The Stata Journal

The *Stata Technical Journal* (ISSN 1536-867X) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

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 teaching, 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

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:

Subscriptions mailed to US and Canadian addresses:

3-year subscription (includes printed and electronic copy)	$153
2-year subscription (includes printed and electronic copy)	$110
1-year subscription (includes printed and electronic copy)	$ 59
1-year student subscription (includes printed and electronic copy)	$ 35

Subscriptions mailed to other countries:

3-year subscription (includes printed and electronic copy)	$225
2-year subscription (includes printed and electronic copy)	$158
1-year subscription (includes printed and electronic copy)	$ 83
1-year student subscription (includes printed and electronic copy)	$ 59
3-year subscription (electronic only)	$153

Back issues of the Stata Journal may be ordered online at

http://www.stata.com/bookstore/sj.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 email sj@stata.com.

The Stata Journal (2004)
4, Number 1, pp. 66–88

Speaking Stata: Graphing distributions

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

Abstract. Graphing univariate distributions is central to both statistical graphics, in general, and Stata's graphics, in particular. Now that Stata 8 is out, a review of official and user-written commands is timely. The emphasis here is on going beyond what is obviously and readily available, with pointers to minor and major trickery and various user-written commands. For plotting histogram-like displays, kernel-density estimates and plots based on distribution functions or quantile functions, a large variety of choices is now available to the researcher.

Keywords: gr0003, graphics, histogram, spikeplot, dotplot, onewayplot, kdensity, distplot, qplot, skewplot, bin width, rug, density function, kernel estimation, transformations, logarithmic scale, root scale, intensity function, distribution function, quantile function, skewness

1 Introduction

The new graphics introduced in Stata 8 has been, by far, the most important step forward in Stata's graphical functionality since early releases in the mid-1980s. It is, therefore, high time that this column turned to discuss graphics directly. I intend to make 2004 a graphic year for *Speaking Stata*, starting with the basic and fundamental issue of graphing univariate distributions. Future columns are intended to discuss graphing categorical and compositional data, comparisons, and model diagnostics. In each case, the aim will be to provide an overview of Stata's provision and to show ways to go beyond what is obviously and readily available. The emphasis will be on graphics commands of potential interest to the largest possible cross-section of Stata users. Thus histograms clearly qualify, but justice cannot be done to details specific to analysis of survival-time distributions.

The core commands for graphing distributions range from `twoway kdensity` and its relative `kdensity` through `twoway histogram` and its relative `histogram` to `graph box` and `graph hbox`. Related but perhaps less-often used commands include `dotplot`, `spikeplot`, and those grouped as diagnostic plots.

2 Histograms, indigenous and exotic

2.1 Number of bins and bin width

With an eye to tradition, including Stata tradition, let us start the discussion with histograms. Up until Stata 7, a histogram was the default graph type if `graph` was fed just one variable. Before Stata 8, such histograms were relatively inflexible and could

not easily be combined with other graph types. Now we have both greater flexibility and easier working with other types. Notable additions include the options to tune both bin width and the start of binning, whereas previously only the number of bins could be controlled directly. The start of binning could be controlled indirectly by tuning `xlabel()` or `xscale()`.

As every good introductory text explains, histogram construction is largely a trade-off problem in which you seek a compromise between detail and generalization or between variance and bias. In doing this, you can tune either the number of bins or the bin width. Theoretical discussions concentrate on the number of bins and its relation to sample size and the kind of distribution being analyzed. However, my guess is that people with their feet in application areas often find it natural to think in terms of a sensible bin width for the variables they have, bearing in mind measurement issues and the magnitude of important or interpretable differences. Whatever your preference, you can now do it either way.

2.2 Varying bin widths

However, one feature that remains wired in histogram commands in Stata 8 is a restriction to bins of equal width. No doubt this is often very sensible whenever the original data are available, but there are occasions on which you might want to break this rule. Let us drill down to some first principles here.

Recall that the idea behind histograms is that the area of each bar represents the fraction of a frequency (probability) distribution within each bin (or class, or interval). Among many books explaining histograms, Freedman, Pisani, and Purves (1998) is an outstanding introductory text that strongly emphasizes the area principle. It is not part of the definition that all bins have the same width, but rather that what is shown on the vertical axis is, or is proportional to, probability density. Frequency density qualifies, as does frequency if all bins have the same width.

In practice, the choice of bin width is often a little arbitrary. If the variable is discrete, a width of 1 is clearly a natural choice. Even then, discrete variables may require some grouping into bins wider than 1. If the variable is number of lifetime sexual partners, the tail (apparently) stretches into very large numbers, and some grouping may be desired. With continuous variables especially, there is always some arbitrariness. Many researchers are most reluctant to compound that by varying the width of the intervals. To do so would complicate the interpretation of the histogram, it might be argued, by any variations in the way the bars were produced. Or, to put it another way, equal widths are relatively simple, and any kind of complexity beyond them needs to be justified.

Despite all that, sometimes the data come grouped into irregular intervals, and the researcher has little or no choice because the raw data may be difficult or impossible to access. Sometimes there is an underlying confidentiality issue. Nevertheless, researchers may still want a histogram, which should be correctly drawn with density, not frequency, on the vertical axis. For example, Altman (1991, 25) gives the ages of 815 road accident casualties for the London Borough of Harrow in 1985:

age	frequency
0–4	28
5–9	46
10–15	58
16	20
17	31
18–19	64
20–24	149
25–59	316
60+	103

In this example and in other similar examples, density can only be calculated for the open-ended class if we specify an upper limit; Altman suggests that 60+ be treated as 60–80.

As usual in statistics, sampling variation is also an issue. If we regard the histogram as a crude estimator of a density function, there is often a case for varying bin width to match the structure of variation, in effect varying how we average probability density locally.

But there is at least one other way to build a histogram in a simple, systematic way: using as limits a set of quantiles equally spaced on a probability scale (e.g., Breiman 1973, 208–209; Scott 1992, 69–70). That way, each bar represents the same area. Unless our data come from something like a uniform distribution, the bin widths will be markedly unequal, but they will reflect the character of the distribution. Breiman points out that the associated error will be approximately a constant multiple of the bar heights, so long as the bin frequencies are not too small.

A related problem is choice of class intervals for a chi-squared test of goodness of fit. Mann and Wald (1942) and Gumbel (1943) urged the merits of choosing classes with equal expected frequencies. That is a simple and definite procedure, which can reduce difficulties arising from low expected frequencies, although data must arrive ungrouped and there may be some loss of sensitivity in the tails of a distribution. Without getting into a wider discussion of the merits of different tests of fit or of tests compared with graphical analysis, it is clear that the equal probability idea is a natural one.

What can be done in Stata? Start with the messier problem in which the data arrive grouped. Much can be done once you know about an undocumented feature of `twoway bar`. We need to enter the lower bin limits and the bin frequencies and one final upper limit as data. For Altman's example, we need to enter data to get

(*Continued on next page*)

```
. list age freq
```

	age	freq
1.	0	28
2.	5	46
3.	10	58
4.	16	20
5.	17	31
6.	18	64
7.	20	149
8.	25	316
9.	60	103
10.	80	.

We then can calculate the densities:

```
. generate density = freq / (815 * (age[_n+1] - age))
```

If you want frequency density rather than probability density, you should omit scaling by the sample size (here 815).

Finally, we can draw the graph, shown in figure 1:

```
. twoway bar density age, bartype(spanning) bstyle(histogram)
```

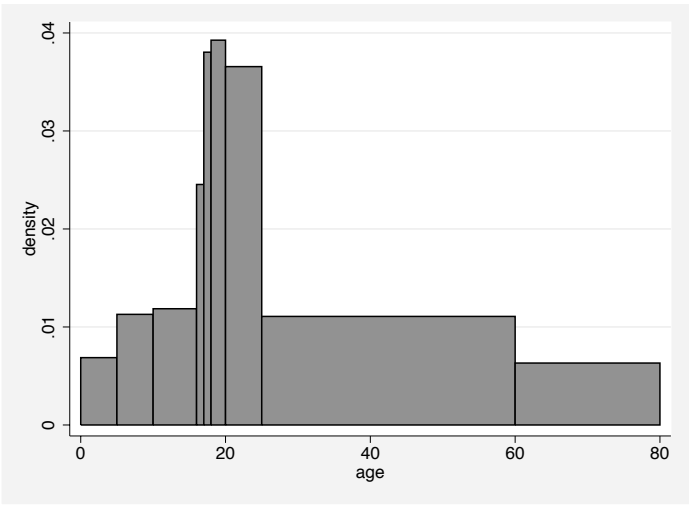

Figure 1: Example of histogram based on data supplied as frequency distribution with varying bin widths.

The "spanning" extends bars to the right until they are curtailed; that is why it is necessary to specify all lower limits and one upper limit for the graph. The data should also be in the correct sort order, as in this example. The option `bstyle(histogram)` is

not compulsory, and you might like to check other possibilities. You might need to add the option `yscale(range(0))` if `twoway bar` does not automatically start bars at 0.

Turning to the more elegant problem, a user-written program for equal-probability histograms can be described and, if desired, downloaded from the Statistical Software Components (SSC) archive by using the `ssc` command; see [R] **ssc**:

```
. ssc describe eqprhistogram
. ssc install eqprhistogram
```

As an illustration, figure 2 is the result of

```
. use http://www.stata-press.com/data/r8/womenwage.dta
. eqprhistogram wage, bin(10) plot(kdensity wage, biweight width(5))
>     legend(ring(0) position(1) column(1))
```

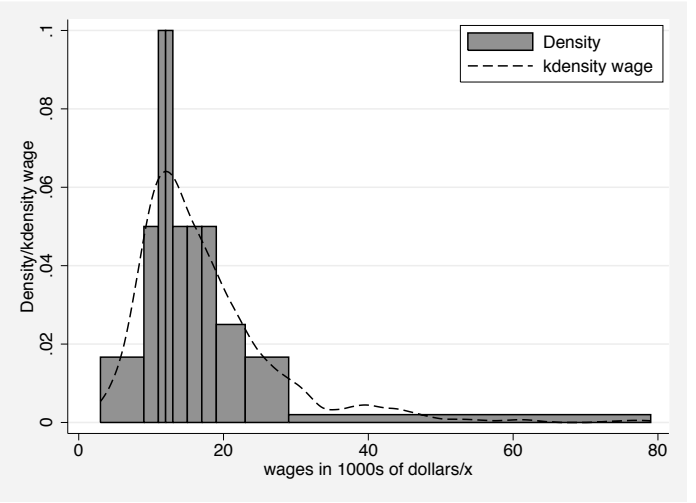

Figure 2: Example of histogram in which bins represent intervals of equal probability. In this case, bin boundaries are deciles, so that the area of each bin represents 1/10 of the distribution. A kernel-density estimate is superimposed.

The bin limits are the deciles, so each bar represents 1/10 of the total probability in the distribution. Note that you can superimpose a density estimate.

Although it may seem a curiosity, the equal-probability histogram has some pedagogic merit. First, it underlines the area principle on which histograms are based. Wider bars are necessarily shorter and narrow bars necessarily taller. Second, it allows a link to be made between histograms and quantile-based methods such as box plots. Arguably, in some datasets it gives a better view of the tails than do the corresponding box plots, especially if within those box plots no values are flagged beyond the quartiles, and so no details are given on structure within the tails. (Box plots are especially poor for U-shaped distributions. In some such cases, no values are identified beyond the

quartiles, and the box plot reduces to a long box and two short tails. Even experienced people can misread this as indicating a unimodal distribution, forgetting that if half the values lie inside the box, then, necessarily, half lie outside it.)

An equal-probability histogram is not suitable for all distributions. Given categorical, discrete, or highly rounded data, quantiles may be tied, especially if the number of bins is large relative to the sample size. If the specified quantiles are tied, `eqprhistogram` refuses to draw the graph. A technical aside: whenever it does this, the exit code is 0. This is in part a diplomatic acknowledgment that inability to draw the graph is either a feature of the data or a limitation of the method, rather than a user error. In addition, it implies that a loop through equal-probability histograms of different variables or groups will not fail merely because a particular graph is impossible.

2.3 Putting a rug underneath

One major merit of histograms is familiarity. All statistically minded people have looked at many histograms, and nonstatistical people who use statistics have also usually come across them. Nevertheless, the basis of histograms, a division of a range into bins, is at best a means to an end, namely easy and effective visualization of a distribution, and at worst a serious distraction. Both psychologically and numerically, densities or frequencies calculated from a set of bins can convey a poor idea of the detailed shape of the distribution of a variable.

One simple way to enhance a histogram by forging a closer link with the raw data is to add a so-called rug, which as the name implies, is almost always placed underneath the histogram. A rug is a very short, long display of point symbols, one for each distinct value. Often a vertical pipe symbol | is used to minimize overlap. Rugs may also be added to other kinds of plots. There are many varied examples in Davison (2003).

Before version 8, Stata had `graph` options to combine rugs, which in Stata were called oneway plots, with box plots and with scatterplots. These options are still accessible under `graph7`. However, they did not make the cut into the new graphics in that or similar form.

Although rugs are not explicitly provided in Stata 8, the procedure for weaving your own rug is straightforward. Starting with a basic histogram for the same wage data,

```
. histogram wage, start(0) width(5)
```

we see that with these choices density varies up to about 0.07 per 1,000 dollars. That leads to a decision to put the rug at about −0.003 on that scale. We need a variable to hold this value:

```
. gen where = -0.003
```

In practice, we can just choose a trial value and then use `replace` to improve upon it. Next, there is no pipe symbol in the `symbolstyle` portfolio, so we must enlist the pipe character as a marker label. Then, the rug is just a scatterplot of `where` against `wage`, suppressing the default marker symbol and placing the marker label exactly on target:

```
. gen pipe = "|"
. histogram wage, start(0) width(5)
>     plot(scatter where wage, ms(none) mlabel(pipe) mlabpos(0)) legend(off)
```

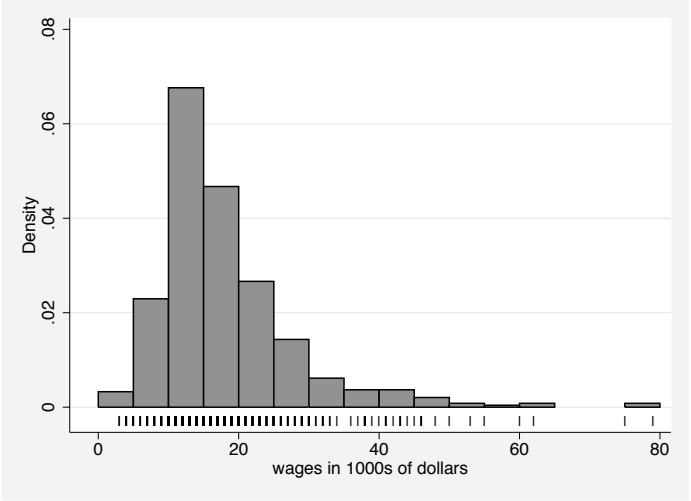

Figure 3: Example of histogram with rug showing distinct values occurring in the data.

In this case, as shown in figure 3, the rug shows rounding of the data, and a tabulation makes it explicit that all values are just multiples of $1,000. In general, a rug is a useful but restrained way of showing some of the fine structure of a distribution.

A rug will take up a lot of bytes in a graph file if any point symbol stands for many repeated values. Clearly, it is unnecessary to overwrite each symbol repeatedly. A solution is to select each distinct value just once. There are two systematic ways to do this, to select the first in each group after sorting or to select the last, and it is immaterial here which you use, so you might as well go

```
. bysort wage: gen tag = _n == 1
```

A canned near-equivalent is

```
. egen tag = tag(wage)
```

The difference is that the `egen` call sorts your data while doing the calculation but then returns it to its original sort order, which may differ. The first method may change your sort order. Having done this, we select points for the rug as `if tag`.

2.4 Horizontal histogram bars

The `histogram` display, by default, has the frequency axis vertical, as is conventional. The manual entry [G] **graph combine** shows how histograms may be placed vertically and horizontally along the margins of a scatterplot. More generally, the `horizontal` option may be used to reverse axes. This may sound merely cosmetic, but there are occasions in which this layout appears more natural. In the environmental sciences, among other fields, height above and depth below some surface are key natural variables. The extra option `yscale(reverse)` would show depths the intuitive way up.

Here is a histogram of the mean elevations of 27,523 glaciers from Central Asia and southern Siberia (figure 4). Data were extracted from the World Glacier Inventory. The tendency to multiple modes is best interpreted as a consequence of lumping together several distinct mountain ranges. The Stata command was

```
. histogram mean_elev, horizontal start(1600) width(100) frequency
>     ylabel(, angle(horizontal))
```

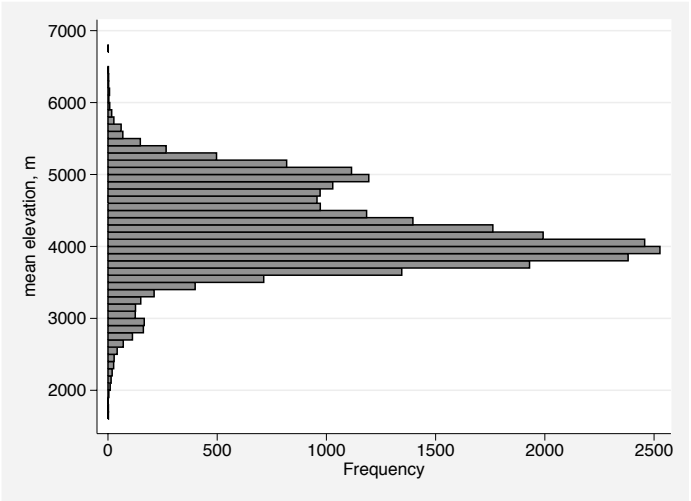

Figure 4: Example of histogram with horizontal bars. In this case, the response variable is altitude.

2.5 Do-it-yourself density calculation

From a simple enhancement of histograms, we turn to the basic underlying calculations. It may be useful to document how to calculate densities yourself, having first chosen a *start* and a *width*. If we are counting so that each bin is defined precisely as *lower_limit* $\leq$ *value* $<$ *upper_limit*, we could use `floor()` to generate lower limits as bin identifiers. With the reverse convention, we could use `ceil()`. See Cox (2003c) for a note on these functions. Then, the frequencies are the counts within each bin:

```
. gen double lower = width * floor((varname - start) / width)
. bysort lower: gen frequency = _N
```

To get fractions and percents, we must be careful to count each value just once:

```
. by lower: gen double sum = frequency * (_n == 1)
. replace sum = sum(sum)
. gen fraction = frequency / sum[_N]
. gen percent = 100 * fraction
```

The density is then

```
. gen density = frequency / (width * _N)
```

Real calculations get messier once you build in selections `if` or `in`, subdivision into groups defined by other variables, or missing values. The messy details are coded up in an `egen` function `density()` in the `egenmore` package on SSC.

3 Relatives of histogram: spikeplot, dotplot, and onewayplot

One common reaction to histograms is to prefer more information in displaying distributions, especially in the tails. The optimistic view is that more details will turn out to be instructive fine structure. The corresponding pessimistic view is, naturally, that such details will be best regarded as noise and, as such, an irreducible nuisance. Most discussions stress the latter view over the former, but there can be real merit in playing deterministic detective rather than stochastic skeptic.

The official commands `spikeplot` and `dotplot` and the user-written command `onewayplot` offer different ways of showing more detail than do equivalent histograms.

`spikeplot`, by default, offers a spike for every distinct value—that is, no binning—and the opportunity to control binning by a `round()` option, which in effect controls bin width. Historically, `spikeplot` offered, before Stata 8, the most obvious official alternative to `graph, histogram` for getting a histogram-like display with more than 50 bins. That role is now lost. However, its discrete representation of a frequency distribution remains available for occasions when you want to emphasize the granularity of data, either as defined in principle (counted variables, in particular) or as measured in practice. The display of the age distribution of Ghana given at [R] **spikeplot** is a good example of what `spikeplot` does best, revealing a fine structure of age preferences, including multiples of 5, even rather than odd ages, and so forth. There is some scope for controlling spike appearance if the default appearance (which is the default of `twoway spike` under the prevailing scheme) appears too exiguous.

`dotplot`, in contrast, is based on the idea (or the ideal) of showing a point symbol for each value; exactly the same description covers rugs and `onewayplot`. Similar plots under a variety of names go back at least as far as van Langren (1644); see Tufte (1997, 15). Wilkinson (1999) gives several further references of historical interest. Chambers et al. (1983) used the term one-dimensional scatterplots. The term oneway plots appears to

have been introduced by StataCorp in its earlier guise as Computing Resource Center (1985). Wild and Seber (2000) show many interesting examples of oneway plots.

`dotplot`, by default, offers, as far as possible, a point symbol for every value and some binning. Binning can be controlled rather indirectly, although in practice, the default is usually adequate, and when desired, the binning can be switched off with the `nogroup` option. The main virtues of dotplots lie in their ability to show some features that might otherwise be obscured by a series of touching bars, especially granularity and details of outliers or other extreme values in the tails. You can also show, for example, median and quartiles by horizontal marks and thus hybridize box plots and histograms.

The considerable flexibility of `histogram`, `spikeplot`, and `dotplot` might seem to leave few important gaps in their territory. Nevertheless, `onewayplot` was written to provide some extra possibilities in this area; it also may be downloaded using `ssc`. As mentioned earlier, `graph, oneway` did not survive as such into Stata 8, although the minor trickery needed to add rugs is just one illustration of how they can be emulated fairly easily. `onewayplot` is essentially a convenience command that bundles together various easy but tedious handles for making your own oneway plots. You can choose between horizontal and vertical layouts, while `stack` and `center` options produce a variant on `dotplot`.[1] There is, by default, no binning of data; binning may be accomplished with the `width()` option.

In figure 5, we show the results of a `onewayplot` using the handle of a regional classification to split the glacier elevations. Both `histogram` and `dotplot` struggle given 18 regions, some with fairly long names.

```
. onewayplot mean_elev, by(region) ytitle("") stack ms(oh) msize(tiny) width(20)
```

(*Continued on next page*)

[1] You can also type `centre`. An undocumented feature of `dotplot` is that `centre` is allowed as well as `center`. This is a convenience for speakers of languages, such as English, which use that spelling, and is emulated by `onewayplot`.

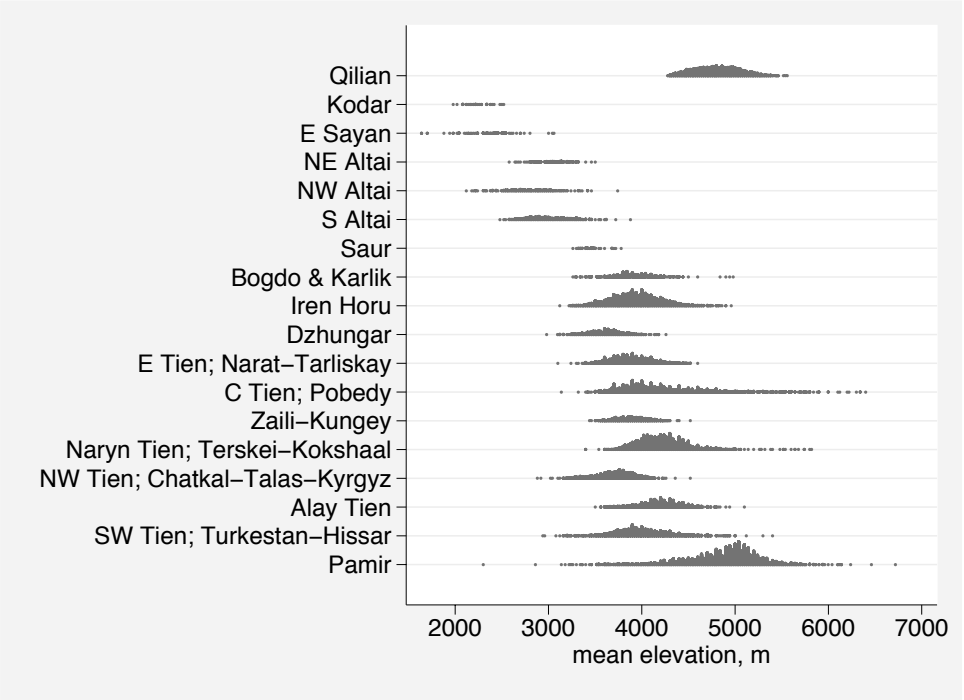

Figure 5: Example of multiple histograms produced by `onewayplot`.

4 Kernel-density estimation

4.1 Available commands

The histogram of a continuous variable is, from one point of view, an estimator of the density function of that variable. Clearly the set of bins used to compute that estimate imposes discontinuities on the estimate, which leads us directly to consider smoother estimates, especially those based on convolution of the data and a symmetric kernel. `twoway kdensity` and `kdensity` are provided in official Stata as basic commands. Recently, users have added variable kernel density estimation commands (Salgado-Ugarte and Pérez-Hernández 2003; Van Kerm 2003).

4.2 Variations on the official theme

Transform before and after estimation

Some simple devices extend the range of applications of Stata's official commands for kernel-density estimation. First is the idea of estimating the density function on a transformed scale and then back-transforming the estimate to one for the raw scale. Two of the most natural transformations here, as elsewhere, are logarithms for positive variables

and logit-like transformations for proportions and other data measured on some interval (a, b). The underlying general principle is that, for a continuous monotone transformation $t(x)$, the densities $f(x)$ and $f\{t(x)\}$ are related by $f(x) = f\{t(x)\}|dt/dx|$. This procedure is mentioned briefly by Silverman (1986, 27–30), although his worked example (page 28) is not very encouraging. Good expositions are given by Wand and Jones (1995, 43–45), Simonoff (1996, 61–64), and Bowman and Azzalini (1997, 14–16).

With a logarithmic transformation of x, we have

$$\text{estimate of } f(x) = \text{estimate of } f(\log x) \times (1/x)$$

given that $d/dx(\log x) = 1/x$. Note in particular, if data are right skewed, that the result of this transformation is more smoothing in the tail and less near the main part of the distribution than in the default method. I have found this to be one of the most valuable ways of going beyond the default. It fits very well both the common finding that positive variables are right-skewed, suggesting a transformation, such as the logarithm, and the common attitude that results on the original scale are of direct scientific or practical interest. To put it another way, the transformation behaves more like a link function than a classical transformation, given that end results are on the scale of the original response. You can get the best of both worlds.

Returning to the wage data, here is an illustrative (and certainly not definitive) example, in which we just use default kernel and width choice.

```
. gen logwage = log(wage)
. kdensity logwage, at(logwage) generate(densitylog)
. gen density = densitylog/wage
. levels wage, local(levels)
. line density wage, sort xtick(‘levels’, tposition(inside))
```

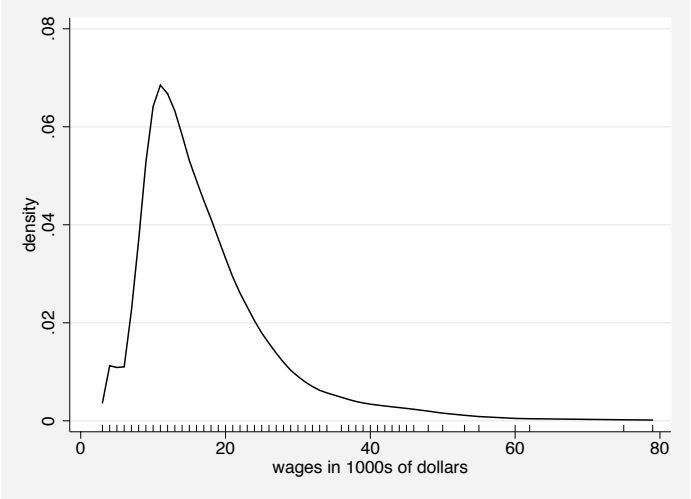

Figure 6: Example of a density function estimated on logarithmic scale and transformed to the original scale of data.

The density function, shown in figure 6, is much smoother in the tails than the equivalent default, which is not shown here. However, the step in the left-hand tail needs investigation: is this some odd artifact or a genuine feature of the data? Incidentally, another technique is used to show a rug by picking up a list of distinct values from `levels` (added to Stata on 16 April 2003). However, this technique is not as general as that previously illustrated, as it hinges on the variable concerned having only integer values. `levels` is not designed to work with noninteger numeric values.

Similarly with a logit-like transformation,

$$\text{estimate of } f(x) = \{\text{estimate of } f(\text{logit } x)\} \frac{(b-a)}{(x-a)(b-x)}$$

where logit $x = \log\{(x-a)/(b-x)\}$, a slight generalization of the usual definition, for which $a = 0, b = 1$. Note that $d/dx(\text{logit } x) = (b-a)/\{(x-a)(b-x)\}$.

Density on a log scale

It can be natural to calculate $f(\log x)$ as a way of getting a better estimate of $f(x)$. It can also be natural to calculate $\log\{f(x)\}$ as way of getting a better visualization of $f(x)$. This seems to be an old idea, periodically rediscovered. One venerable reference is the work of the soldier, explorer, and scientist R. A. Bagnold (1937, 1941), who worked on size distributions—especially those of sand—while at present the idea is widely used in fields ranging from statistical physics (Bardou et al. 2002) to statistical finance (Hazelton 2003). There is clearly no barrier also to looking at $\log\{f(\log x)\}$ or using some other transformation before density estimation if it seems appropriate.

The highly original contribution of Bagnold deserves some explanation, as it appears to be little known within the statistical sciences. Born in England in 1896[2], he joined the British army from school and served in the First World War. He then took an engineering degree at Cambridge. Remaining in the army, he used leaves to travel and explore, particularly on pioneering long trips into the deserts of Egypt, Sudan, and Libya using specially adapted cars. This provoked an interest in the physics of blown sand, leading ultimately to a now-classic monograph (Bagnold 1941). Wind transport of loose particles is highly size selective, as ordinary experience confirms: very coarse material will not move, while very fine material may easily be lofted high into the atmosphere and carried over vast distances. Thus, the particle size distribution of a deposit (say from a sand dune) is of central interest. Bagnold found plots of log density versus log grain diameter the most helpful way to show his data. It seems clear from his very readable autobiography (1990), published just after his death, that the crucial first step of plotting densities on a log scale owed most to an engineer's feeling of a sensible thing to do. By thinking for himself, he was not inhibited by ideas on what was or was not standard statistical practice. Much later, Bagnold returned to the question and contributed to the development of log-hyperbolic distributions by Ole Barndorff-Nielsen. There is a full bibliography of his publications in Thorne et al. (1988).

[2]His sister was Enid Bagnold, later a novelist, dramatist, and poet, and best remembered for the children's classic *National Velvet*.

Several properties are simple on a log-density scale. One exploited by Bagnold is that a normal (Gaussian) density plots as a parabola

$$\log f(x) = -\log\left(\sigma\sqrt{2\pi}\right) - \frac{(x-\mu)^2}{2\sigma^2}$$

while exact or approximate exponential or power-law decay of density will show exact or approximate linear patterns, the latter requiring also a logarithmic scale for the variable.

Let us illustrate with log of wage from the wage data considered above:

```
. gen logdensitylog = log(densitylog)
. qui summarize logwage
. local mean = r(mean)
. local sd = r(sd)
. scatter logdensitylog logwage
>         || function normal =
>         -log(‘sd’ * sqrt(2 * _pi)) - ((x - (‘mean’))^2 / (2 * ‘sd’^2)),
>         ra(logwage) ytitle(log density) xtitle(log wage)
>         legend(off) subtitle(log density plot)
```

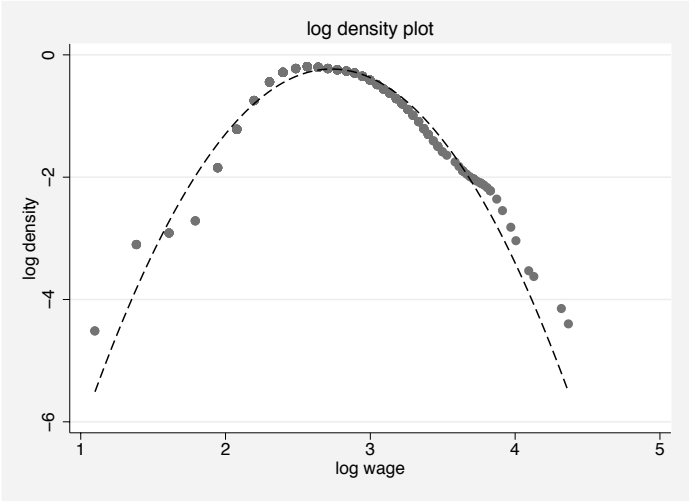

Figure 7: Example of plot using log density. The parabola shows a normal density function with the same mean and standard deviation as log wage.

The results in figure 7 suggest a good but not spectacularly good fit to a lognormal. The slightly fat tails seem suggestive. At the same time, the density estimates, especially in the tails, are, as always, subject to sampling variation and sensitive to kernel bandwidth; note also that neighboring density estimates are necessarily highly dependent.

What is implemented above is just a first stab. Hazelton (2003) suggests various refinements, including robust estimation of the mean and standard deviation and, given a density estimation bandwidth h, fitting a normal with variance $\text{sd}^2 + h^2$ to correct for side-effects of using a kernel.

Density on a root scale

There would also be some advantages to a square-root scale, given that densities behave a bit like counts, for which a root transformation is often the first to be tried. Also, the square root of a Gaussian shape is another Gaussian shape. So we can have our cake and eat it too: hunt for a Gaussian yet benefit from stabilized sampling fluctuation. Check the assertion with

```
. twoway function sqrt(normden(x)), range(-4 4)
```

There is a `root` option in `spikeplot` for a similar reason. Tukey (1977, chapter 17) worked through a bundle of related ideas, which seem to have been little explored since.

Intensities, too

Those interested in data on events, considered as the result of a point process in one dimension (most obviously, time or space), should note that Stata's kernel-density commands can readily be used to estimate the intensity function (say, frequency per unit of time or space). Suppose that a variable contains dates of earthquakes, eruptions, strikes, honors for a sports team, or whatever else is of interest. To get results on an intensity scale, just multiply 'density' by the number of observed data points. A key detail is that intensities will be smoothed beyond the beginning and the end of the interval in question; whether this is tolerable or further surgery is desired is a question for the user.

5 Quantile plots and distribution plots

Another key approach eschews any kind of binning or smoothing and starts with the idea of directly showing the pattern of the quantiles (the ordered values). Formally, we order n data values for a variable x and label them such that $x_{(1)} \leq \cdots \leq x_{(n)}$. `quantile` has long been available as an official Stata command for quantile plots, in which the $x_{(i)}$ are plotted against $(i - 0.5)/n$. See [R] **diagnostic plots**. The term *quantile plot* appears in Chambers et al. (1983) and Cleveland (1993, 1994). Modern use of quantile plots and their relatives stems largely from the path-breaking paper of Wilk and Gnanadesikan (1968). Examples of antecedents from the nineteenth century can be found in Quetelet (1827) and Galton (1875); see Stigler (1986, 167, 270).

Essentially the same information can be shown in a plot of cumulative distribution functions or of survival (a.k.a., survivor, reliability, complementary, or reverse distribution) functions, in which we plot either probabilities or frequencies of values being $\leq x$

or $> x$. In many biomedical or engineering applications, the survival function appears closer to the practical problem, the name being suitably evocative if data are indeed times to patient death, component failure, or something similar.

According to Hald (1990, 108), the first graph of (the complement of) a distribution function appears in a 1669 letter from Christiaan Huygens (1629–1695) to his brother Lodewijk (1631–1699). He plotted a survival function from data from the life table of John Graunt (1620–1674). Huygens made numerous contributions to mathematics, astronomy, and physics, studying, among many other matters, games of chance, the collision of elastic bodies, the rings of Saturn, the pendulum clock, and the wave theory of light.

In Stata 8, the graphics of `quantile` were revised to match the new graphics, but the functionality was unchanged. A broader command is `qplot` (Cox 2004), which in most respects is a generalization of `quantile`; just one detail is omitted, the reference line. It supersedes the previous program `quantil2` (Cox 1999b, 2001).

Stata already has an official graph command for survival functions, `sts graph`. If your data really are survival times and you have any of the complications that are the stuff of survival analysis, such as censoring or subjects entering at different times, you should use `sts graph`. However, it is not and does not purport to be a general purpose command for all kinds of distribution.

In addition, the official command `cumul` ([R] **cumul**) is available to calculate the cumulative distribution function for a single variable, after which the function may be plotted using `twoway`. The user with several variables to be compared or with an interest in survival functions thus needs to repeat the `cumul` command or take the further step of calculating survival functions from cumulative distribution functions. A command `distplot` that bundles calculation and graphing steps together is, however, available (Cox 1999a, 2003a,b).

`qplot` and `distplot` are, in effect, siblings. The choice between them is most obviously one of choice of axes and thus, in a sense, trivial, but different conventions may seem natural for different problems and even different fields or traditions. In particular, there seems to be a growth of interest in quantile functions as responses, which makes `qplot` a possible choice (see, for example, Gilchrist 2000).

`qplot` and `distplot` have in common

1. Support for graphing several variables.
2. Support for graphing several groups, through a `by()` option.
3. Choice of `twoway` plottypes, from `area`, `bar`, `connected`, `dot`, `dropline`, `line`, `scatter`, or `spike`. These are not in general equally useful or attractive, but there is at least much choice, courtesy of `twoway`'s generous design.
4. Support for reversing the sort order so that values decrease from top left.
5. Support for alternative transformed scales.

In addition, `qplot` has support for choice of a in a general rule for plotting position $(i - a)/(n - 2a + 1)$ for $i = 1, \ldots, n$. The default is $a = 0.5$, giving $(i - 0.5)/n$. Other choices include $a = 0$, giving $i/(n + 1)$, and $a = 1/3$, giving $(i - 1/3)/(n + 1/3)$. The choice is often immaterial, but some authorities have strong opinions on the best choice on various grounds, some even statistical. For more discussion and references, see Cox (1999b).

6 Skewness plots

The skewness of a variable is often of interest, perhaps especially as an indicator of potential problems in subsequent analysis. Commonly a single measure is used, whether the moment-based measure produced by `summarize, detail` or other measures (which, in most cases, are readily calculated from the output of `summarize`). Graphically, skewness may be assessed with varying degrees of ease and efficiency from the plots mentioned so far, but there is also a case for a customized design.

Various possibilities are based on the quantiles (Gnanadesikan 1977, 1997). The quantiles may be paired as lower and upper quantiles $x_{(1)}$ and $x_{(n)}$, $x_{(2)}$ and $x_{(n-1)}$, etc., and a median may be calculated in the usual way.

Stata supports `symplot`, a plot of (upper quantile − median) versus (median − lower quantile), for which the reference situation of symmetry or lack of skew plots as a line of equality. See [R] **diagnostic plots**. However, `symplot` will show only a single group of data and thus cannot be used for comparisons, while a plot with a sloping reference line is more difficult to deal with than the plot now to be described, which has horizontal reference lines.

`skewplot` produces, by default, a plot of the *midsummary* versus the *spread* for the variables supplied, also known as the mid-versus-spread plot. With the `skew` option, it produces a plot of the *skewness function* versus the *spread function*. Such plots convey both the general character and the fine structure of the symmetry or skewness of datasets and can be used to compare distributions or to assess whether transformations are necessary or effective.

There are some little-used terms here, so we need a few definitions. In a perfectly symmetric set of data, the midsummaries $(x_{(1)} + x_{(n)})/2$, $(x_{(2)} + x_{(n-1)})/2$, etc., would all be identical and equal to the median. A plot of each midsummary (or mean of lower and upper quantiles) $(x_{(i)} + x_{(n-i+1)})/2$ versus each difference or spread of lower and upper quantiles $x_{(n-i+1)} - x_{(i)}$ would, thus, yield a horizontal straight line. Conversely, skewness in sets of data will be reflected by departures from horizontality. In particular, right skewness would be shown by rising lines and left skewness by falling lines.

Apart from the divisor of 2, this plot was suggested by J. W. Tukey (Wilk and Gnanadesikan 1968). See also Gnanadesikan (1977, 1997, chapter 6.2) or Fisher (1983). The form used here and the name *mid-versus-spread plot* are found in Hoaglin (1985). It is usual to plot only that half of the sample results for which spread is ≥ 0.

The `skew` option produces an alternative form promoted by Benjamini and Krieger (1996, 1999). Consider the identity, which introduces their terminology,

$$\begin{aligned} x_{(n-i+1)} &= \text{median} + (x_{(n-i+1)} - x_{(i)})/2 + (x_{(i)} + x_{(n-i+1)} - 2 \times \text{median})/2 \\ &= \text{median} + \textit{spread function} + \textit{skewness function} \end{aligned}$$

for $x_{(i)}$ in the lower half of a sample. This leads to a plot of the skewness function versus the spread function, known as the skewness versus spread plot. Note that the skewness function is (midsummary − median) and so will be constant and zero for a perfectly symmetric distribution and that the spread function is half the spread of the mid versus spread plot. In short, the `skew` option does not change the configuration of the plot but merely the labeling of the axes.

In addition, the ratio of the skewness and spread functions or

$$\frac{x_{(i)} + x_{(n-i+1)} - 2 \times \text{median}}{x_{(n-i+1)} - x_{(i)}}$$

is a measure of skewness (in the traditional sense) originally suggested for quartiles by Bowley (1902) and generalized to this form by David and Johnson (1956). Another incarnation is as the p-skewness index (Gilchrist 2000, 54, 72).[3] It varies between -1 and 1. A similar general measure was used by Parzen (1979). Graphically this measure is the slope of the line connecting $(0, 0)$ and each data point if the `skew` option is used.

See Benjamini and Krieger (1996, 1999) and Groeneveld (1998) for concise reviews tracing such ideas from late 19th-century antecedents to recent work and further details on the interpretation of the skewness-versus-spread plot.

Let us close with an example for data on 158 glacial cirques from the English Lake District (Evans and Cox 1995). Glacial cirques are hollows excavated by glaciers that are open downstream, bounded upstream by the crest of a steep slope (wall), and arcuate in plan around a more gently sloping floor. More informally, they are sometimes described as "armchair-shaped". Glacial cirques are common in mountain areas that have or have had glaciers present. Three among many possible measurements of their size are length, width and wall height, and the distribution of all in the area studied is shown by

```
. skewplot length width wall_height, legend(ring(0) position(5) column(1))
```

to be markedly right skew (figure 8). Logarithmic transformation seems an obvious possibility, after which

[3] Gilchrist calls the special case for quartiles Galton's skewness (pages 8, 25, 53, and 72), but there is no evidence that Galton used it.

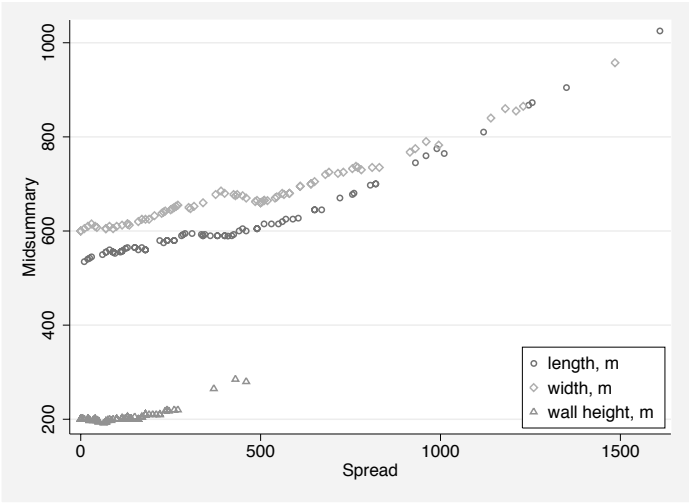

Figure 8: Skewness plot for three variables. The systematic upward drift indicates marked right skewness.

```
. skewplot log_length log_width log_wall_height, legend(ring(0) position(3)
>     column(1))
```

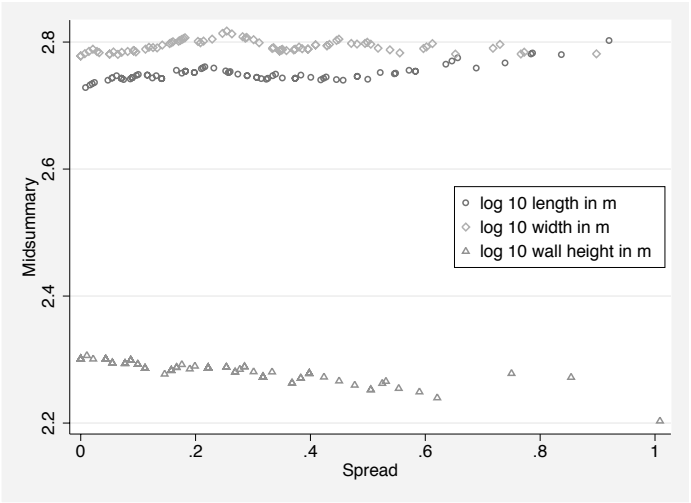

Figure 9: Skewness plot for three log-transformed variables. Approximately horizontal patterns indicate that transformations have yielded near symmetry of distributions.

shows approximate lack of skew (although a hint of mild overtransformation in wall height, which is best left alone for simplicity) (figure 9). An important feature of such plots is that the effects of outliers are localized.

7 Conclusions

With one command or another, users can now plot univariate distributions in many different ways. You can choose between several depictions of the density function or several depictions of the distribution function or its inverse, the quantile function. You can choose discrete or continuous representations and vertical or horizontal alignments. Less obviously, it is straightforward to add details (for example, rugs of distinct data values) or exploit the inbuilt flexibility of `graph` (for example, by looking at density estimates on a log scale or by constructing your own histogram with varying bin width).

The theme of distributions will continue into the next column but with a focus on categorical data. Distributions of categorical variables may be shown in a variety of displays: the survey will range from old staples to less well-known plots, with emphasis on the important special cases of graded data and of three variables with constant sum.

8 Acknowledgments

Ian Evans provided the Lake District cirques data, pointed me to the World Glacier Inventory data, and participated in many discussions on statistical graphics over more than 30 years. Marcello Pagano persistently urged the merits of equal-probability histograms and parenthetically underlined the connection with chi-squared tests. Vince Wiggins specifically alerted me to spanning bars and generally advised on strategies and tactics for using the new graphics. Elizabeth Allred, Ronán Conroy, Philip Ender, and Roger Harbord made helpful comments during development of various predecessors or versions of some programs discussed here. Richard Groeneveld kindly tracked down the Bowley reference.

9 References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. London: Chapman & Hall.

Bagnold, R. A. 1937. The size-grading of sand by wind. *Proceedings of the Royal Society Series A* 163: 250–264.

—. 1941. *The Physics of Blown Sand and Desert Dunes*. London: Methuen.

—. 1990. *Sand, Wind, and War: Memoirs of a Desert Explorer*. Tucson: University of Arizona Press.

Bardou, F., J.-P. Bouchaud, A. Aspect, and C. Cohen-Tannoudji. 2002. *Lévy Statistics and Laser Cooling: How Rare Events Bring Atoms to Rest*. Cambridge: Cambridge University Press.

Benjamini, Y. and A. M. Krieger. 1996. Concepts and measures for skewness with data-analytic implications. *Canadian Journal of Statistics* 24: 131–140.

—. 1999. Skewness—concepts and measures. In *Encyclopedia of Statistical Sciences Update*, ed. S. Kotz, C. B. Read, and D. L. Banks, vol. 3, 663–670. New York: John Wiley & Sons.

Bowley, A. L. 1902. *Elements of Statistics*. 2d ed. London: P. S. King.

Bowman, A. W. and A. Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Applications*. Oxford: Oxford University Press.

Breiman, L. 1973. *Statistics: With a View towards Applications*. Boston: Houghton Mifflin.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Cleveland, W. S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.

—. 1994. *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Computing Resource Center. 1985. *STATA/Graphics User's Guide*. Los Angeles, CA: Computing Resource Center.

Cox, N. J. 1999a. gr41: Distribution function plots. *Stata Technical Bulletin* 51: 12–16. In *Stata Technical Bulletin Reprints*, vol. 9, 108–112. College Station, TX: Stata Press.

—. 1999b. gr42: Quantile plots, generalized. *Stata Technical Bulletin* 51: 16–18. In *Stata Technical Bulletin Reprints*, vol. 9, 113–116. College Station, TX: Stata Press.

—. 2001. gr42.1: Quantile plots, generalized: update to Stata 7.0. *Stata Technical Bulletin* 61: 10–11. In *Stata Technical Bulletin Reprints*, vol. 10, 55–56. College Station, TX: Stata Press.

—. 2003a. Software update: gr41_1: Distribution function plots. *Stata Journal* 3(2): 211.

—. 2003b. Software update: gr41_2: Distribution function plots. *Stata Journal* 3(4): 449.

—. 2003c. Stata tip 2: Building with floors and ceilings. *Stata Journal* 3(4): 446–447.

—. 2004. Software update: gr42_2: Quantile plots, generalized. *Stata Journal* 4(1): 97.

David, F. N. and N. L. Johnson. 1956. Some tests of significance with ordered variables. *Journal of the Royal Statistical Society, Series B* 18: 1–20.

Davison, A. C. 2003. *Statistical Models*. Cambridge: Cambridge University Press.

Evans, I. S. and N. J. Cox. 1995. The form of glacial cirques in the English Lake District, Cumbria. *Zeitschrift für Geomorphologie* 39: 175–202.

Fisher, N. I. 1983. Graphical methods in nonparametric statistics: a review and annotated bibliography. *International Statistical Review* 51: 25–38.

Freedman, D., R. Pisani, and R. Purves. 1998. *Statistics*. New York: W. W. Norton.

Galton, F. 1875. Statistics by intercomparison, with remarks on the law of frequency of error. *Philosophical Magazine*, 4th series, 49: 33–46.

Gilchrist, W. G. 2000. *Statistical Modelling with Quantile Functions*. Boca Raton, FL: Chapman & Hall/CRC.

Gnanadesikan, R. 1977. *Methods for Statistical Data Analysis of Multivariate Observations*. New York: John Wiley & Sons.

—. 1997. *Methods for Statistical Data Analysis of Multivariate Observations*. 2d ed. New York: John Wiley & Sons.

Groeneveld, R. 1998. Skewness, Bowley's measure of. In *Encyclopedia of Statistical Scienes Update*, ed. S. Kotz, C. B. Read, and D. L. Banks, vol. 2, 619–621. New York: John Wiley & Sons.

Gumbel, E. J. 1943. On the reliability of the classical chi-square test. *Annals of Mathematical Statistics* 14: 253–263.

Hald, A. 1990. *A History of Probability and Statistics and their Applications before 1750*. New York: John Wiley & Sons.

Hazelton, M. L. 2003. A graphical tool for assessing normality. *American Statistician* 57: 285–288.

Hoaglin, D. C. 1985. Using quantiles to study shape. In *Exploring Data Tables, Trends, and Shapes*, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, 417–460. New York: John Wiley & Sons.

van Langren, M. F. 1644. *La Verdadera Longitud po Mar y Tierra*. Antwerp.

Mann, H. B. and A. Wald. 1942. On the choice of the number of class intervals in the application of the chi-square test. *Annals of Mathematical Statistics* 13: 306–317.

Parzen, E. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association* 74: 105–131.

Quetelet, A. 1827. Recherches sur la population, les naissances, les décès, les prisons, les dépôts de mendicité, etc., dans le Royaume des Pays-Bas. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-lettres de Bruxelles* 4: 117–192.

Salgado-Ugarte, I. H. and M. A. Pérez-Hernández. 2003. Exploring the use of variable bandwidth kernel density estimators. *Stata Journal* 3(2): 133–147.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer.

Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Thorne, C. R., R. C. MacArthur, and J. B. Bradley, ed. 1988. *The Physics of Sediment Transport by Wind and Water: A Collection of Hallmark Papers by R. A. Bagnold*. New York: American Society of Civil Engineers.

Tufte, E. R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

Van Kerm, P. 2003. Adaptive kernel density estimation. *Stata Journal* 3(2): 148–156.

Wand, M. P. and M. C. Jones. 1995. *Kernel Smoothing*. London: Chapman & Hall.

Wild, C. J. and G. Seber. 2000. *Chance Encounters: A First Course in Data Analysis and Inference*. New York: John Wiley & Sons.

Wilk, M. B. and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55: 1–17.

Wilkinson, L. 1999. Dot plots. *American Statistician* 53: 276–281.

About the Author

Nicholas Cox is a statistically minded geographer at the University of Durham. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also co-authored fourteen commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is Executive Editor of the *Stata Journal*.