

Graphics (and numerics) for univariate distributions

Nicholas J. Cox

Department of Geography

Durham University, UK

Klein and mine

Felix Klein (1848–1925) wrote a classic:
1908, 1925, 1928. *Elementarmathematik
vom höheren Standpunkte aus*.
Leipzig: Teubner; Berlin: Springer.



In this talk I look at elementary statistical graphics
from an *intermediate* standpoint.

Why is Stata graphics so complicated?

It offers

canned convenience commands for common tasks (e.g. histograms, survival functions)

a framework for creating new kinds of graphs, vital for programmers

cosmetic control of small details such as colours, text and symbols

How to learn about Stata graphics?

The radical solution:

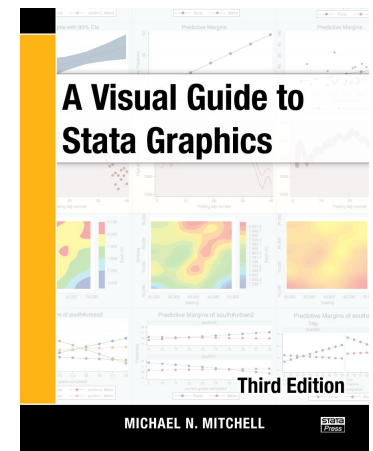
Read the documentation.

The friendly solution:

Read Michael Mitchell's books.

Another solution:

Follow Statalist and the *Stata Journal*.



This talk

I will give a rag-bag of tips and tricks, including

some examples for official Stata commands

some examples of my own commands,

from the *Stata Journal* or SSC

(use `net` or `ssc` to install)

Code and datasets will be downloadable shortly.

Distributions

Most examples will show (fairly) raw data, but there is plenty of scope to show distributions of residuals, estimates, figures of merit, P -values, and so forth.

Categorical variables will get short shrift, but my best single tip is to check out `catplot` and `tabplot` from SSC.

Small distributions with names

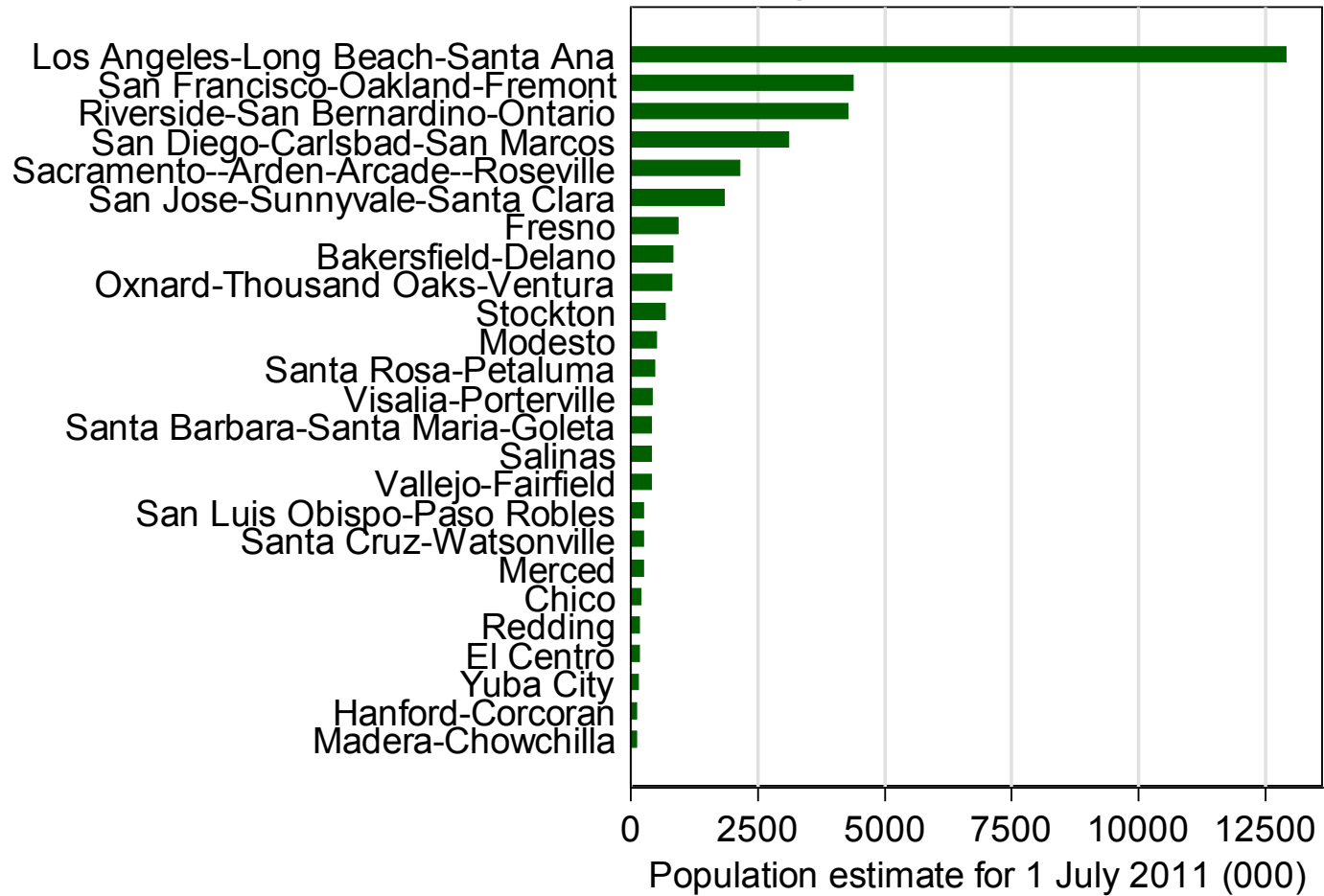
Bar charts need no introduction here.

`graph hbar` is a basic graph for showing distributions with informative names attached.

`hbar` allows names to be written left to right.

20 or so values can be so shown fairly well, more if the medium allows (e.g. whole-page figure, poster).

25 largest cities in California



Small distributions with names

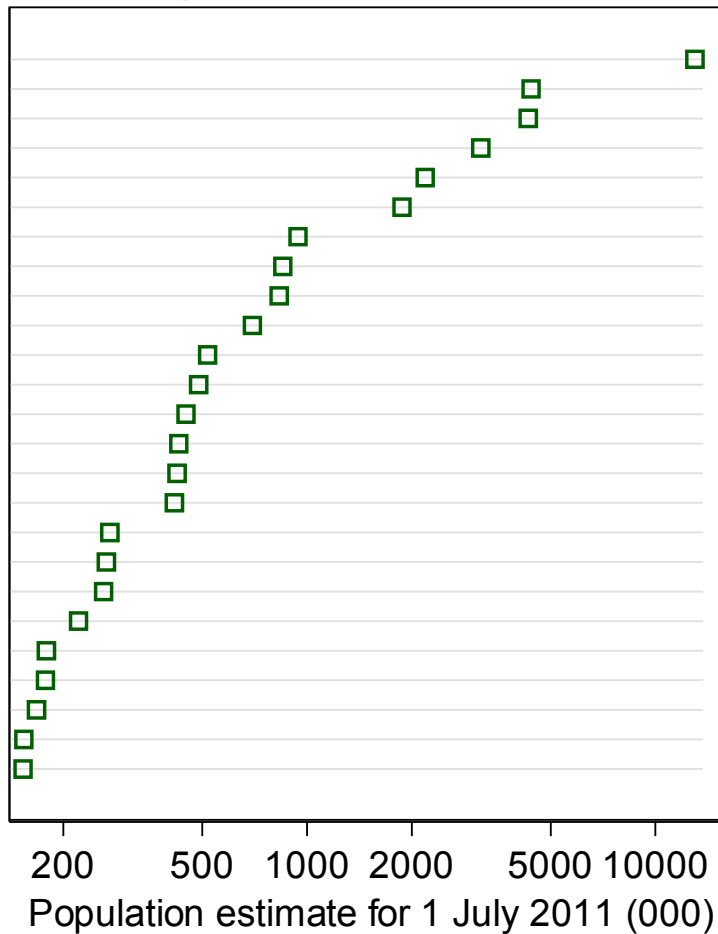
Less well known, `graph dot` is also a basic graph for showing distributions with informative names attached.

`graph dot` also allows names to be written left to right.

Unlike bar graphs, `graph dot` also extends naturally to cases in which logarithmic scales are desired.

25 largest cities in California

Los Angeles-Long Beach-Santa Ana
 San Francisco-Oakland-Fremont
 Riverside-San Bernardino-Ontario
 San Diego-Carlsbad-San Marcos
 Sacramento--Arden-Arcade--Roseville
 San Jose-Sunnyvale-Santa Clara
 Fresno
 Bakersfield-Delano
 Oxnard-Thousand Oaks-Ventura
 Stockton
 Modesto
 Santa Rosa-Petaluma
 Visalia-Porterville
 Santa Barbara-Santa Maria-Goleta
 Salinas
 Vallejo-Fairfield
 San Luis Obispo-Paso Robles
 Santa Cruz-Watsonville
 Merced
 Chico
 Redding
 El Centro
 Yuba City
 Hanford-Corcoran
 Madera-Chowchilla



graph dot

This kind of graph is often called a dot chart or dot plot.

There is scope for confusion, as the same name has been applied to a different plot, on which more later.

It is often named for William S. Cleveland, who promoted it in various articles and books, as a Cleveland dot chart.

graph dot small tips

Guide lines are best kept thin and a light colour.

MS Word users beware: dotted lines don't transfer well.

There is an undocumented `vertical` option, not often needed but there if you really want it.

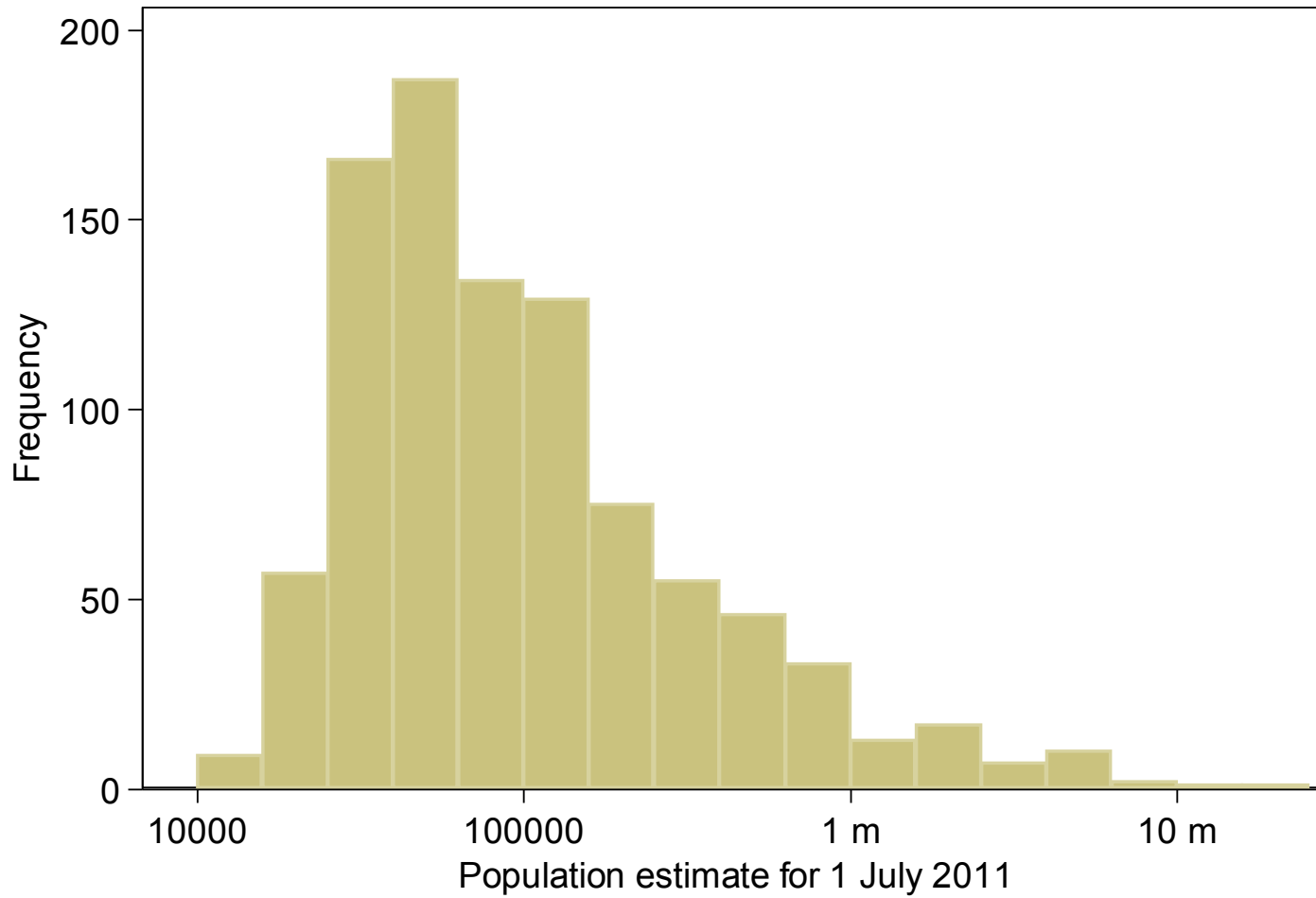
Larger distributions: histograms

At some point with larger distributions we have to abandon naming every observation on a plot, even if the names are known and informative.

Histograms remain very popular, despite the possibility of graphic artefacts arising from choices of bin width and bin origin.

Note that `histogram` and `twoway histogram` are related but distinct commands.

Core based statistical areas, USA



Transformations and histograms

A twist in this example is use of a logarithmic scale.

Transform the variable first, here with `log10()` .

Draw the histogram on a transformed scale.

Fix labels, e.g. 4 "10000", in `xlabel()` .

Note that `xsc(log)` won't do this for you.

Dividing histograms

Frequencies can be added.

So, for two subsets:

- Lay down the frequency histogram for all.

- Put the frequency histogram for a subset on top.

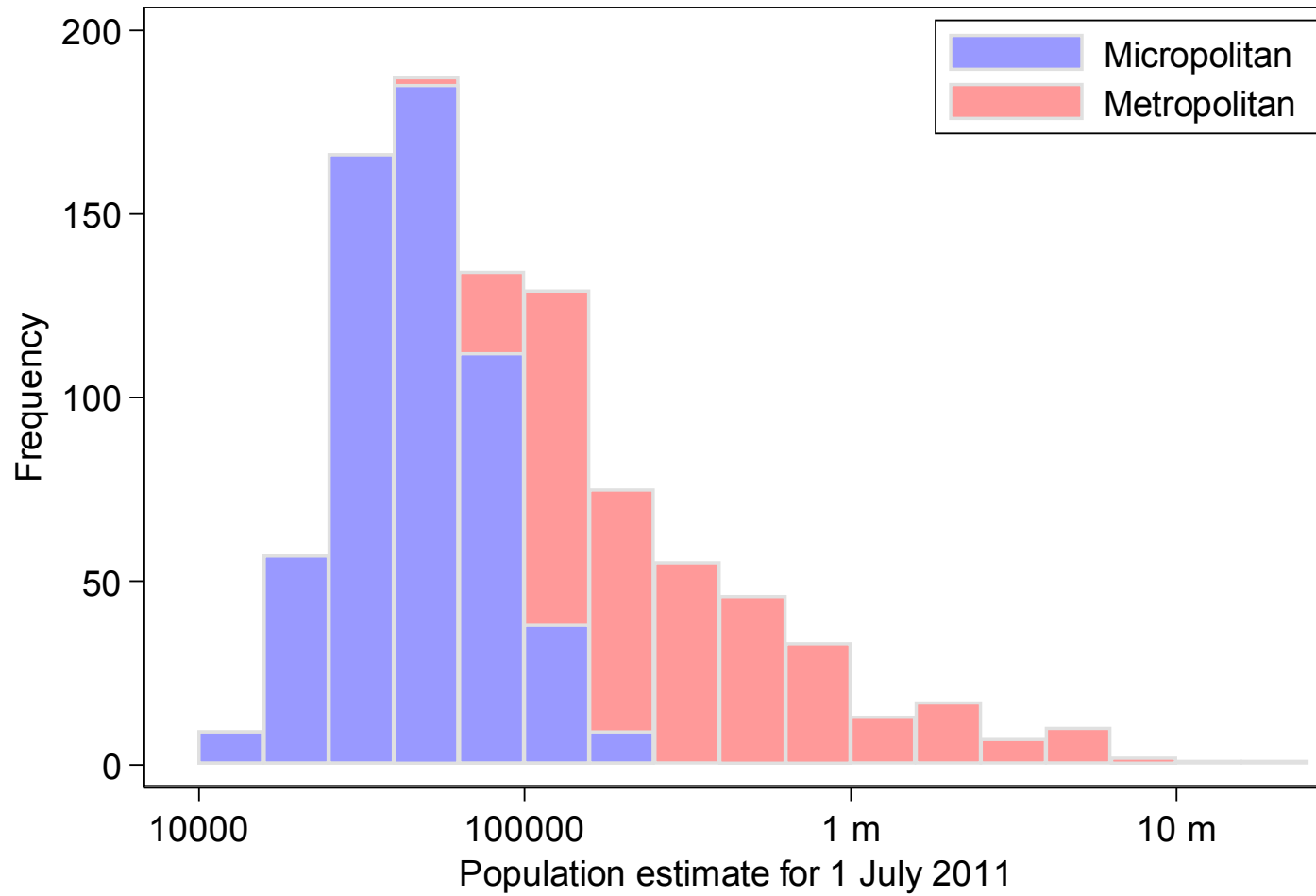
- The difference is the other subset.

- Use different colours, but the same bins.

- Use the same (light) colour for `blcolor()`.

This can be extended to three or more subsets.

Core based statistical areas, USA



Densities

If you really want to plot densities, `kdensity` is the natural place to start.

Note that `kdensity` and `twoway kdensity` are related but distinct commands.

Density estimation on transformed scales

A longstanding but under-used idea is to estimate densities on a transformed scale.

This will ensure that estimates are positive only within the natural support and should help stabilise estimates where data are thin on the ground.

See *Stata Journal* 4: 66–88 (2004) for some references.

Density estimation on transformed scales

For density functions f of a variable x and a monotone transform $t(x)$,

estimate for $f(x)$ =
estimate for $f(t(x)) \cdot | dt/dx |$.

For example, estimate $f(x)$ by $f(\ln x) \cdot (1/x)$.

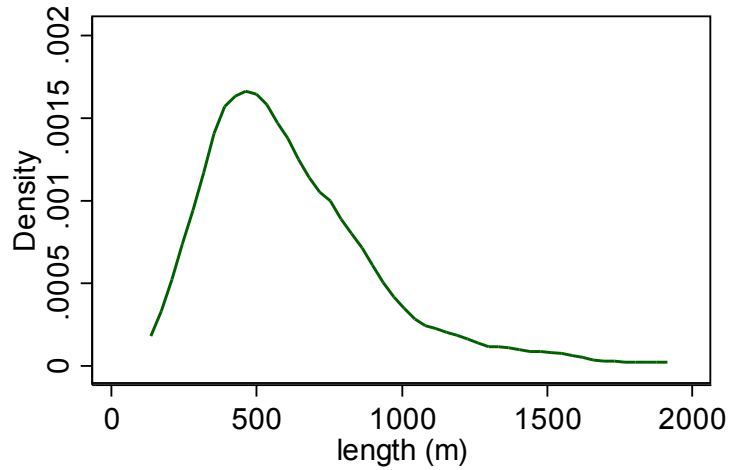
Example and code

Example data are lengths and widths of 158 glacial cirques in the English Lake District. See more at *Earth Surface Processes and Landforms* 32: 1902–1912 (2007).

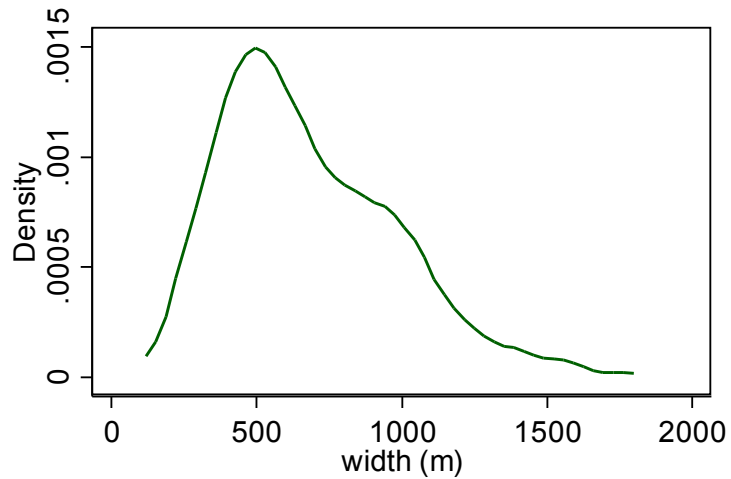
`tkdensity` (forthcoming) is a convenience command that does the estimation and graphing in one.

[A paper with some photos of Romanian cirques](#)

kdensity

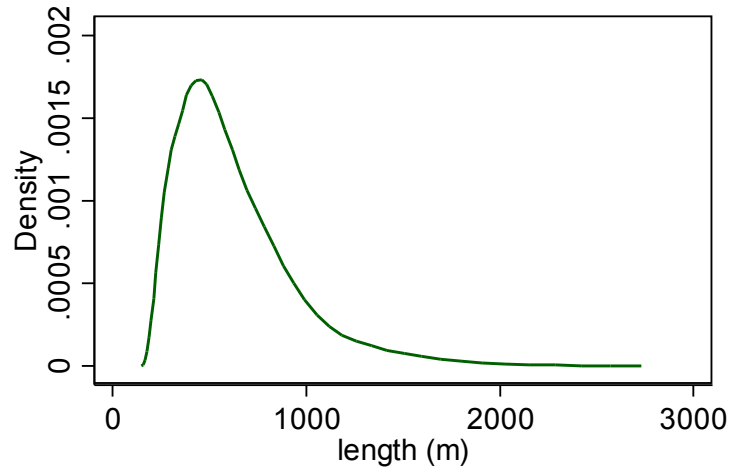


kernel = epanechnikov, bandwidth = 83.6216

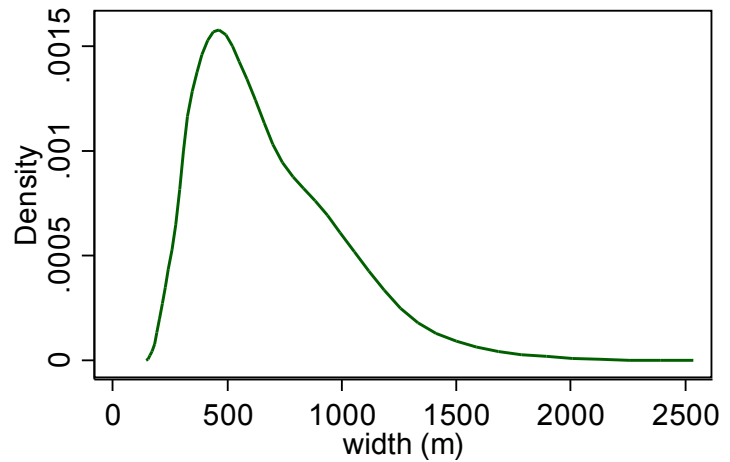


kernel = epanechnikov, bandwidth = 96.1975

tkdensity with ln



biweight on ln scale .4



biweight on ln scale .4

Dot plots or strip plots

The main idea is to show each data point by one marker symbol on a magnitude scale.

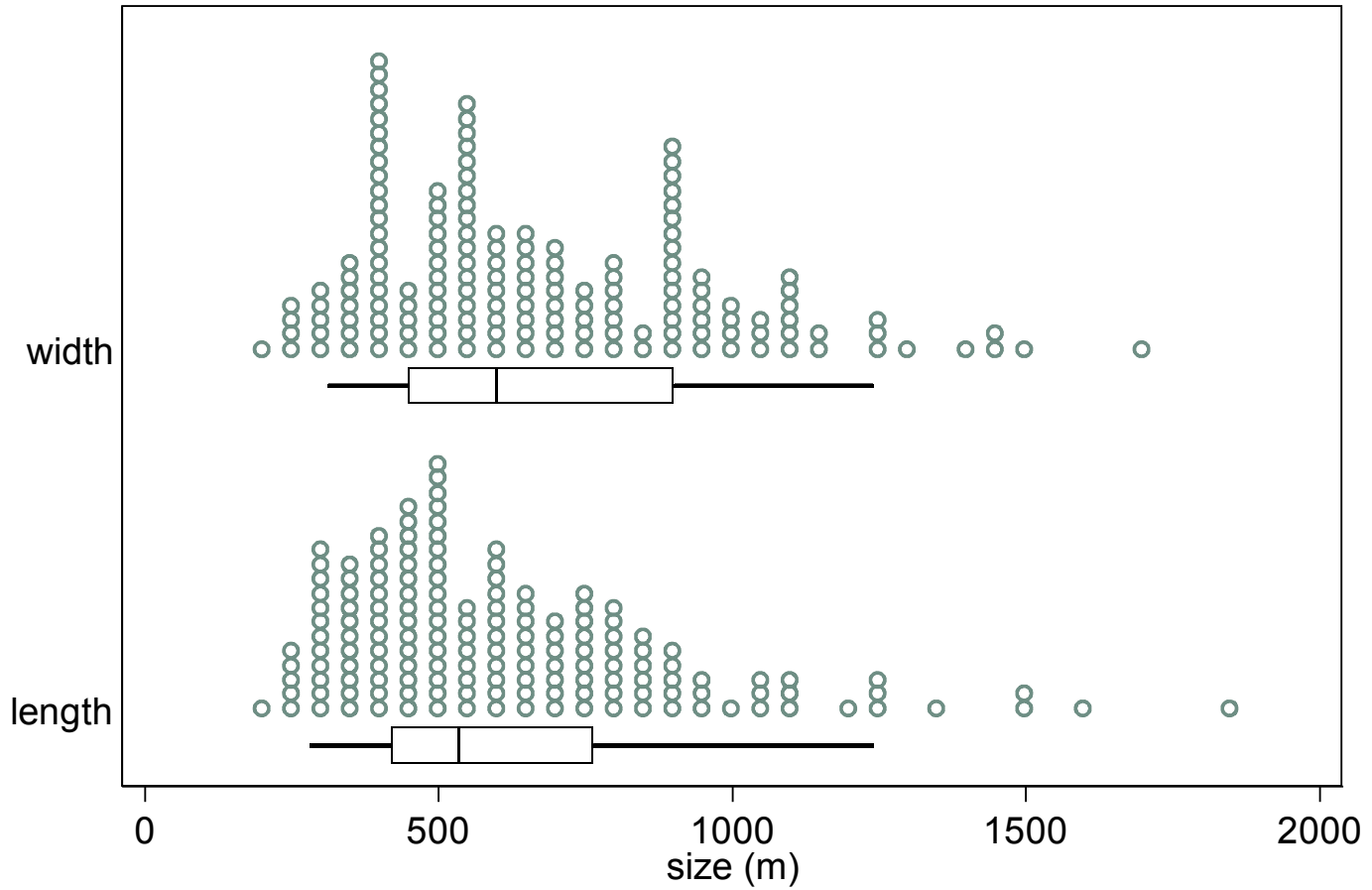
Usually, although not necessarily, there is binning too and tied values are jittered or stacked to show relative frequency.

In official Stata the command is `dotplot`.

`stripplot` from SSC is much more versatile.

First we look at some examples using the default horizontal alignment.

50 m bins



boxes show 5 25 50 75 95 percentiles

Marginal box plots

`stripplot` (for that matter `dotplot` too) can add box plots.

That way box plots do what they arguably do best: summarize.

The fine structure of the data remains visible.

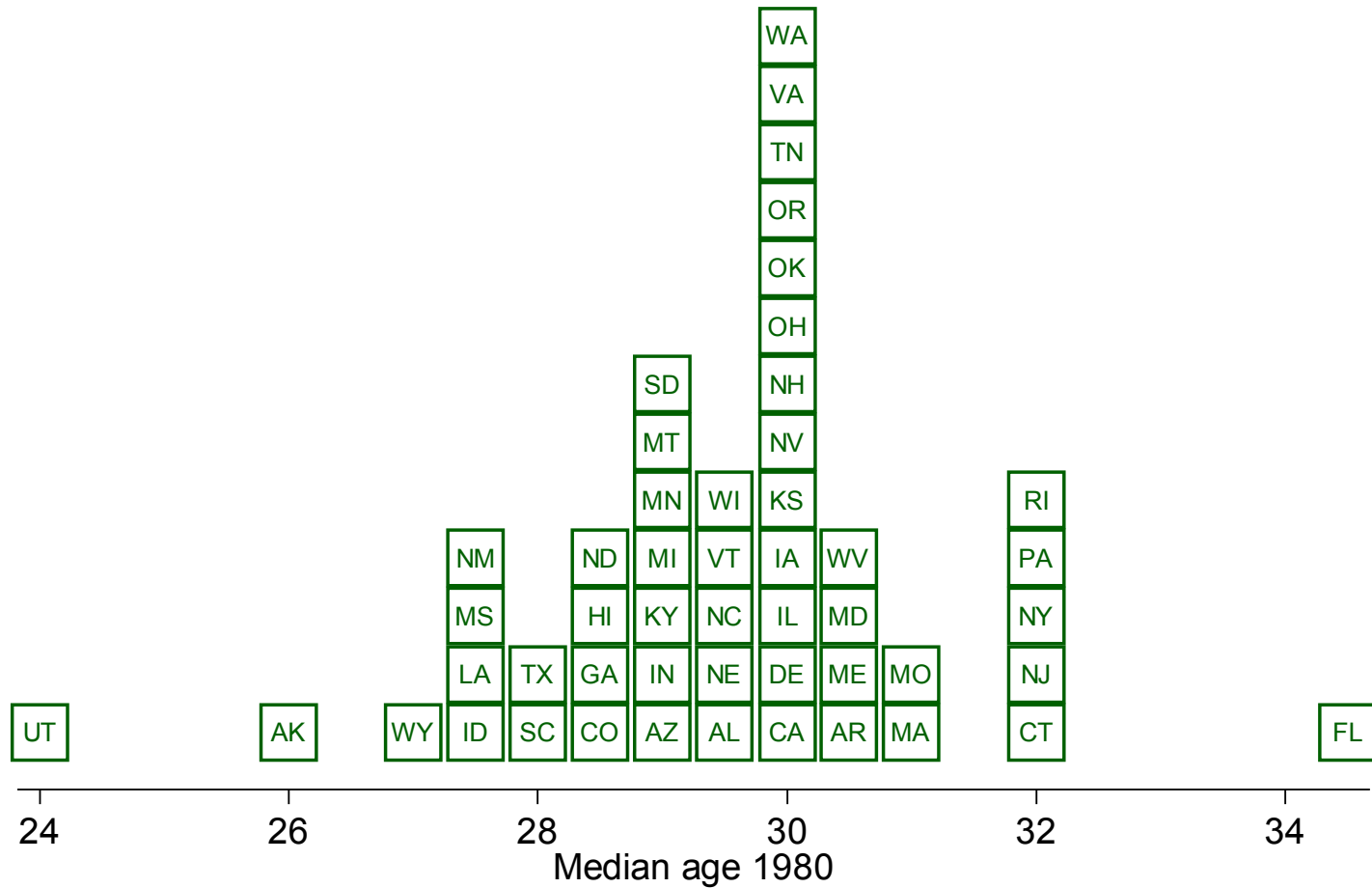
`stripplot` allows box plots with whiskers drawn to specified percentiles, as well as those following the Tukey rule that whiskers span data points within 1.5 IQR of the nearer quartile.

An aside on box plots

If you like box plots, you will know that `graph box` and `graph hbox` get you there...
... except in so far as they don't.

Suppose you want to do something a bit different, such as add points for means, or join medians.

See *Stata Journal* 9: 478-496 (2009) for details on how to do box plots from first principles.



0.5 year bins

How was that done?

This plot of median age in 1980 for US states also used `stripplot`.

The main trick is very simple: make marker symbols big enough and marker labels small enough so they jointly act as small text boxes.

OH yes, I agree: 50 US states with two-letter abbreviations AR an easy case, but WY not?

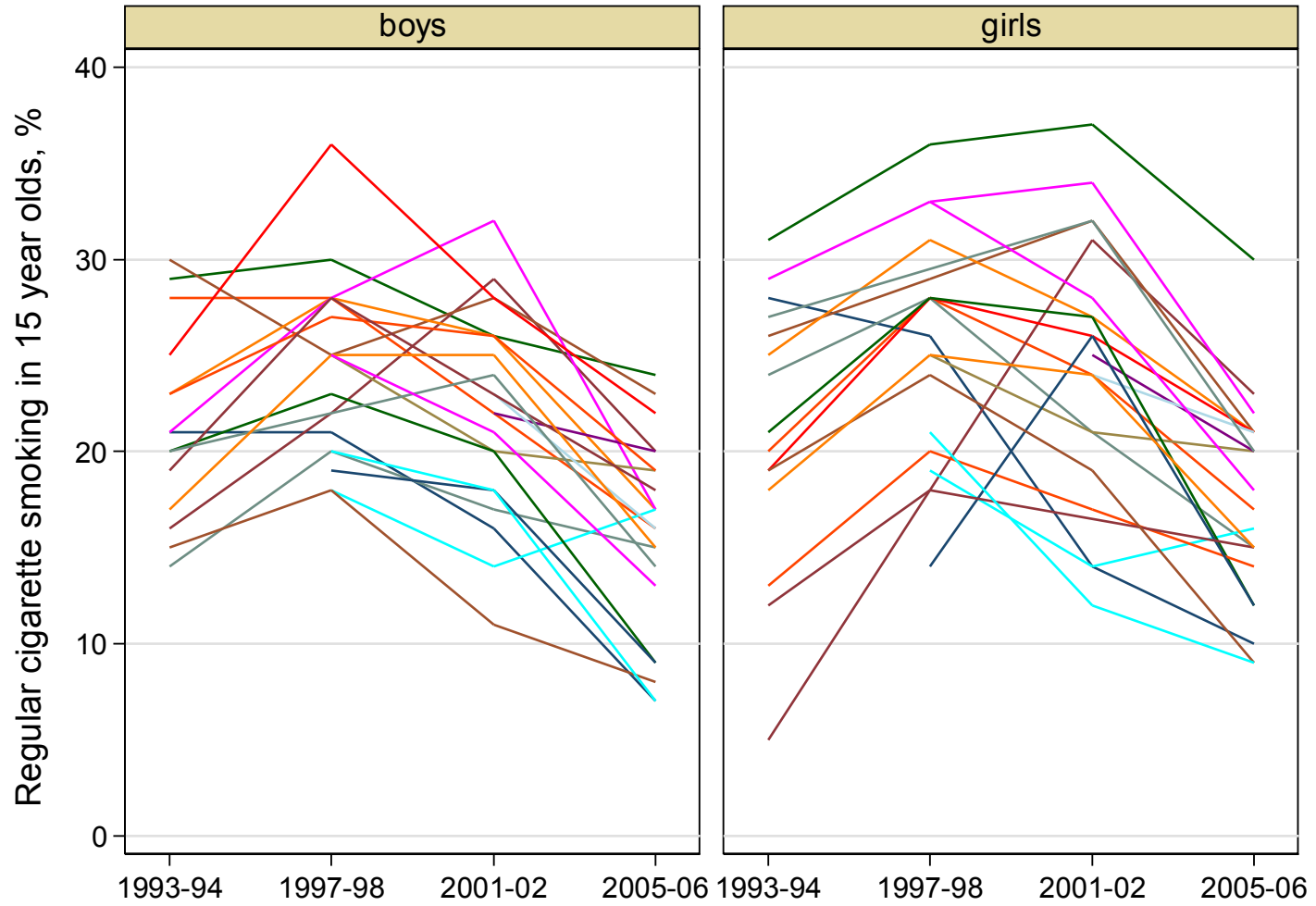
Panel or longitudinal data

Let's change tack for a different kind of example, with panel data.

The dataset is small: OECD data on percent regular cigarette smoking at age 15 for 24 countries, 4 time periods and 2 genders.

Panel data can be seen as a series of distributions.

The distribution can serve as context for any interesting case, just as a test score is reported as a percentile rank.



Spaghetti plot, or a graphical pastiche

The usual kind of multiple time series plot (here using `line`) is the usual kind of mess. I suppressed the legend naming the countries.

There are ways of improving it as a time series plot, such as using a `by()` option or some other device for splitting out subsets.

But we will stick with the distribution theme. First, look at a `striplot` in which the USA is highlighted.

Regular cigarette smoking in 15 year olds



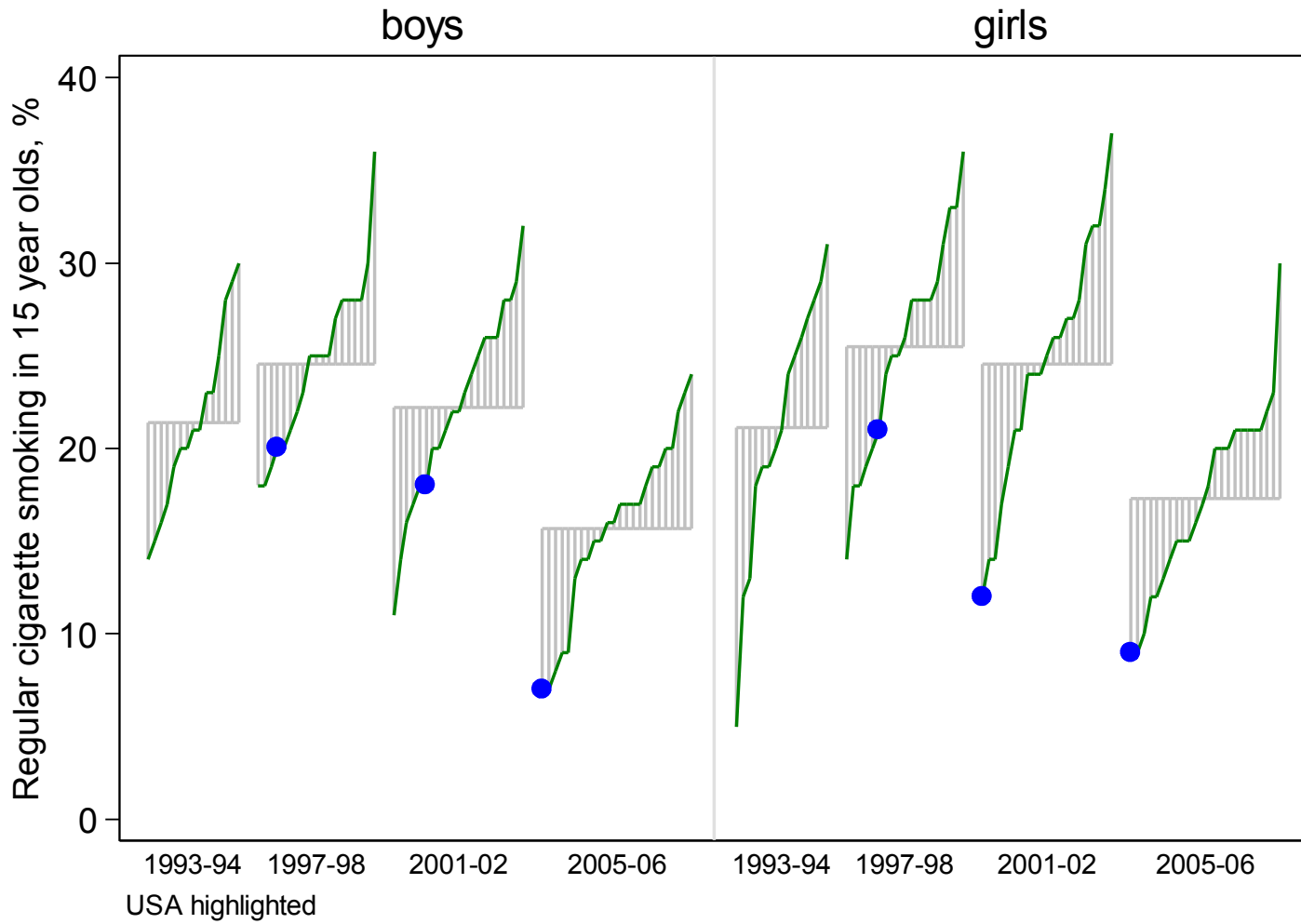
USA highlighted

stripplot for panel data

In principle, we lose some information on individual trajectories.

In practice, a multiple time series plot is likely to be too unattractive to invite detailed scrutiny.

As before, we could add box plots, or bars with means and confidence intervals.



devnplot

The previous graph was from `devnplot` (SSC).
`devn` here stands for deviation.

The values for each group are plotted as quantiles
or order statistics.

A subset of cases may be highlighted (here just one
panel).

A backdrop shows values as deviations from group
means.

devnplot

The choices shown are the defaults.

Other plotting orders are possible.

The backdrop can be removed, or tuned to show deviations from any specified set of levels.

`devnplot` was first written with the aim of showing data and summaries in `anova` style, but I mostly use it to show sets of quantiles.

devnp1ot

A small but sometimes useful detail is that `devnp1ot` adjusts the width for each group according to the number of its values.

This can help if groups are of very different sizes.

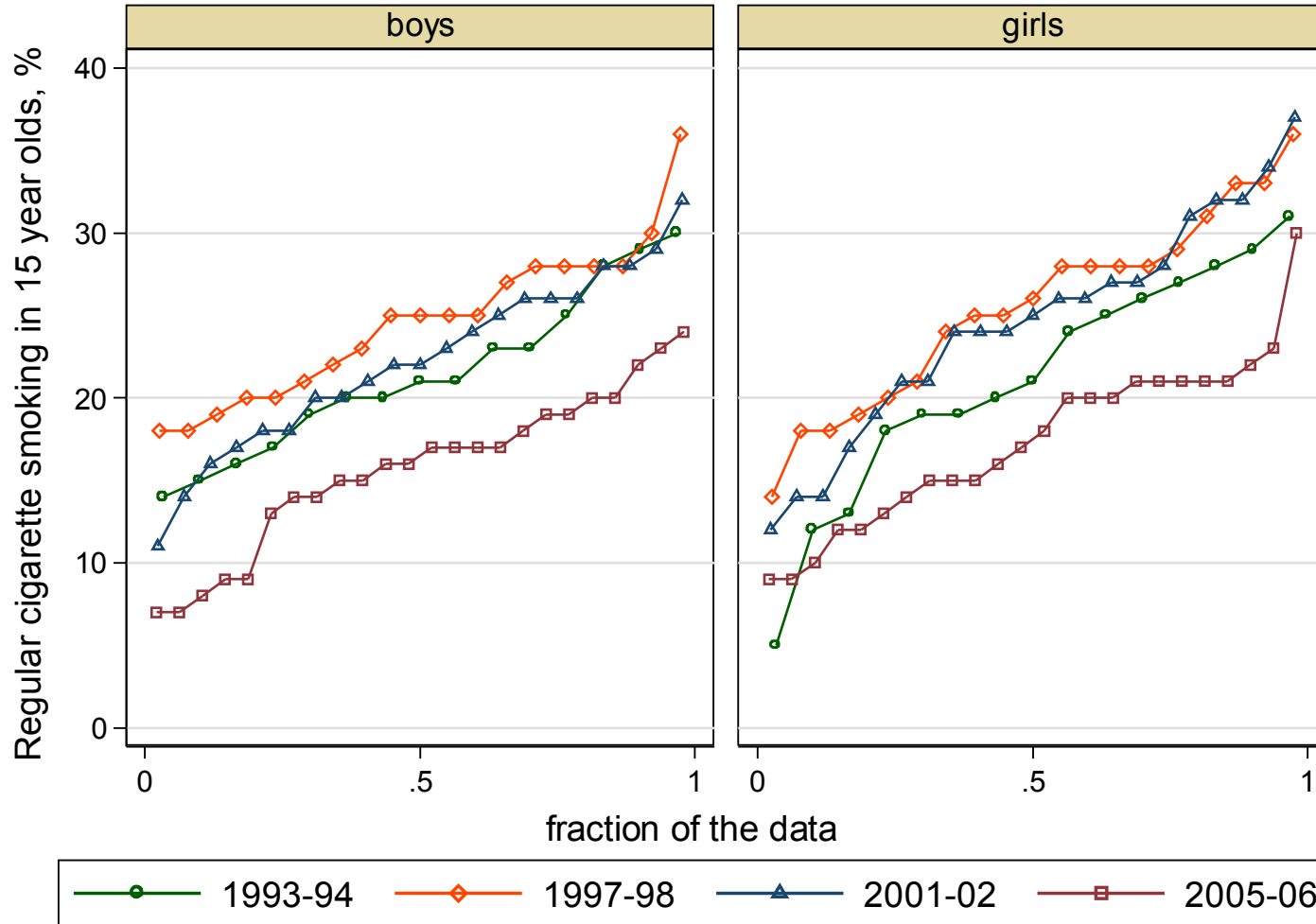
Quantile plots

Quantile functions are also known as inverse (cumulative) distribution functions.

They are the order statistics, given as a function of cumulative probability or fraction of the data.

For ranks $i = 1, \dots, n$, use a plotting position such as $(i - 0.5)/n$ as abscissa.

In official Stata the main command is `quantile`.
`qqplot` from *SJ* is much more versatile.



qqplot options

What about smoothing?

qqplot supports `over()` and `by()` options, to plot quantiles by distinct groups within and between graph panels.

In this example, some of the irregularity stems from reporting values as integers. None of the irregularity is easy to interpret.

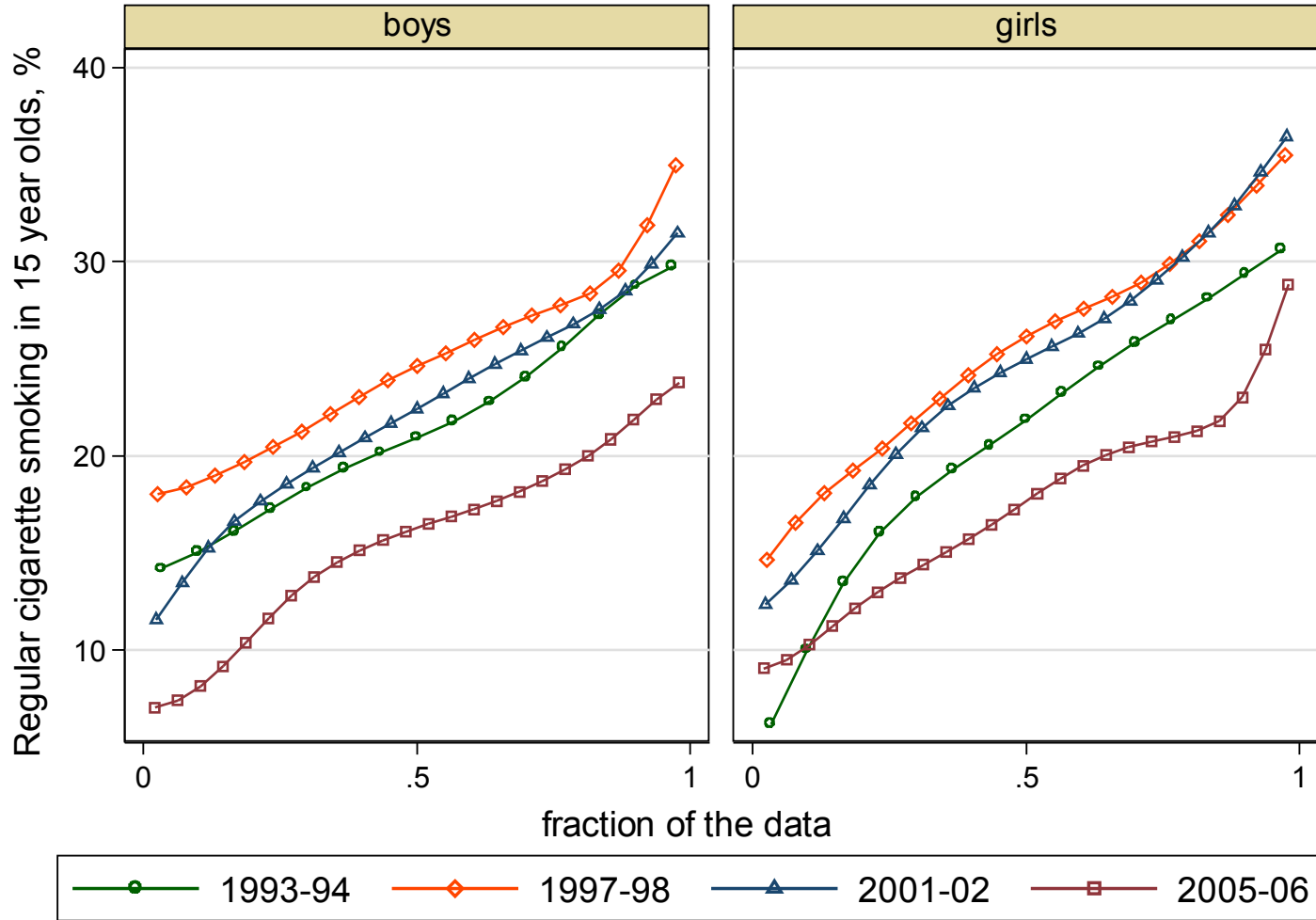
So why not smooth the quantile functions?

Quantile smoothing

Quantile smoothing is less well known than kernel density estimation.

The method of Harrell, F.E. and C.E. Davis. 1982.
A new distribution-free quantile estimator.
Biometrika 69: 635–640 turns out to be an exact bootstrap estimator of the corresponding population quantile.

`hdquantile` (SSC) offers an implementation.

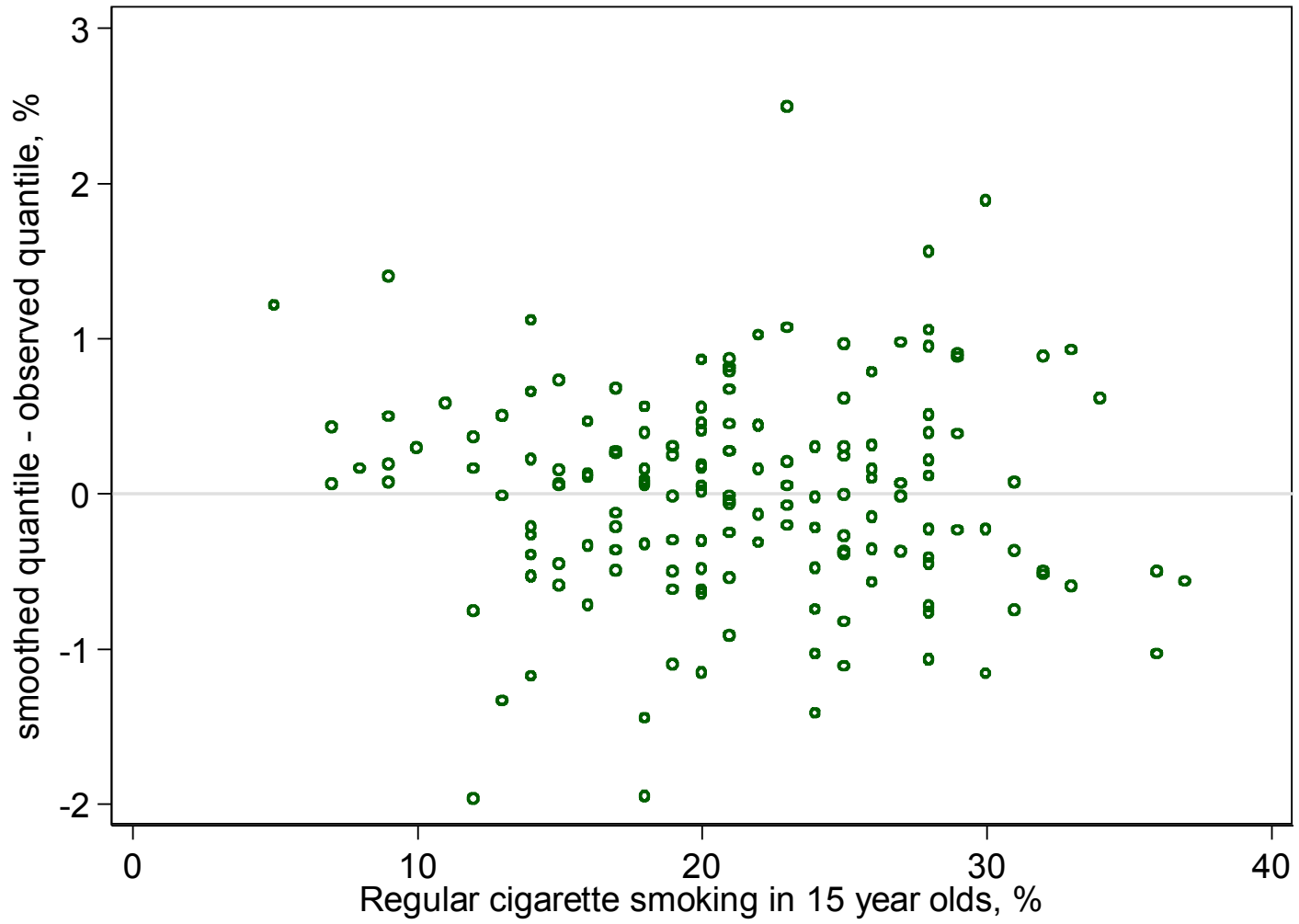


How much difference does quantile smoothing make?

Quantile smoothing is conservative.

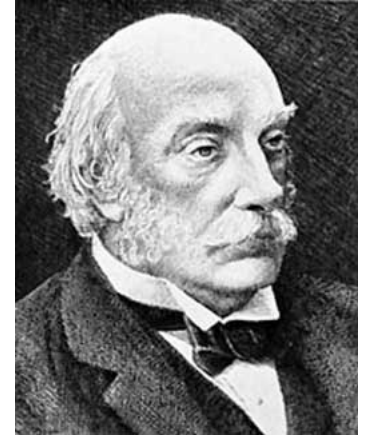
Here the difference between smoothed and observed quantiles is $\sim 1\%$.

So, smoothing mostly takes out noise, which is its job.



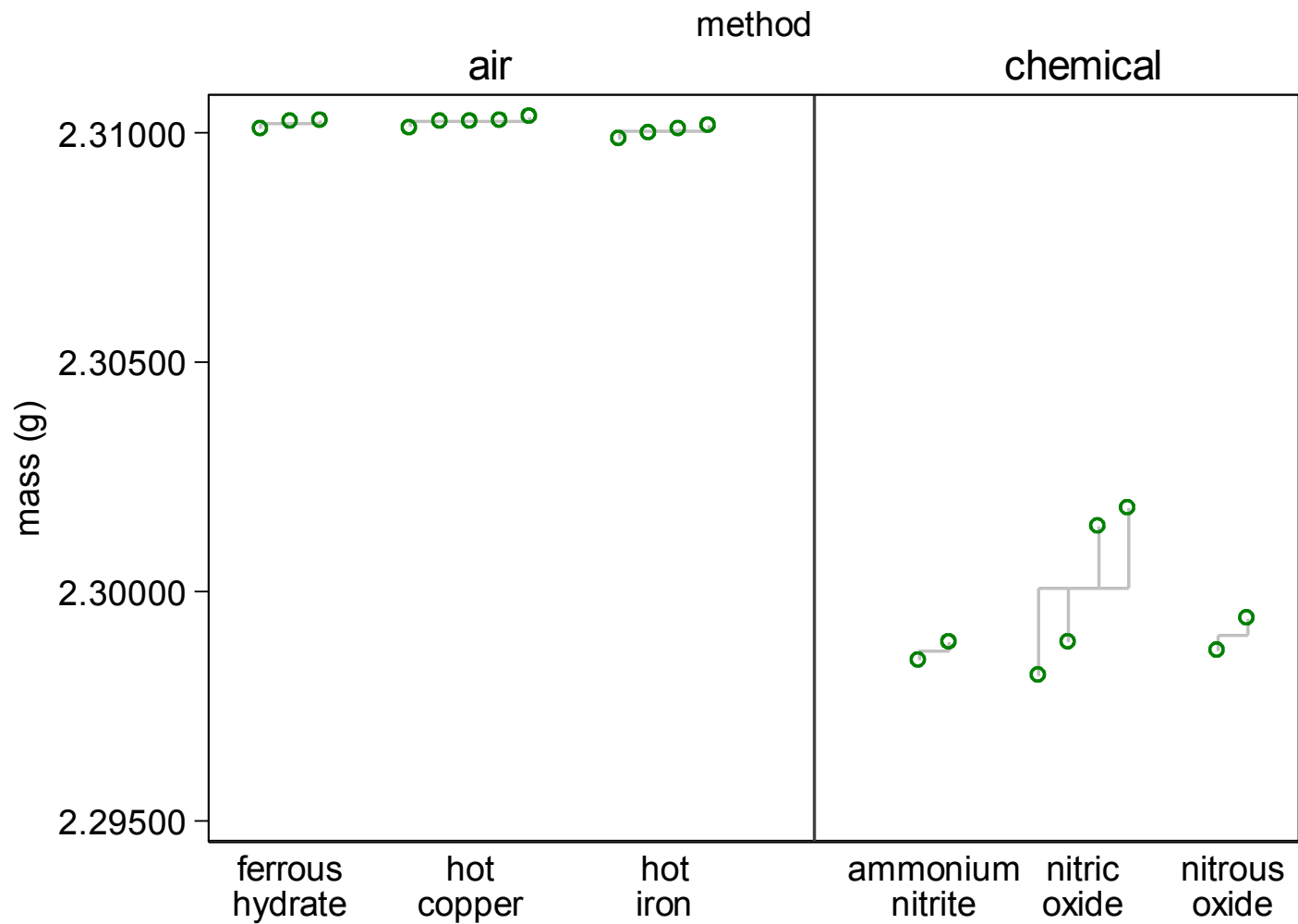
Lord Rayleigh discovering argon

John William Strutt, Lord Rayleigh (1842–1919) compared the mass of nitrogen obtained by different methods from a given container.



The marked difference led to the discovery of argon with Sir William Ramsay and the award to Rayleigh of the Nobel Prize for Physics in 1904.

The Rayleigh distribution is named for the same Rayleigh.



Which plot?

`devnplot` works well here to show fine structure in the data.

`stripplot` doesn't work so well and a boxplot just suppresses detail unnecessarily.

(Rayleigh was reporting extremely careful experimental results to a resolution of 10 μg .)

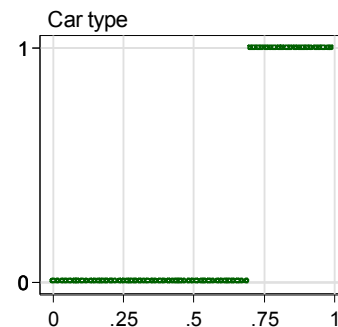
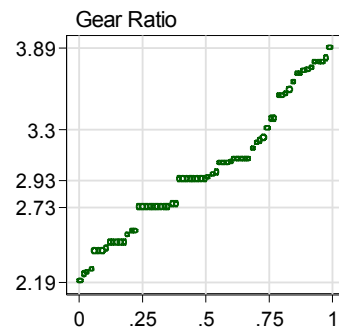
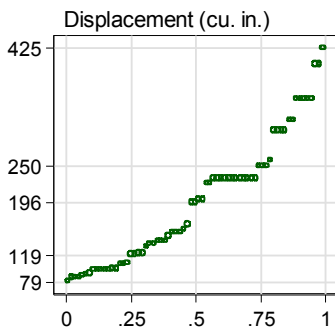
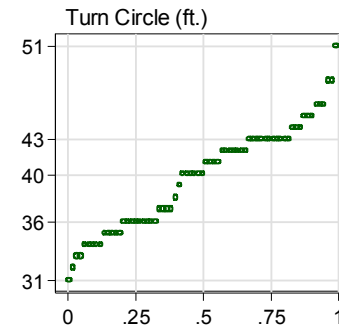
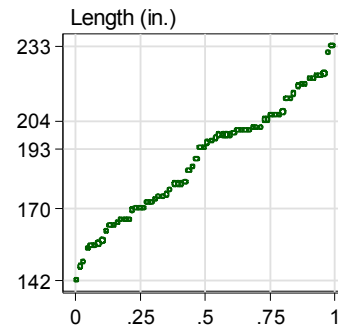
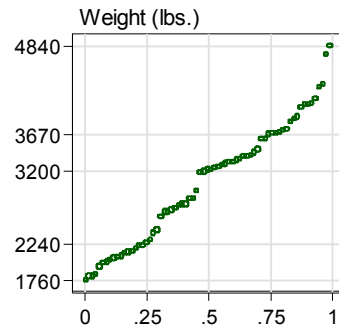
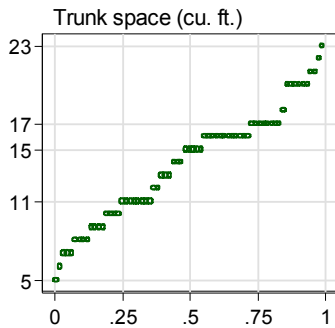
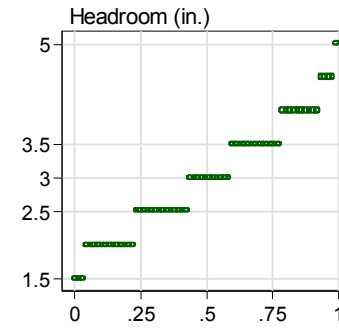
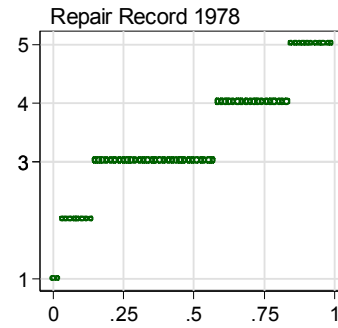
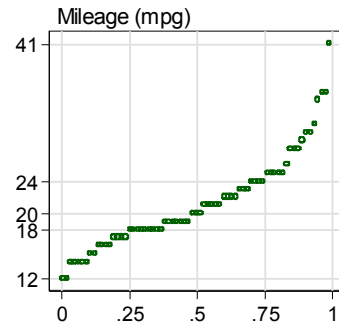
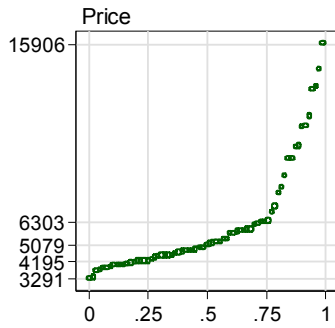
Multiple quantile plots

For exploring a bundle of numeric variables, likely to have very different ranges and units, `multqplot` (forthcoming) is offered.

The recipe is just to produce a `qplot` for each variable and then use `graph combine`.

A graph for each variable puts a premium on space. The variable's details go on top.

Values of selected quantiles are shown (by default 0(25)100, giving a box plot flavour).



Features of quantile plots

Show well any outliers, gaps or granularity.

Scale well over a large range of possible sample sizes.

Entail a minimum of arbitrary choices.

Signal behaviour that might be awkward in modelling.

Behave reasonably with ordinal or binary variables.

For more propaganda: *Stata Journal* 5: 442–460 (2005).

Distribution and survival plots

Those who prefer distribution plots with axes interchanged will find a command in `distplot` (*SJ*).

The convention is to plot cumulative probability against magnitude.

Those who plot survival functions are likely to be working already with `sts` graph.

When to write a new graphics command?

Sometimes you want a graph that is new to you. After checking that no command exists, most often you will plan to construct a graph using twoway commands.

Less often, it will be an application for graph dot, bar, hbar, box or hbox.

But play with do-files first.

Most advice is to plan program writing, but for small projects it makes as much sense to see what grows easily and naturally out of play.

Principles of laziness

Let official Stata do as much as possible.

Let other programs do as much as possible.

Don't generalise programs or add features too readily.

Don't trust what you didn't create.

Don't plan too much; play and see what works.

Assessing normal probability plots

Suppose you are assessing fit to a normal or Gaussian distribution.

`qnorm` is a dedicated official command for normal probability plots (which are in fact quantile-quantile plots).

How much departure from a straight line is acceptable?

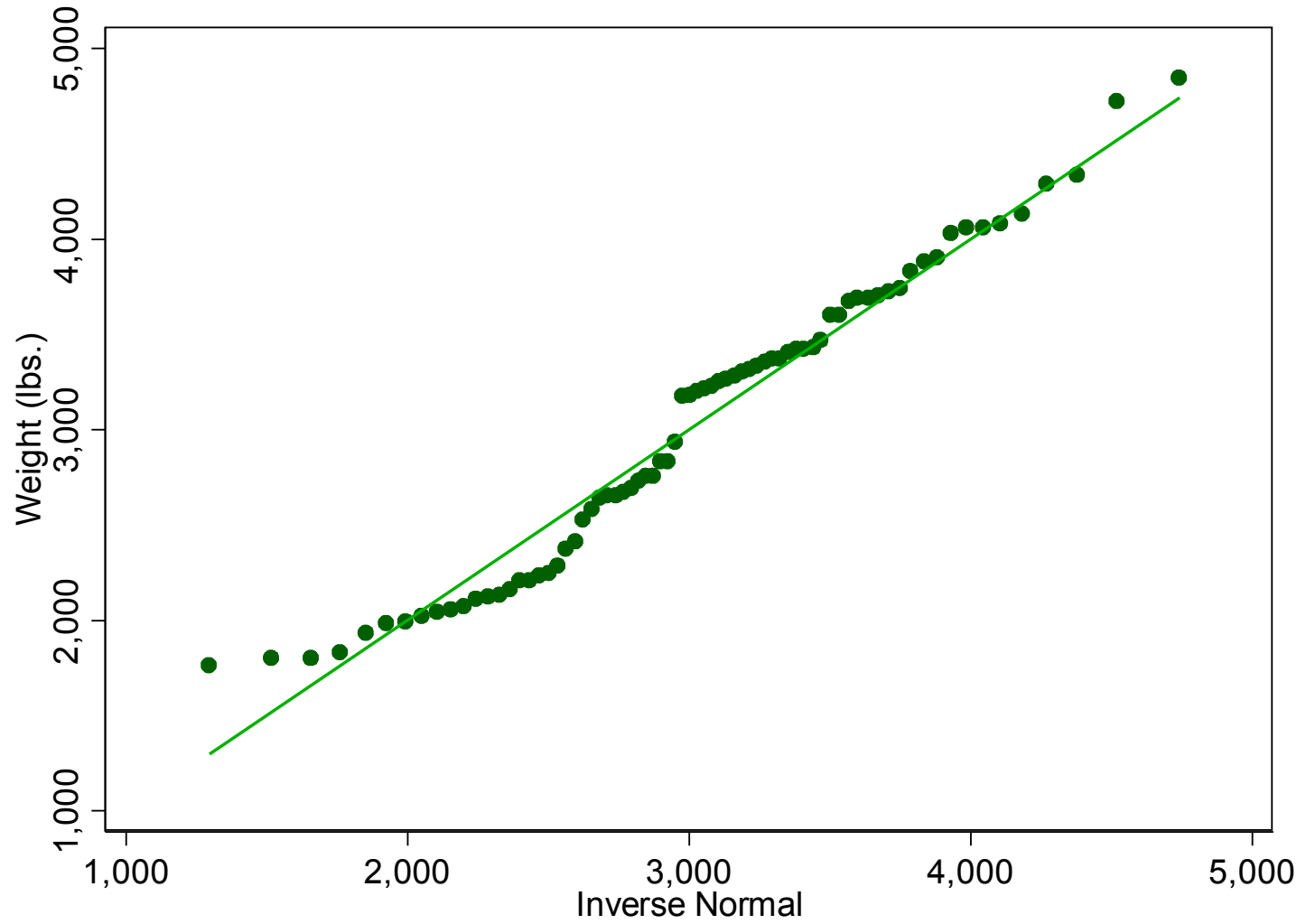
(If you really want a formal test, there are plenty on offer.)

A plot is a sample statistic

Even in exploration, the attitude that a plot from a sample is a sample statistic, just like a sample mean or a slope estimate, is always salutary.

So we should be worrying about how the plot that we do have — from our one sample — lies within a sampling distribution of possible plots for different samples.

The auto dataset gives a sample of 74 car weights.



Envelope curves

One recipe suggested is to get envelopes by

simulating several samples of the same size from a Gaussian with the same mean and SD

sorting each sample from smallest to largest

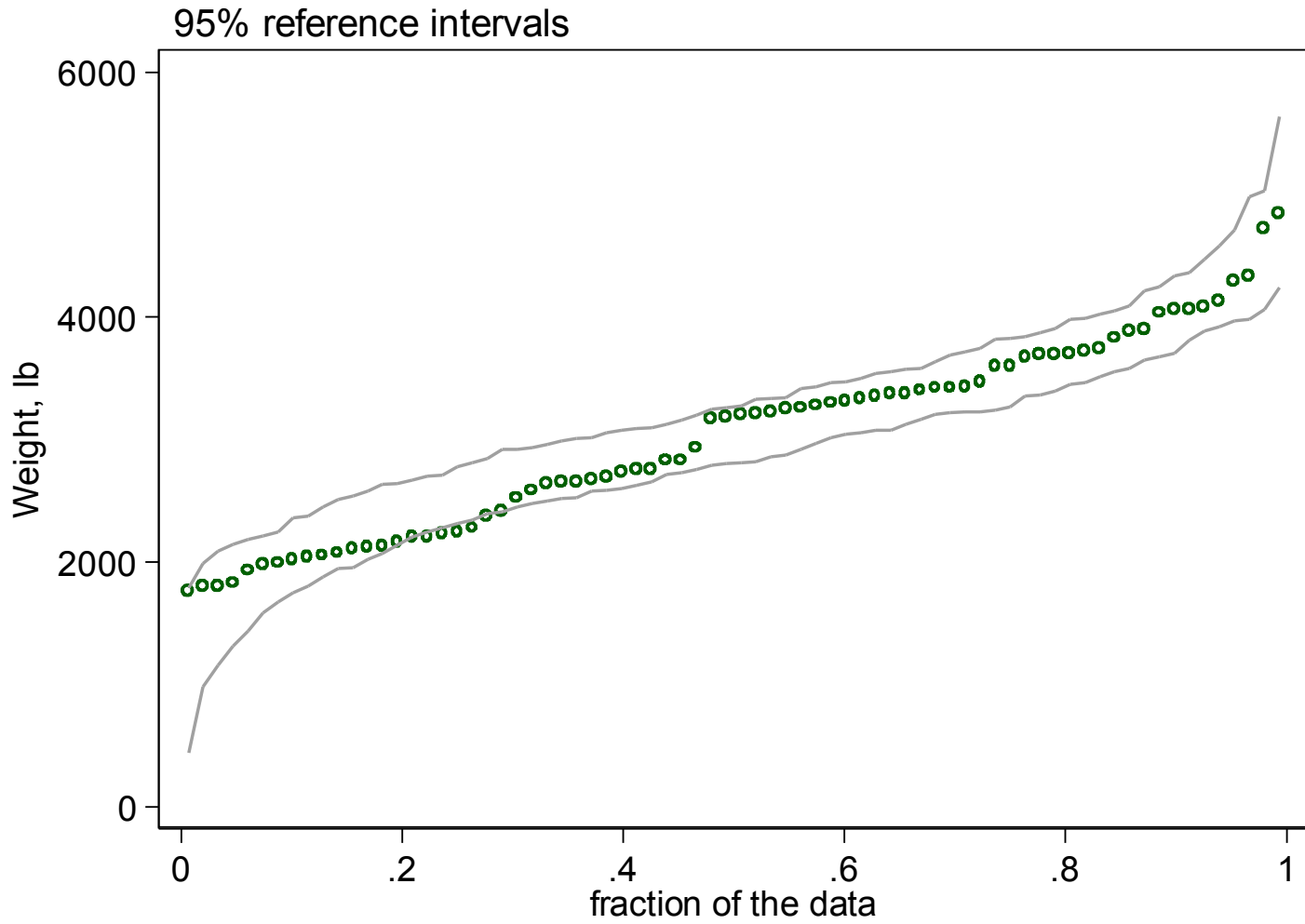
reporting results for each order statistic as an interval (e.g. spanning 95% of results)

One solution (mine)

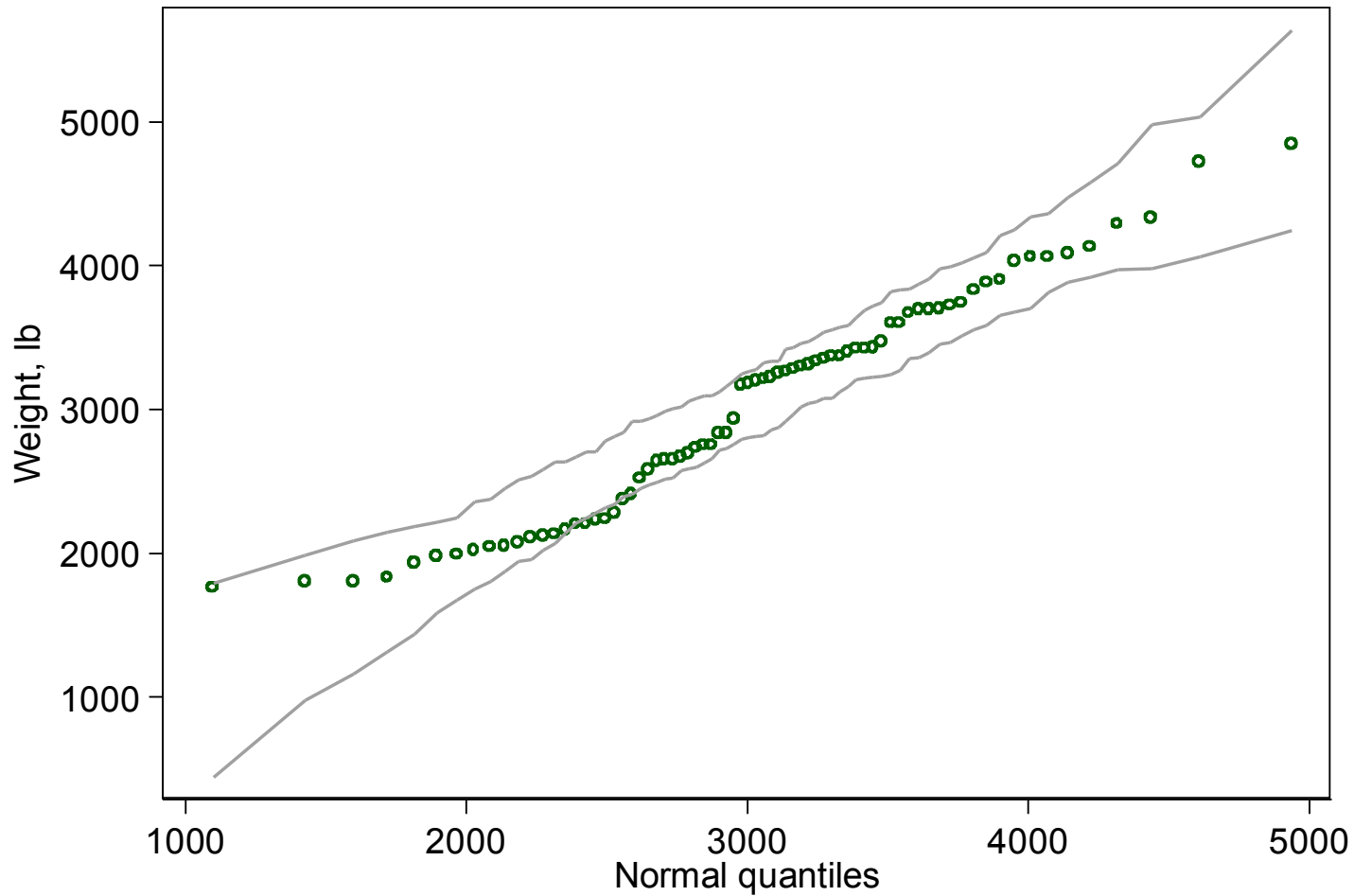
Write a helper program, `qenvnormal` (forthcoming), that calculates the envelopes. Mata is the work-horse.

`qplot` is already general enough to plot the envelopes too, so there is no need for an extra graphics program.

Stifle the urge to extend the first program to include all my favourite distributions (gamma, lognormal, etc.)



95% reference intervals



qqplot solutions

`qqplot` is able to plot observed quantiles and the envelopes, against both fraction of the data and normal quantiles.

By the way, `qenvnormal` warns if there is a reversal of order in the generated envelopes, best taken to mean that the number of replications is too small.

Going gray

A detail in several graphs worth flagging is the usefulness of gray colours for less important elements such as grid lines.

For more, see *Stata Journal* 9: 499–503 and 9: 648–651 (2009).

The aim

... delight lies somewhere between boredom and confusion.

Sir Ernst Gombrich (1909–2001)

1984. *The sense of order: A study in the psychology of decorative art.*

Oxford: Phaidon, p.9.

Bits and pieces follow

The ratings to follow are subjective ratings based on how often I see such graphs compared with how often they seem about the best possible.

A really good graph used when appropriate would come in the middle on such a rating.

Graphs for measured variables

underrated

quantile plots

strip plots

distribution plots

survival plots

density plots

histograms

box plots

overrated

Graphs for categorical variables

underrated

dot charts

multi-way bar charts

side-by-side bar charts

stacked bar charts

spineplots

mosaic plots generally

pie charts

overrated

Presentation notes for font freaks and similar strange people

The main font is Georgia.

Stata syntax is in `Lucida Console`.

The graphs use Arial. I nearly used Gill Sans MT.

The Stata graph scheme is `s1color`.