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.

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

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- Science Citation Index Expanded (also known as SciSearch®)
- CompuMath Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

The Stata Journal (2008)
8, Number 1, pp. 105–121

Speaking Stata: Spineplots and their kin

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

Abstract. The term *spineplot* has been applied over the last decade or so to a type of bar chart used particularly for showing frequencies, proportions, or percentages of two cross-classified categorical variables. The principle is that the areas of rectangular tiles are proportional to the frequencies in the cells of a contingency table. Often both coarse and fine structure are easy to see, including departures from independence. The main idea has, in fact, been rediscovered repeatedly over at least the last 130 years. In its most general form, it has been widely publicized under the name *mosaic plots*. This column introduces, discusses, and exemplifies a Stata implementation of spineplots. It is noted that a restriction to two variables is more apparent than real, as either axis of a spineplot can show a composite variable defined by cross combinations of two or more variables.

Keywords: gr0031, spineplots, mosaic plots, bar charts, graphics, categorical data

1 Introduction

The recent history of categorical data analysis within statistical science has been marked by increasing convergence with what might reasonably be dubbed continuous data analysis. Even a generation ago, categorical data analysis was little more to practitioners in many quantitative fields than a ragbag of chi-squared tests and measures of bivariate association. (Indeed, even now many introductory texts appear to offer little more.) In stark contrast, those looking at continuous response variables could exploit a steadily more coherent and powerful toolbox based on regression and ANOVA, seen as members of a family of linear models. However, much greater focus in categorical data analysis over the last few decades on models of various kinds, including log linear and logit models and their several relatives, has greatly lessened the contrasts between the two major parts of statistical practice (see, for example, Agresti [2002]).

One facet of categorical data analysis which continues to receive uneven attention is the use of graphical methods. It is often argued (for example, by Tufte [2001]) that tabular displays, whether of data or summaries or model results, may be more effective or informative than graphs for many categorical problems. Nevertheless, various plotting methods have been suggested for such problems. Friendly (2000) surveys many recent innovations, but none yet appear to challenge bar charts as the most popular graphical method for categorical data.

Bar charts provoke a range of reactions from statistically minded people. Some charts showing only a few frequencies may strike readers as a waste of space in any

outlet supposedly aimed at intelligent adults or as too elementary or trivial to deserve much coverage in professional literature. Yet there are many reasons for thinking that bar charts may complement tables helpfully, particularly when the bar charts are well designed and well chosen.

In a previous column, I reviewed some ways of producing such charts in Stata for categorical data (Cox 2004). In this column, I focus on what are now widely known as *spineplots*, discussing the main ideas of spineplots and showing a Stata implementation. The term may be new to you, but the idea may yet be familiar; in any case, it will not appear strange. Spineplots grow out of the basic graphical notion that area may usefully encode frequency, which underlies several other standard forms, including histograms.

2 Spineplots

Names should not matter, but they do. Labels should matter much less than the underlying ideas. A wind rose or a stem-and-leaf plot by any other name is just as sweet, or as prickly, an idea. Yet across times and places and disciplines, all sorts of minor and major confusions can arise when the same name is used for different things, different names are used for the same thing, or authors unthinkingly assume that readers have had the same education and experience and possess the same terminology. Explaining what is, and what is not, a spineplot—or more precisely what is and is not done by the Stata program `spineplot`—thus requires attention to usages in the literature.

The name *spineplot* is credited to Hummel (1996). The term is gaining in popularity but appears already to be differently understood. In the strictest definition, spineplots are one-dimensional, horizontal stacked bar charts, but many discussions and implementations allow vertical subdivision (e.g., by highlighting) into two or possibly more categories. Some literature treats spineplots, as understood here, under the heading of *mosaic plots* (or *mosaicplots*), variously with and without also using the term *spineplot*.

The Stata implementation `spineplot` discussed here adopts a broad interpretation of the term. It works on two categorical variables—not one—and conveys the frequencies shown in a two-way contingency table. (One-dimensional, horizontal stacked bar charts have long been possible in Stata; in Stata 8 the official command `graph hbar` became available.) Conversely, the implementation here does not purport to be a general mosaic plot program capable of producing mosaic plots given three or more categorical variables.

Textbooks and monographs with examples of spineplots and related plots include Friendly (2000); Venables and Ripley (2002); Robbins (2005); Unwin, Theus, and Hofmann (2006); Young, Valero-Mora, and Friendly (2006); and Cook and Swayne (2007). Among several papers, Hofmann's (2000) discussion is clear, concise, and well illustrated.

Mosaic plots, including spineplots as a special case, have been reinvented several times under different names. Hartigan and Kleiner (1981, 1984) introduced them, or reintroduced them, into mainstream statistics. Friendly (2002) cites earlier examples, including the work of Georg von Mayr (1877), Karl G. Karsten (1923), and Erwin J.

Raisz (1934). Hofmann (2007) discusses a mosaic by Francis A. Walker (1874). Other early examples are those of Willard C. Brinton (1914, quoting earlier work), Berend G. Escher (1924), and Hans Zeisel (1947, 1985).[1] Further, independent reinventions of the idea continue to appear (e.g., Bertin [1983]; Feinstein and Kwoh [1988]; and Feinstein [2002]).

3 First examples

Examples will convey the essence far better than a word description. With a nod to Stata tradition, fire up Stata with the `auto` data, and look at the cross classification of two categorical variables: whether cars are foreign (from outside the United States) `foreign` and their 1978 repair record `rep78`. Repair record may be considered to be a response variable; hence, as with scatter plots and the Stata command `scatter`, it is named first to `spineplot` as the variable to be shown on the y axis. `spineplot` does not try to be smart about colors, nor does it know whether a categorical variable is ordered (ordinal) or not (nominal). Thus we here skip the default and move directly to specifying an ordered series of gray scales for bar colors (figure 1):

```
. sysuse auto
(1978 Automobile Data)

. spineplot rep78 foreign, bar1(bcolor(gs14)) bar2(bcolor(gs11))
> bar3(bcolor(gs8)) bar4(bcolor(gs5)) bar5(bcolor(gs2))
```

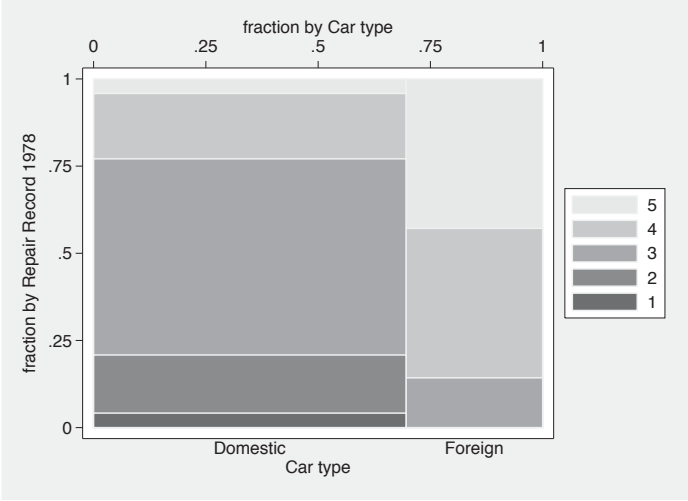

Figure 1. Spineplot of repair record and whether foreign for 74 cars, as produced by `spineplot`

1. See Anonymous (1967), Robinson (1970), Sills (1992), Anderson (2001), and Hertz (2001) for biographical pieces on several of these pioneers. Karl Karsten has been credited with the idea of hedge funds. Berend Escher is now better known as a brother of Maurits C. Escher, whose own mosaics are immensely more intricate and intriguing than any to be discussed here.

As you might guess, options like `bar1()` and `bar2()` override defaults for the first, second, and subsequent bars. Counting is from the top downward. Here the darkest gray scales show poor repair records. Adopting the reverse choice, or indeed any other choice of colors, is naturally at your discretion. Whatever the choice, the spineplot makes clear that foreign and domestic cars had very different distributions of repair record in 1978.

The graph structure is similar to the structure of a standard two-way contingency table, such as the one `tabulate rep78 foreign` would produce. One detailed difference is that high response values are in the last rows but toward the top of the y axis, reflecting table and graph conventions, respectively. Another detailed difference is that cells with zero frequency are represented in the spineplot by tiles of zero area, that is, not at all.

For interpretation of spineplots, note that cross classification of independent variables would yield tiles that align consistently, as the resulting conditional distributions would be identical. Conversely, departures from independence, or relationships between variables, are shown by failure of alignment. The fine structure of such departures is open to inspection, although limits are imposed by the low visibility of cells with low frequencies and thus low tile areas. Spineplots are especially useful when considering a null hypothesis of independence.

However, in some cases where independence is highly implausible, spineplots may not be particularly effective. A common example is assessing categorical agreement of observers or methods, the problem which to many users is that addressed by the `kappa` command ([R] **kappa**). Here the usual expectation is that the diagonal or near-diagonal cells of the contingency table would show much higher frequencies than those near the opposite corners. Such a pattern would indeed be obvious on a spineplot, but the coloring used in `spineplot` does not make further scrutiny especially helpful.

Be that as it may, let us consider how this spineplot differs from more conventional bar charts. Surprising although it may seem, official Stata offers no direct and obvious command for bar charts of categorical data. Two user-written commands, `catplot` and `tabplot`, are among the alternatives (Cox 2004). Both may be downloaded from the Statistical Software Components archive by using the `ssc` command (see [R] **ssc** for further information).

With `catplot`, there is considerable choice of format. Two close relatives of the spineplot are particularly pertinent. The first shows frequencies (figure 2):

```
. tabulate rep78 foreign
  (output omitted)
. catplot bar rep78 foreign, asyvars stack bar(1, bcolor(gs2))
> bar(2, bcolor(gs5)) bar(3, bcolor(gs8)) bar(4, bcolor(gs11))
> bar(5, bcolor(gs14)) legend(pos(3) col(1))
```

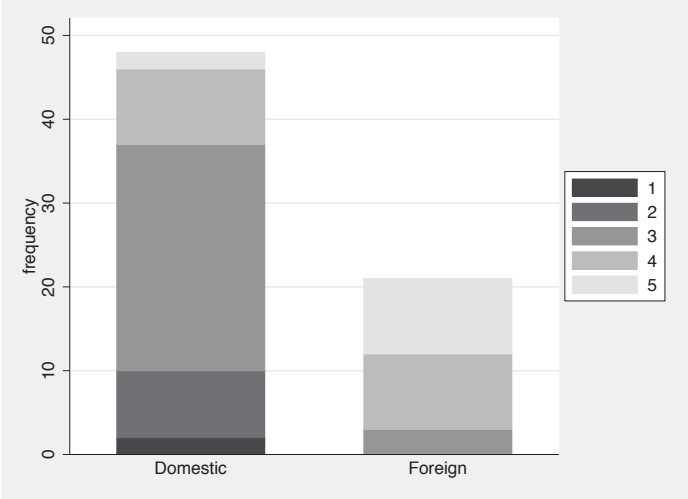

Figure 2. Bar chart of repair record and whether foreign for 74 cars, as produced by `catplot`

The second shows stacked percentages (figure 3):

```
. catplot bar rep78 foreign, asyvars stack percent(foreign) bar(1, bcolor(gs2))
> bar(2, bcolor(gs5)) bar(3, bcolor(gs8)) bar(4, bcolor(gs11))
> bar(5, bcolor(gs14)) legend(pos(3) col(1))
```

(*Continued on next page*)

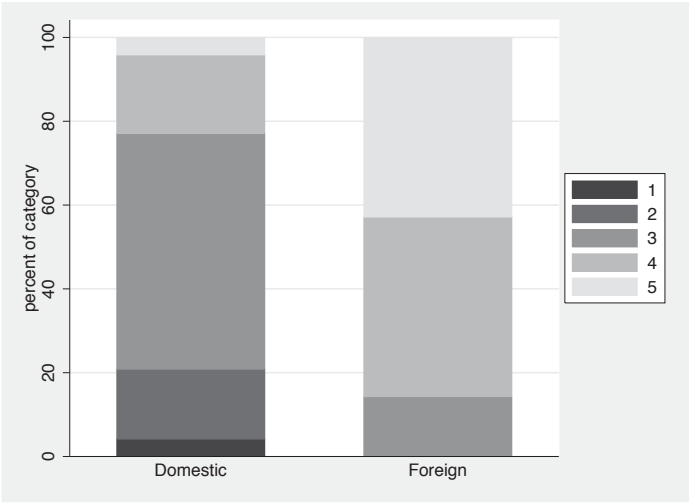

Figure 3. Bar chart of repair record and whether foreign for 74 cars, as produced by `catplot`, showing column percentages

With `catplot`, therefore, as with most bar chart software, it is easy to get a display of stacked frequencies. In that display, proportions or percentages are tacit and so often difficult to read off precisely. It is also easy to get a display of stacked percentages. In that display, the underlying frequencies are not in view. (In this case, `catplot` is a wrapper for `graph bar`, which might suggest the use of the `blabel()` option. But `blabel()` shows numerically what is being shown graphically, and we would want to show something else, so `blabel()` would not help.)

`tabplot` is another possibility. Here the percentage breakdown is shown in figure 4. Omitting the `percent()` option would yield a display of frequencies instead.

```
. tabplot rep78 foreign, percent(foreign) showval(format(%2.1f))
```

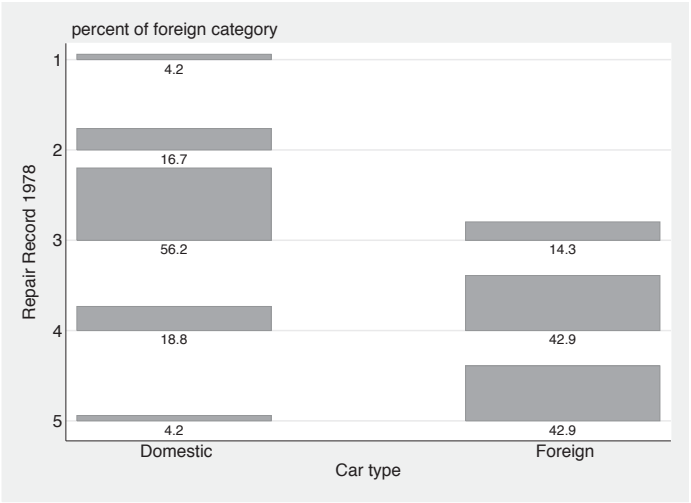

Figure 4. Tabular bar chart of repair record and whether foreign for 74 cars, as produced by `tabplot`, showing column percentages

This plot echoes the structure of a two-way contingency table even more clearly than does a spineplot. A glance at the code shows that much of the work within `tabplot` is done by a call to `twoway rbar`. But again there is a choice between showing frequencies and showing percentages. There is no scope for showing both simultaneously.

In sum: Spineplots show conditional distributions on both axes simultaneously. We can easily add information on absolute frequencies using the `text()` option (figure 5):

```
. by foreign rep78, sort: gen N = _N
. spineplot rep78 foreign, bar1(bcolor(gs14)) bar2(bcolor(gs11))
> bar3(bcolor(gs8)) bar4(bcolor(gs5)) bar5(bcolor(gs2)) text(N)
```

(*Continued on next page*)

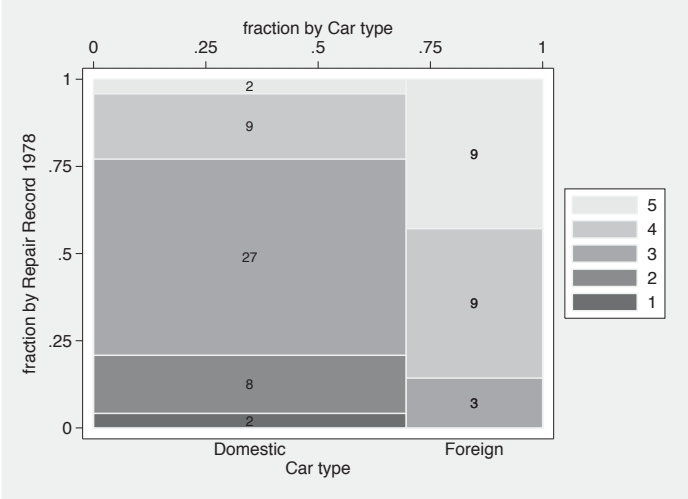

Figure 5. Spineplot of repair record and whether foreign for 74 cars, as produced by `spineplot`, with cell frequencies shown

Missing values in either of the two variables do not perturb the frequencies produced by the `generate` command above. The resulting frequencies are assigned but then ignored by `spineplot`. Conversely, empty cells of the contingency table do not, by definition, correspond to any observations, so counts of zero will not be shown. Combining the count with an `if` or `in` condition would require more care, but the details need not detain us now. Plotting something else, such as standardized residuals given some model, is another possibility. It would often be a good idea to impose a particular numeric format before display, say, by `string(`*residual*`, "%4.3f")`.

Most implementations of spineplots (and, more generally, mosaic plots) in other software omit axes and numerical scales and convey a recursive subdivision according to what may be several categorical variables by a hierarchy of gaps of various sizes. As the graphs produced by `spineplot` are restricted to two variables, axes and numerical scales are kept as defaults. The distinction between categories is conveyed by bar boundaries rather than explicit gaps. Naturally, there is scope for omitting graph elements not desired using standard `graph` options, or, in Stata 10 upward, the Graph Editor. Similarly, users may vary the thickness of bar boundaries, although thick boundaries would distort the relative sizes of what are perceived as bar areas.

The examples already seen raise other small matters of presentation.

First, note the possibility of using `plotregion(margin(zero))` to place axes alongside the plot region. Having a margin is often useful for scatterplots and their kin but is perhaps distracting for spineplots.

As with scatterplots, response variables are usually better shown on the y axis of spineplots. But as with scatter plots, there can be reasons for overriding that convention. (In the Earth or environmental sciences, plotting height above or depth below the land surface on the vertical axis is common and indeed often expected.) If one variable is binary, it is often better to plot that one on the y axis. The `foreign` variable is a case in point. Even though `foreign` is arguably a predictor of `rep78` rather than vice versa, I suggest that the spineplot with `foreign` on the y axis is more congenial. See figure 6 and judge for yourself. Notice that ordering of colors is now less of an issue, as any two distinct colors are ordered one way or the other.

```
. spineplot foreign rep78
```

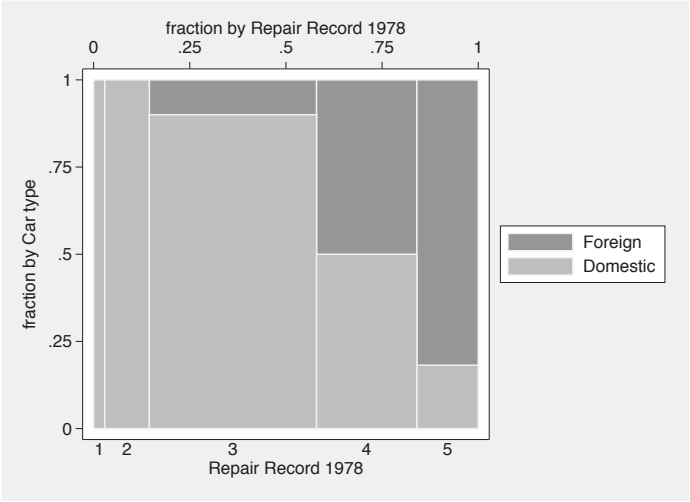

Figure 6. Spineplot of whether foreign and repair record for 74 cars, as produced by `spineplot`, with cell frequencies shown

Even more mundane, but very possibly troublesome in practice, is that if one or more cells have very small frequencies, then a squeeze of some sort is inevitable with `spineplot`. There is no way to show the corresponding tiles, or descriptive labels, or added text, without some difficulty. There are no easy solutions to this problem. You may decide to amalgamate cells; or to use the Graph Editor to ease crowding by moving text, adding arrows, and so forth; or just to use some other kind of graph. Manifestly, all kinds of graphs have some limitations on what they can show easily and effectively, and spineplots are no exception.

4 Discrimination at Berkeley?

A now classic problem among categorical analysts concerns the success or failure of applications for admittance as graduate students at the University of California, Berkeley. The problem was first discussed by Bickel, Hammel, and O'Connell (1975) and since then worked over in various ways in many articles and texts (e.g., Freedman, Pisani, and Purves [1978; 2007]; Friendly [2000]; and Agresti [2002]). Here we use a subset of the data presented by Friendly (2000) and Agresti (2002). The response is decision—admitted or rejected—and the covariates are intended major (masked by identifiers A, B, C, D, E, F) and sex of applicants. The data are available with the files for this column as `berkeley.dta`. They come as frequencies of the various cross combinations, so we must specify weights when we call up `spineplot`. (Alternatively, `expand`ing the dataset on the frequencies so that every individual application became an observation would make that unnecessary; see [D] **expand** for more.)

```
. use berkeley, clear
. spineplot decision sex [fw=frequency]
```

Figure 7. Spineplot of decision versus sex for admissions to various Berkeley graduate majors. At first sight, substantial discrimination against females is evident.

A spineplot of decision versus gender shows apparent discrimination against females (figure 7). However, majors are by no means equally easy to get into (figure 8). A corresponding `tabulate` shows that admission rates vary from 64% for A to 6.4% for F.

```
. spineplot decision major [fw=frequency]
```

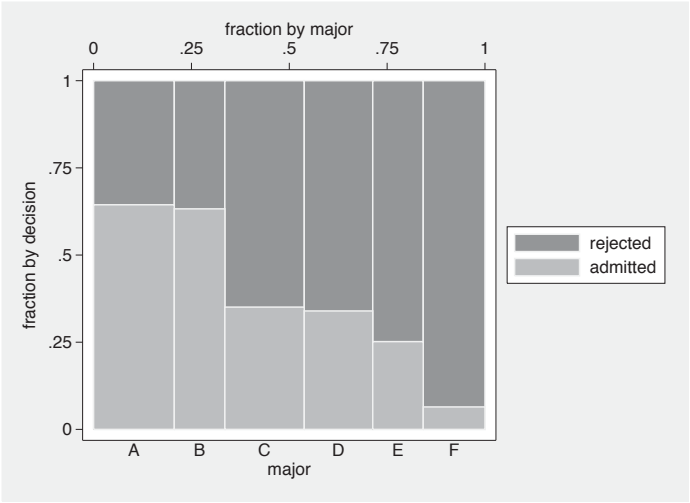

Figure 8. Spineplot of decision versus major for admissions to various Berkeley graduate majors. Acceptance rates vary over a tenfold range.

These are just two-dimensional representations of three-dimensional data. We need to see what structure may exist in three dimensions, including whether there are interactions between the covariates. How can we do that with a two-dimensional display? The answer lies in a composite categorical variable, defined by the cross combinations of two or more categorical variables (Cox 2007). Although not the only method, `egen`'s `group()` function is fine for this purpose:

```
. egen group = group(major sex), label
```

The `label` option is essential for graphs and tables to make sense. Without it, the resulting groups would just show as groups 1 to 12. Further, the order of variables fed to the function is crucial. `group(major sex)` aligns male and female for each major. `group(sex major)` would align majors for each sex. The first is what we need here. In other problems, experimentation with group order may be needed to see what works best.

```
. spineplot decision group [fw=frequency], xlabel(, angle(v) axis(2))
> xtitle("", axis(2)) xtitle(fraction by major and sex, axis(1))
```

Figure 9 shows the result. Vertical axis labels are the lesser of two evils, as there is far too little room for horizontal labels to be legible. Some readers may prefer to try a compromise, say, an angle of 45°. The default title for the bottom x axis would be the variable label for `group`, `group(major sex)`, which we prefer to blank out. Similarly, the title for the top x axis improves on the default.

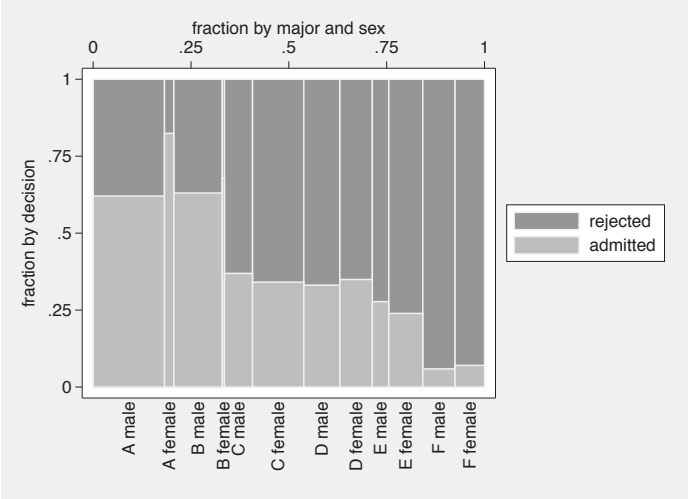

Figure 9. Spineplot of decision versus major and sex for admissions to various Berkeley graduate majors. Females are admitted proportionally more than males to four majors and proportionally less to the remaining two.

The fine structure of the display allows focus on the key question. Major by major, a higher proportion of females than males is admitted to A, B, D, and F, and a lower proportion to C and E. (Admittedly, the comparison for B is not clear on the graph given the small frequencies concerned; for that result, a peek at a table is needed.) Hence, the appearance of discrimination against females appears very much an artifact of the sex and major composition of the applicants or, in other terminology, an example of the amalgamation paradox often named for E. H. Simpson, despite its earlier elucidation by G. U. Yule and several others (Agresti 2002).

A lesson for other examples is that the restriction of spineplots to two variables is more apparent than real given the scope for creating composite variables. Compare what Hofmann (2001) calls "double-decker plots" (for binary responses) and what Wilkinson (2005) calls "region trees".

5 Spineplot details

5.1 Syntax

`spineplot` *yvar xvar* [*if*] [*in*] [*weight*] [`,`
`bar1(`*twoway_bar_options*`) ... bar20(`*twoway_bar_options*`)`
`barall(`*twoway_bar_options*`) missing percent`
`text(`*textvar* [`,` *marker_label_options*]`)` *twoway_options*]

`fweights` and `aweights` may be specified; see [U] **11.1.6 weight**.

5.2 Description

`spineplot` produces a spineplot for two-way categorical data. The fractional breakdown of the categories of the first-named variable *yvar* is shown for each category of the second-named variable *xvar*. Stacked bars are drawn with vertical extent showing fraction in each *yvar* category given each *xvar* category and horizontal extent showing fraction in each *xvar* category. Thus the areas of tiles formed represent the frequencies, or more generally totals, for each cross combination of *yvar* and *xvar*.

5.3 Options

`bar1(`*twoway_bar_options*`) ... bar20(`*twoway_bar_options*`)` allow specification of the appearance of the bars for each category of *yvar* using options of `twoway bar`.

`barall(`*twoway_bar_options*`)` allows specification of the appearance of the bars for all categories of *yvar* using options of `twoway bar`.

`missing` specifies that any missing values of either of the variables specified should also be included within their own categories. The default is to omit them.

`percent` specifies labeling as percentages. The default is labeling as fractions.

`text(`*textvar* [`,` *marker_label_options*]`)` specifies a variable to be shown as text at the center of each tile. *textvar* may be a numeric or string variable. It should contain identical values for all observations in each cross combination of *yvar* and *xvar*. A simple example is the frequency of each cross combination. To show nothing in particular tiles, use a variable with missing values (either numeric missing or empty strings) for those tiles. A numeric variable with fractional part will typically look best converted to string as, for example, `string(`*residual*`,"%4.3f")`. The user is responsible for choice of tile colors so that text is readable. `text()` may also include *marker_label_options* for tuning the display.

twoway_options refers to options of `twoway`; see [G] ***twoway_options***. By default there are two x axes, `axis(1)` on top and `axis(2)` on bottom, and two y axes, `axis(1)` on right and `axis(2)` on left.

5.4 Inside the program

You may wish to know more about how the program works. The code, naturally, is open for inspection in your favorite text editor.

The program works by calculating cumulative frequencies. The plot is then produced by overlaying distinct graphs, each being a call to `twoway bar, bartype(spanning)` for one category of *yvar*. By default, each bar is shown with `blcolor(bg) blw(medium)`, which should be sufficient to outline each bar distinctly but delicately. By default also, the categories of *yvar* will be distinguished according to the graph scheme you are using. With the default `s2color` scheme, the effect is reminiscent of canned fruit salad (which

may be fine for exploratory work). For a publishable graph, you might want to use something more subdued, such as various gray scales or different intensities, as in this column.

Options `bar1()` to `bar20()` are provided to allow overriding the defaults on up to 20 categories, the first, second, etc., shown. The limit of 20 is plucked out of the air as more than any user should really want. The option `barall()` is available to override the defaults for all bars. Any `bar#()` option always overrides `barall()`. Thus if you wanted thicker `blwidth()` on all bars, you could specify `barall(blwidth(thick))`. If you wanted to highlight the first category only, you could specify `bar1(blwidth(thick))` or a particular color.

Other defaults include `legend(col(1) pos(3))`. At least with `s2color`, a legend on the right implies an approximately square plot region, which can look quite good. A legend is supplied partly because there is no guarantee that all *yvar* categories will be represented for extreme categories of *xvar*. However, it will often be possible and tasteful to omit the legend and show categories as axis label text.

6 Conclusion

Spineplots offer an alternative to more conventional bar charts for showing the data in a two-way contingency table. Their particular merit arises from the fact that frequencies are encoded by tile areas so that, in principle, spineplots convey the information in both marginal and conditional distributions. Departure from independence is shown by failure of tiles to align, which is easily seen. Spineplots can also be extended to higher-order contingency tables, in so far as two or more categorical variables may be combined to form a single composite variable to be shown on either axis.

However, what is a key feature of spineplots can also be a limitation. Cells with small frequencies will be represented by small tiles, and cells with zero frequencies will not be represented at all, so the fine structure associated with such cells may be difficult to discern. Hence, other kinds of bar charts remain complementary for showing the structure of contingency tables.

7 Acknowledgments

Matthias Schonlau, Scott Merryman, and Maarten Buis provoked the writing of the `spineplot` command through challenging Statalist postings. A suggestion from Peter Jepsen led to the `text()` option. Private emails from Matthias Schonlau and Antony Unwin highlighted different senses of spineplots and the importance of sort order. Antony suggested standardizing on "spineplot" rather than "spine plot". Maarten verified for me that the spineplot in my copy of Escher (1934) also appears in Escher (1924). Vince Wiggins originally told me about the undocumented `bartype(spanning)` option.

8 References

Agresti, A. 2002. *Categorical Data Analysis*. 2nd ed. Hoboken, NJ: Wiley.

Anderson, M. J. 2001. Francis Amasa Walker. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 216–218. New York: Springer.

Anonymous. 1967. In memoriam Prof. Dr. B. G. Escher. *Geologie en Mijnbouw* 46: 417–422.

Bertin, J. 1983. *Semiology of Graphics: Diagrams, Networks, Maps*. Madison: University of Wisconsin Press.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187: 398–404.

Brinton, W. C. 1914. *Graphic Methods for Presenting Facts*. New York: Engineering Magazine Company.

Cook, D., and D. F. Swayne. 2007. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*. New York: Springer.

Cox, N. J. 2004. Speaking Stata: Graphing categorical and compositional data. *Stata Journal* 4: 190–215.

———. 2007. Stata tip 52: Generating composite categorical variables. *Stata Journal* 7: 582–583.

Escher, B. G. 1924. *De Methodes der Grafische Voorstelling*. Amsterdam: Wereldbibliotheek.

———. 1934. *De Methodes der Grafische Voorstelling*. 2nd ed. Amsterdam: Wereldbibliotheek.

Feinstein, A. R. 2002. *Principles of Medical Statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Feinstein, A. R., and C. K. Kwoh. 1988. A box-graph method for illustrating relative size relationships in a 2×2 table. *International Journal of Epidemiology* 17: 222–224.

Freedman, D., R. Pisani, and R. Purves. 1978. *Statistics*. New York: W. W. Norton.

———. 2007. *Statistics*. 4th ed. New York: W. W. Norton.

Friendly, M. 2000. *Visualizing Categorical Data*. Cary, NC: SAS Institute.

———. 2002. A brief history of the mosaic display. *Journal of Computational and Graphical Statistics* 11: 89–107.

Hartigan, J. A., and B. Kleiner. 1981. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, ed. W. F. Eddy, 268–273. New York: Springer.

———. 1984. A mosaic of television ratings. *American Statistician* 38: 32–35.

Hertz, S. 2001. Georg von Mayr. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 219–222. New York: Springer.

Hofmann, H. 2000. Exploring categorical data: Interactive mosaic plots. *Metrika* 51: 11–26.

———. 2001. Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics* 10: 628–640.

———. 2007. Interview with a centennial chart. *Chance* 20(2): 26–35.

Hummel, J. 1996. Linked bar charts: Analysing categorical data graphically. *Computational Statistics* 11: 23–33.

Karsten, K. G. 1923. *Charts and Graphs: An Introduction to Graphic Methods in the Control and Analysis of Statistics*. New York: Prentice-Hall.

Raisz, E. J. 1934. The rectangular statistical cartogram. *Geographical Review* 24: 292–296.

Robbins, N. M. 2005. *Creating More Effective Graphs*. Hoboken, NJ: Wiley.

Robinson, A. H. 1970. Erwin Josephus Raisz, 1893–1968. *Annals of the Association of American Geographers* 60: 189–193.

Sills, D. L. 1992. In memoriam: Hans Zeisel, 1905–1992. *Public Opinion Quarterly* 56: 536–537.

Tufte, E. R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Unwin, A., M. Theus, and H. Hofmann. 2006. *Graphics of Large Datasets: Visualizing a Million*. New York: Springer.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer.

von Mayr, G. 1877. *Die Gesetzmässigkeit im Gesellschaftsleben*. München: Oldenbourg.

Walker, F. A. 1874. *Statistical Atlas of the United States Based on the Results of the Ninth Census 1870*. New York: Census Office.

Wilkinson, L. 2005. *The Grammar of Graphics*. 2nd ed. New York: Springer.

Young, F. W., P. M. Valero-Mora, and M. Friendly. 2006. *Visual Statistics: Seeing Data with Dynamic Interactive Graphics*. Hoboken, NJ: Wiley.

Zeisel, H. 1947. *Say It with Figures*. New York: Harper.

———. 1985. *Say It with Figures*. 6th ed. New York: Harper & Row.

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 15 commands in official Stata. He wrote several inserts in the *Stata Technical Bulletin* and is an editor of the *Stata Journal*.