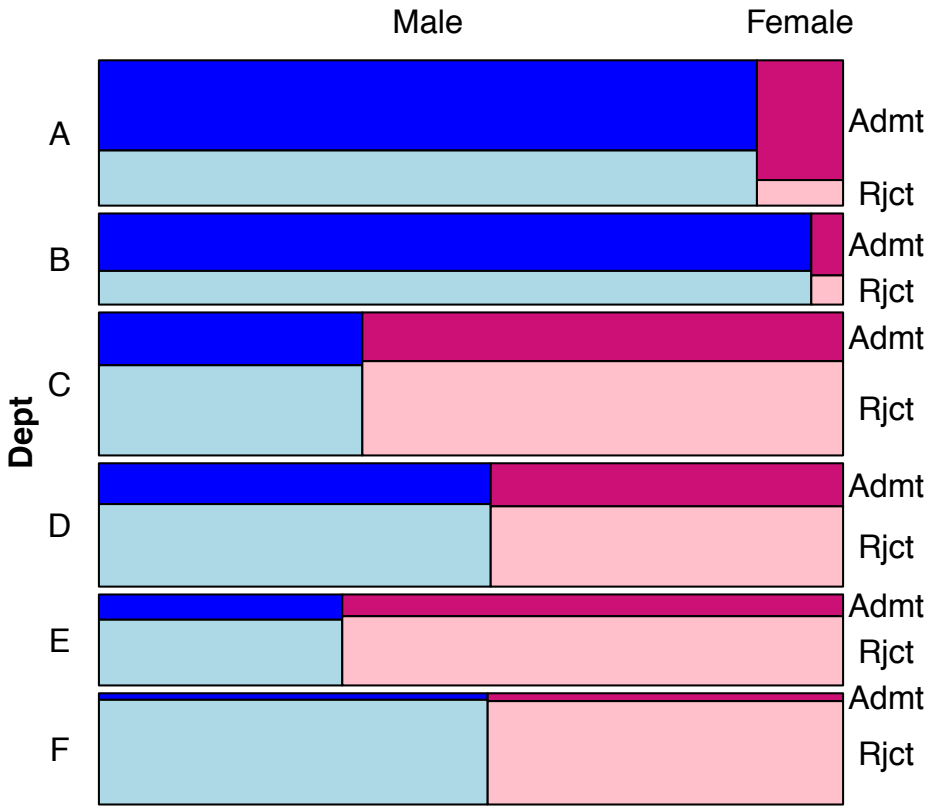
Tufte theory: Berkeley 1973 data

... where conditioning matters;



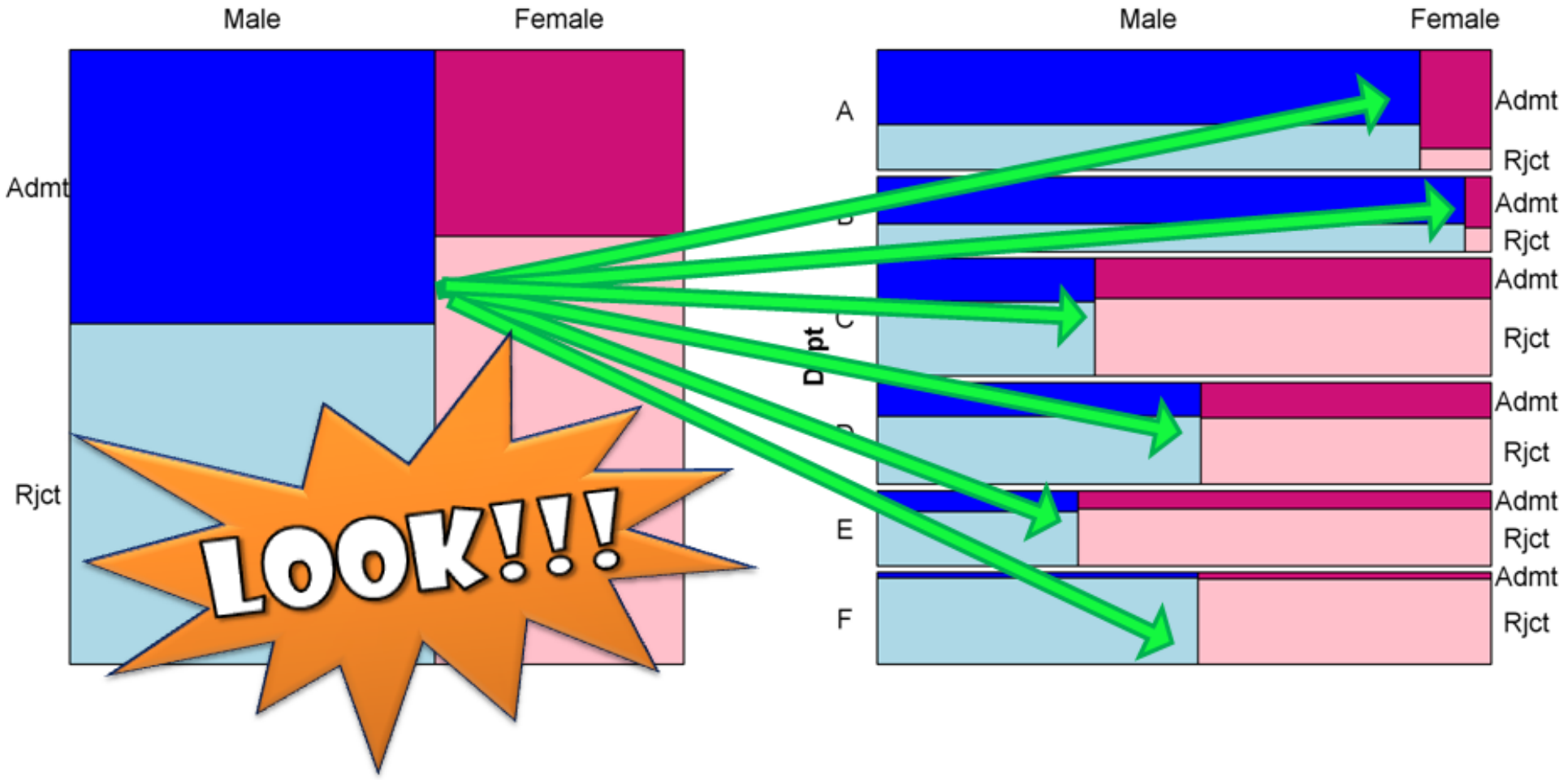
Tufte theory: Berkeley 1973 data

Broken down by department...



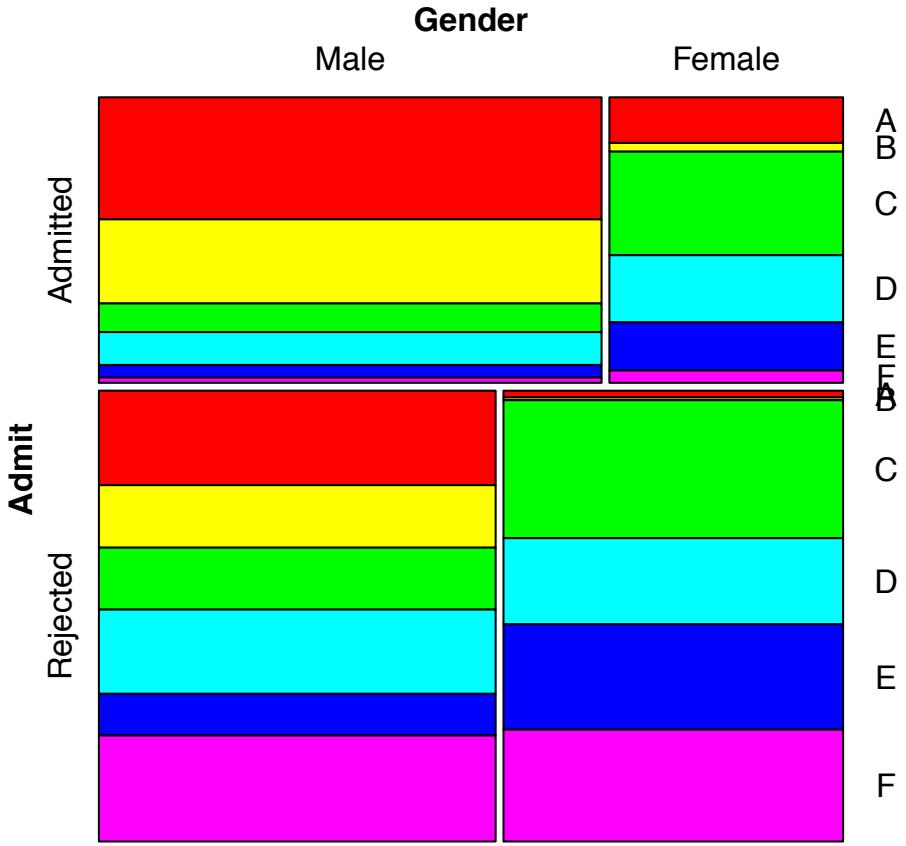
Tufte theory: Berkeley 1973 data

... in a talk, one can dramatize the difference;



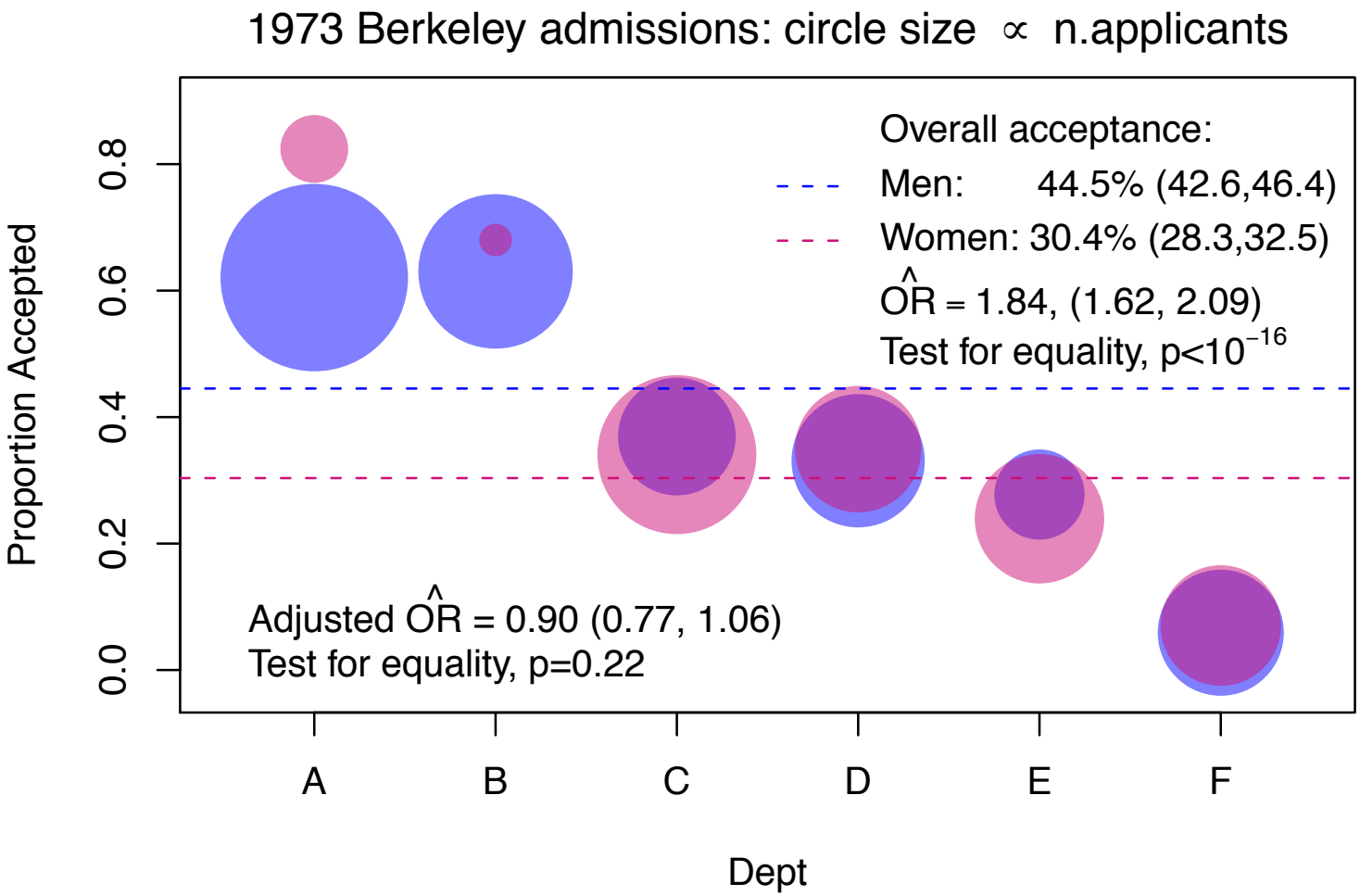
Tufte theory: Berkeley 1973 data

... but this is hard, on a single plot;



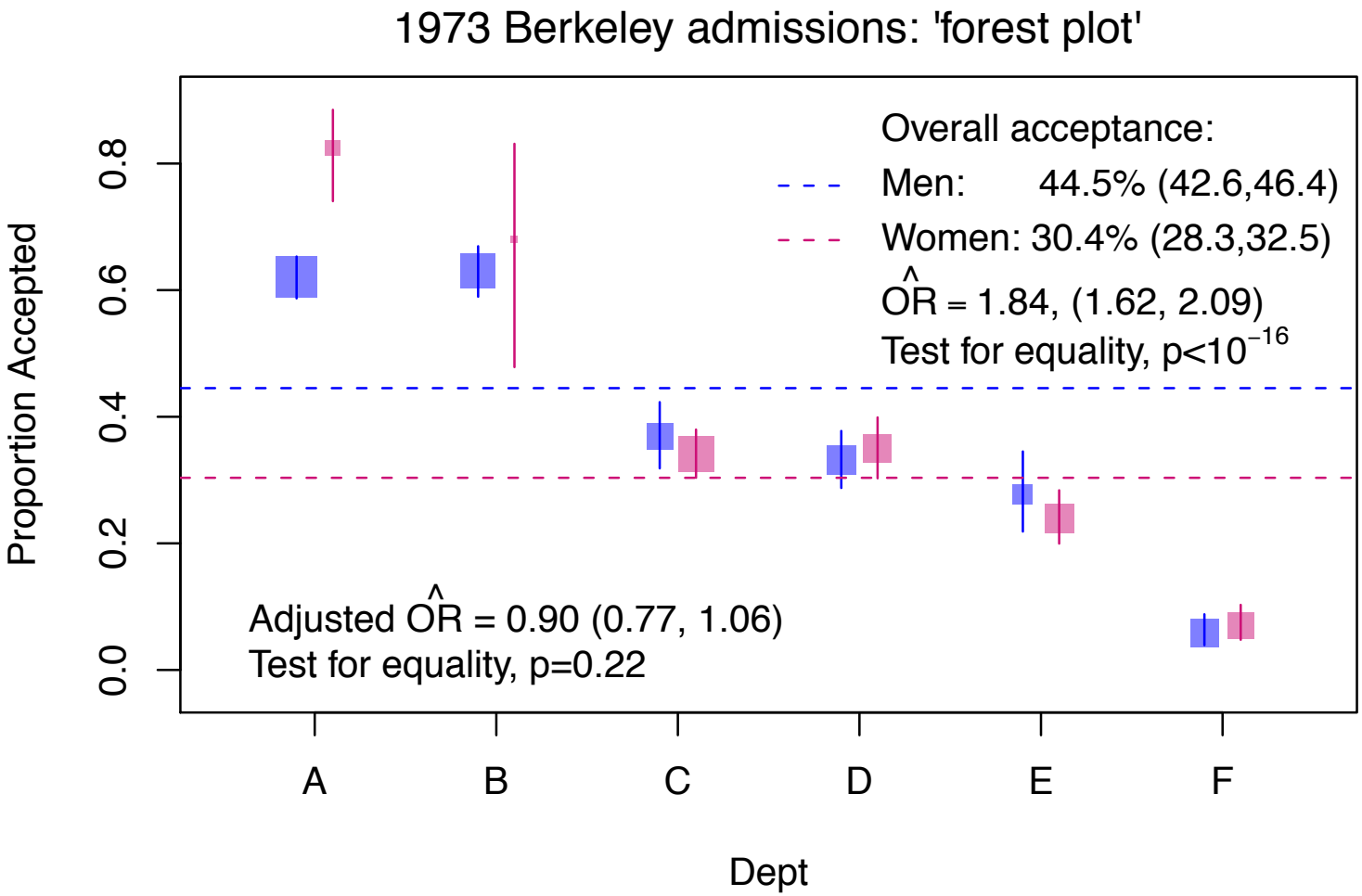
Tufte theory: Berkeley 1973 data

Recall 'position on a common scale'/Tufte;



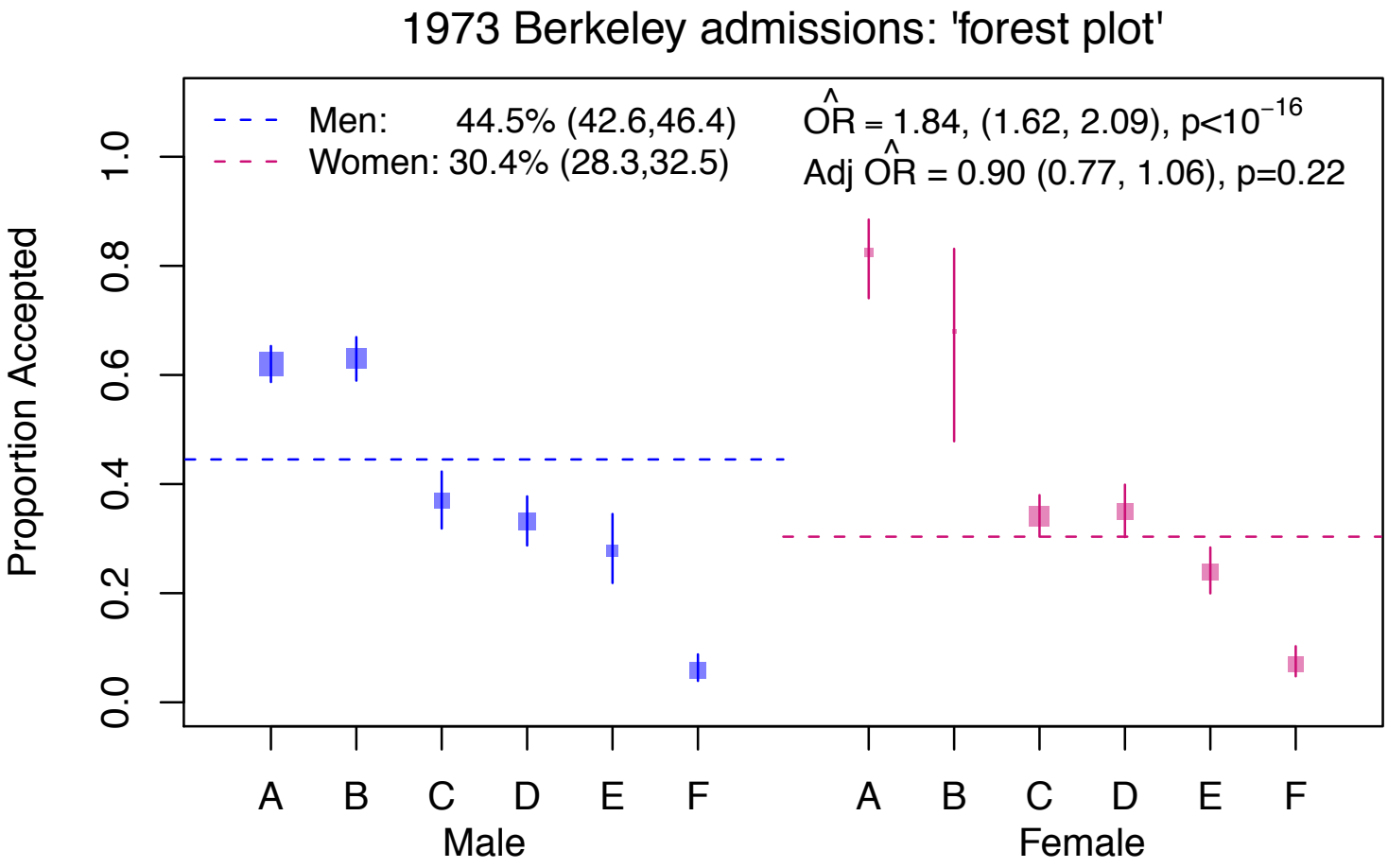
Tufte theory: Berkeley 1973 data

Less ink – confounding less obvious



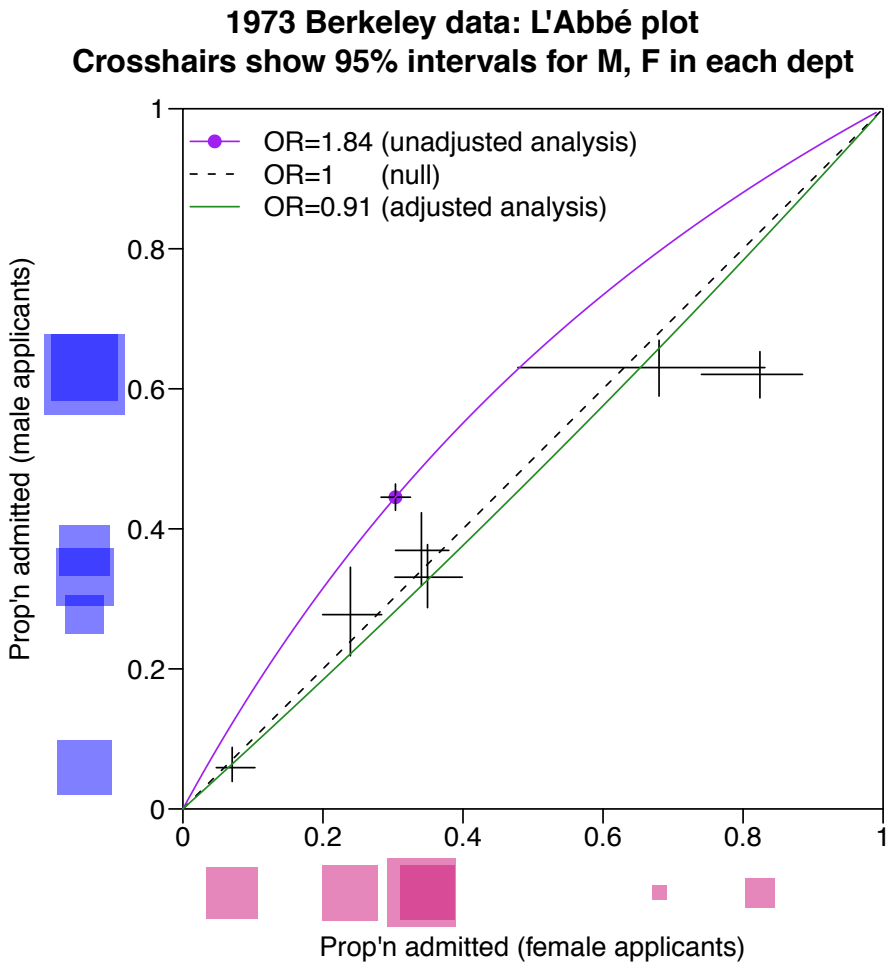
Tufte theory: Berkeley 1973 data

Berkeley-wide comparison of admittance;



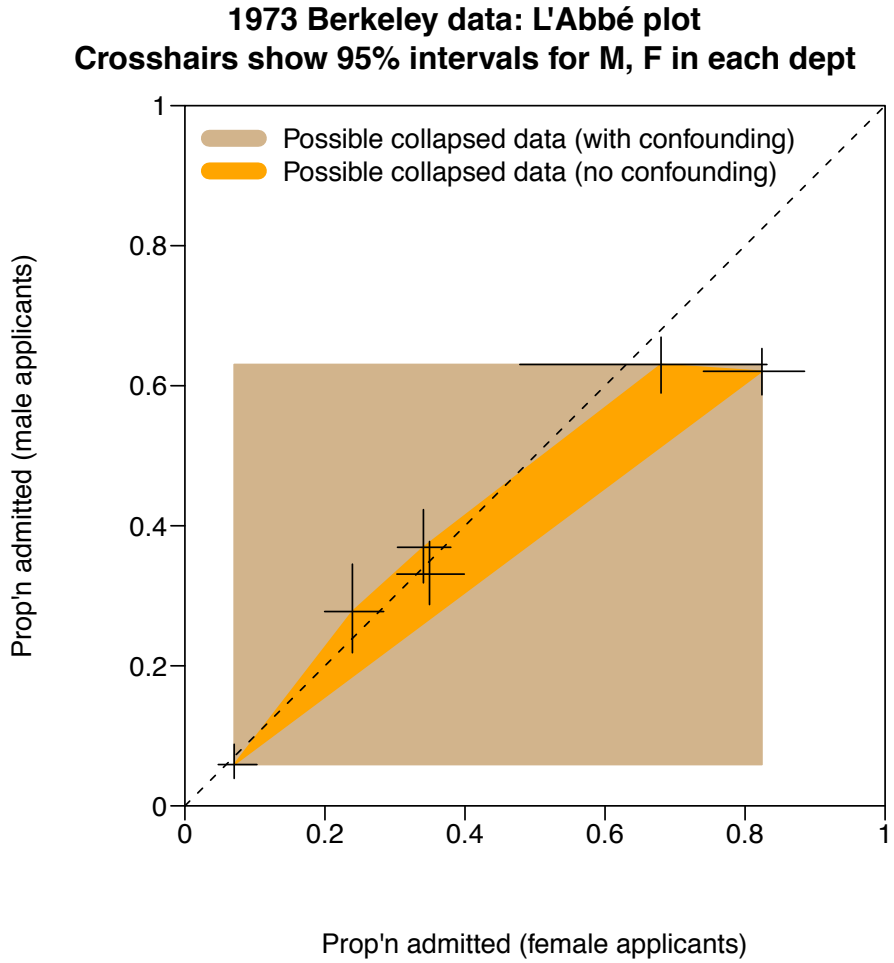
Tufte theory: Berkeley 1973 data

Remove irrelevant A/B/C ordering;



Tufte theory: Berkeley 1973 data

Best for confounding/collapsibility discussion;



Different points

Scatterplots can be enhanced by using a selection of plotting symbols; Lewandowsky and Spence (JASA, 1989) rank options as follows

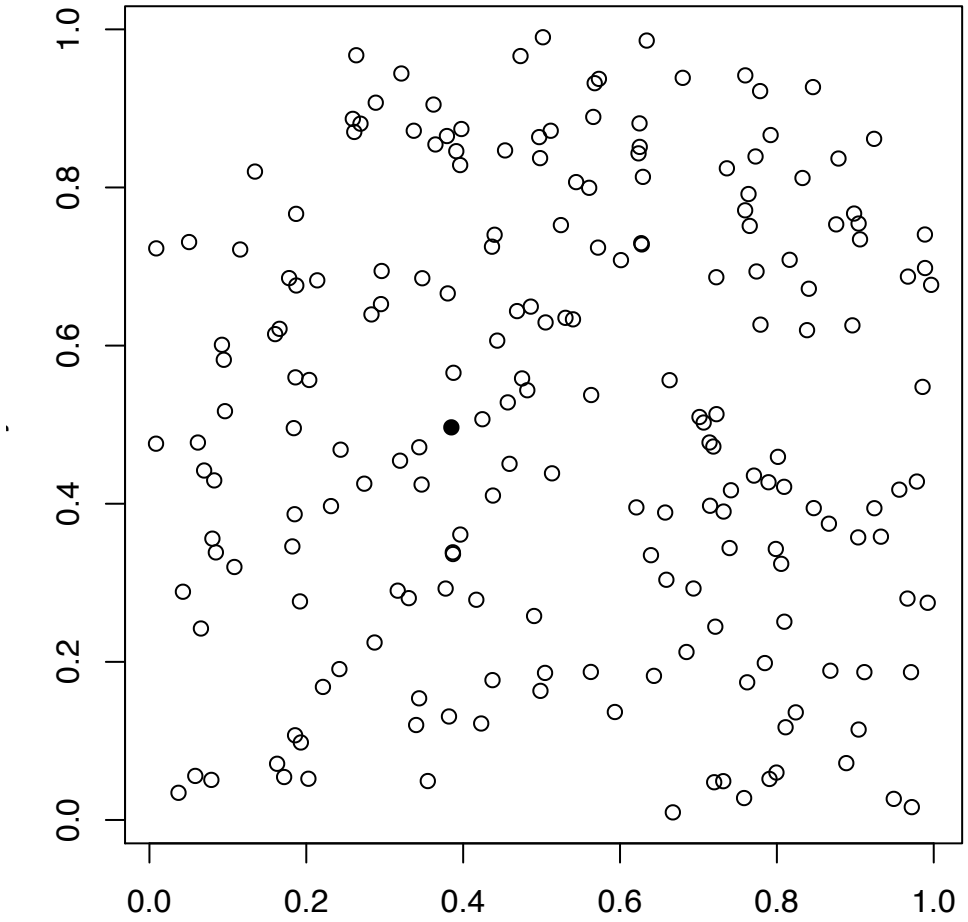
- color, and transparency (best)
- degree of filling – e.g. `symbols(..., thermometers)`
- shapes/size
- letters (worst)

– of course, you must have a legend

- Combinations of the above are possible, but this rapidly gets confusing
- ... combinations of the above
- It's very easy to overuse *all* of them...

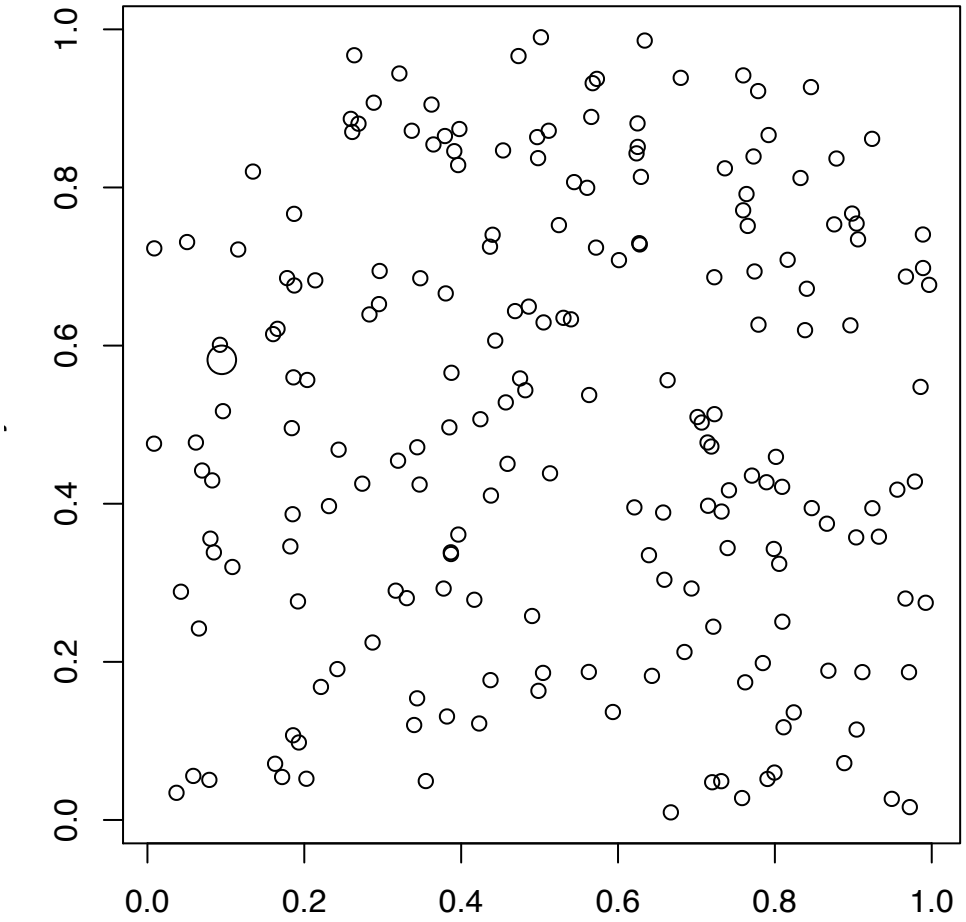
Different points

One of these points is not like the others...



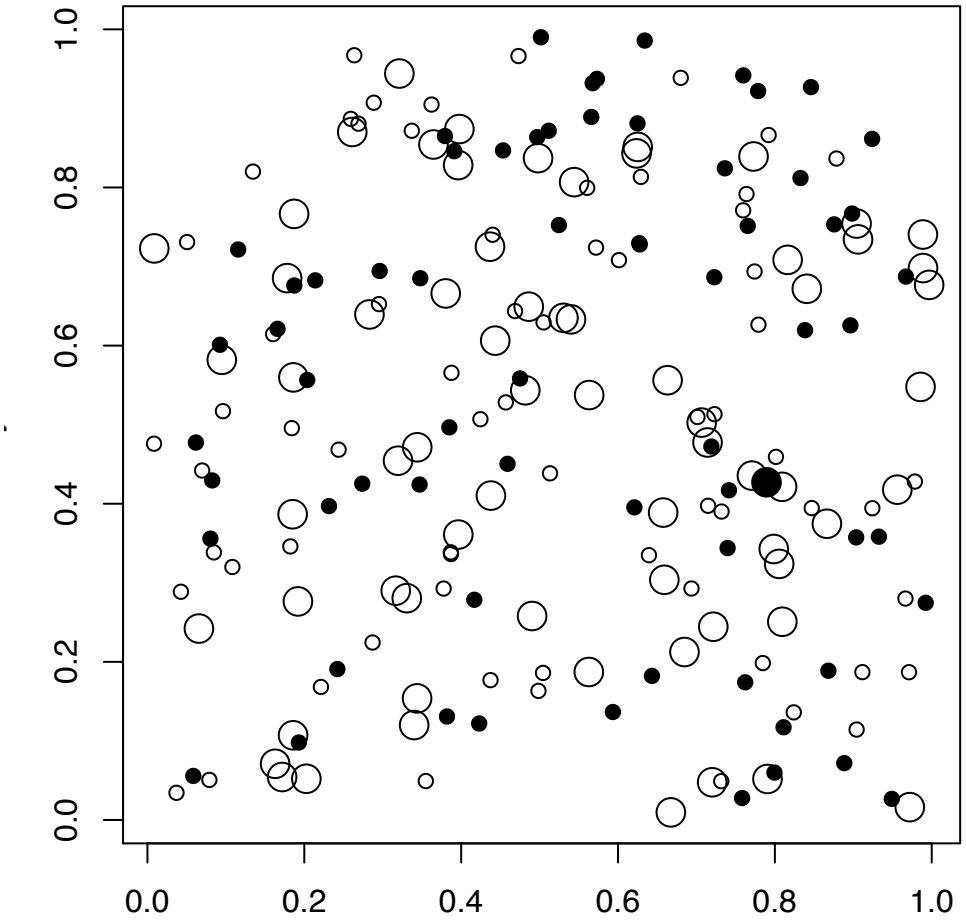
Different points

One of these points is not like the others...



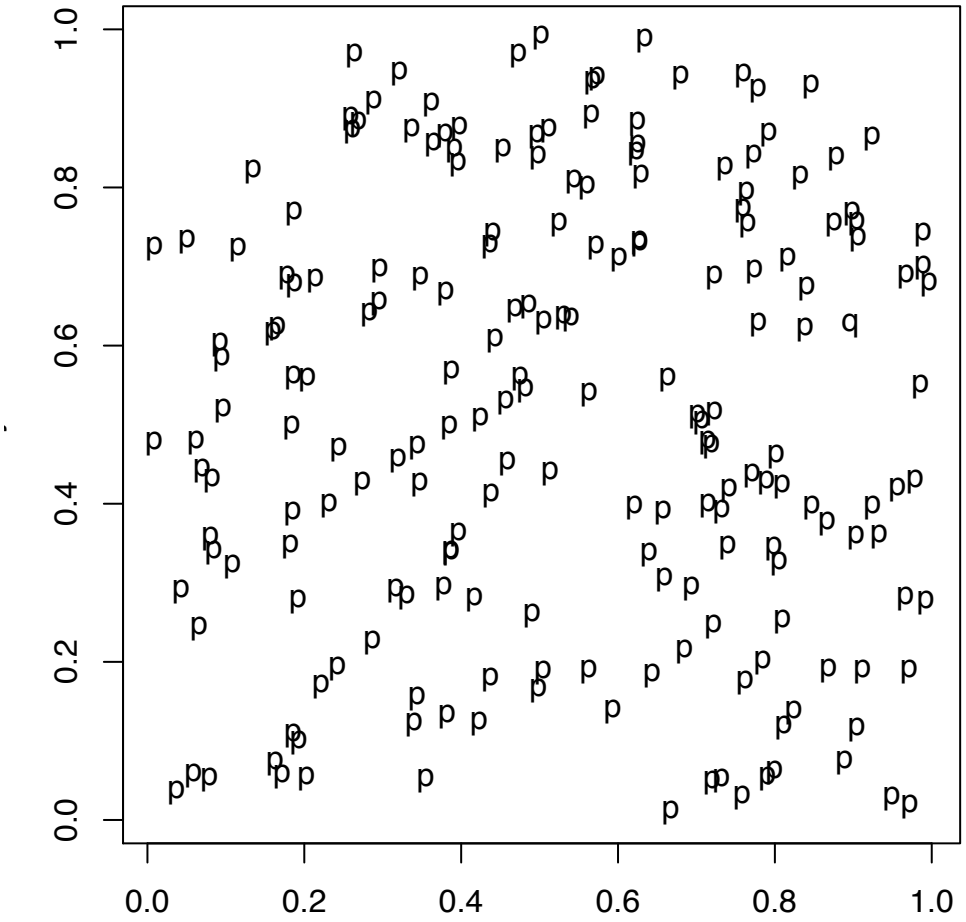
Different points

One of these points is not like the others...



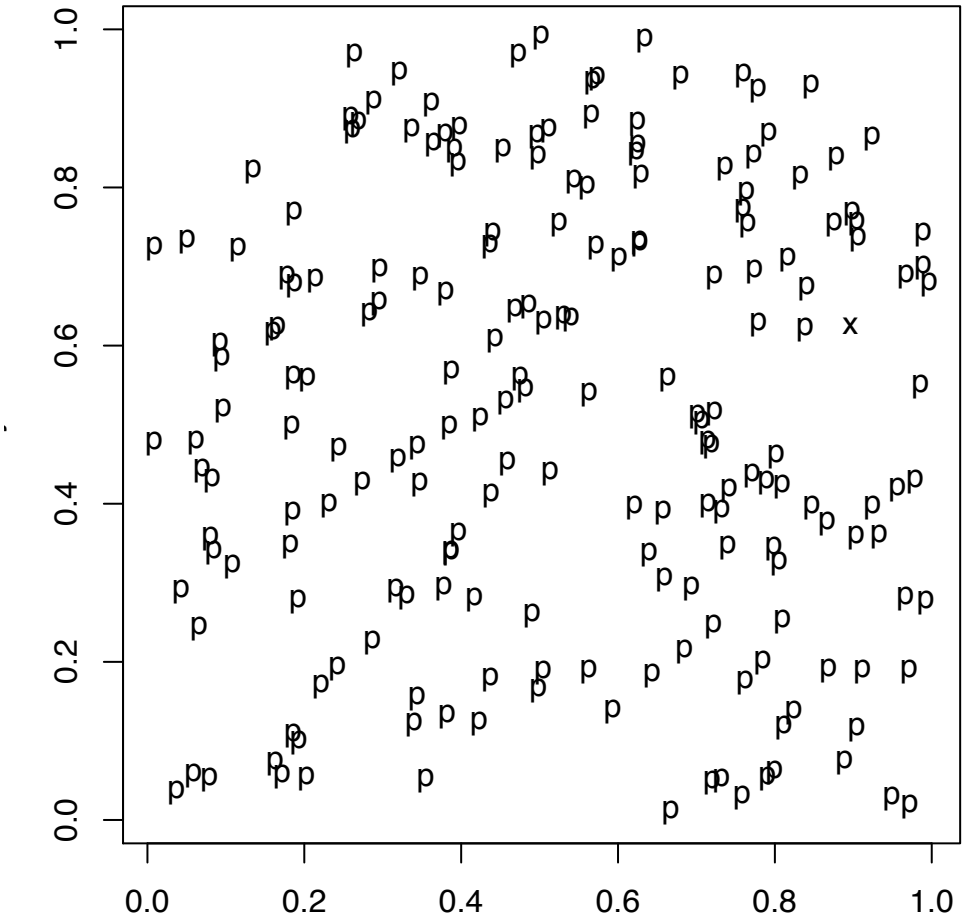
Different points

One of these points is not like the others...



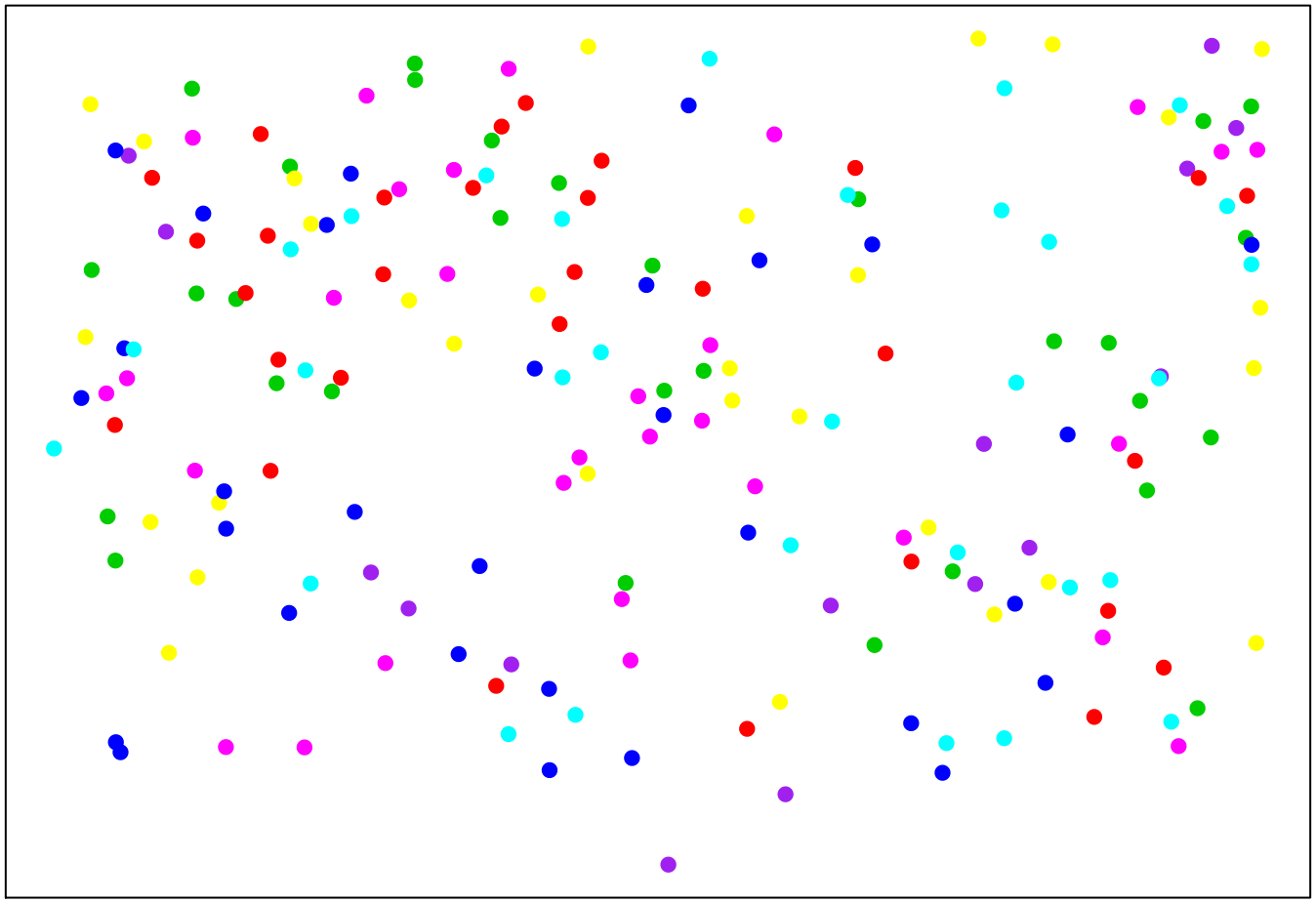
Different points

One of these points is not like the others...



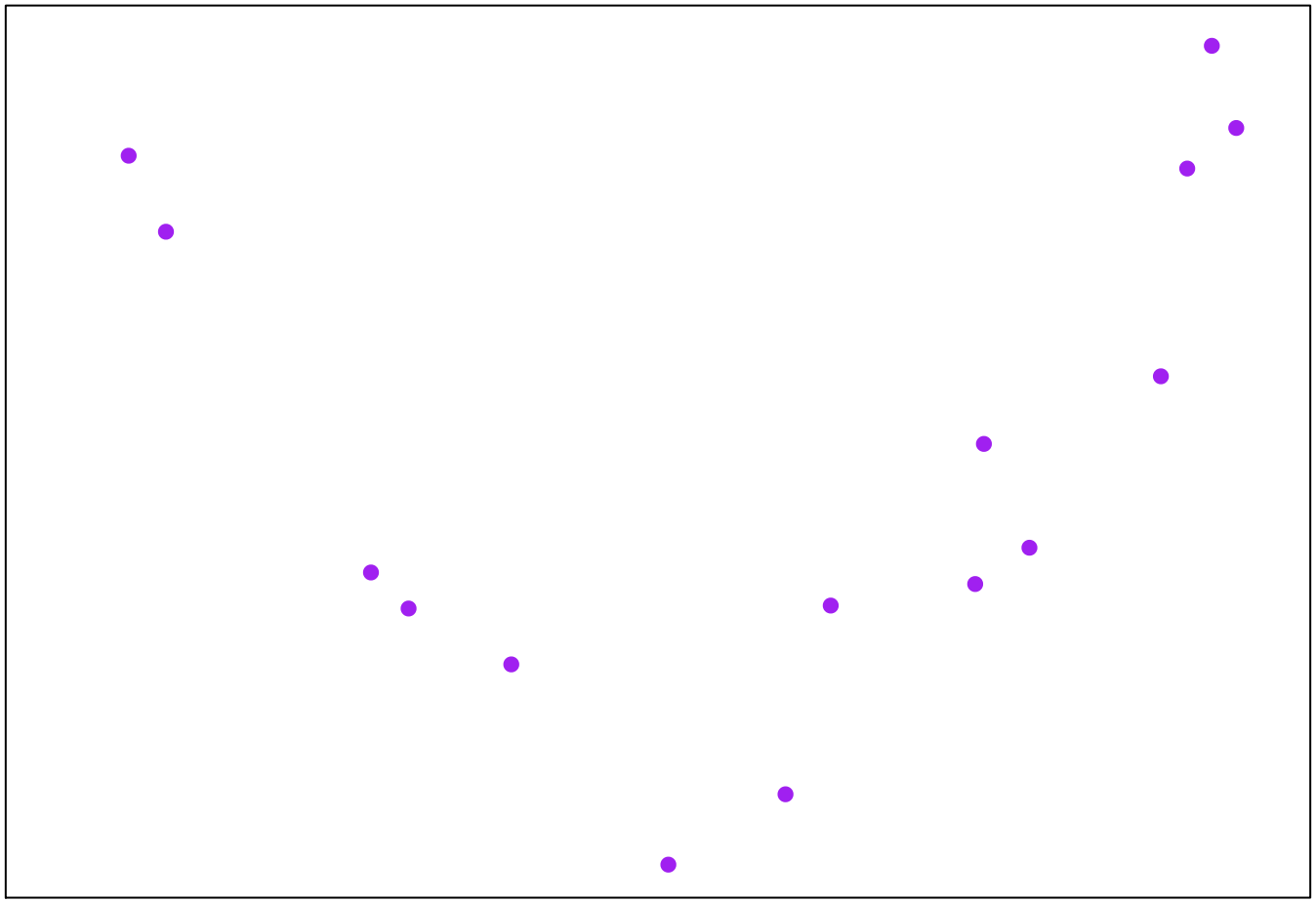
Different points

Some of these points are not like the others...



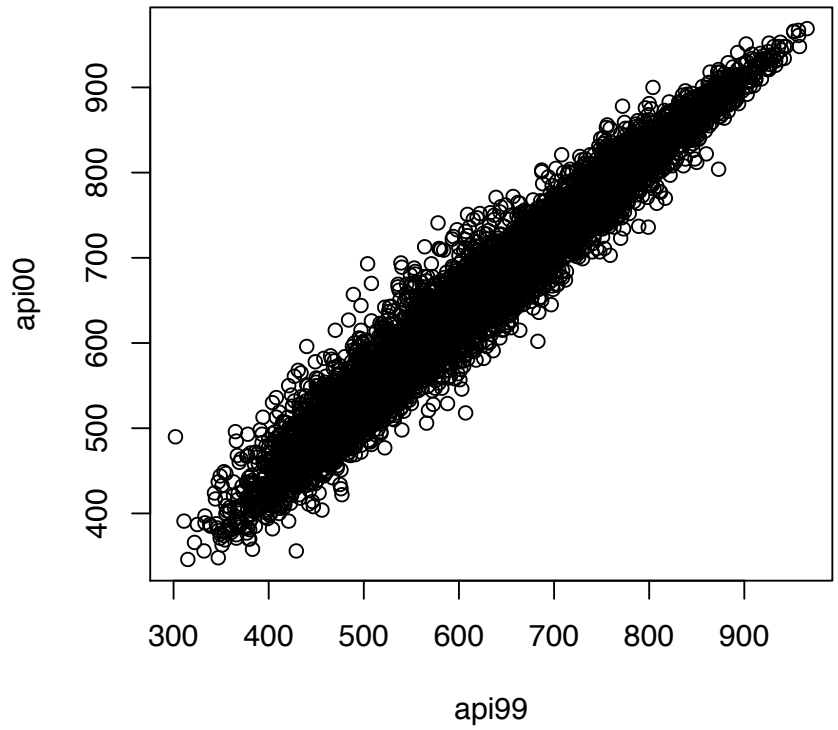
Different points

Some of these points are not like the others...



Different points

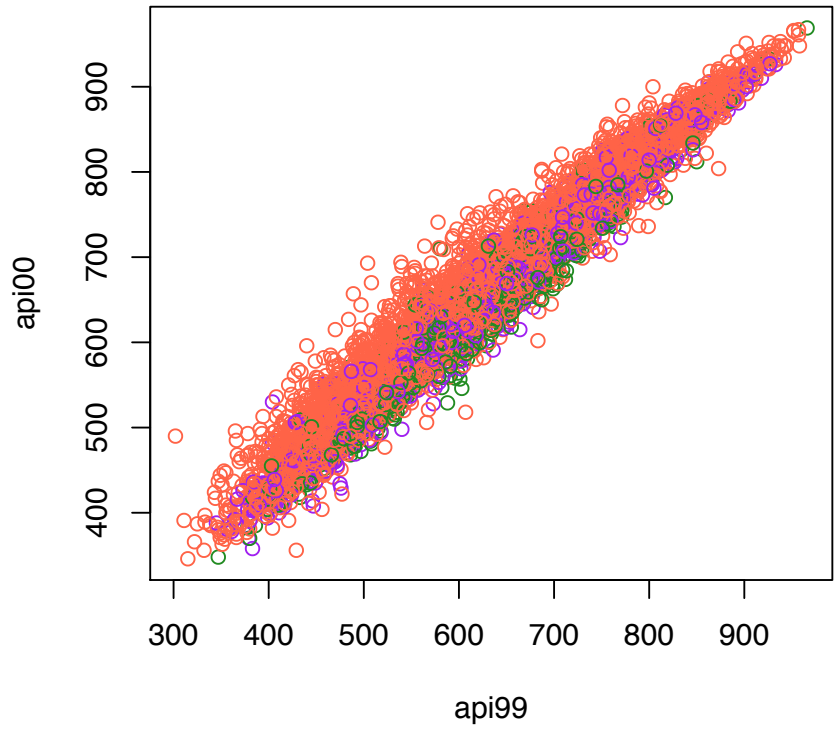
For large(ish) data, 'overlap' is a fundamental problem...



(California Academic Performance Index on 6194 schools)

Different points

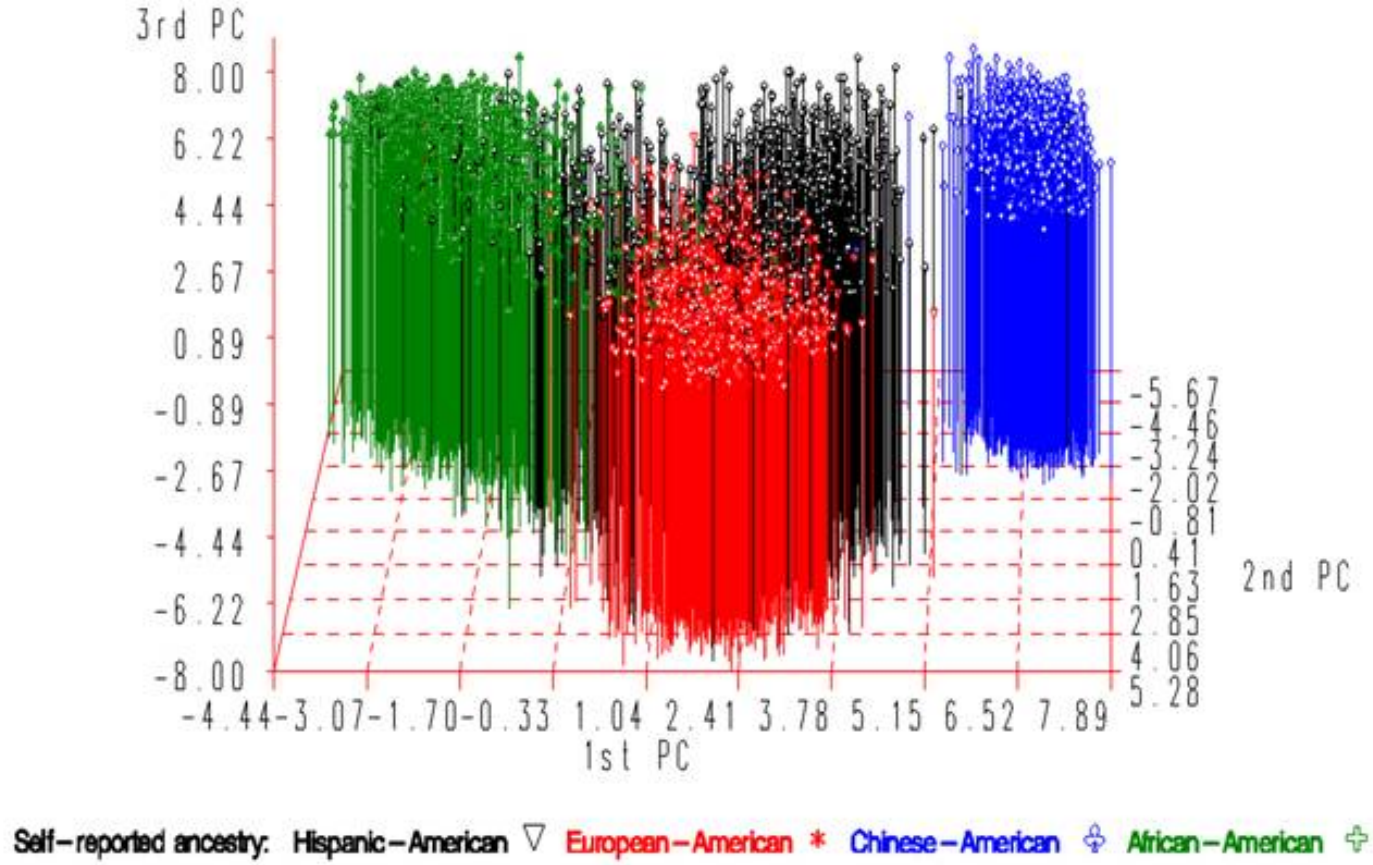
... which remains, when we color-code.



Colors denote Elementary, Middle & High Schools

Different points

With three dimensions + color-codes, this can happen;



(R does have `persp()`, for occasional use)

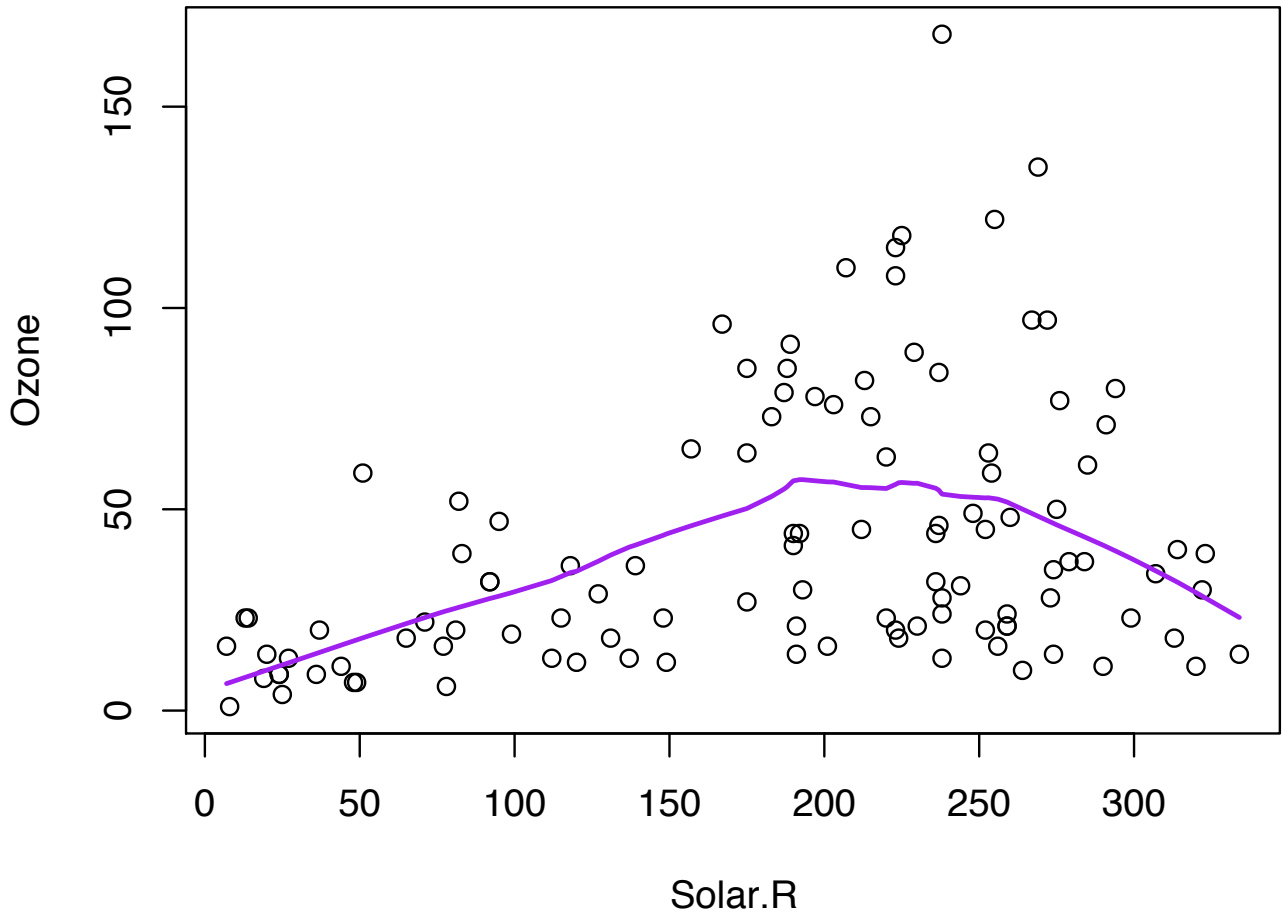
Conditioning plots

A typical goal for measuring Z is to see whether the $Y - X$ relationship changes at different values of Z. For example, we might want to see if a Blood Pressure/genotype association varies by Body Mass Index (weight/height^2)

In this case, it's useful to show plots of Y against X conditioned on the value of Z, i.e. Y versus X for all data with Z in a small range. This is known as a *conditioning plot*, and can be produced with `coplot()`.

Ozone is a *secondary pollutant*, it is produced from organic compounds and atmostpheric oxygen in reactions catalyzed by nitrogen oxides and powered by sunlight. But looking at ozone concentrations in NY in summer (Y) we see a non-monotone relationship with sunlight (X) ...

Conditioning plots

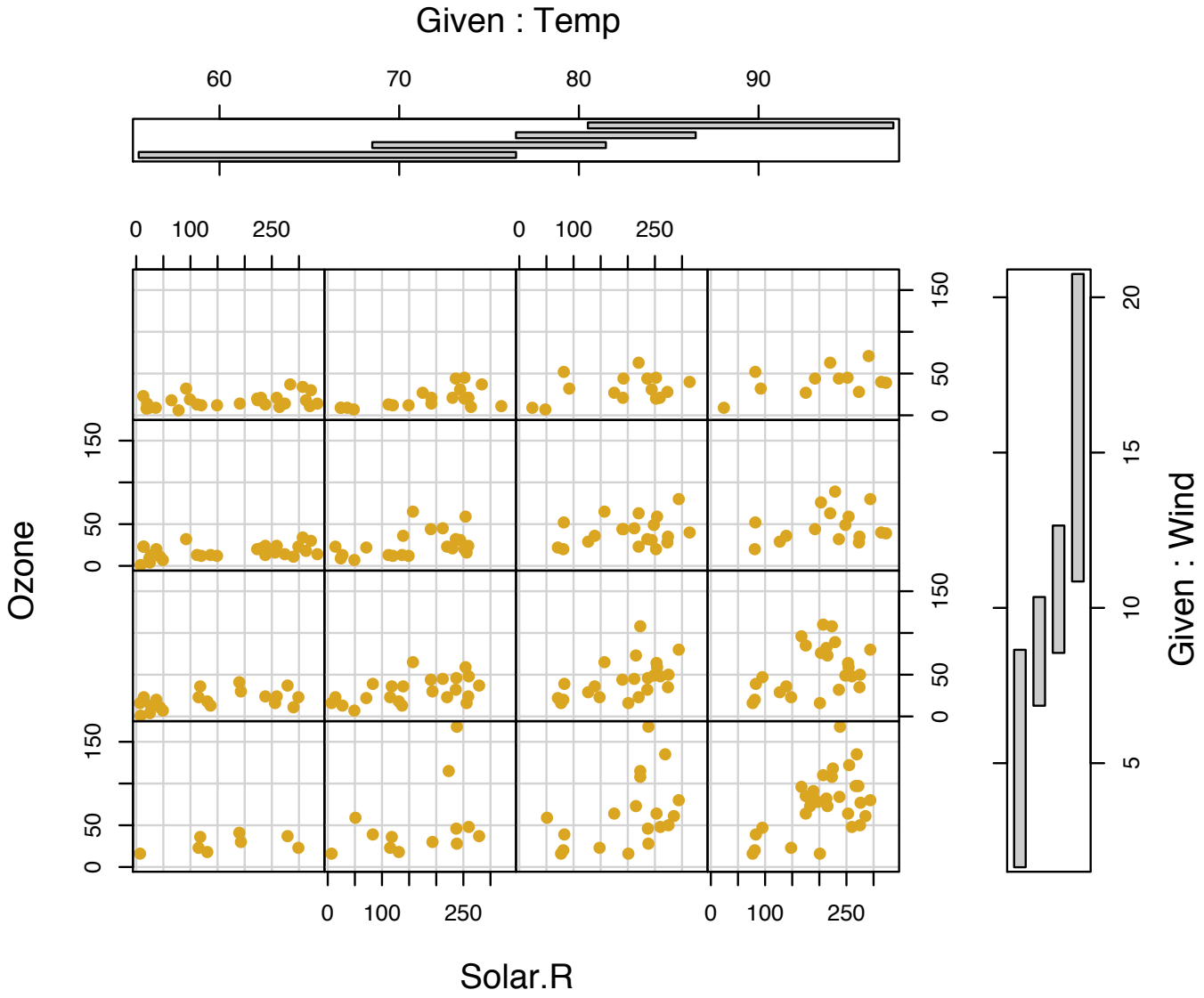


Conditioning plots

Now draw a scatterplot of `Ozone` vs `Solar.R` for various subranges of `Temp` and `Wind`. (For more examples like this, see the commands in the `lattice` package.)

```
data(airquality)
coplot(Ozone ~ Solar.R | Temp * Wind, number = c(4, 4),
     data = airquality,
    pch = 21, col = "goldenrod", bg = "goldenrod")
```

Conditioning plots



Conditioning plots

- A 4-D relationship is illustrated; the Ozone/sunlight relationship changes in strength depending on both the Temperature and Wind

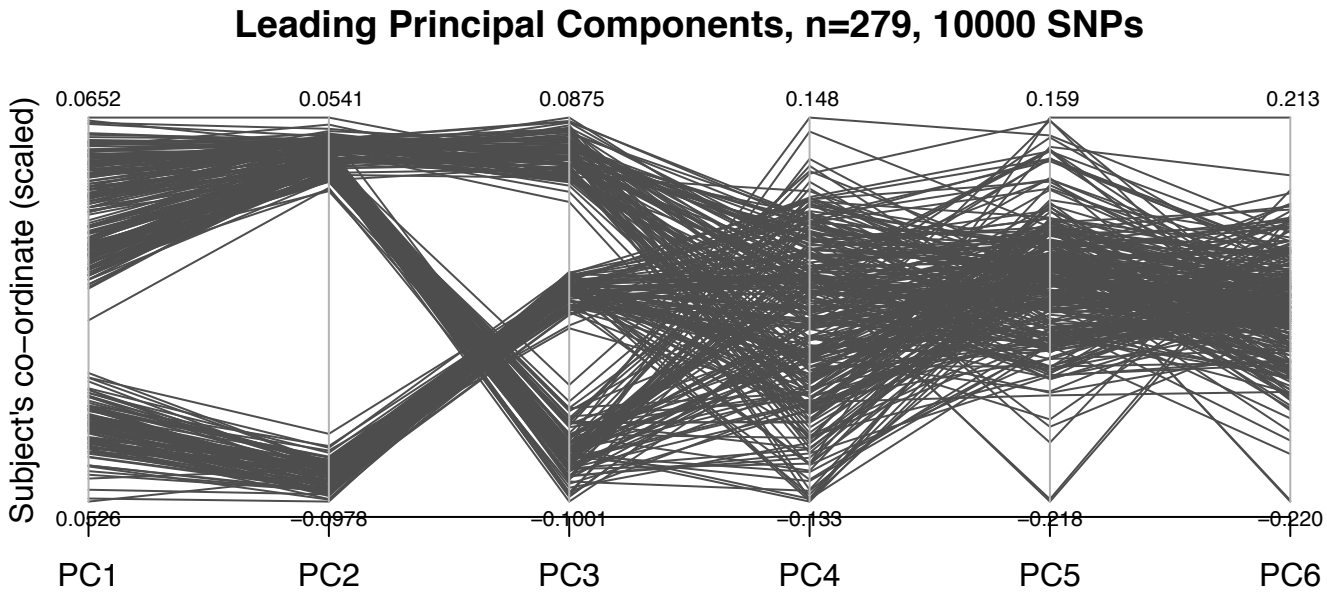
- The horizontal/vertical 'shingles' tell you which data appear in which plot. The overlap can be set to zero, if preferred

- `coplot()`'s default layout is a bit odd; try setting `rows`, `columns` to different values

- For more plotting commands that support conditioning, see `library(help="lattice")`

Parallel Coordinate Plots

For even higher-dimensional data, scatterplots can not provide adequate summaries. For data where the dimensions can be ordered, the *parallel co-ordinates plot* is useful;

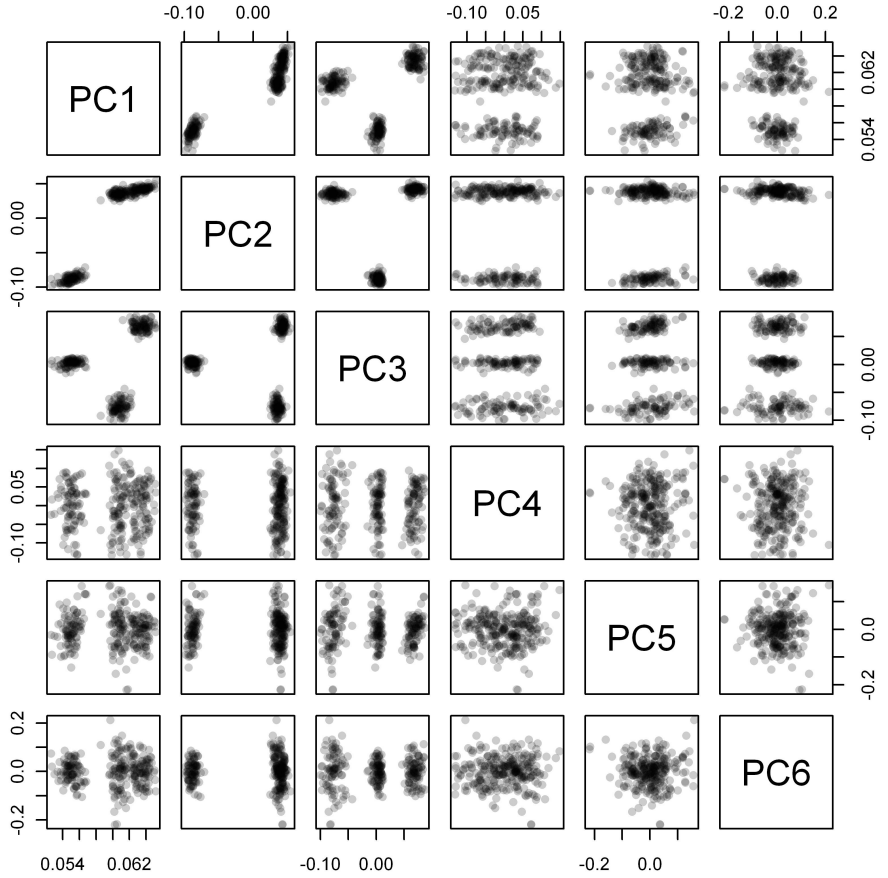


Parallel Coordinate Plots

- Each multi-dimensional data point (i.e. each person) is represented by a line – not a point
- `parcoord()` in the `MASS` package is one simple implementation – writing your own version is not a big job
- Coloring the lines also helps (example later)
- Scaling of axes, and their vertical positions are arbitrary
- Doing 'Principal Components Analysis' is just choosing axes for your data so that their variance is maximized on axis 1, then axis 2, ...

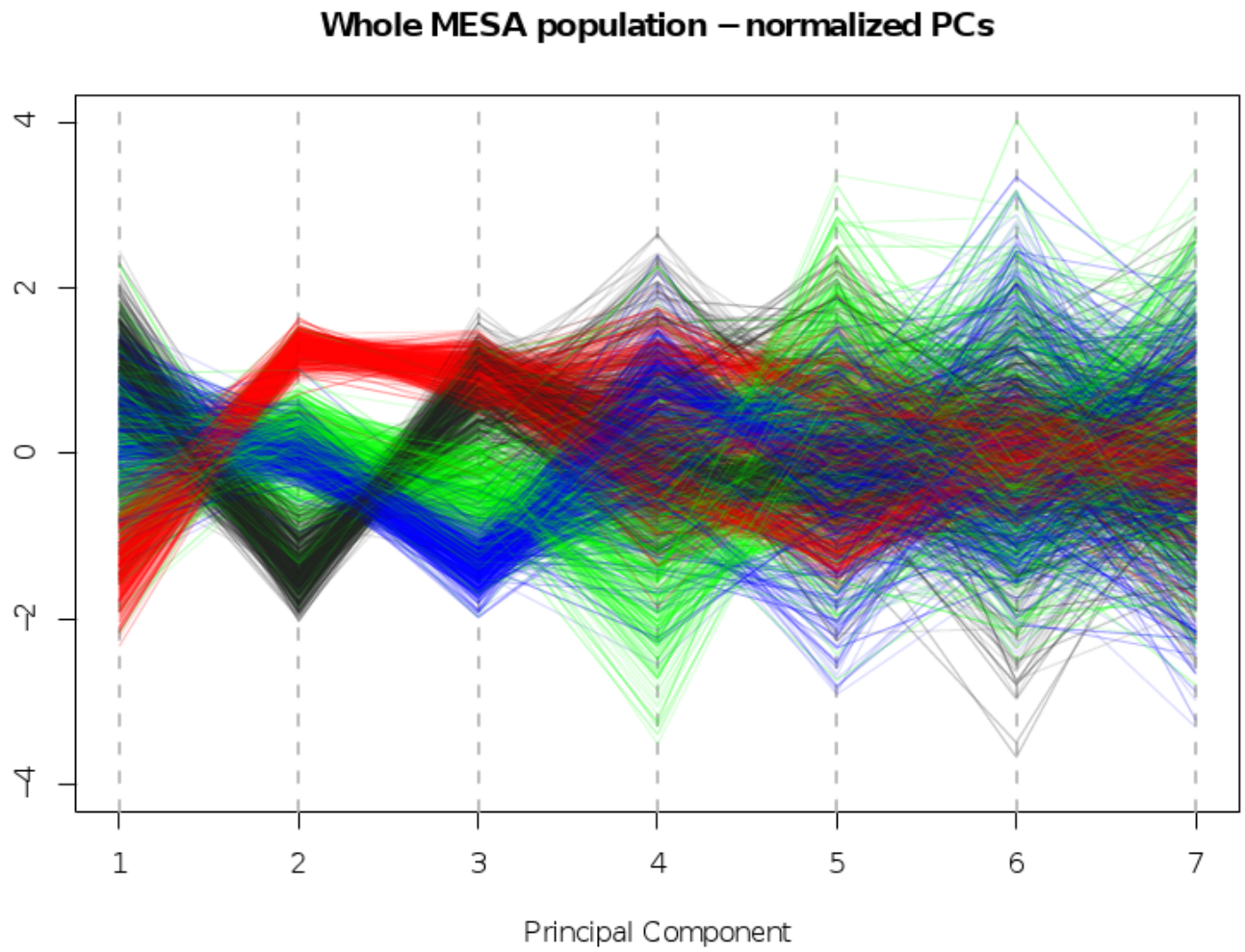
Parallel Coordinate Plots

A `pairs()` plot of the same thing; (nasty!)



Parallel Coordinate Plots

The pin cushion data++ : colors indicate self-report ancestry



Transparency

The colors in the last examples were *transparent*. As well as specifiying e.g. `col=2` or `col="red"`, you can also specify

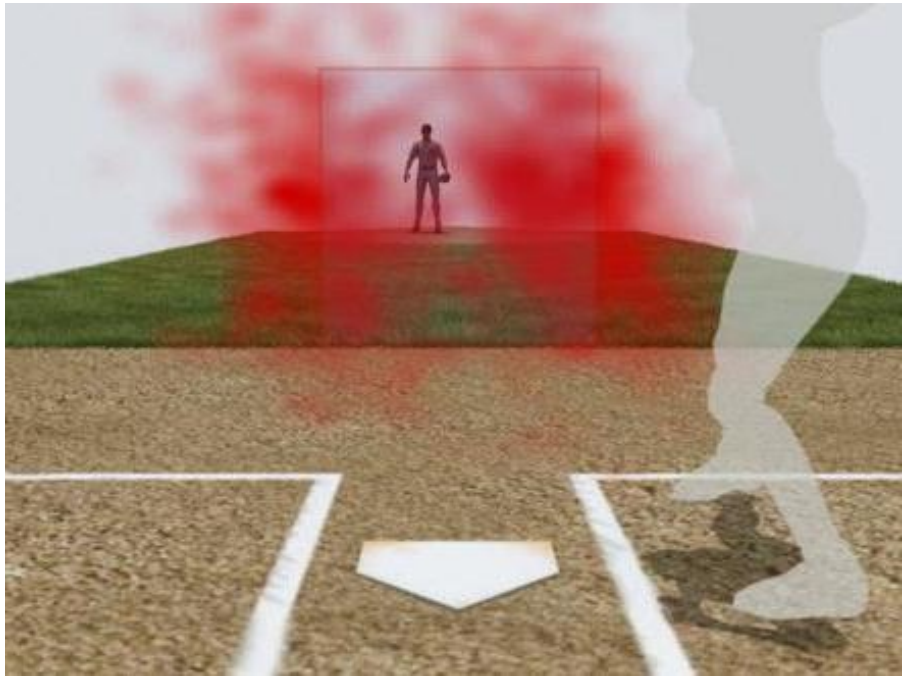
```
col="#FF000033"
```

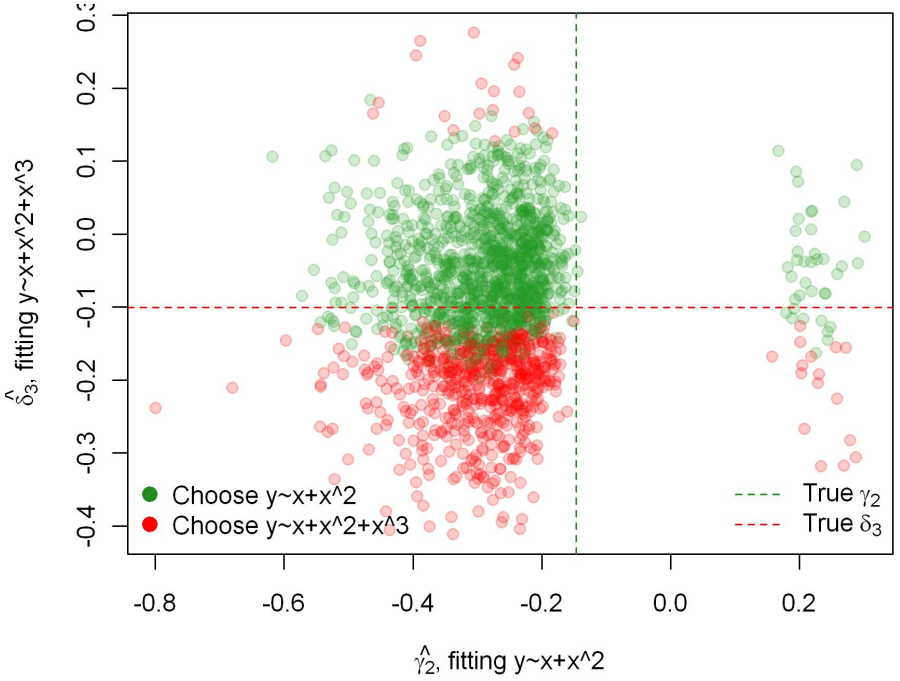
– coded as RRGGBB in hexadecimal, with transparency 33 (also hexadecimal). This is a 'pale' red – 33/FF $\approx$ 20%.

Get from color names to RGB with `col2rgb()`, and from base 10 to base 16 using `format(as.hexmode(11), width=2)`

Transparency

A couple more examples;





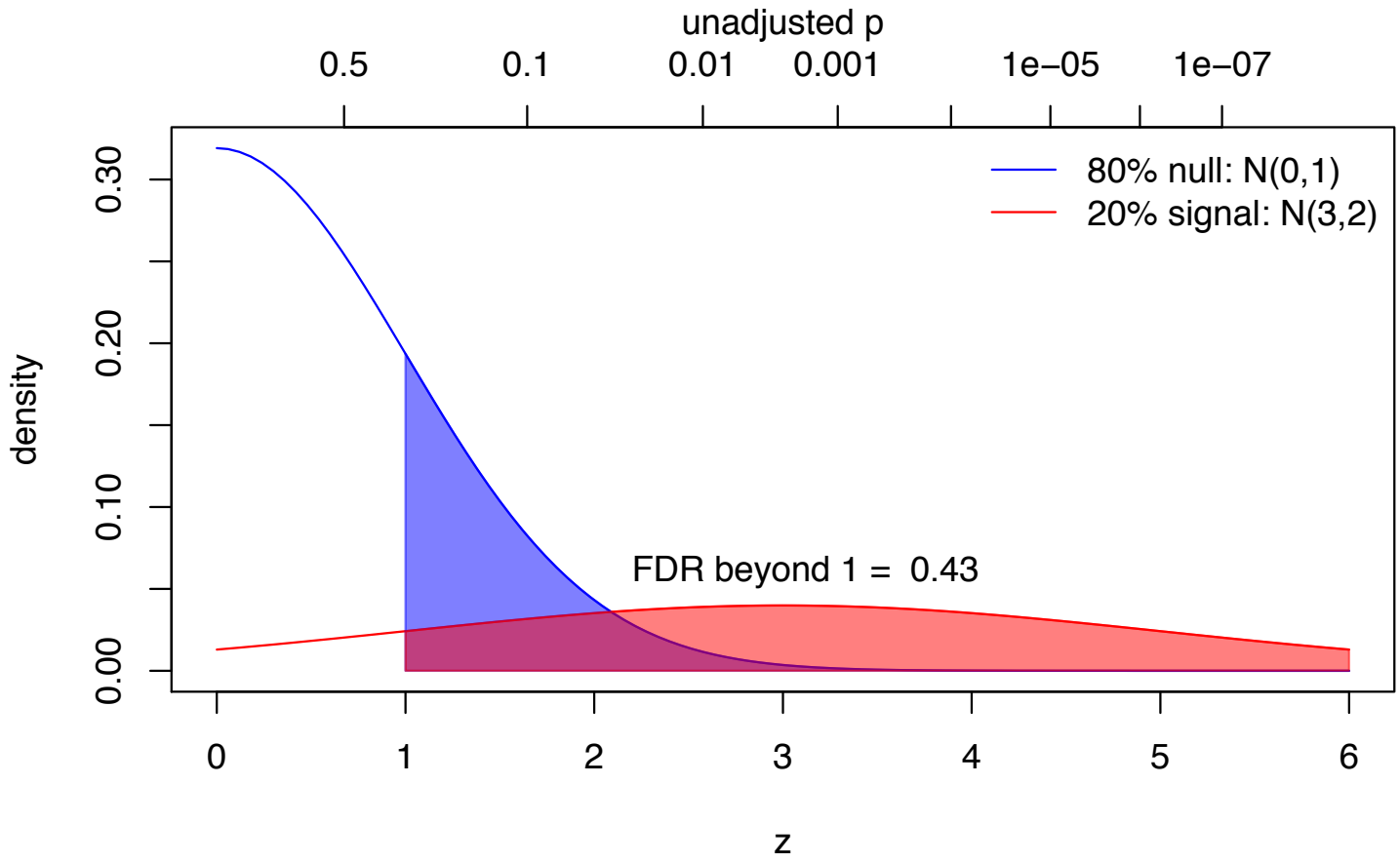
Transparency

R code for another; (also shows other graphics commands)

```
curve(0.8*dnorm(x), 0, 6, col="blue", ylab="density", xlab="z")
curve(0.2*dnorm(x,3,2), 0, 6, col="red", add=T)

xvals <- seq(1, 6, l=101)
polygon(
c(xvals,6,1), c(0.8*dnorm(xvals), 0,0),
density=NA, col="#0000FF80" ) # transparent blue
polygon(
c(xvals,6,1), c(0.2*dnorm(xvals,3,2), 0,0),
density=NA, col="#FF000080" ) # transparent red

legend("topright", bty="n", lty=1, col=c("blue","red"),
c("80% null: N(0,1)", "20% signal: N(3,2)"))
axis(3, at=qnorm(c(0.25, 0.5*10^(-1:-7)), lower=F), c(0.5, 10^(-1:-7)) )
mtext(side=3, line=2, "unadjusted p")

text(2.2, 0.07, adj=c(0,1), paste("FDR beyond 1 = ",
round(0.8*pnorm(1,lower=F)/(0.8*pnorm(1,lower=F) + 0.2*pnorm(1,3,2,lower=F)),3)))
```

Transparency

Here's the output;



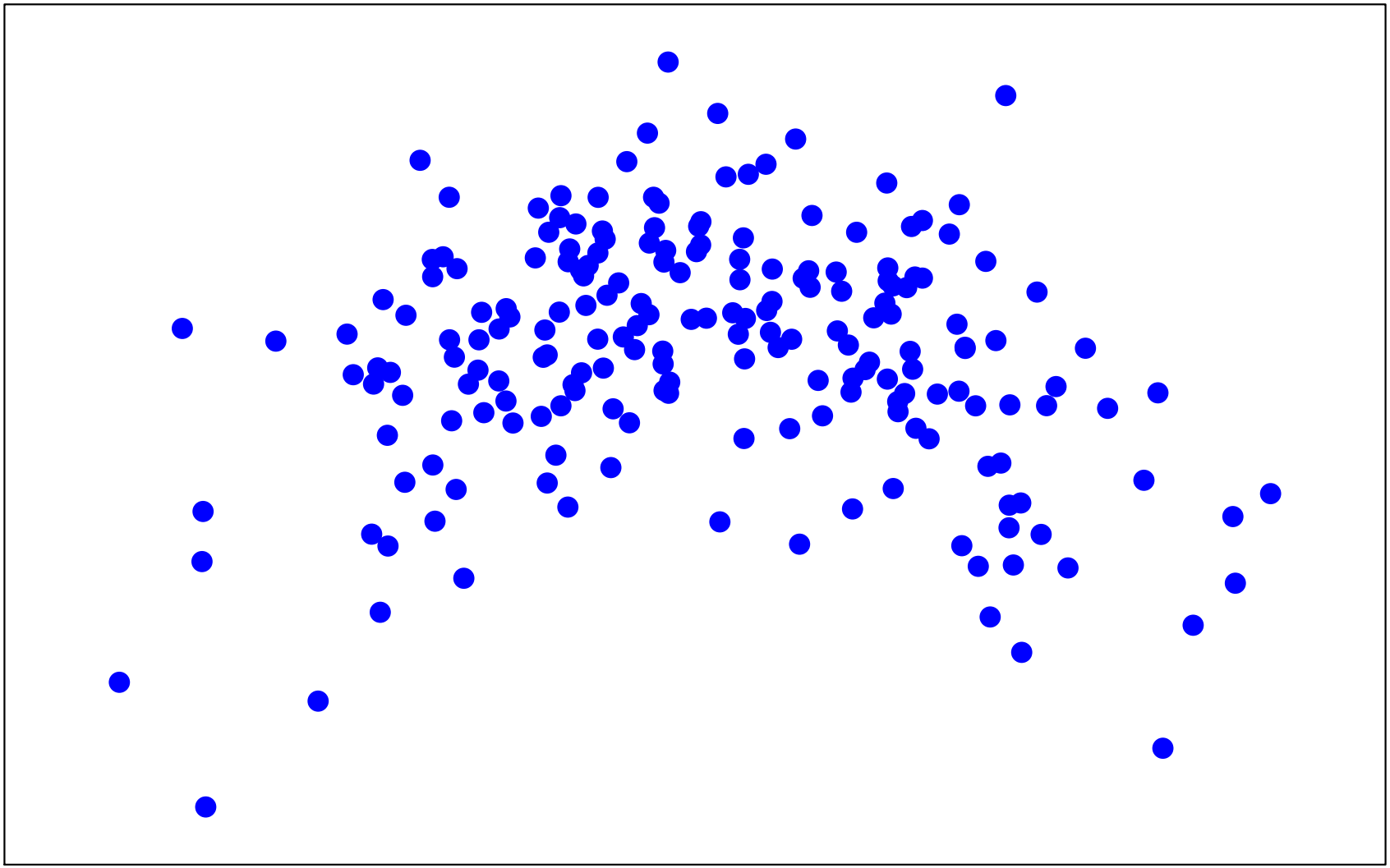
Hexagonal binning

Using transparent plotting symbols is a quick-and-dirty way to adapt scatterplots for use with large datasets.

A better method is ‘hexagonal binning’; this is a 2D analog of a histogram – where you would count the number of data in one area, and then draw a bar with height proportional to that count.

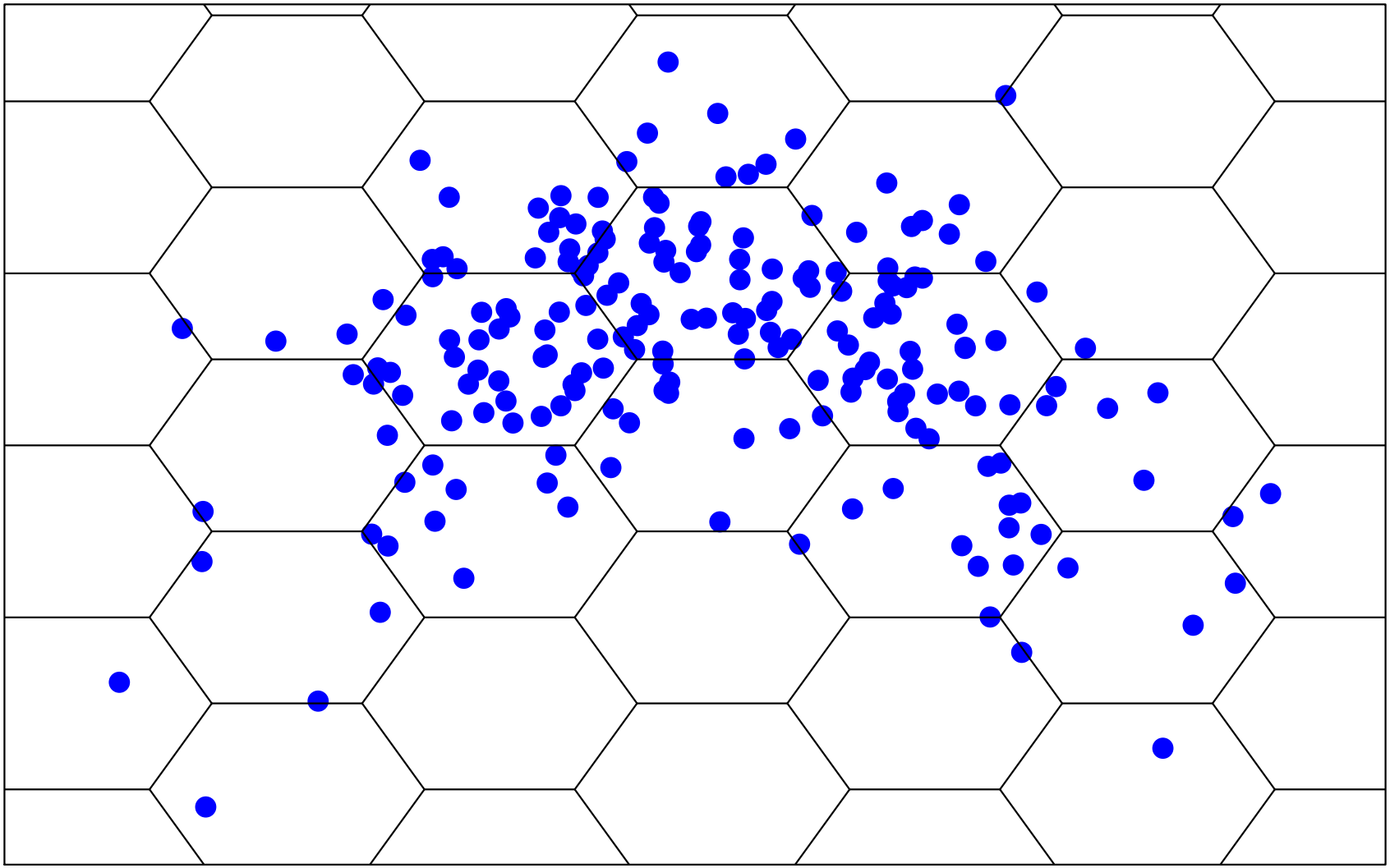
Hexagonal binning

Binning in two dimensions;



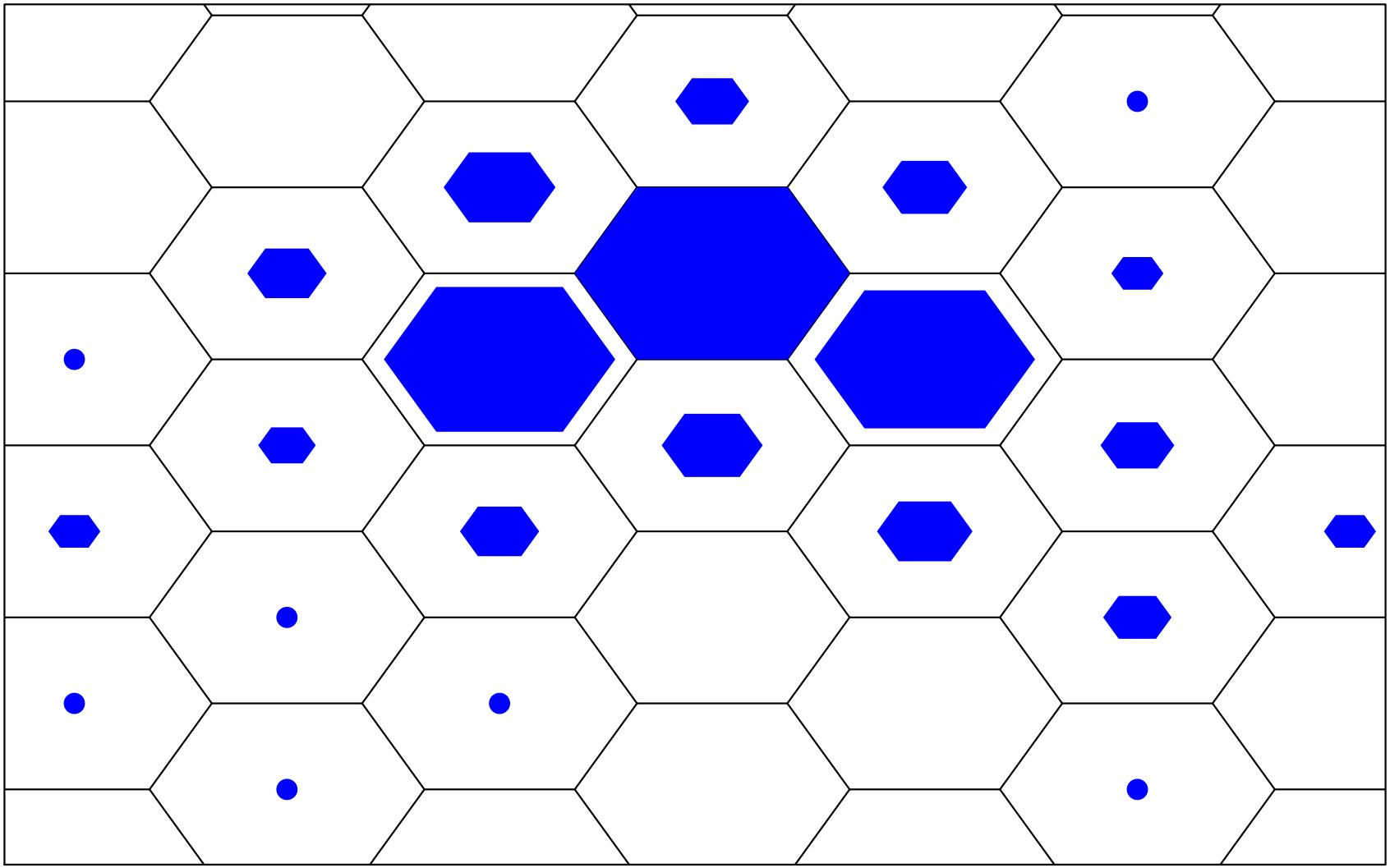
Hexagonal binning

Binning in two dimensions;



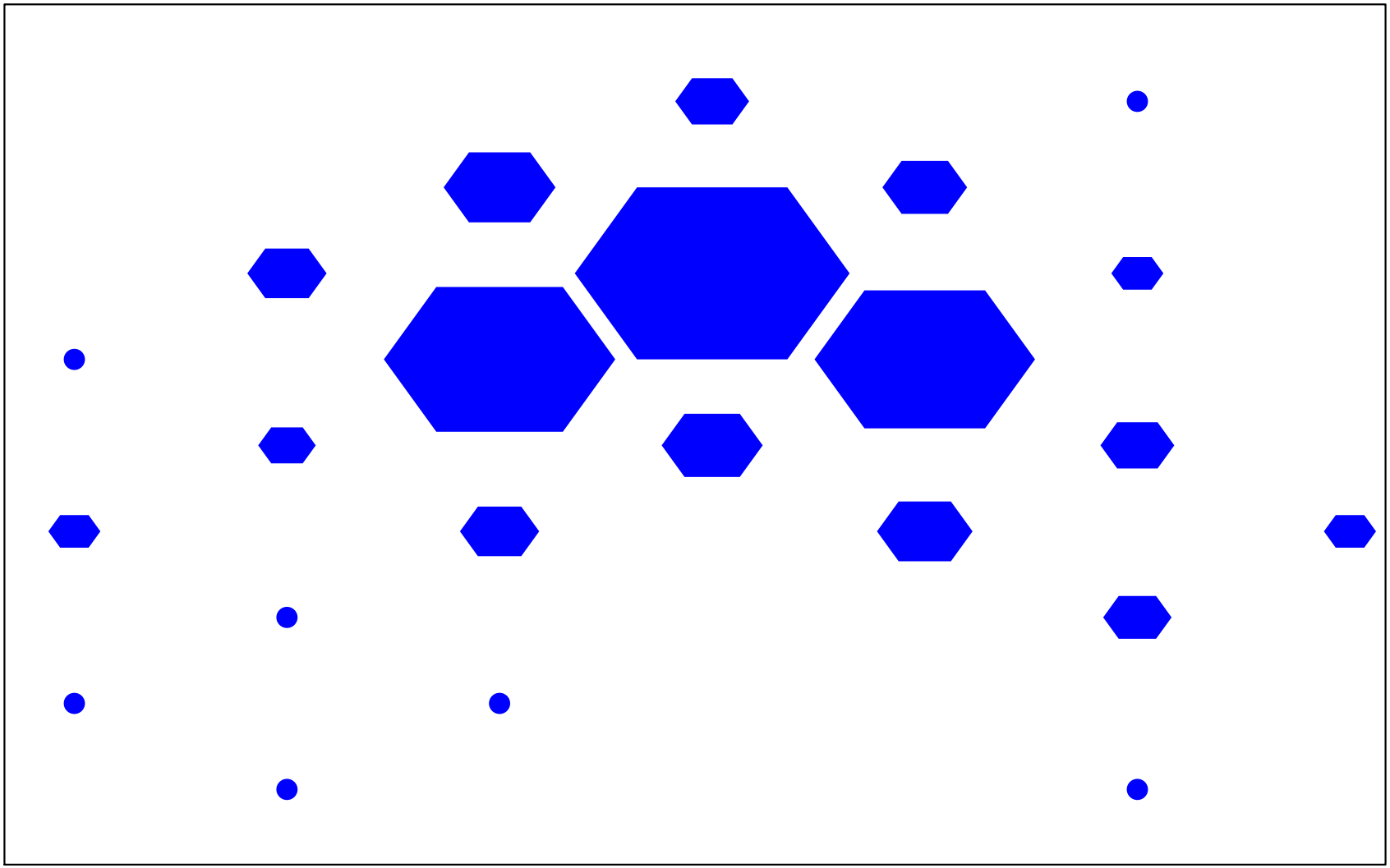
Hexagonal binning

Binning in two dimensions;



Hexagonal binning

Binning in two dimensions;

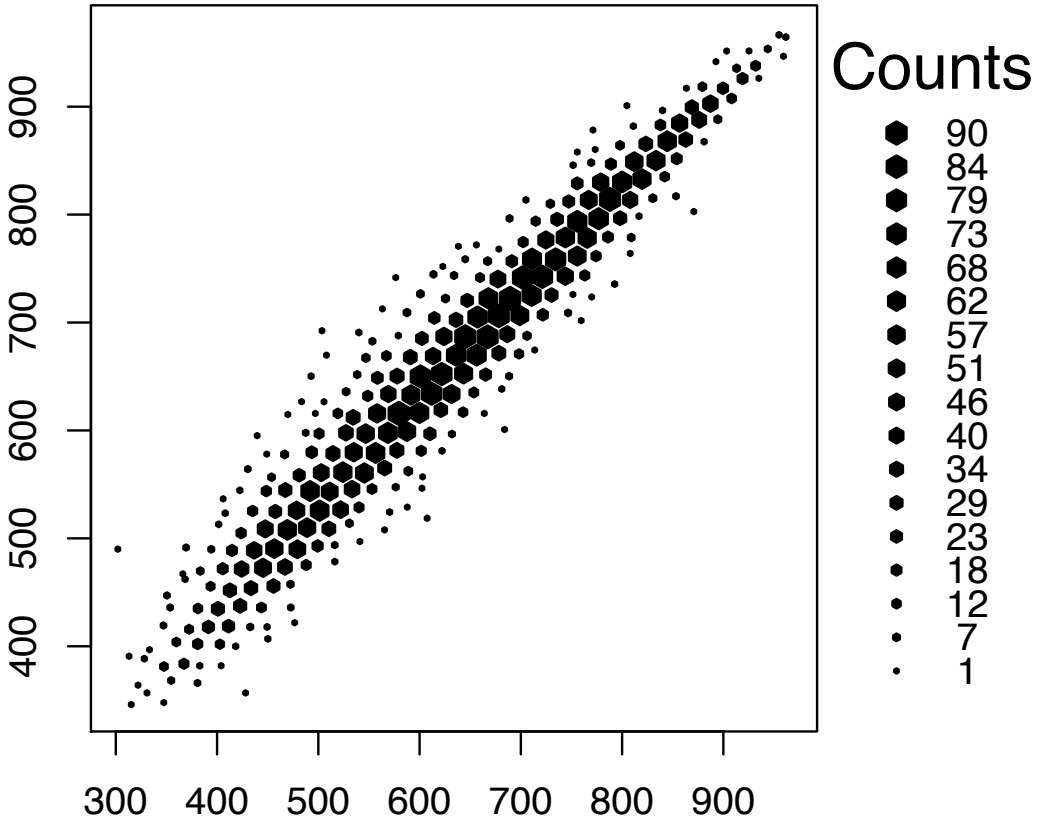


Hexagonal binning

The `hexbin()` package does all the bin construction, and counting. It has a `plot` method for its `hexbin` objects;

```
install.packages(c("hexbin","survey"))
library("hexbin")
library("survey")# for apipop data frame

with(apipop, plot(hexbin(api99,api00), style="centroids"))
```

Hexagonal binning



Hexagonal binning

Hexbin is used when you don't *really* care about the exact location of every single point

- Singleton points are plotted 'as usual'; you do (perhaps) care about them
- `hexbin` centers the 'ink' at the cell data's 'center of gravity'
- `style="centroids"` gives the center-of-gravity version; the default `style` is `colorscale` – usually grayscale. See `?gplot.hexagons` for more options

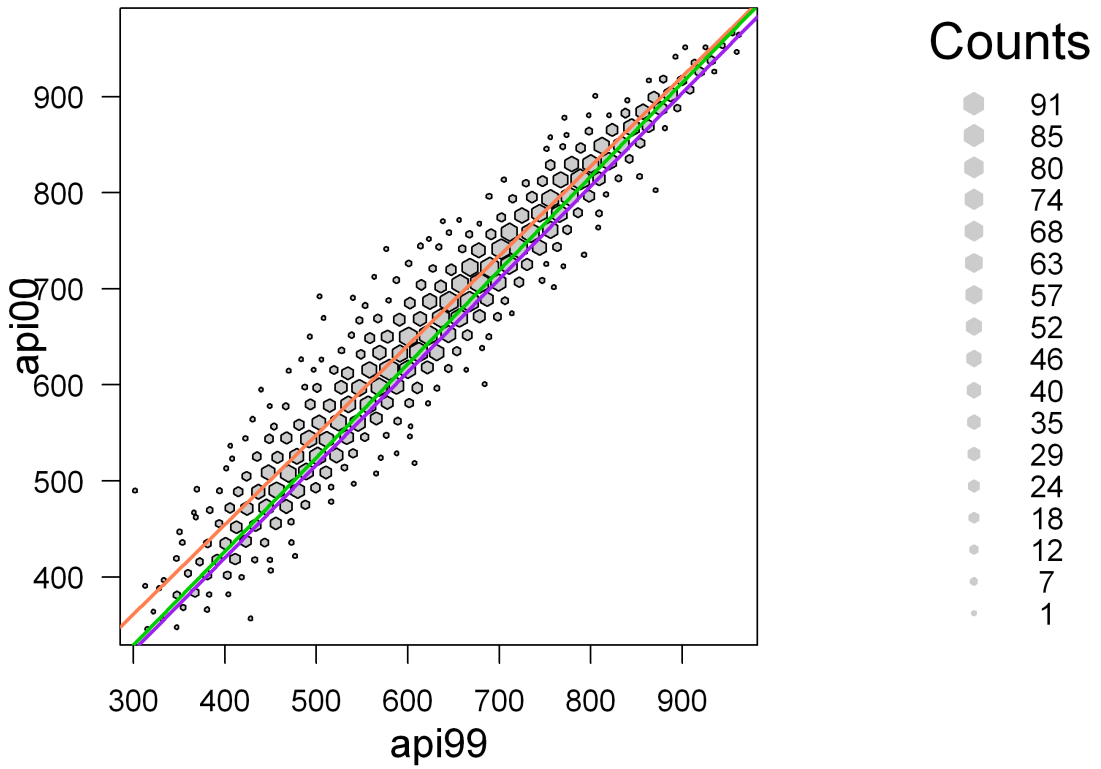
Hexagonal binning

For keen people: the `hexbin` package doesn't use the standard R graphics plotting devices; instead, it operates through the `Grid` system (in the `grid` package) which defines rectangular regions on a graphics device; these `viewport` regions can have a number of coordinate systems. To add lines to a hexbin plot, the options are;

- Use `hexVP.abline()` to add these directly
- Move everything into 'standard' graphics – not `Grid` graphics (see `?Grid`. This system lets you alter graphics *after* plotting them
- Write your own plot method for hexbin objects, with standard R graphics commands

Hexagonal binning

An example; color-coded lines of best fit, by school type;




```
lm.e <- coef(lm(api00~api99, data=apipop, subset=stype=="E"))
lm.m <- coef(lm(api00~api99, data=apipop, subset=stype=="M"))
lm.h <- coef(lm(api00~api99, data=apipop, subset=stype=="H"))

hexVP.abline(vp1$plot.vp, lm.e[1], lm.e[2], col="coral")
```

Graphics for simulation studies

Simulations studies are very common in methods work

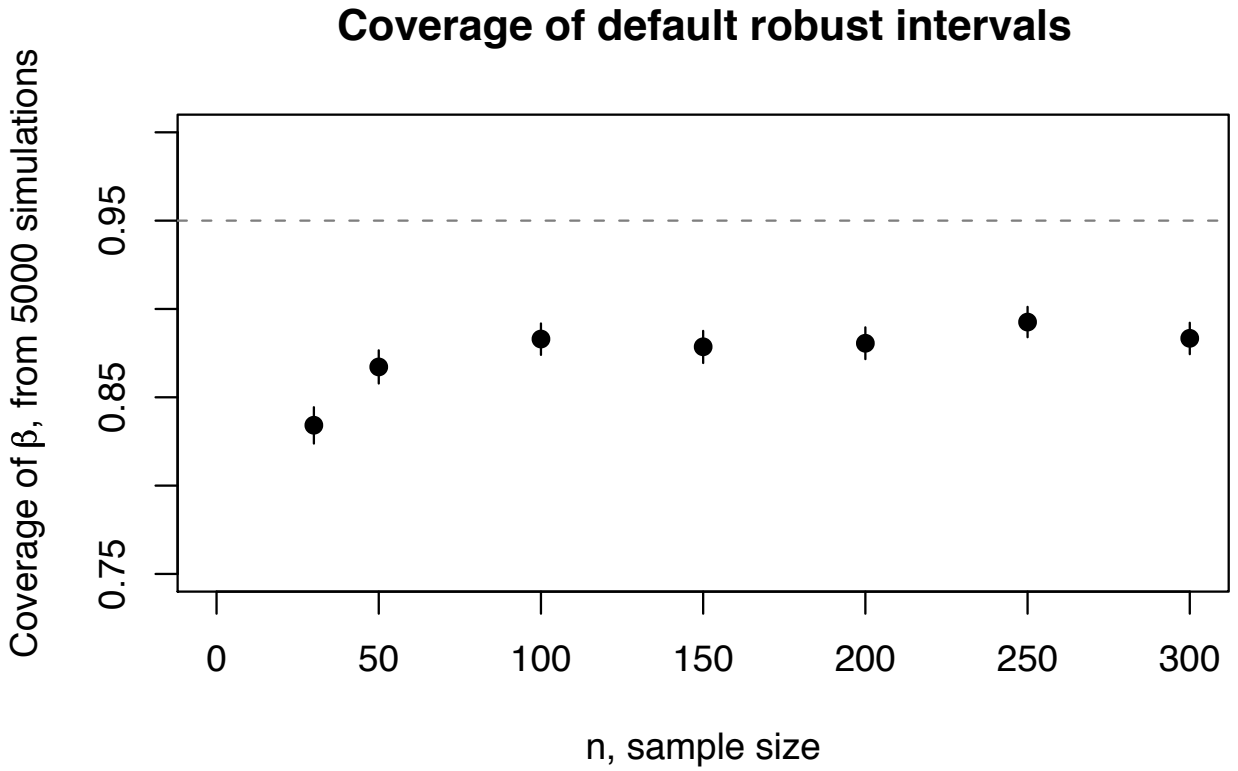
- Tables of estimated coverages (all near 95%) are very common
- Tables of estimated coverages (all near 95%) are *immensely* boring
- Graphics are easier to comprehend – but
- Show the Monte Carlo error!

A game for seminars; before the speaker tells you, decide whether they will say;

- "Look how **different** these lines are – **and** mine is best!"
- "Look how **similar** these lines are – **but** mine is best!"

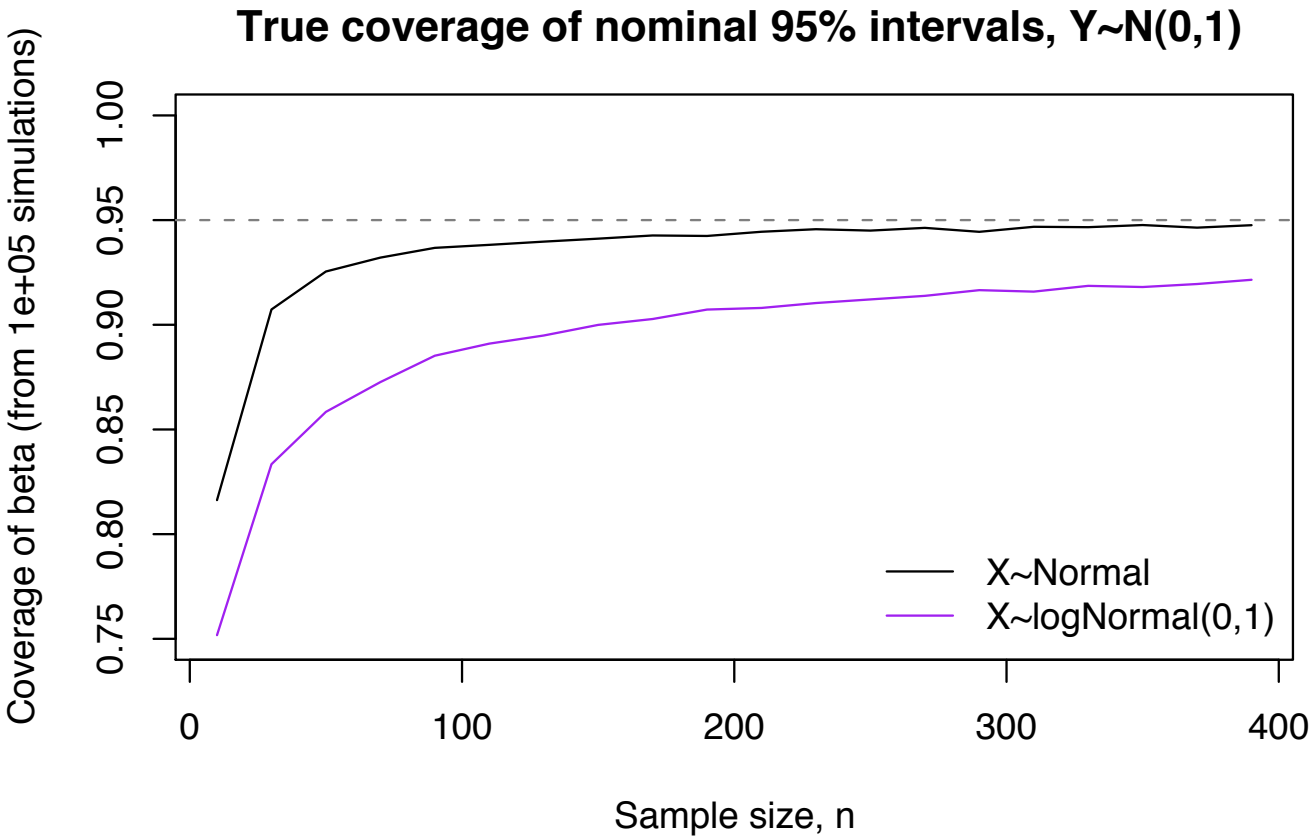
Graphics for simulation studies

From 533; (violating regularity conditions)



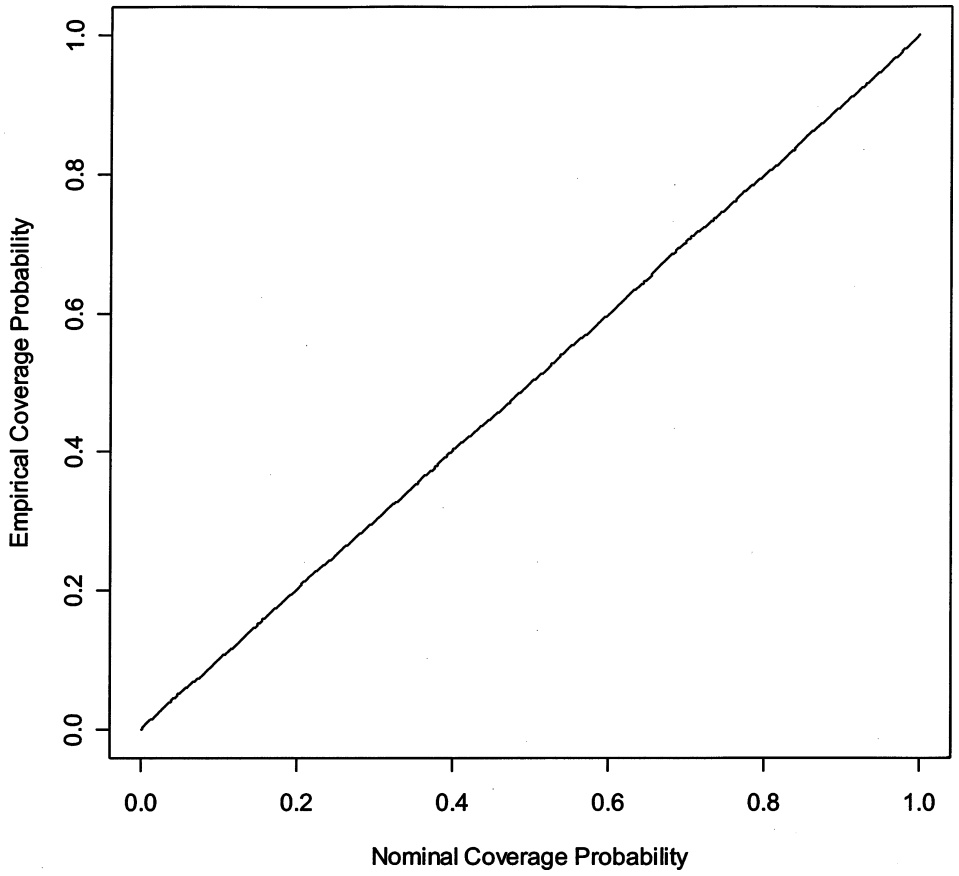
Graphics for simulation studies

From 533; (negligible Monte Carlo error)



Graphics for simulation studies

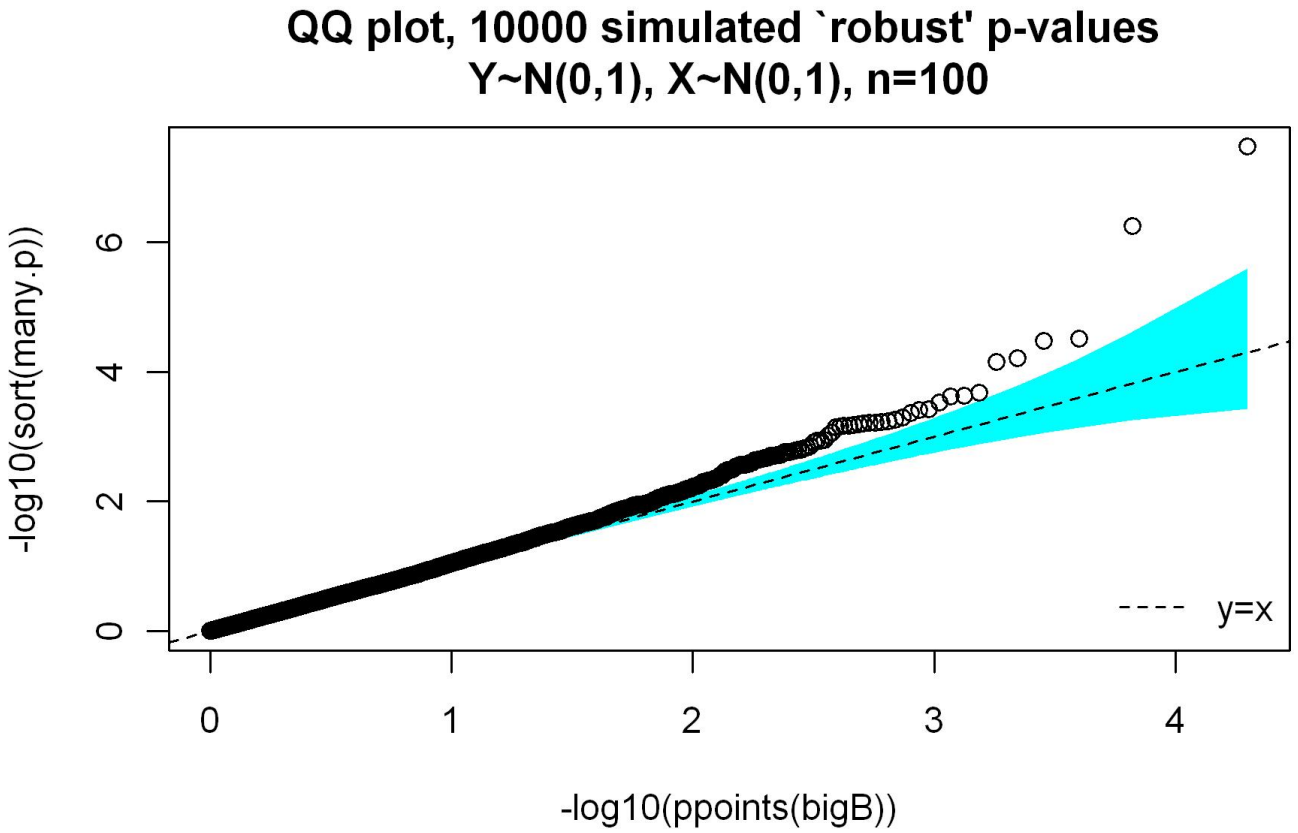
Here's a very bad display of many p-values;



Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am Jrnl Hum Genet* 73:1316-1329

Graphics for simulation studies

Here's a better one;



Behavior of small p-values is of interest (recall the favorite color example)

File formats

Ultimately, we want to output the graph in an appropriate file format. (Cut-and-paste is possible, but not recommended)

R knows more about font sizes and spacing than most users – so first design the graph at the size it will end up, eg:

```
## on Windows
windows(height=4,width=6)
## on Unix
x11(height=4,width=6)
```

... and, when that's done, write a version to a file

File formats

For example, for a 6×4 PDF file;

```
pdf("myprettypic.pdf", height=4, width=6) # inches
... plotting commands here ...
dev.off()                 # close the file
```

Some other formats: (see `?Devices` for a full list)

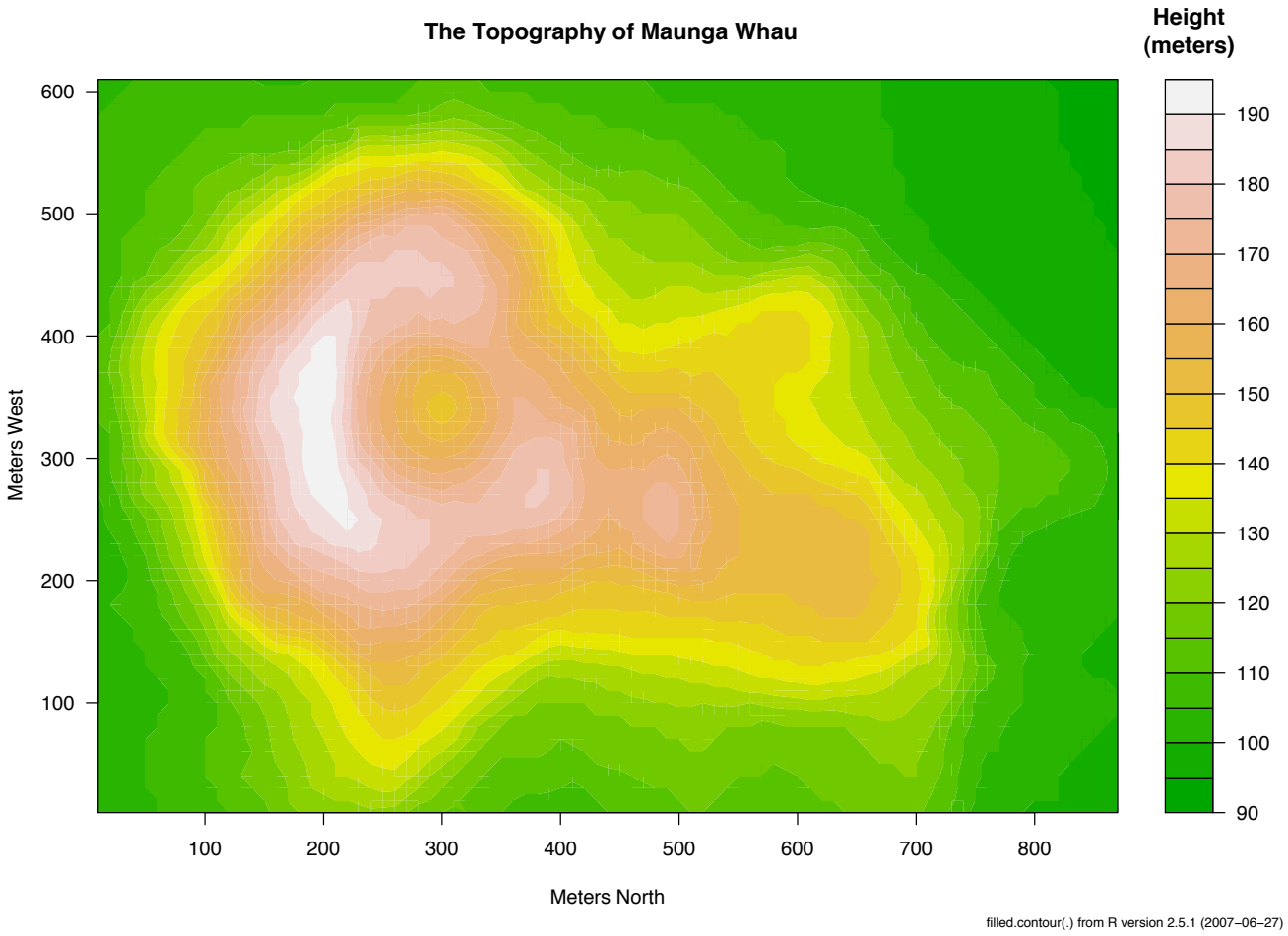
- `jpeg("mypic.jpg", w=6*288, h=4*288, res=288)` – lossy
- `png("mypic.png", w=6*288, h=4*288, res=288)` – lossless

– point size of text can also be manipulated, which can be useful when making posters

PowerPoint, or Word, or LaTeX can all rescale graphs. But when the graph gets smaller, so do the axis labels...

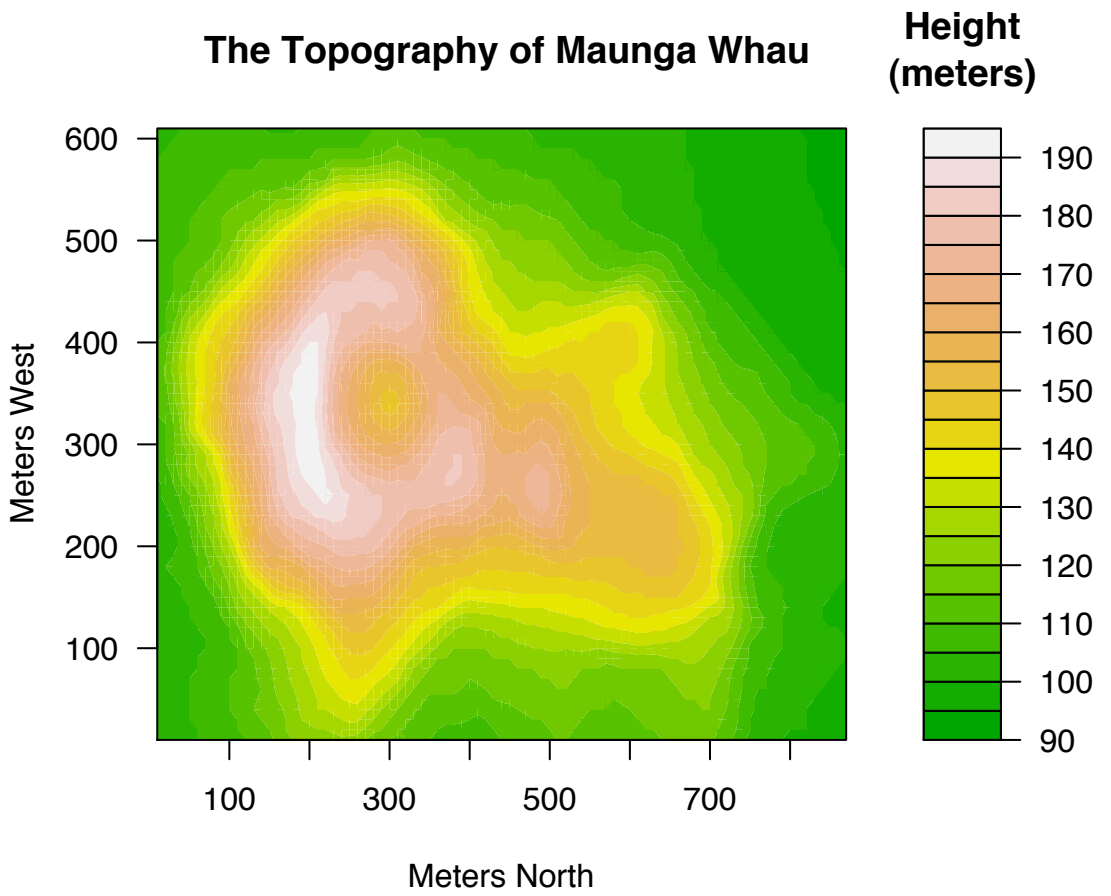
File formats

Created at full-page size (11×8.5 inches)



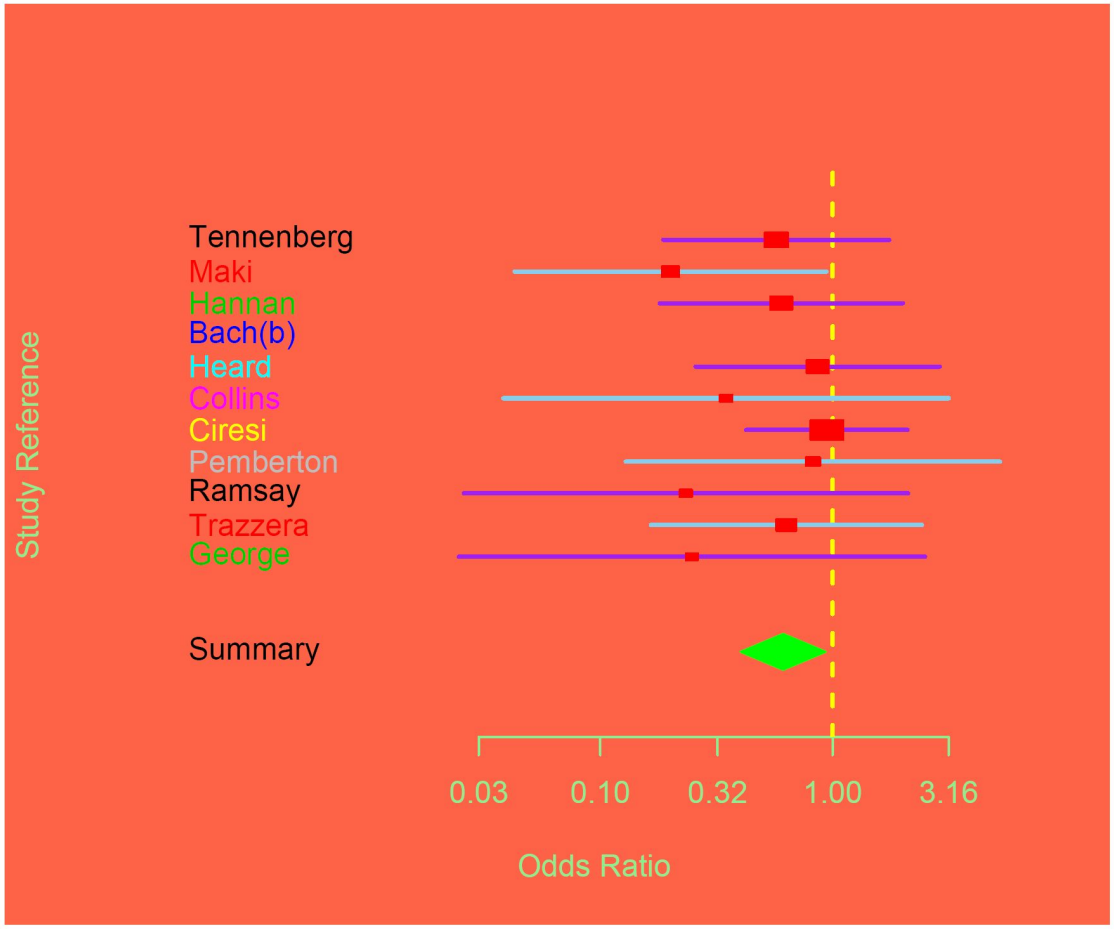
File formats

Created at 6×5 inches



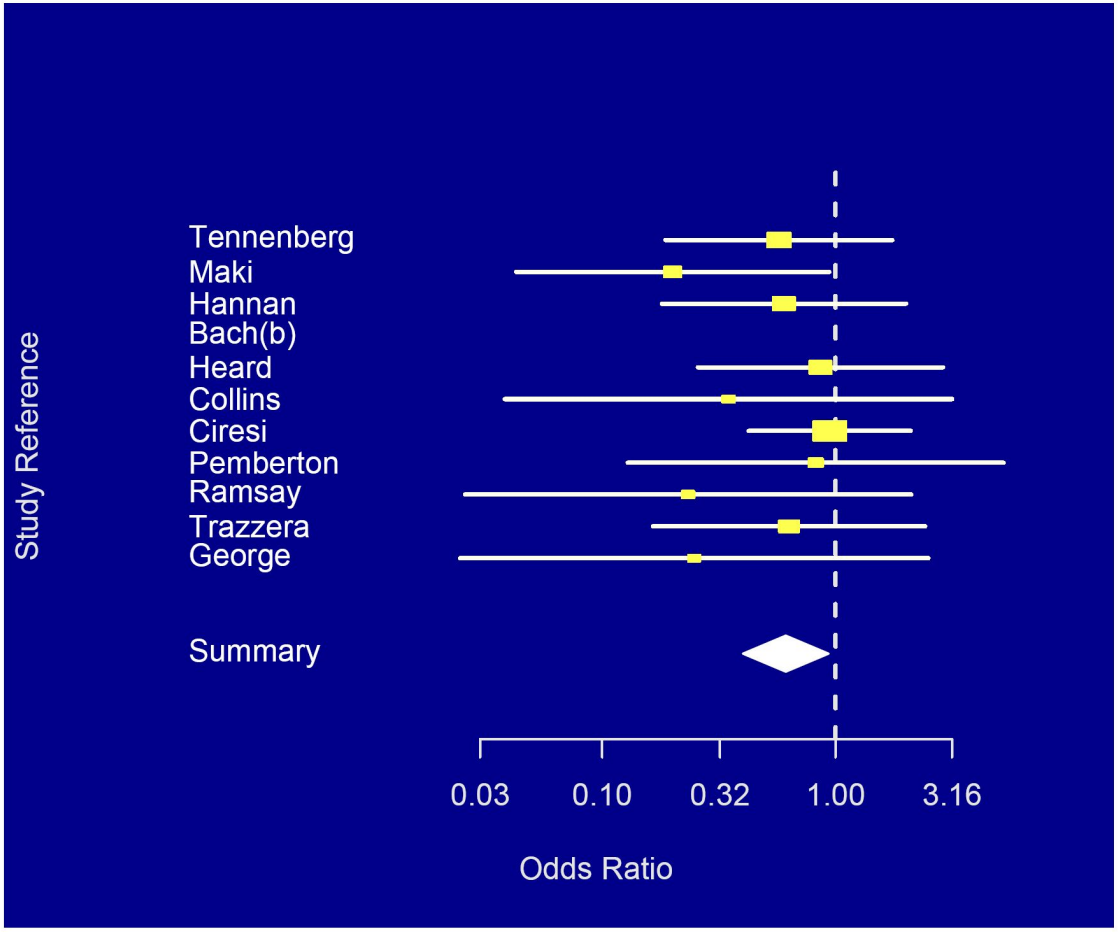
Color schemes

The choice is not just 'does it look cool'?



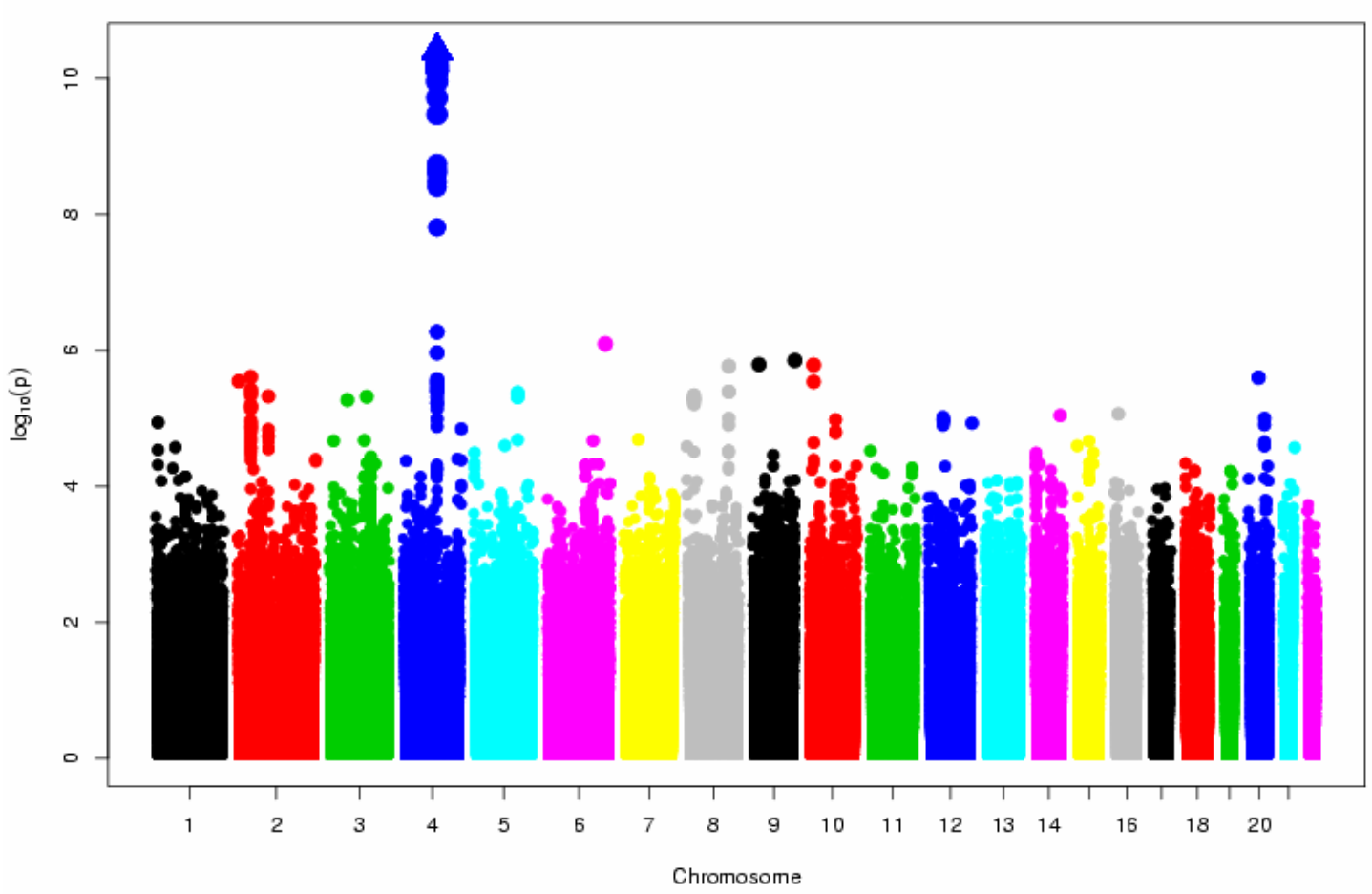
Color schemes

The choice is not just ‘does it look cool’?



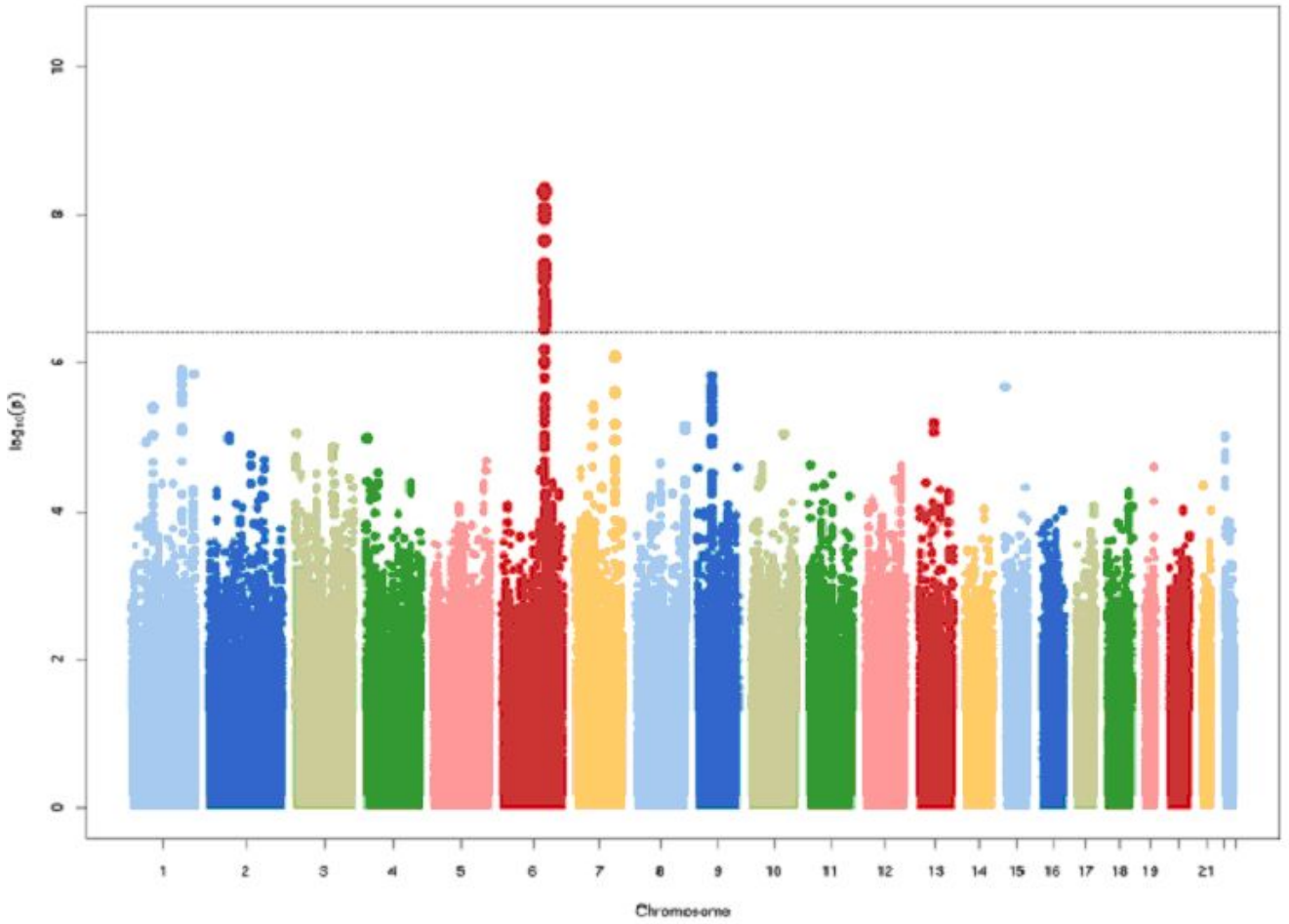
Color schemes

Which blobs of color stand out?



Color schemes

Which blobs of color stand out?



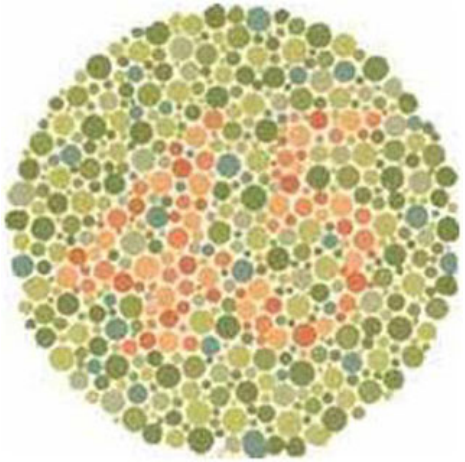
Color schemes

Color choice is best left to experts, or people with taste.

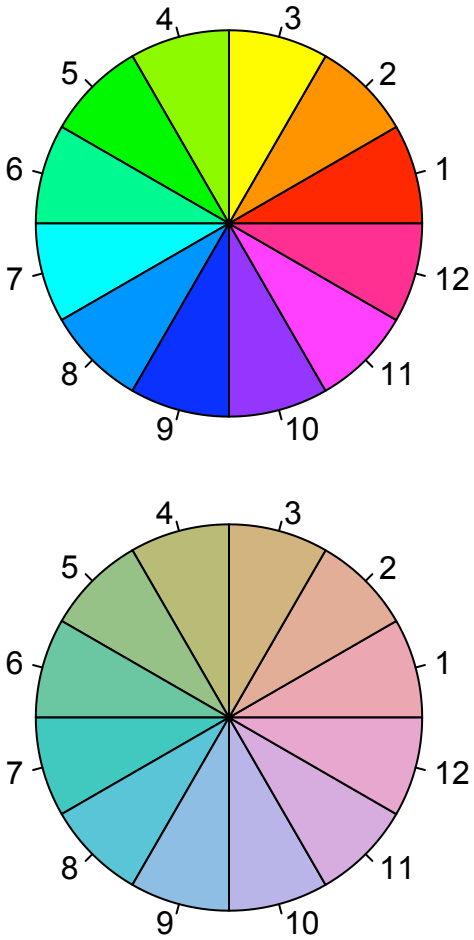
`http://www.colorbrewer.org` has color schemes designed for the National Cancer Atlas, also in package `RColorBrewer`

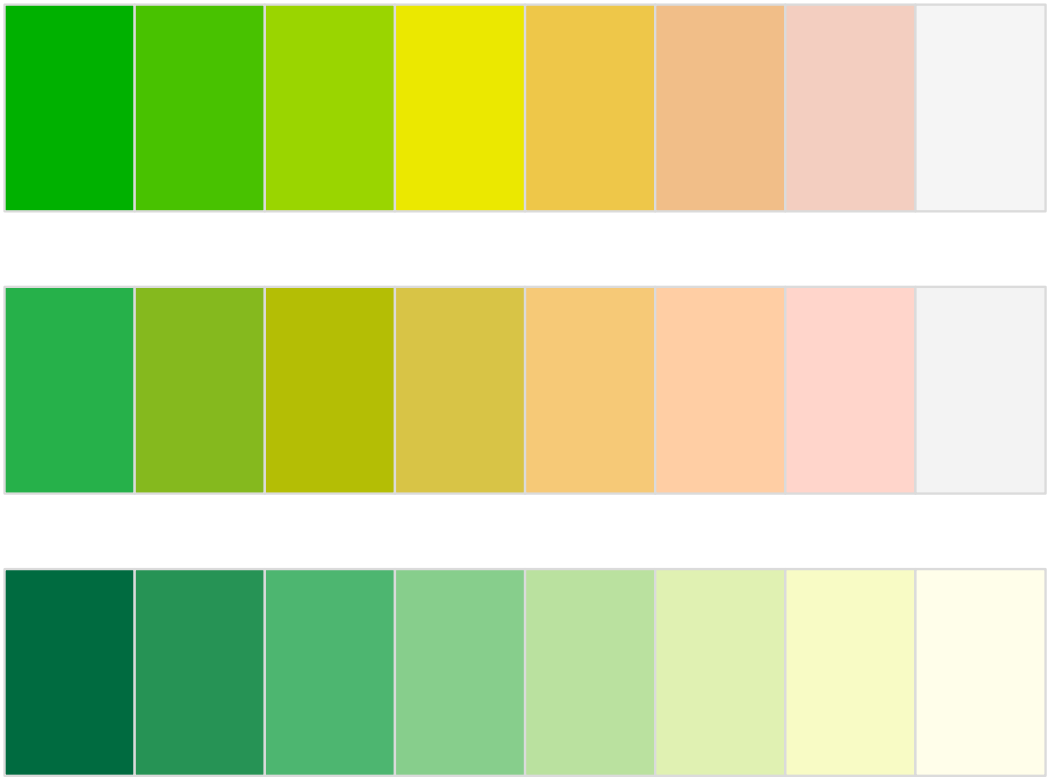
`colorspace` package has color schemes based on straight lines in a perceptually-based color space (rather than RGB).

`dichromat` package attempts to show the impact of red:green color blindness on your R color schemes.

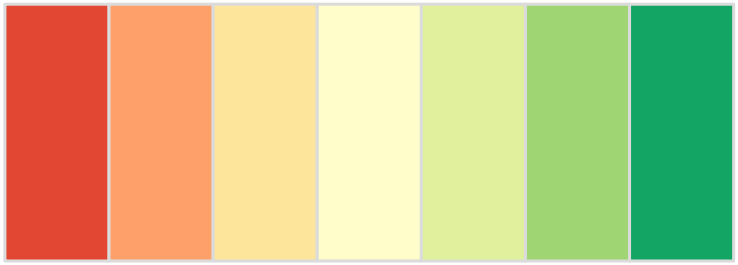


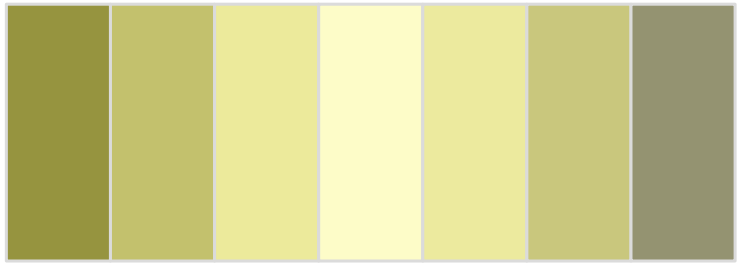
Color choice

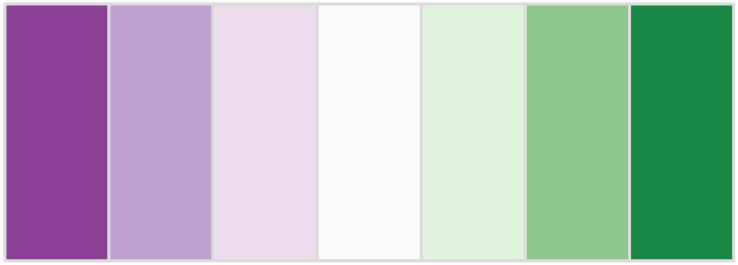


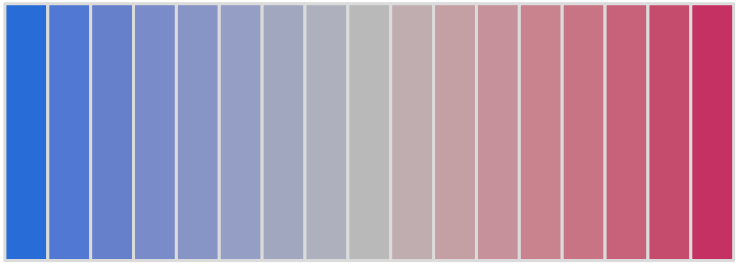


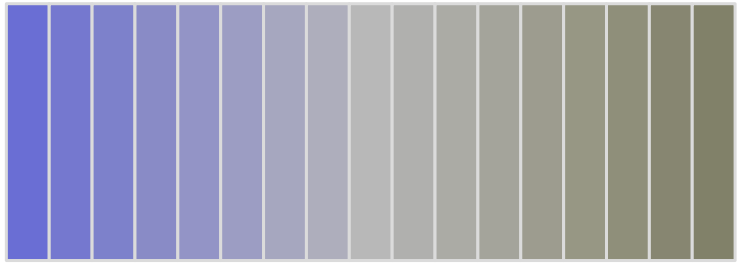
Color blindness









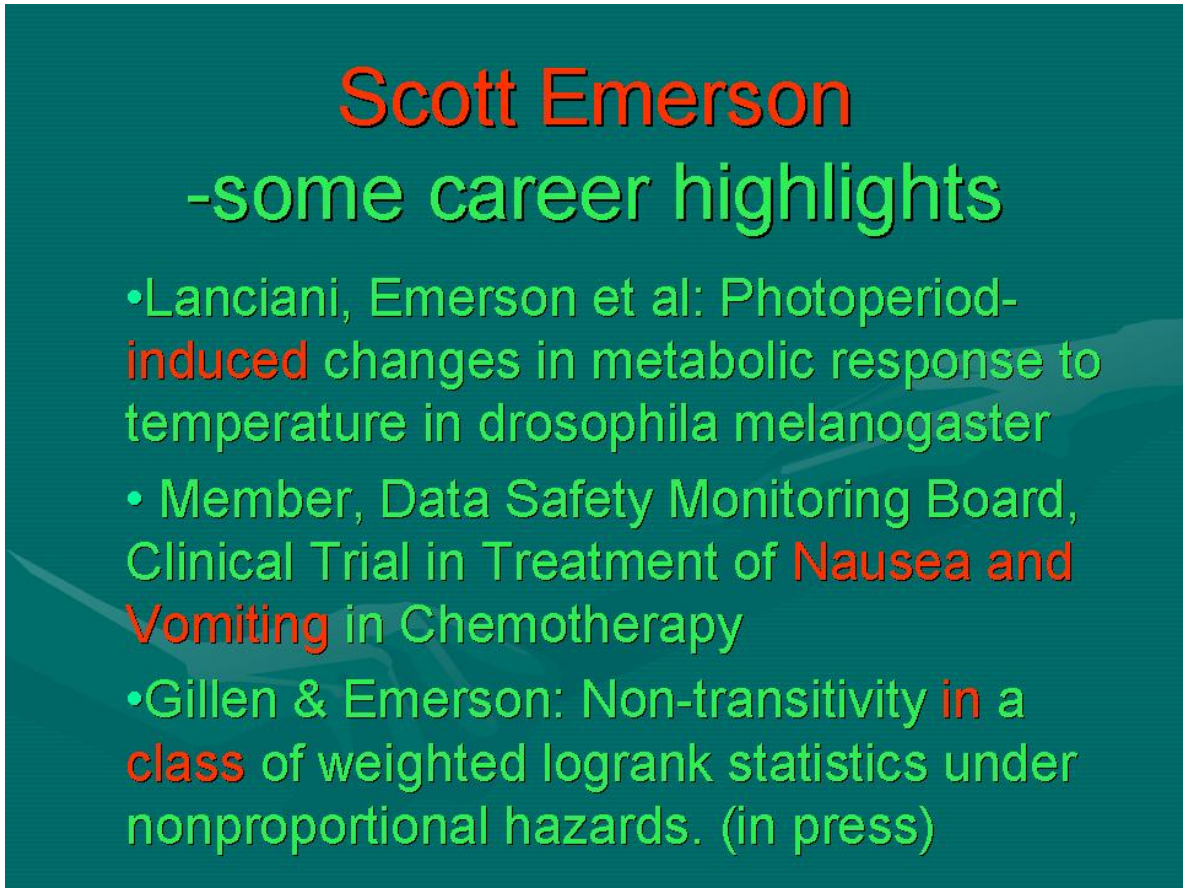


Color blindness

Color blindness is more common in men (5–10% of adults)

Color blindness

Color blindness is more common in men (5–10% of adults)



Resources

These slides are on the 572 course site;

`courses.washington.edu/biost572`

... along with the papers mentioned, and a few other resources

- See also Thomas Lumley's course – Specials Topics, on Visual Display of Quantitative Information
- Look around! Use other people's good ideas
- I collect horrible graphs – all donations gratefully received