The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 4, pp. 991–996

Stata tip 121: Box plots side by side

Nicholas J. Cox
Department of Geography
Durham University
Durham, UK
n.j.cox@durham.ac.uk

Box plots are a standard plot type in statistical graphics and, as such, are popular with Stata users. The official Stata commands `graph box` and `graph hbox` are identical except that `graph box` draws box plots with the response (or outcome) scale on the vertical axis and `graph hbox` draws plots with the response scale on the horizontal axis. Contrary to the usual mathematical convention, the response axis is always regarded as the y axis for these commands so that options such as `ytitle()`, `ylabel()`, and `yscale()` always apply to the axis with the response variable. The manual entry [G-2] **graph box** gives much more detail and pertinent references. For a wider discussion of box plots, including how to draw box plots and related plots with `graph twoway`, see Cox (2009, 2013).

The greatest value of box plots is for comparing distributions of related variables or distributions of single variables for different groups of observations. This tip focuses on how and which data are plotted side by side. I explain the default appearance and structure of side-by-side box plots and how to tune or even to reverse that default.

To make this question concrete, we read in some data and then plot some graphs. As often happens, the code here is a cleaned-up version of what was done in preparing the tip, with afterthoughts and second guesses turned into anticipations of useful ideas.

```
. set scheme sj
. sysuse citytemp
. local title "Mean temperatures ({&degree}F)"
. label var tempjan "January"
. label var tempjul "July"
```

The `citytemp` dataset distributed with Stata contains temperature data for various U.S. cities. These are given in degrees Fahrenheit, a scale on which water freezes at 32° F and water boils at 212° F. Most countries of the world use the Celsius (formerly centigrade) scale ° C, for which water freezes at 0° C and boils at 100° C. The degree symbol can be shown as a text symbol, as explained in the help for `text`. If this functionality is not available in your Stata, you can use the trick explained in Cox (2004). One way or another, we create a local macro indicating units of measurement for use in later graphs. Because we plan to use a graph title explaining that we are showing temperatures, the month names suffice as variable labels.

Some examples of box plots are shown in figures 1 and 2. Figures 1a and 2a show vertical box plots for January and July temperatures in various regions of the United States, while figures 1b and 2b show corresponding horizontal box plots. The commands are as follows:

```
. graph box tempjan tempjuly, over(region) ytitle(`title´)
> ylabel(14 32 50 68 86, angle(h))
. graph hbox tempjan tempjuly, over(region) ytitle(`title´) ylabel(14(18)86)
. graph box tempjan tempjuly, by(region, rows(1) compact note(""))
> ytitle(`title´) ylabel(14(18)86, angle(h))
. graph hbox tempjan tempjuly, by(region, cols(1) compact note(""))
> ytitle(`title´) ylabel(14(18)86)
```

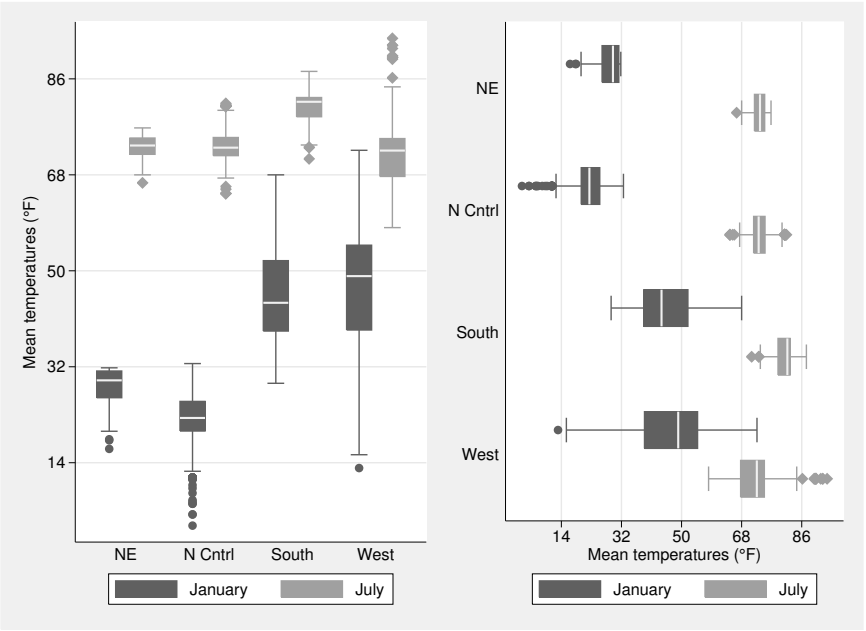

Figure 1. Box plots for January and July temperatures of various U.S. cities using the `over()` option to compare different regions

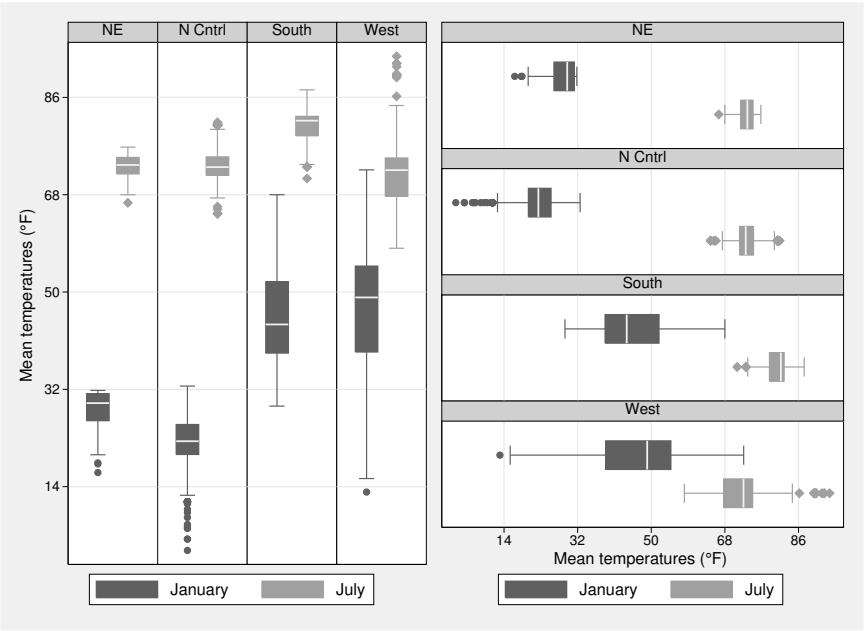

Figure 2. Box plots for January and July temperatures of various U.S. cities using the `by()` option to compare different regions

The axis labels 14(18)86 may seem a strange choice to U.S. readers, but 32° F is a key threshold, while differences of 18° F between labeled ticks match differences of 10° C. Figure 1 uses the `over()` option to compare different regions, while figure 2 uses the `by()` option to compare regions. In broad terms, the `by()` option is more flexible but produces more scaffolding. The scaffolding is sometimes helpful in indicating the subdivisions of the graph clearly but sometimes less helpful in that it may take up valuable space. Users aware of both syntaxes can make an informed choice.

What is less well known is that the `by()` option can be tuned so that results resemble those of the `over()` option. This trick may be applied more widely than just to box plots. Appropriate incantations tweak the position and appearance of the subtitles of the component graphs. It is convenient, but not essential, to define those incantations with local macros for repeated use in later commands. Note the clock notation for position, which places subtitles for vertical box plots at 12 o'clock and those for horizontal plots at 9 o'clock. Figure 3 shows the results.

```
. local incant1 subtitle(, position(12) ring(1) nobexpand bcolor(none)
> placement(n))
. local incant2 subtitle(, position(9) ring(1) nobexpand bcolor(none) placement(e))
. graph box tempjan tempjuly, by(region, rows(1) compact note(""))
> ytitle(`title´) ylabel(14(18)86, angle(h)) `incant1´
. graph hbox tempjan tempjuly, by(region, cols(1) compact note(""))
> ytitle(`title´) ylabel(14 32 50 68 86) `incant2´
```

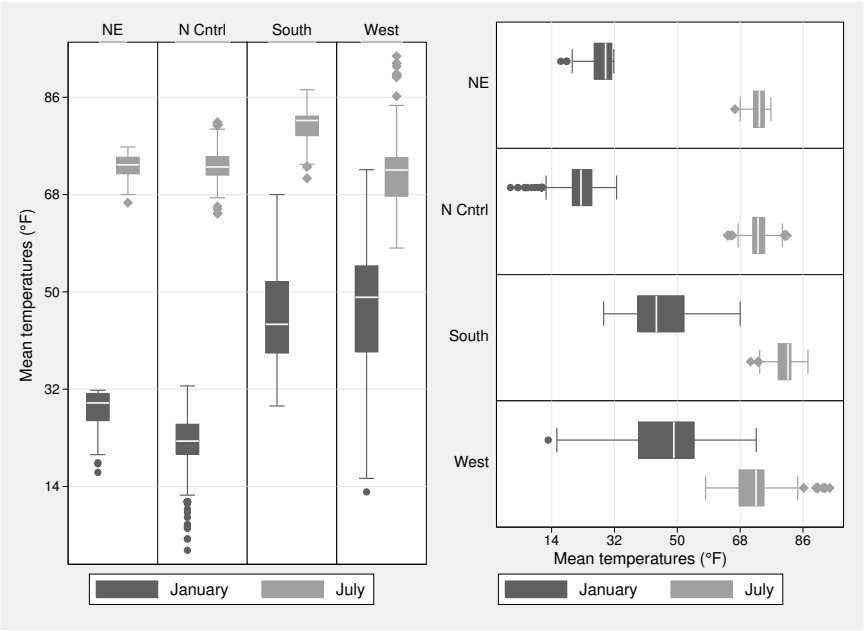

Figure 3. Box plots for January and July temperatures of various U.S. cities using the `by()` option to compare different regions, but with panel titles shown differently

Despite these minor variations, the design common to all the plots so far is that different variables are placed closest (on the inside, as it were) and groups of observations, as defined by the distinct values of the variable specified in `over()` or `by()`, are placed more broadly. What is to be done if the opposite order is wanted? Suppose that the contrast between January and July (Northern Hemisphere winter and summer) is thought less interesting than the contrasts between different regions. We then need regions, not months, to be next to each other.

For the opposite order, we need a different data structure, which can be obtained through the `reshape` command. If `reshape` is new to you, refer to the online help and manual entry. In this example, `reshape` stacks different variables into one variable that is subdivided by a group variable indicating where the groups came from. This is an easy change of data structure to envisage and one that is often needed.

The `citytemp` data lack an identifier variable naming the observations, here cities. We do need an identifier for `reshape`, but the observation number will work well.

```
. generate id = _n
```

In a very large dataset, we would make such an identifier of `long` storage type. Some judicious renaming of variables can also be a good idea:

```
. rename (tempjan tempjul) (tempJanuary tempJuly)
. reshape long temp, i(id) j(month) string
```

Now the combined variable `temp` can be grouped by `region`, as before, and also by `month`, a new variable created by `reshape`. We can choose which variable goes on the inside. In this example, we already suspect that comparing temperatures by region may be more interesting than comparing by month. With many other datasets (for example, medical results compared by sex and age group), you may need to experiment to see what works best. Comparisons between subtle effects of interest and starker but well-known effects often recur. Figure 4 shows the results of this example.

```
. graph box temp, over(region) by(month, rows(1) compact note(""))
> ytitle(`title´) ylabel(14(18)86, angle(h) grid) `incant1´
. graph hbox temp, over(region) by(month, cols(1) compact note(""))
> ytitle(`title´) ylabel(14(18)86, grid) `incant2´
```

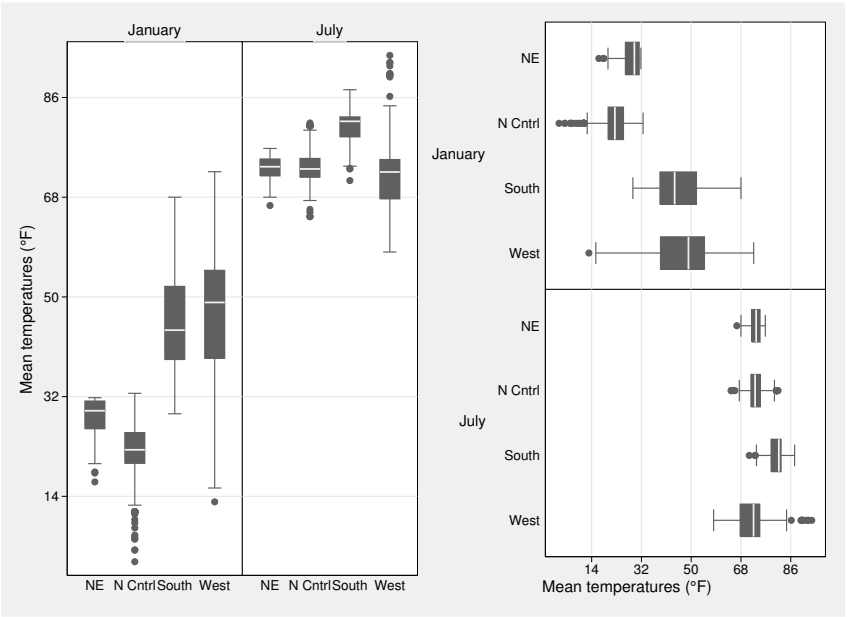

Figure 4. Box plots for January and July temperatures of various U.S. cities using both `over()` and `by()` options to compare different regions and months

References

Cox, N. J. 2004. Stata tip 6: Inserting awkward characters in the plot. *Stata Journal* 4: 95–96.

———. 2009. Speaking Stata: Creating and varying box plots. *Stata Journal* 9: 478–496.

———. 2013. Speaking Stata: Creating and varying box plots: Correction. *Stata Journal* 13: 398–400.