Graphics (and numerics) for comparison

Nicholas J. Cox
University of Durham
n.j.cox@durham.ac.uk

Synopsis

Most statistical data analysis, and thus most graphical data analysis, is directed towards modelling of relationships, but many statistical problems have a different flavour: their focus is comparison, and the key question is assessing agreement or disagreement between two or more data sets or subsets with variables measured in the same units.

Here I survey (some of) the range of official and user-written graphical programs available in Stata 8 for such problems, with emphasis on making use of all the information in the data.

Themes

Recurrent themes include

◇ use reference lines, especially horizontal reference lines, to indicate benchmark cases

◇ don't juxtapose, superimpose!

◇ what methods work well at a range of sample sizes?

◇ stand on giant's shoulders: write wrappers around existing Stata commands

◇ what summary statistics are appropriate for comparison?

The biggest message of all is that

Previously...

I talked in London last year on graphics before (and especially after) modelling, particularly about a suite of commands which can follow a variety of modelling commands, not just `regress` or `anova`.

That project involved graphics for Stata 7, now obsolete. There is an ongoing project revising programs for Stata 8 and some results were shown at the Boston meeting.

Overheads for the Boston talk can be found at
`http://fmwww.bc.edu/repec/nasug2003/`
`CoxNASUG2003.pdf`

Homilies I

A good graph is a good answer to a good question…

The question could be **general**, often one of

◇ what are these data like?

◇ what patterns, trends, similarities, differences?

◇ what interesting, informative, puzzling detail?

⇒ we want graphs to provide **summary** and **exposure** and to show **both coarse and fine structure**
The best general graphs allow both:
e.g. dot plots or scatter plots, but arguably less so e.g. box plots or histograms.

Homilies II

A good graph is a good answer to a good question…

The question could be **specific**, as for example

◇ are these variables or subsets identical (as a reference case)?

◇ is difference constant (e.g. 0)?
is ratio constant (e.g. 1)?

◇ has a transformation done what we wanted? is distribution more nearly symmetric, Gaussian, homoscedastic?

⇒ we want graphs to answer the question directly without posing too many challenges (e.g. decoding or mental rotation)

An emerging issue

No dialogs* have been written for the user-written commands discussed in this talk.

Ars longa, vita brevis…

How many users (really) want dialogs? How many user-programmers will write dialogs?

A possible and valuable new role in the Stata user community – a new ecological niche – would be writing dialogs for other people's programs.

* dialogues, if you prefer

General: Distribution functions

cumulative probability or frequency as y
is plotted versus
value of variable as x

OS* `cumul` will calculate

SSC† `distplot` will graph for

◇ several variables
◇ **or** several subsets
◇ reverse (a.k.a. survival) functions
◇ transformed vertical scale
◇ different plot types

* OS means ‘official Stata’

† SSC means Boston archive: use `ssc`

General: Quantile functions

quantile as y
is plotted versus
'plotting position' as x

OS `quantile` is limited

STB-61* `quantil2` for Stata 7 is more flexible

`qplot` (under development) will graph

- ◇ several variables
- ◇ **or** several subsets
- ◇ reverse order
- ◇ transformed horizontal scale
- ◇ ranks
- ◇ different plot types

* STB means *Stata Technical Bulletin*

Two novel details

Different plot types may be obtained by a syntax of
command subcommand ...
Here we can exploit the structure of
`graph twoway`, as for example with
`distplot line` ... or
`qplot spike` ...
The menu is `area`, `bar`, `connected`, `dot`,
`dropline line`, `scatter`, `spike` for these two commands.

Transformed scales can be shown on any user-specified scale supplied e.g.
`tscale(invnorm(@))`
with `@` as placeholder for transformand

SSC `mylabels` is a utility to produce labels in the original scale
The idea goes back to Patrick Royston (STB-34, 1996)

Specific: Symmetry of distribution

OS `symplot` plots
(upper quantile − median) as y
versus
(median − lower quantile) as x

Symmetry implies that $y = x$
i.e. reference case plots as a sloping line

SSC `skewplot` plots
$\frac{1}{2}$ (upper quantile + lower quantile) as y
versus
(upper quantile − lower quantile) as x

Symmetry implies that $y =$ median
i.e. reference case plots as a horizontal line

You can plot several variables **or** several subsets

`skewplot` shows more detail than box plots!

Specific: Paired observations I

We often ask specific questions:
can the means be considered identical?

With scatter plots, it is easy to forget the precise question, especially as linearity $y = a + bx$ is more general than equality $y = x$ or constant difference $y = a + x$ or constant ratio $y = bx$.

The concordance correlation coefficient measures agreement (is $y = x$?). In many problems correlation is used as a measure of agreement, whereas it is a measure of linearity (is $y = a + bx$?).

See *Stata Journal* 2(2) (Tom Steichen and NJC) for `concord`. A Stata 8 version is in preparation.

Specific: Paired observations II

It is not easy to read off $y - x$ or y/x from a scatter plot.

John Tukey (and Doug Altman and Martin Bland) urged plotting difference versus mean, so that equality plots as a horizontal line.

SSC `pairplot` gives this as one of various choices. Others are ratio versus geometric mean, differences or ratios in specified sort order, etc.

`pairplot` is a wrapper around `twoway rspike || function`.

If differences or ratios answer the question of interest, show them directly!

General: Frequencies of categorical data

OS `histogram` is geared towards continuous variables, and must be coaxed when given categorical data.
OS `graph bar`, `hbar`, `dot` offer more scope, but for some basic problems (e.g. showing percents) you must calculate the variables to be shown beforehand.
This cries out for a wrapper.

SSC `catplot`* was originally written for teaching and can be used for 1, 2 or 3 categorical variables.
A structure of *command subcommand* allows `bar`, `hbar`, `dot` with almost identical syntax. Horizontal displays are preferable unless value labels are very short.

* Names can be difficult! A liking for cats was decisive here.

Specific: Frequencies of graded data

Graded data (e.g. opinion scales) are often plotted as stacked bars. A refinement is to split grades into those shown as positive and those shown as negative, i.e. to slide each bar so that it straddles zero.

Calculating frequencies or percents, adjusting some to negative values and stacking in a sensible order are trivial but also fiddly, so a wrapper saves the busy work.

SSC* `slideplot` does the graphics. Subcommands are `hbar` and `bar`.

A twin program, SSC* `majority`, shows tables of majorities, calculated from ‘votes’ deemed positive and negative.

* soon!

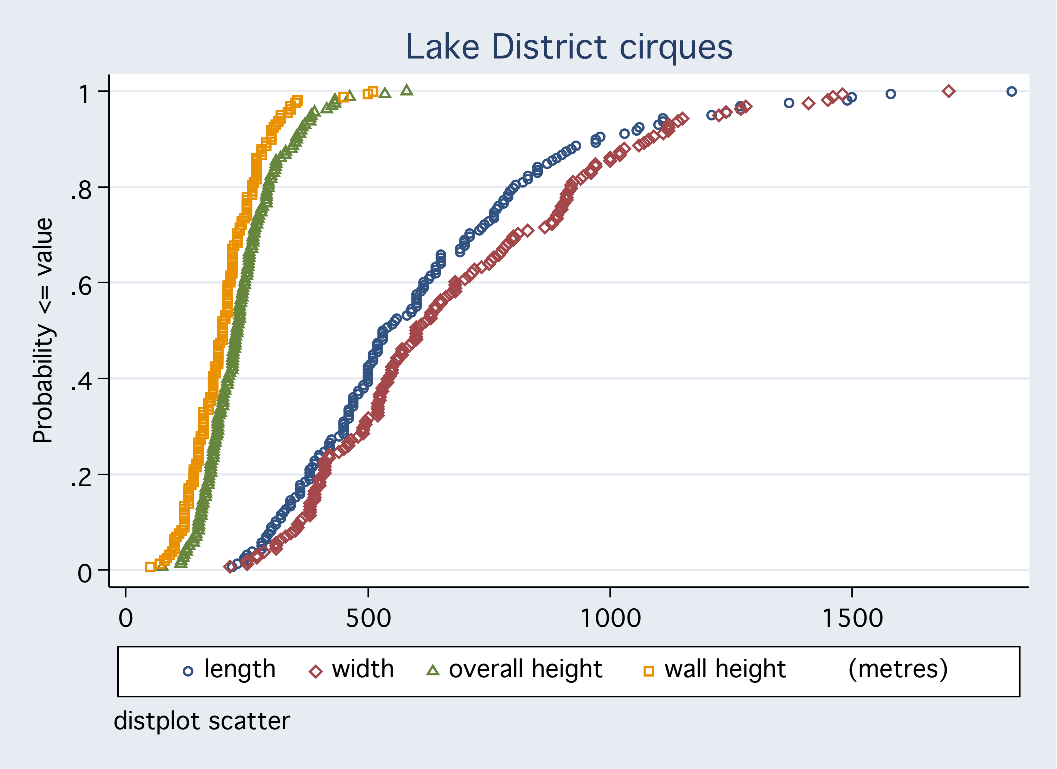
Specific: Confidence intervals

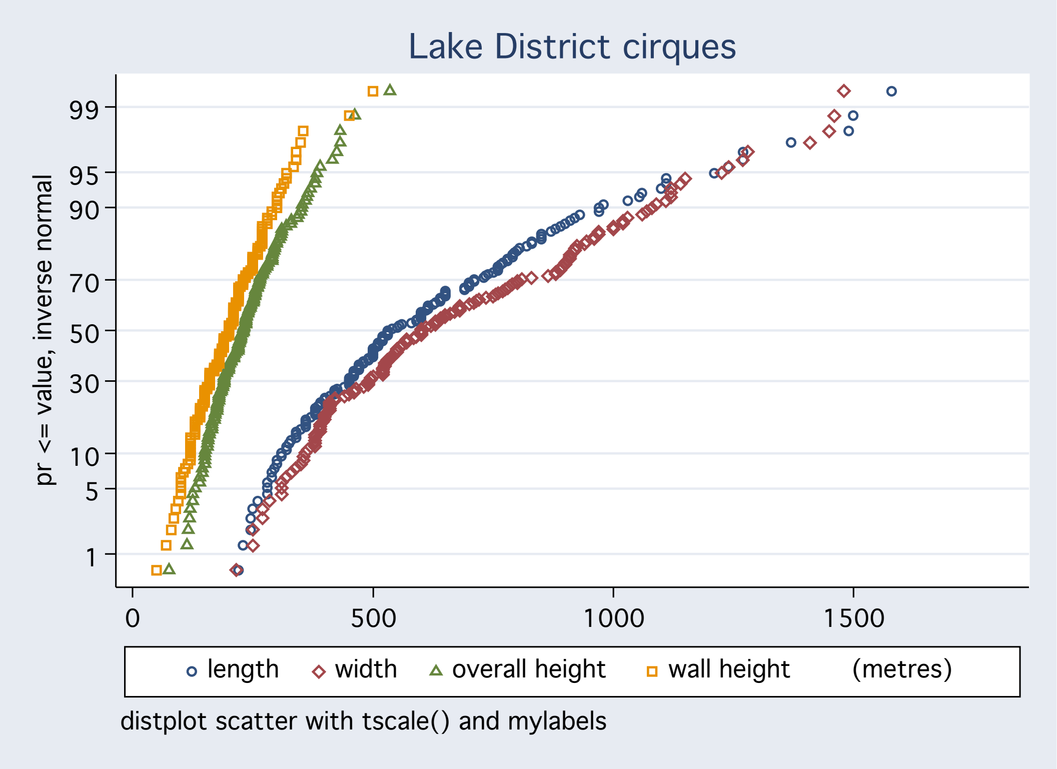
Plotting a bundle of confidence intervals vertically or horizontally is a standard tool.

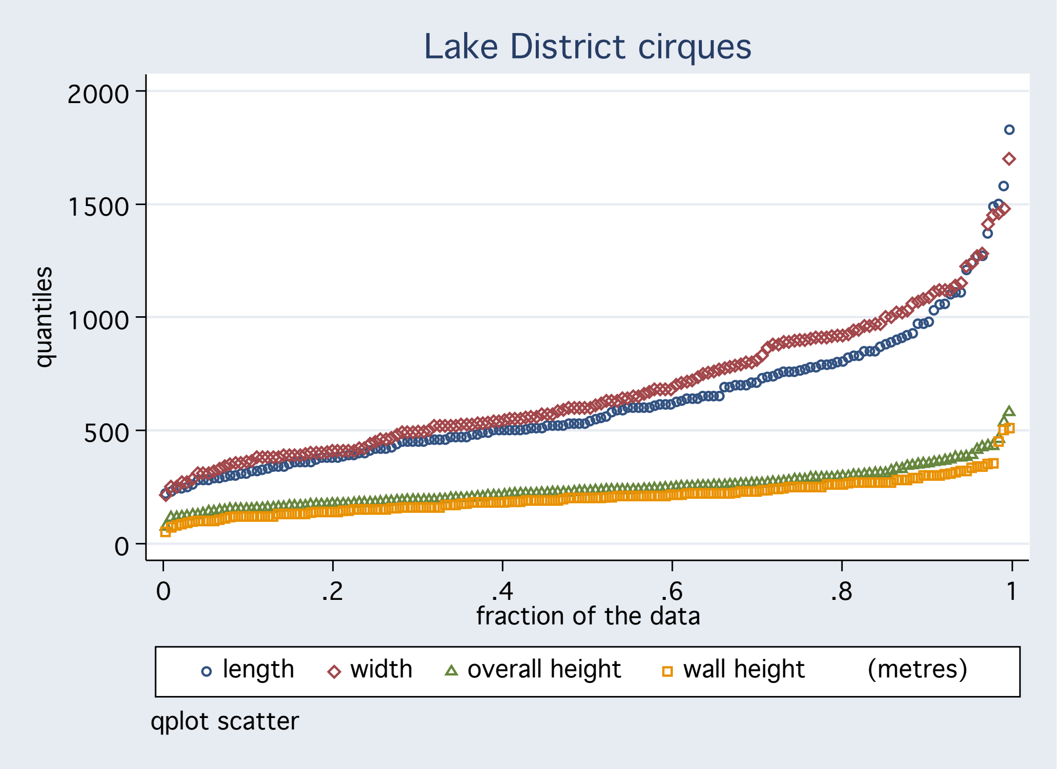
SSC `ciplot` allows you to fire up OS `ci` and plot intervals for several subsets and/or several variables, all in one. Thus `ciplot` is a wrapper for `ci` as well as for `graph`.

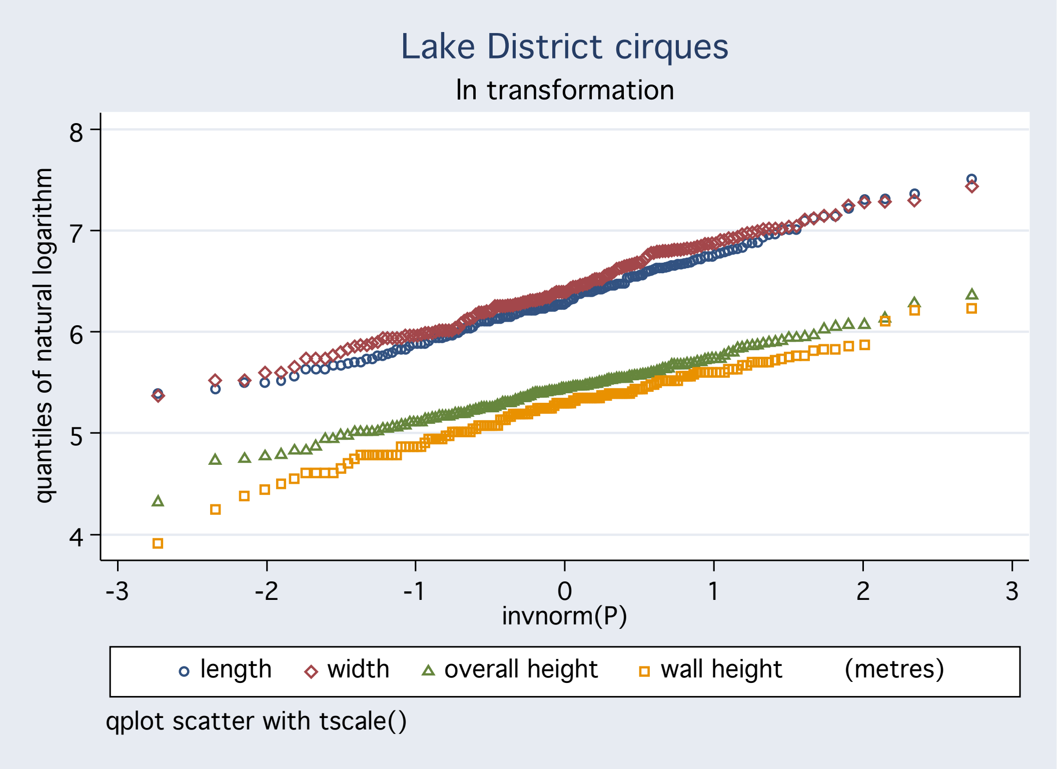
A single option `horizontal` flips axes, courtesy of a handle provided by `graph twoway rcap`.

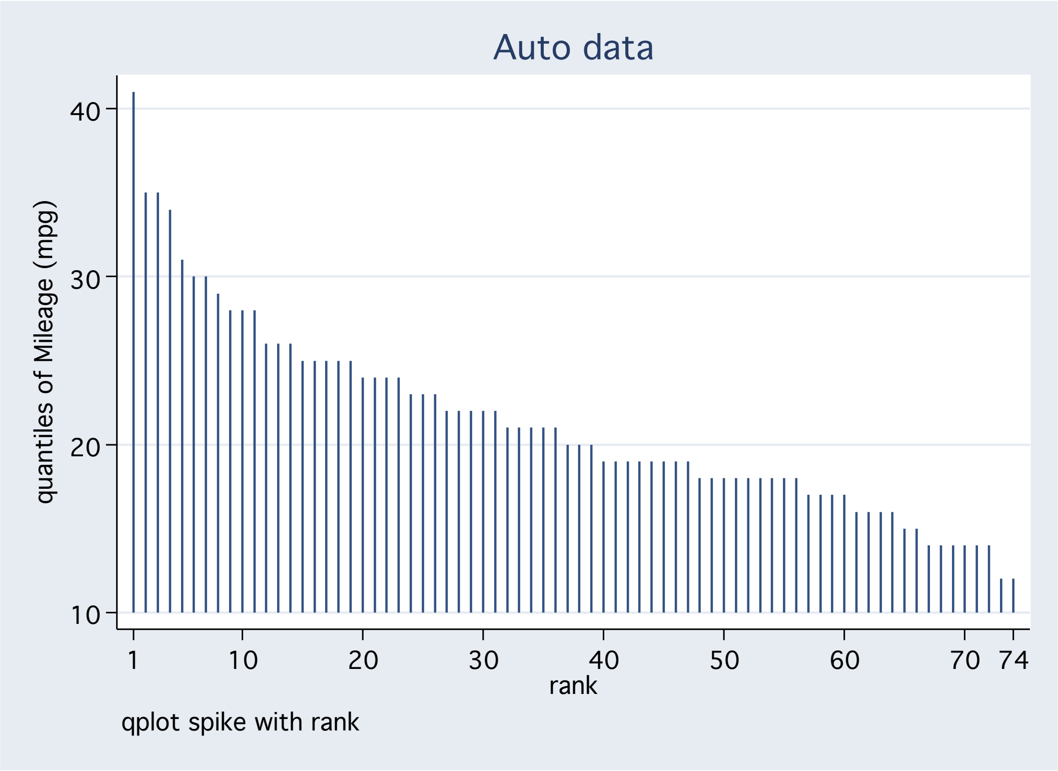
Roger Newson is developing a more general and very much complementary approach, plotting confidence intervals from a reduced data set prepared beforehand.

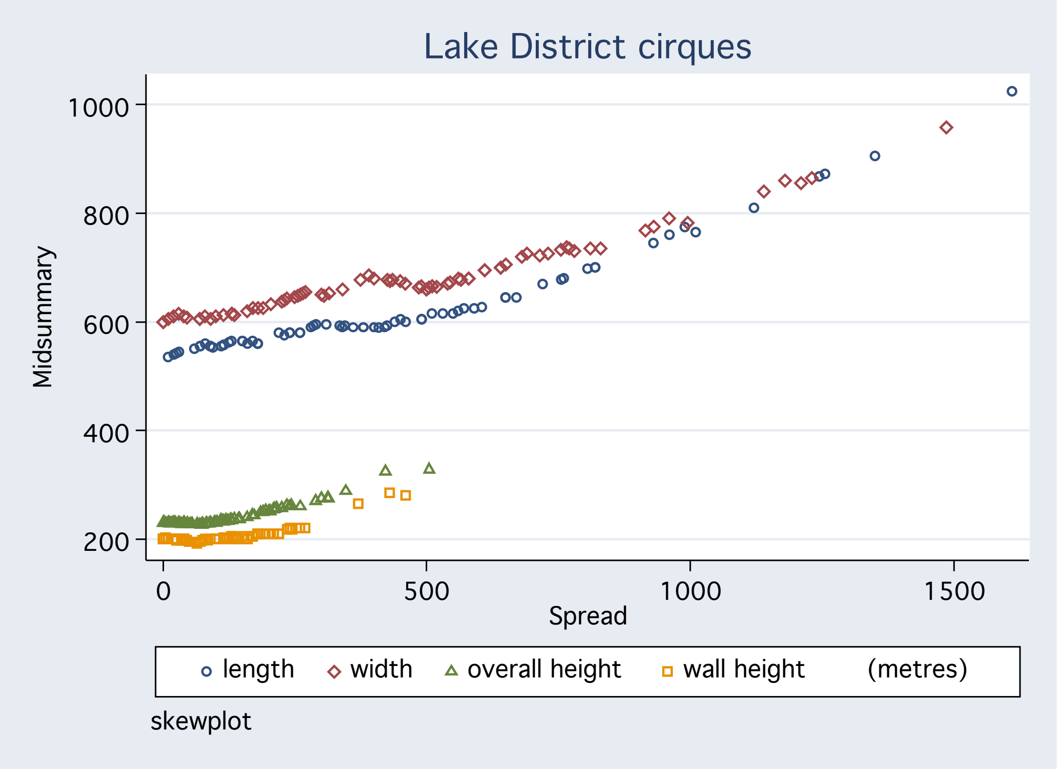

Lake District cirques
Probability <= value
0
.2
.4
.6
.8
1
0
500
1000
1500
○ length
◇ width
△ overall height
□ wall height
(metres)
distplot scatter

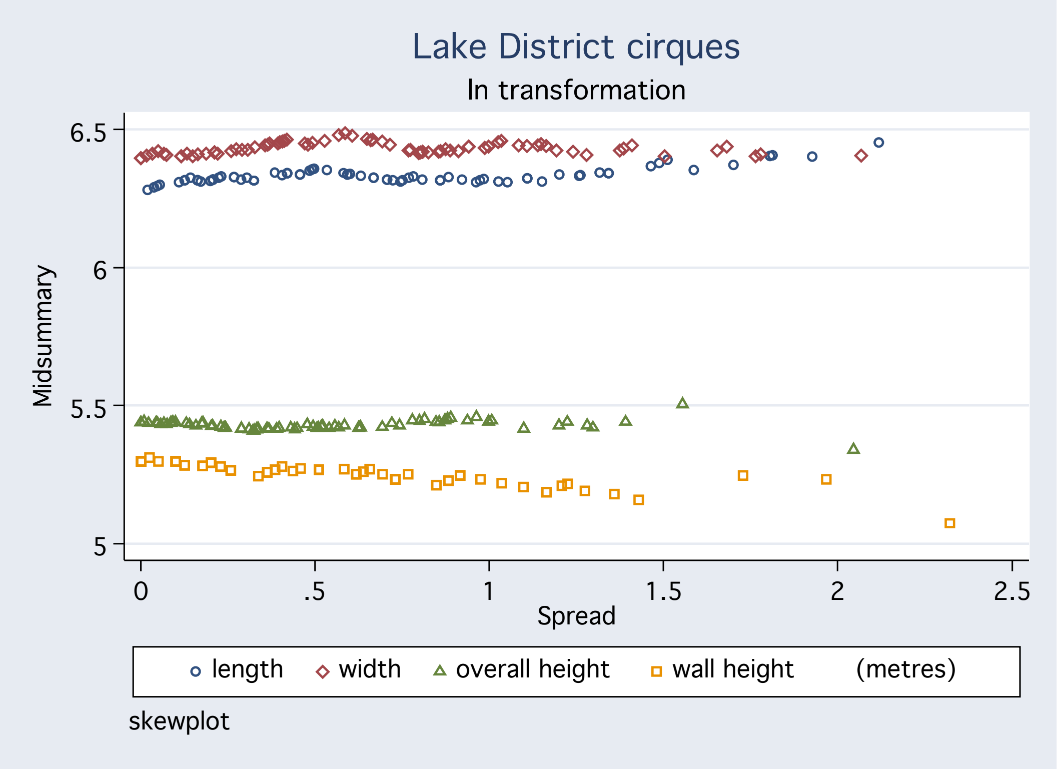

Lake District cirques
pr <= value, inverse normal
99
95
90
70
50
30
10
5
1
0
500
1000
1500
○ length
◇ width
△ overall height
□ wall height
(metres)
distplot scatter with tscale() and mylabels

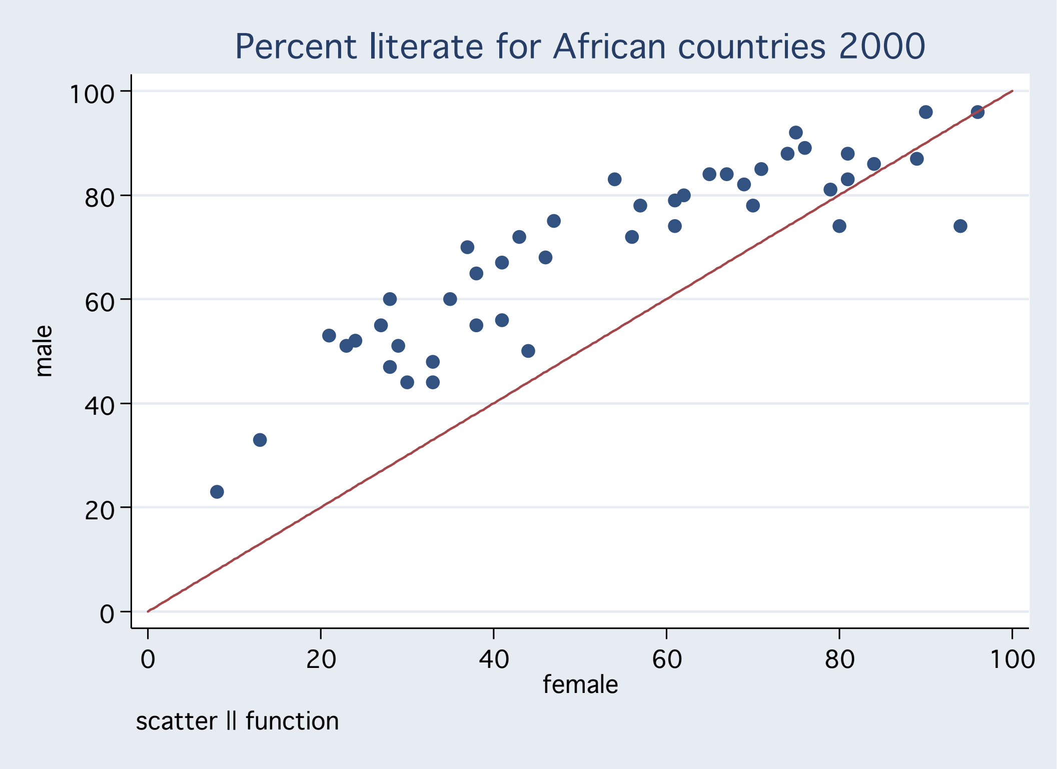

Lake District cirques
quantiles
2000
1500
1000
500
0
0
.2
.4
.6
.8
1
fraction of the data
○ length
◇ width
△ overall height
□ wall height
(metres)
qplot scatter

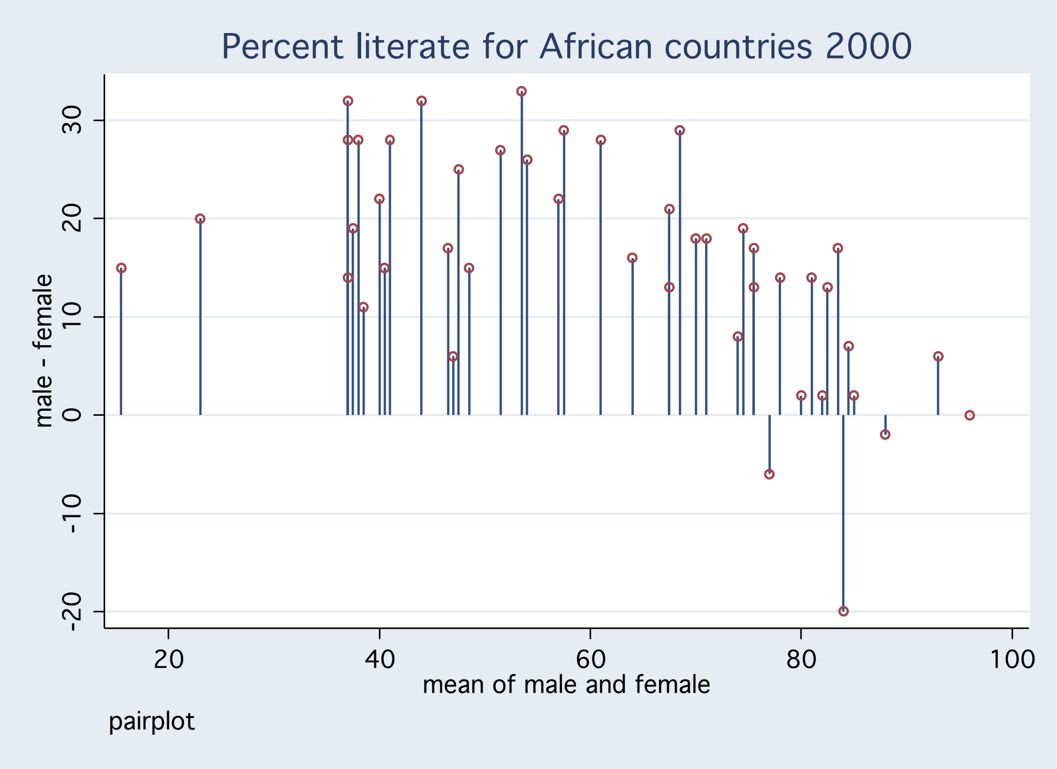

Lake District cirques
ln transformation
quantiles of natural logarithm
8
7
6
5
4
-3
-2
-1
0
1
2
3
invnorm(P)
○ length
◇ width
△ overall height
□ wall height
(metres)
qplot scatter with tscale()

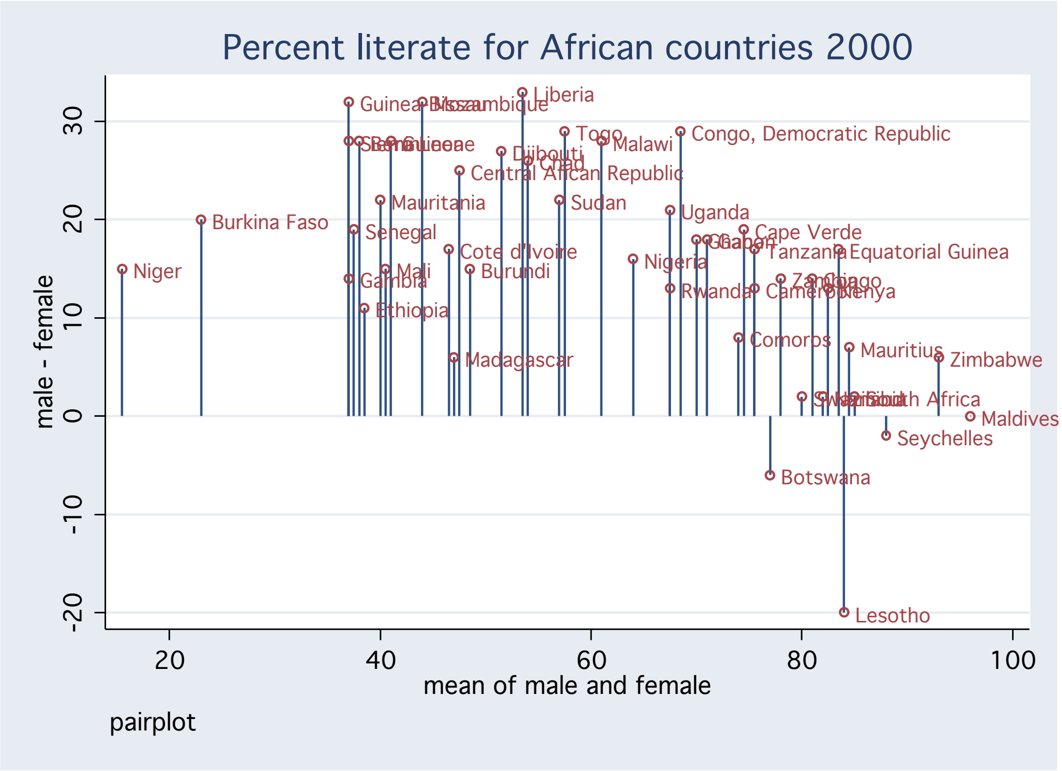

Auto data
quantiles of Mileage (mpg)
40
30
20
10
1
10
20
30
40
50
60
70
74
rank
qplot spike with rank

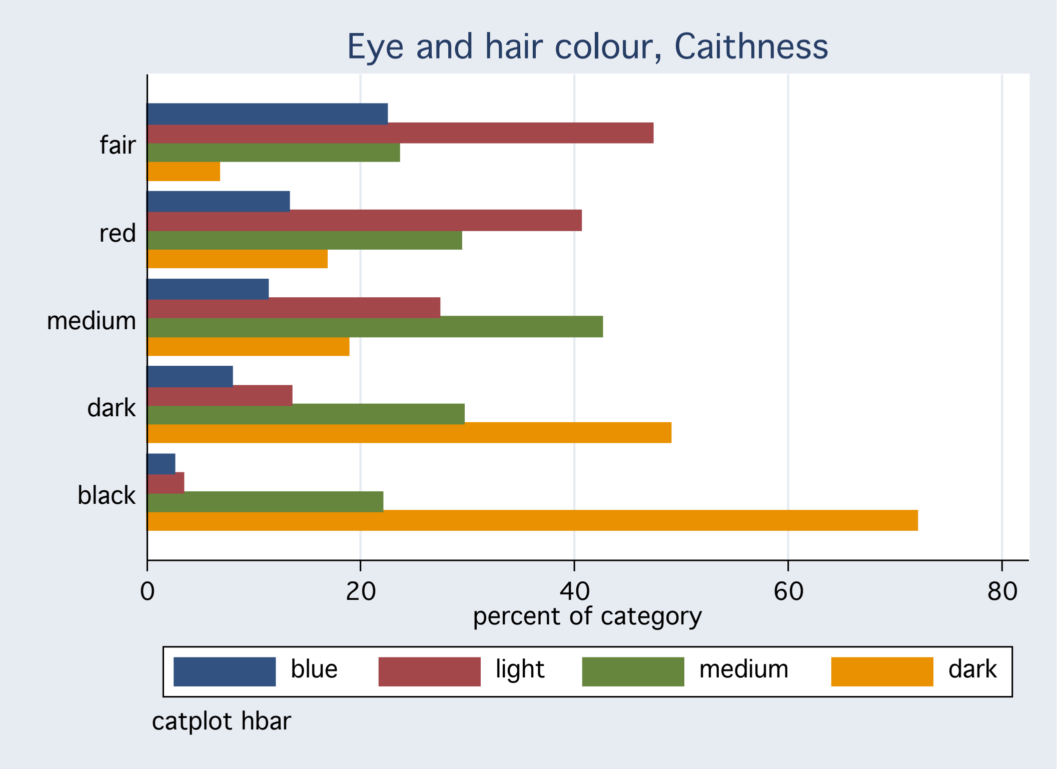

Lake District cirques
Midsummary
1000
800
600
400
200
0
500
1000
1500
Spread
○ length
◇ width
△ overall height
□ wall height
(metres)
skewplot

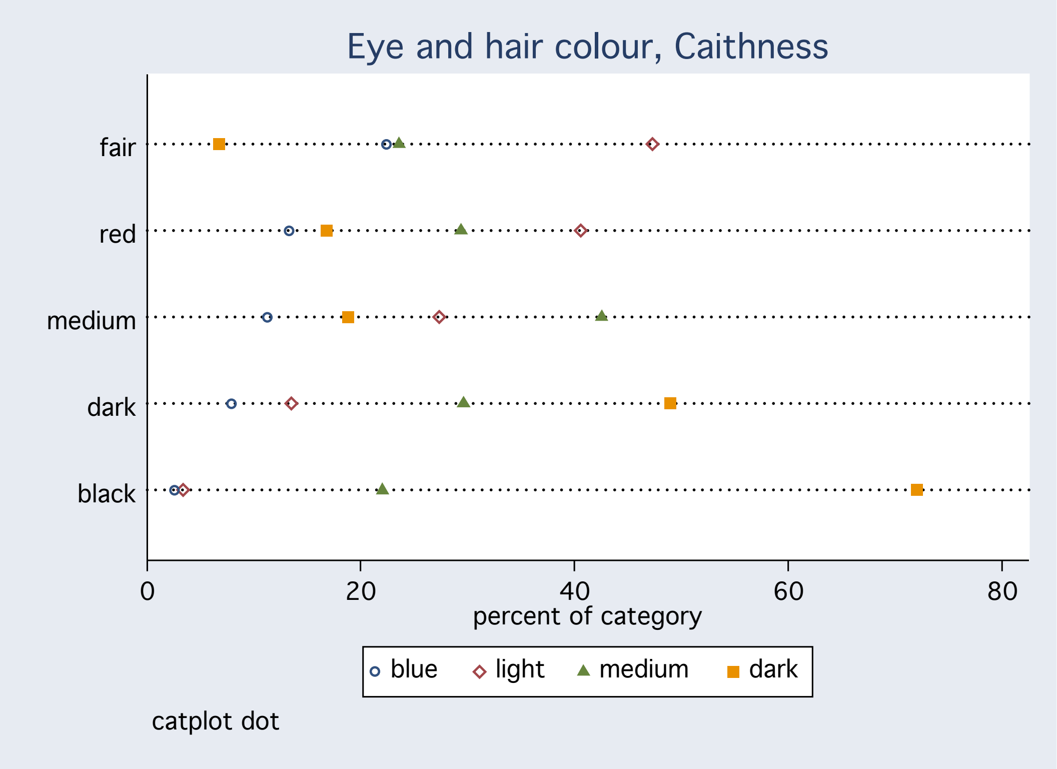

Lake District cirques
ln transformation
Midsummary
6.5
6
5.5
5
0
.5
1
1.5
2
2.5
Spread
○ length ◇ width △ overall height □ wall height (metres)
skewplot

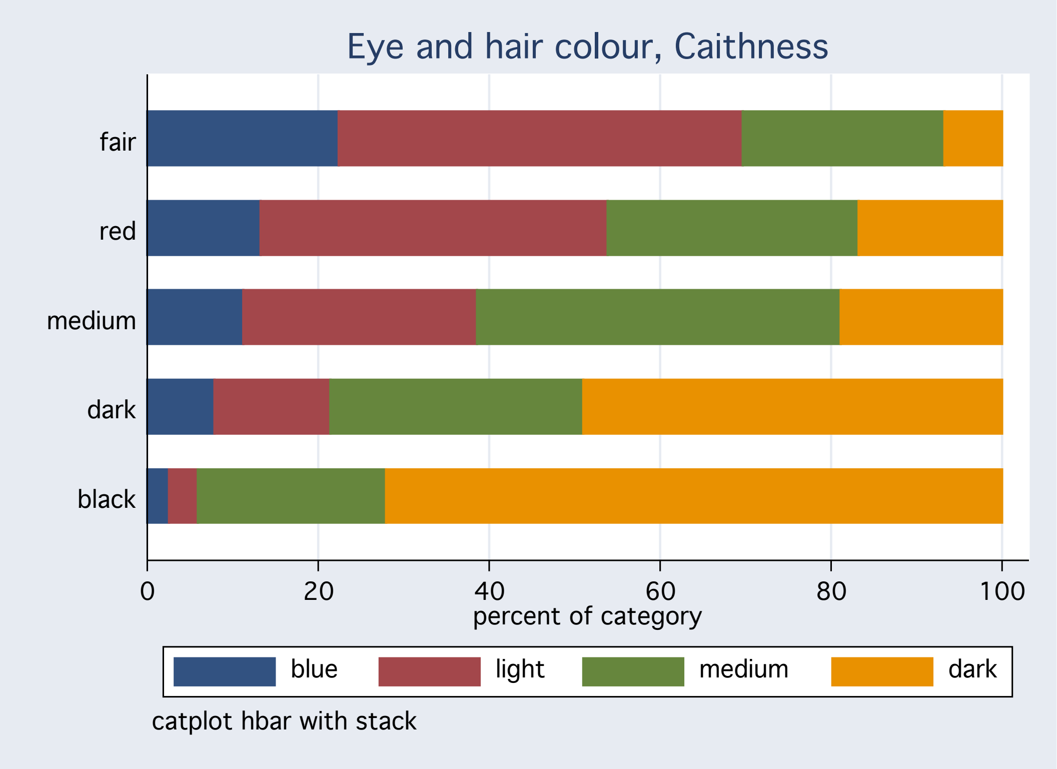

Percent literate for African countries 2000
male
0
20
40
60
80
100
0
20
40
60
80
100
female
scatter || function

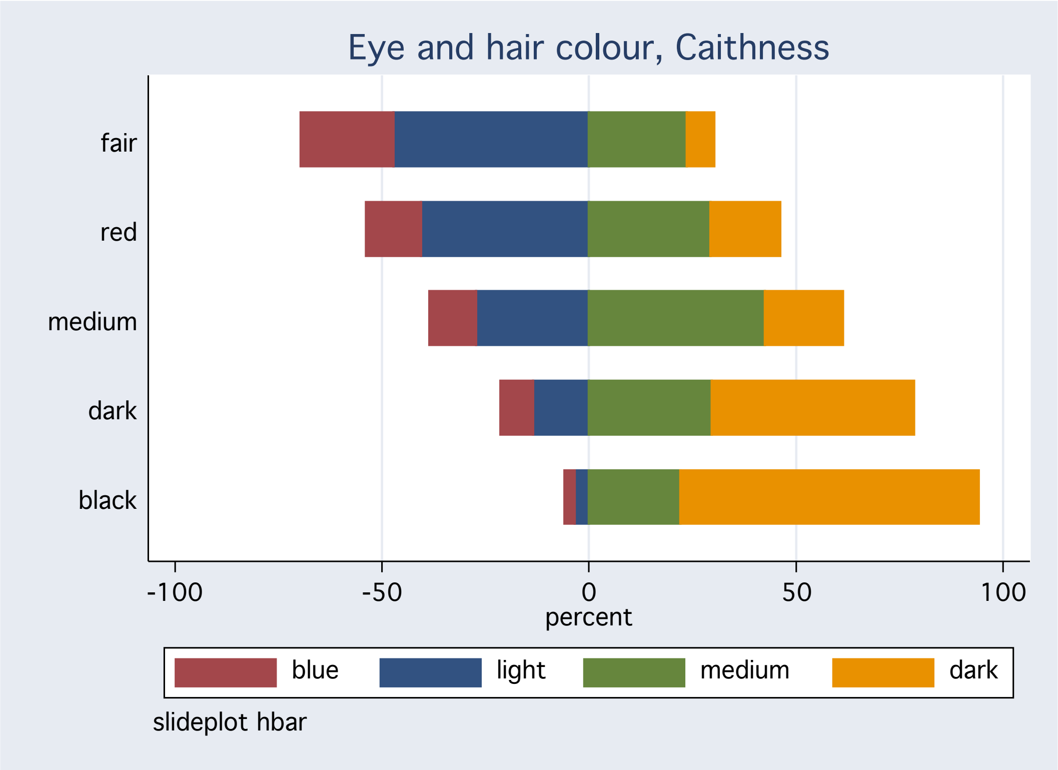

Percent literate for African countries 2000
male - female
30
20
10
0
-10
-20
20
40
60
80
100
mean of male and female
pairplot

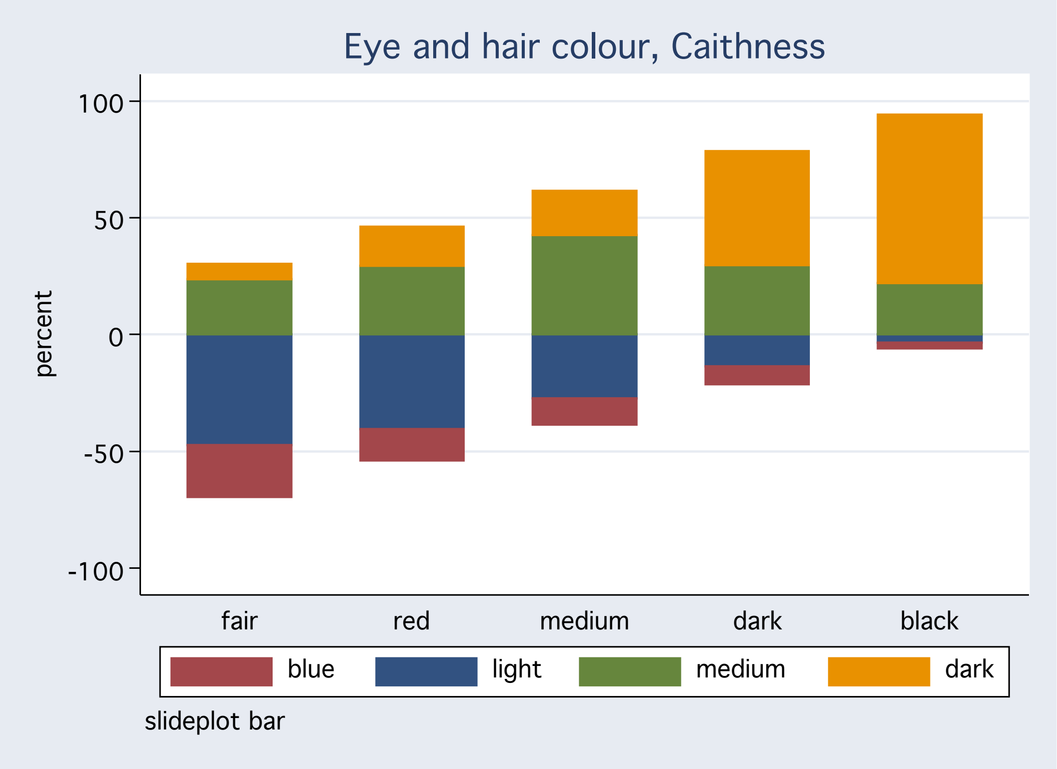

Percent literate for African countries 2000
male - female
30
20
10
0
-10
-20
20
40
60
80
100
mean of male and female
pairplot
Niger
Burkina Faso
Guinea-Bissau
Mozambique
Liberia
Sierra Leone
Benin
Guinea
Togo
Malawi
Congo, Democratic Republic
Djibouti
Chad
Central African Republic
Mauritania
Sudan
Uganda
Senegal
Cape Verde
Ghana
Gabon
Cote d'Ivoire
Tanzania
Equatorial Guinea
Nigeria
Mali
Burundi
Gambia
Zambia
Congo
Rwanda
Cameroon
Kenya
Ethiopia
Comoros
Mauritius
Madagascar
Zimbabwe
Swaziland
Namibia
South Africa
Maldives
Seychelles
Botswana
Lesotho

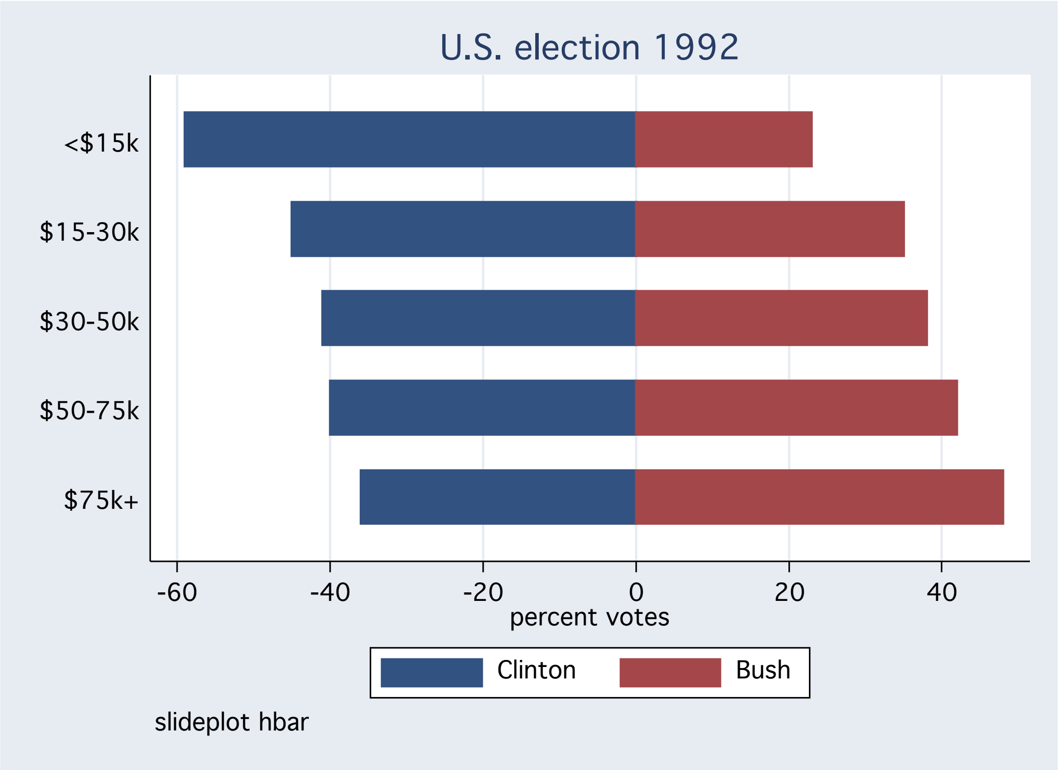

Eye and hair colour, Caithness
fair
red
medium
dark
black
0
20
40
60
80
percent of category
blue
light
medium
dark
catplot hbar

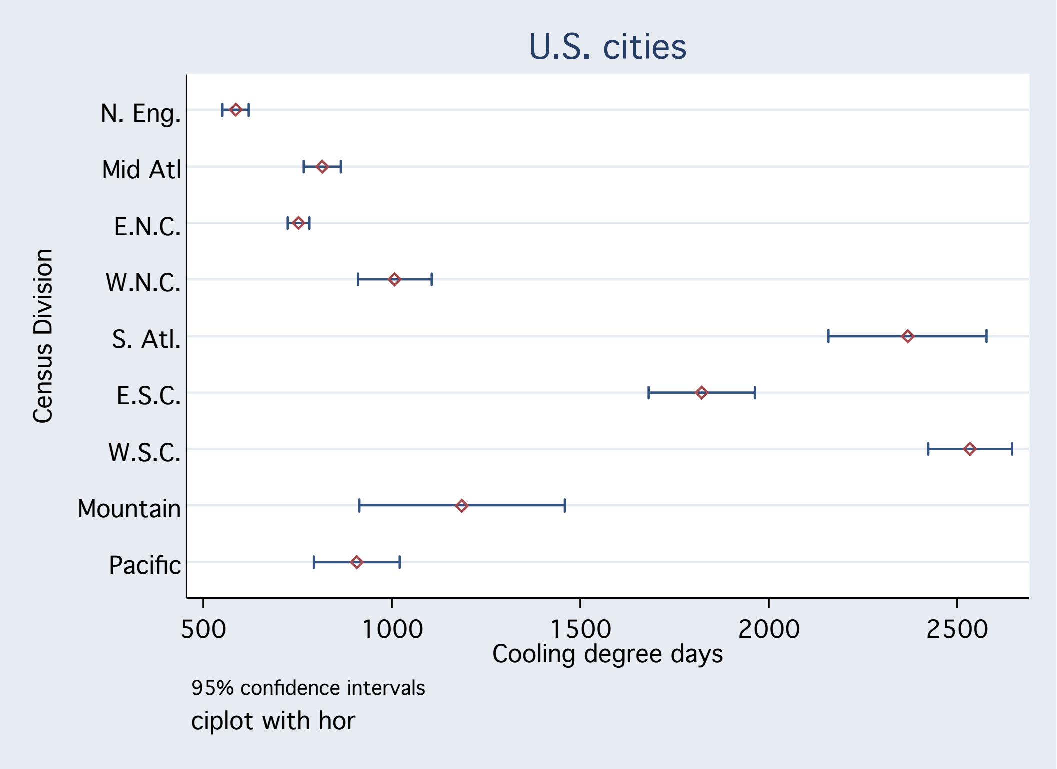

Eye and hair colour, Caithness
fair
red
medium
dark
black
0
20
40
60
80
percent of category
○ blue
◇ light
▲ medium
■ dark
catplot dot


Eye and hair colour, Caithness
fair
red
medium
dark
black
0
20
40
60
80
100
percent of category
blue
light
medium
dark
catplot hbar with stack


Eye and hair colour, Caithness
fair
red
medium
dark
black
-100
-50
0
50
100
percent
blue
light
medium
dark
slideplot hbar


Eye and hair colour, Caithness
percent
100
50
0
-50
-100
fair
red
medium
dark
black
blue
light
medium
dark
slideplot bar


U.S. election 1992
<$15k
$15-30k
$30-50k
$50-75k
$75k+
-60
-40
-20
0
20
40
percent votes
Clinton
Bush
slideplot hbar


U.S. cities
Census Division
N. Eng.
Mid Atl
E.N.C.
W.N.C.
S. Atl.
E.S.C.
W.S.C.
Mountain
Pacific
500
1000
1500
2000
2500
Cooling degree days
95% confidence intervals
ciplot with hor