The Stata Journal (2024)
24, Number 4, pp. 766–776

DOI: 10.1177/1536867X241297949

Speaking Stata: Getting by without the by() option: Some graphics for unequal groups

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

Abstract. The `by()` option of the `graph` command is often used to show groups or subsets of some data in separate panels or facets of a graphical display. If groups are unequal in size, the result may seem awkward or inefficient in use of space. Devices to allow such groups to be shown directly without using a `by()` option are explained and exemplified for `graph dot` and its siblings and for `graph twoway`.

New variables to show rank within groups and (if needed) separation of groups are easily constructed. Group summaries such as medians may easily be added. Graph types shown are dot charts, quantile plots, and displays using spikes to show differences between variables. Data examples are for ocean salinity and changes in weight of anorexic girls.

Keywords: gr0098, graphics, distributions, groups, dot charts, quantile plots, paired data, change, comparisons, `by()` option

1 The problem

Graphical comparison in Stata of data or results for two or more groups often hinges on showing those groups side by side using a `by()` option. If those groups have unequal numbers of observations, the resulting display may seem at best a little awkward and at worst inefficient in its use of space. The problem is that the panels or facets created with `by()` have the same size and shape, regardless of their contents.

This column focuses on two solutions for simple descriptive or exploratory plots using some slight trickery.

First, `graph dot`, `graph bar`, and `graph hbar` are more flexible than they may seem. These commands are evidently focused on showing some outcome or summary as it changes with one or more categorical predictors. So the trick is just to create a predictor that helps with what you want.

Second, using `twoway` to show groups side by side is eased if an axis variable, usually to be shown horizontally, is created for the purpose of showing additional space between groups. That done, extra details such as axis titles, labels, and ticks may be added simply using standard syntax.

The value of these small tricks is greatest with smallish datasets, say, whenever the number of observations is of the order of tens or hundreds, not thousands or millions.

But with datasets that small, researchers not only can but also should use graphical designs that show all the fine structure in the data and allow an overview of their broad structure. Too many reports reach for histograms, box plots, or more modish designs that suppress or obscure fine details that may be interesting or important.

Tufte (2020, 100–101) urged the point vigorously: "Detailed data moves closer to the truth. No more binning, less cherry picking, less truncation. ... To improve learning from data, credibility, and integrity, show the data."

Two reservations should be flagged at the risk of emphasizing the obvious. Comparisons side by side or juxtaposed use just one graphical style: showing data or results superimposed in the same space is often possible or even preferable. The examples here are just indicative and not exhaustive of the possibilities.

Writing this column arose, as often, from explaining *how* to use Stata to get better results. But, as always, *what* data you have, *why* you are drawing a graph, and *who* will be reading your graph are also crucial.

2 Exploiting over() options in graph dot, graph bar, or graph hbar

The sibling commands `graph bar`, `graph hbar`, and `graph dot` support `over()` options for different subsets of your data. These options adjust display automatically according to the number of distinct elements shown.

2.1 Salinity dataset

A small sandbox dataset of salinity measurements in three water masses of the Bimini lagoon of the Bahamas was given by Till (1974, 104) and reproduced by Hand et al. (1994, 203). There are 12, 8, and 10 measurements from the water masses. Till (1974, 104–113) showed no graphs; his main concern was one-way analysis of variance. A Stata-readable copy of the dataset is included with the media for this issue.

```
. set scheme stsj
. use salinity
(Till 1974. Statistical Methods for the Earth Scientist ... p.104. See notes.)
```

The data for each water mass are presented in what appears to be an arbitrary order. Without any information on the meaning of that order, we lose nothing by sorting before graphics. The same applies to the water mass identifiers, so we can calculate medians for each water mass (or any other summary if you prefer) as a preliminary to later sorting.

```
. bysort water_mass (salinity): generate rank = _n
. egen median = median(salinity), by(water_mass)
```

We could have used `egen`'s `rank()` function for the ranking, but if you choose to do so, make sure in this graphical context that you use its `unique` option to get unique or distinct integer ranks.

```
. graph dot (asis) salinity, over(rank, label(nolabels))
> over(water_mass, sort(median) gap(*3)) nofill vertical exclude0
> ylabel(36/41) b2title("Water mass") linetype(line)
> lines(lwidth(vthin) lpattern(solid)) name(GB1, replace)
```

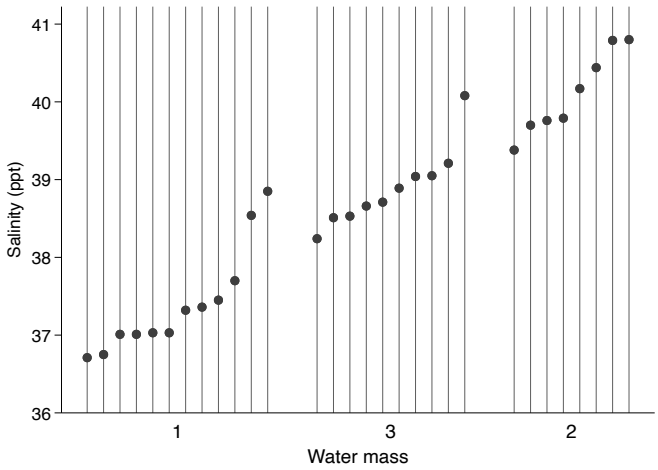

Figure 1. Dot chart of salinity measurements in three water masses, Bimini, Bahamas

Figure 1 is a dot chart of these data, showing a simple contrast in salinity levels between water masses. Some choices here arise from personal taste or preference. Other details deserve emphasis.

The new variable `rank` provides an identifier of measurements within each water mass, but we are not obliged to show its values as axis labels.

We have reordered the water masses. Note that you can tune the gap between the outer groups. Naturally, you may choose your own alternative to 3 or accept the default spacing.

The `nofill` option is essential for tidiness.

`vertical` is an undocumented option but is often useful. (Incidentally, `blabel(bar)` is another option that works fine with `graph dot` but is not documented in the manual entry.)

The salinity values vary over a narrow range. Showing zero not only is unnecessary but also would make comparisons much harder. Here and elsewhere, the best advice is usually as stated by Cleveland (1994, 93): "in science and technology *assume the viewer will look at the tick mark labels and understand them*."

The display has produced side-by-side quantile plots. See, for example, Cox (2005a, 2024) and those articles' references if you are curious for more discussion.

We will come back to this dataset in the next section. For the moment, save the revised dataset:

```
. save salinity2, replace
file salinity2.dta saved
```

2.2 Anorexia data

A larger and more challenging example concerns weights before and after various treatments of anorexic girls. The data were provided by the statistician B. S. Everitt and are given in Hand et al. (1994, 229). The weights are said to be in kilograms, but—as also pointed out by McNeil (1996, 57)—must be in pounds. McNeil's treatment includes some excellent graphs and inspired the choice of this example.

The treatments were given as cognitive behavioral treatment (29 observations), a control group (26 observations), and family therapy (17 observations). There seems to be no reason to follow this alphabetical order in presentation. For previous discussion and several graphs, see Cox (2009). Unequal group size was not mentioned then but is the twist now addressed directly.

The focus is the effectiveness of each positive treatment compared with the control of no treatment, so direct display of weight change will be of interest. For a first stab, we look at the data as provided but use the order of group medians of weight change. As before, use any different criterion if you prefer, especially if that makes more sense for your own dataset. In particular, any interest in analysis of variance should be matched by use of means.

```
. use anorexia, clear
(Hand et al. 1994. A Handbook of Small Data Sets. London: Chapman & Hall, p.229)
. generate weight_change = after - before
. format weight_change %3.0f
. egen median = median(weight_change), by(treatment)
. bysort treatment (before): generate rank = _n
. graph dot (asis) before after, over(rank, label(nolabels))
> over(treatment, sort(median) gap(*5)) nofill marker(1, msymbol(Oh))
> marker(2, msymbol(+)) vertical exclude0 ylabel(65(5)105)
> ytitle("Weights before and after treatment (pounds)")
> linetype(line) lines(lwidth(vvthin)) legend(pos(12)
> order(1 "before" 2 "after")) b2title("Treatment") name(GB2, replace)
```

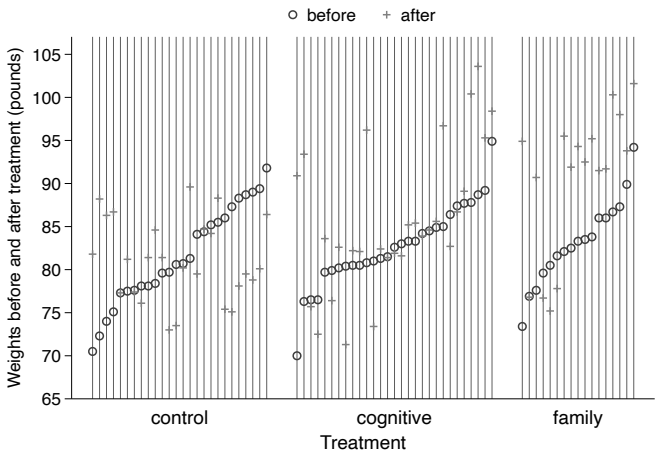

Figure 2. Dot chart or quantile plot of weights before and after treatment for various anorexic girls

Figure 2 is a dot chart of weights before and after treatment. Given sorting within treatments, it is also a quantile plot. A broad contrast emerges fairly clearly: the control group shows about equal frequencies of weight gain and weight loss and hints strongly at a regression effect too. Of the two treatments intended positively, family therapy appears more effective. There are the usual reservations about sample size, so many readers would be inclined to proceed to more formal modeling and testing.

To get a feeling for magnitudes, readers in most countries may like to know that 30 (40, 50) kilograms are about 66 (88, 110) pounds, spanning the range shown.

What is different from the previous example? Group sizes are larger, and we are plotting two outcome variables rather than one. Because some weight changes were very small, there is some risk of overplotting. Although it is currently unfashionable in statistical graphics, I repeat the advice of Cleveland (1994, 164) that open circle and plus markers are a good pair because they are more easily visible even for identical values.

As with the previous example, we will return to this dataset and so save it for now:

```
. save anorexia2, replace
file anorexia2.dta saved
```

3 Use of twoway with a customized axis variable

We now switch to seeing how far we can get with `twoway` graphs. There is one major trick: producing a customized axis variable. Other minor tricks follow in its wake.

3.1 Salinity data

We read in the data again.

```
. use salinity2, clear
(Till 1974. Statistical Methods for the Earth Scientist ... p.104. See notes.)
```

In case water masses tie on medians, we insist on sorting on the lower median first.

```
. sort median water_mass rank
```

The observations are now in the order we desire. It will help perception if we insert extra space between each group (here each water mass). The customized axis variable increases by 1 each time we see a new observation and by 2 each time we see a new group. Choose your own alternative to 2 if you need a bigger or smaller space. We are going to add grid lines wherever there is a data point. `levelsof` gives us all the locations and puts their values in a single local macro.

```
. generate axis = sum(2 * (water_mass != water_mass[_n-1])) +
> sum(rank != rank[_n-1])
. levelsof axis, local(levels)
3 4 5 6 7 8 9 10 11 12 13 14 17 18 19 20 21 22 23 24 25 26 29 30 31 32 33 34 35 36
```

We need to label each group clearly. Optionally, you could add dividing lines between groups, thus getting closer to the default style with the `by()` option. We could work out label and line positions using simple arithmetic but prefer to automate the calculations. (As usual, automation could go even further. If you can identify a style you like and will use repeatedly, writing a more general do-file or program would be a good idea.)

The labels belong opposite mean positions for each group. Any dividing lines belong after the last (maximum) position for all groups except the last group.

```
. forvalues g = 1/3 {
  2. summarize axis if water_mass == `g', meanonly
  3. local pos`g' = r(mean)
  4. local line`g' = r(max) + 1.5
  5. }
```

Because we are now dealing with `twoway`, it is simple to add other `twoway` calls for the same graph space. One possibility is to show the medians we calculated. However, if you try that with some variant on `line median axis`, you will get spurious connections between each group. The remedy is to apply `separate` to get separate variables containing the medians, one for each group. Although it has no visible effect here, we flag the `veryshortlabel` option of `separate` (Cox 2005c), which is often useful for graphical work.

For present purposes, we insist that all medians are shown in the same way. For a presentation, you might go in the opposite direction, choosing different line colors and patterns and different marker properties. However, some consistency of colors is a very good idea. That is, the median line and the markers for each group are best shown in the same color.

```
. separate median, by(water_mass) veryshortlabel
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
median1         float   %9.0g                 1
median2         float   %9.0g                 2
median3         float   %9.0g                 3
. line median? axis, lcolor(black ..) lpattern(dash ..) || scatter salinity axis,
> msymbol(O) xline(`levels', lpattern(solid) lwidth(vthin))
> xlabel(`pos1' "1" `pos2' "2" `pos3' "3", tlcolor(none))
> ytitle("Salinity (ppt)") ylabel(36/41) xtitle("Water mass") legend(off)
> note("Group medians shown") name(GB3, replace)
```

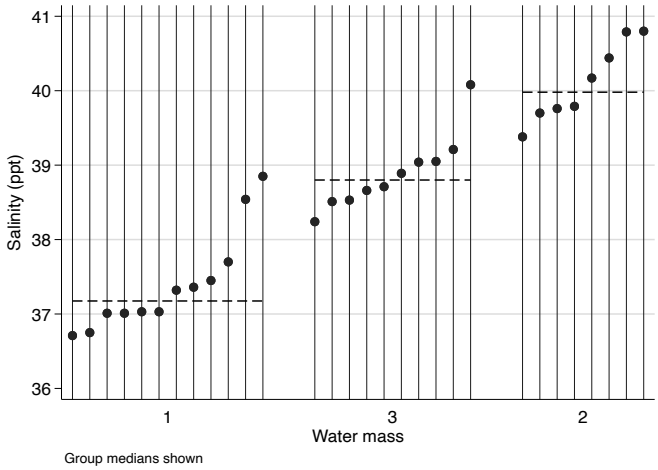

Figure 3. Dot chart or quantile plot of salinity measurements in three water masses, Bimini, Bahamas. Medians are shown for each water mass.

Figure 3 therefore goes beyond figure 1, especially in the sense that medians are shown too. Horizontal line segments for medians as well as dividing lines help to bind the groups visually. Again, we stress that other summaries could be better for your own data, depending. Just means? Or something customized such as geometric or trimmed means?

3.2 Anorexia data

For our last burst, we revisit the anorexia data. The same basic tricks are used, starting with a customized axis variable.

```
. use anorexia2, clear
(Hand et al. 1994. A Handbook of Small Data Sets. London: Chapman & Hall, p.229)
. sort median treatment rank
. generate axis = sum(3 * (treatment != treatment[_n-1])) +
> sum(rank != rank[_n-1])
. forvalues g = 1/3 {
  2. summarize axis if treatment == `g', meanonly
  3. local text`g' : label (treatment) `g'
  4. local pos`g' = r(mean)
  5. local line`g' = r(max) + 2
  6. }
```

One clean display (see Cox [2009] for exploration and discussion) sorts on weights before treatment and represents change by vertical spikes. Some readers may prefer arrows (Cox 2005b).

```
. twoway rspike before after axis || scatter before axis, xline(`line1' `line2')
> xlabel(`pos1' "`text1'" `pos2' "`text2'" `pos3' "`text3'", tlcolor(none)
> tlength(0.5)) ylabel(65(5)105)
> ytitle("Weights before and after treatment (pounds)") msymbol(O)
> legend(pos(12)) xtitle("Treatment") legend(order(2 "before" 1 "change"))
> name(GB4, replace)
```

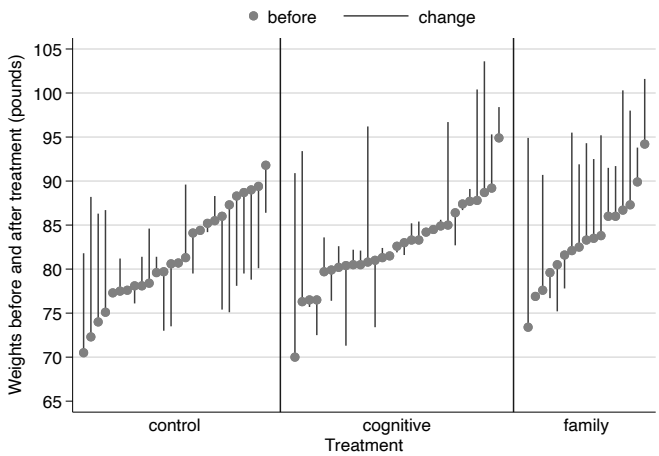

Figure 4. Weights before and after treatment for various anorexic girls. Subjects are ordered according to their weights before treatment. Upward and downward spikes show change after treatment.

Figure 4 is thus another take on weight change, which is now more nearly explicit.

We finally take this further by showing weight change directly. The plots are again quantile plots. The sting here is that the weights before and after treatment are no longer shown.

We need new rank and axis variables. The `separate` trick may be repeated.

```
. bysort median treatment (weight_change): generate rank2 = _n
. generate axis2 = sum(3 * (treatment != treatment[_n-1])) +
> sum(rank2 != rank2[_n-1])
. separate median, by(treatment) veryshortlabel
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
median1         float   %9.0g                 cognitive
median2         float   %9.0g                 control
median3         byte    %9.0g                 family
. line median? axis2, lcolor(black ..) lpattern(dash ..) || scatter
> weight_change axis2, msymbol(+) xline(`line1' `line2') ylabel(-10(5)20)
> xlabel(`pos1' "`text1'" `pos2' "`text2'" `pos3' "`text3'", tlcolor(none)
> tlength(*0.5)) ytitle("Weight change after treatment (pounds)")
> xtitle("Treatment") xscale(titlegap(*3)) legend(off)
> note("Group medians shown") name(GB5, replace)
```

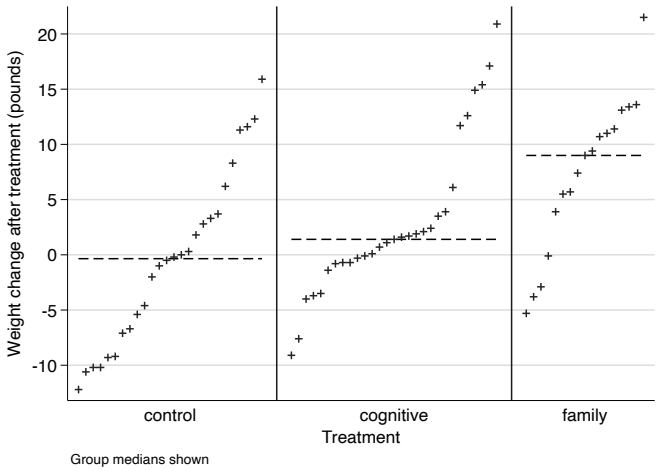

Figure 5. Quantile plots of weight changes experienced by anorexic girls by treatment. Medians for each treatment are shown by horizontal lines.

Figure 5 is our final take on weight changes. It does clarify the contrasts between the control group and the two positive treatments. For any journal or thesis publication, fitting some suitable model might now be considered.

As with all graphs here, all kinds of small and large variations remain possible, including emphasis on zero change as a crucial dividing line between weight gain and weight loss or just use of different colors.

4 Conclusion

In a previous tip (Cox 2020), the focus was on how the `by()` option can be used as an alternative to `graph combine`. This column has complementary intent: to show how the `by()` option is not always needed, even for side-by-side comparison of groups. In detail, using `graph dot` and its siblings and using `twoway` allow different trickery. The obvious motive for their use is having unequal group frequencies, but nothing prevents their use even if groups are in fact equal in size.

The aim is to show some useful devices without too much sales pitch. Your own datasets might need something similar yet also something different. Although superficially contrasting, the two datasets used as examples have in common that identifiers are not available. For the salinity data, the reason is that the original details (location? time?) were never published. For the anorexia data, the primary reason is presumably maintaining confidentiality of sensitive data. In any case, identifiers would not help analysis. Other way round, the rank identifiers constructed to make the graphs possible at all could be suppressed without loss. But data for known entities, identified in, say, time or space, might benefit from display of identifiers, which might, in turn, imply swapping axes to supply enough space for readable labels.

5 References

Cleveland, W. S. 1994. *The Elements of Graphing Data*. Rev. ed. Summit, NJ: Hobart.

Cox, N. J. 2005a. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460. https://doi.org/10.1177/1536867X0500500312.

———. 2005b. Stata tip 21: The arrows of outrageous fortune. *Stata Journal* 5: 282–284. https://doi.org/10.1177/1536867X0500500214.

———. 2005c. Stata tip 27: Classifying data points on scatter plots. *Stata Journal* 5: 604–606. https://doi.org/10.1177/1536867X0500500412.

———. 2009. Speaking Stata: Paired, parallel, or profile plots for changes, correlations, and other comparisons. *Stata Journal* 9: 621–639. https://doi.org/10.1177/1536867X0900900408.

———. 2020. Stata tip 139: The by() option of graph can work better than graph combine. *Stata Journal* 20: 1016–1027. https://doi.org/10.1177/1536867X20976341.

———. 2024. Speaking Stata: Quantile–quantile plots, generalized. *Stata Journal* 24: 514–534. https://doi.org/10.1177/1536867X241276114.

Hand, D. J., F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, eds. 1994. *A Handbook of Small Data Sets*. London: Chapman and Hall. https://doi.org/10.1201/9780429246579.

McNeil, D. 1996. *Epidemiological Research Methods*. Chichester, UK: Wiley.

Till, R. 1974. *Statistical Methods for the Earth Scientist: An Introduction*. London: Macmillan.

Tufte, E. R. 2020. *Seeing with Fresh Eyes: Meaning, Space, Data, Truth*. Cheshire, CT: Graphics Press.

About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 16 commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is Editor-at-Large of the *Stata Journal*. His "Speaking Stata" articles on graphics from 2004 to 2013 have been collected as *Speaking Stata Graphics* (2014, College Station, TX: Stata Press). He is the Editor of *Stata Tips, Volumes I and II* (2024, also Stata Press).