



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal (2016)
16, Number 4, pp. 1058–1071

Speaking Stata: Letter values as selected quantiles

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

Abstract. Letter values were introduced and named by J. W. Tukey in the 1970s as selected quantiles. The idea is to choose a small set of quantiles to characterize an ordered sample of values, quantiles that are each defined as either individual order statistics or the mean of two such statistics. The procedure is to start with the median, to continue with quartiles, and then with first and last octiles, and so on. At each step, approximate medians are identified of successively smaller tail fractions until the extremes, the minimum and maximum, are reached. As a historical aside, the same idea can be identified in work by Francis Galton from 1880.

Letter values are supported by Stata through the official command `lv`, but that command is geared to letter-value displays and (arbitrarily) will compute no more than 21 letter values. A new command, `lvalues`, is introduced that supports calculation of letter values without such limits and is designed to save results in new variables for as many variables and distinct groups as are specified. Results may then be easily listed and (especially) plotted in pursuit of identification and comparison of distribution level, spread, and shape. Examples are given with emphasis on quantile plots.

Keywords: `st0465`, `lvalues`, letter values, order statistics, quantiles, distributions, level, spread, shape, skewness, exploratory data analysis, graphics

1 Introduction

How to summarize a set of values for a single variable is among the first problems met in any introduction to statistics. It remains an early challenge for researchers in any statistical project, particularly as datasets that are large and heterogeneous become more common, even if the major goal is to move beyond summary to some kind of predictive modeling. Good strategy is explained by many teachers and texts: seek a parsimonious summary in terms of selected key measures, but plot the data carefully, not least to look out for any awkward or pathological features. Nevertheless, there is still much room for variations and innovations in technique.

This article revisits a proposal by John W. Tukey from the 1970s to use what he called “letter values” as one basis for summary and display. Letter values are selected quantiles, starting with the median (tagged M), hinges or fourths (approximate quar-

tiles) (tagged *H* or *F*), eighths (approximate first and last octiles) (tagged *E*), sixteenths (tagged *D*), and so on until we reach the extremes, the minimum and maximum (tagged just 1). That last tag is explained by the fact that the extremes both have a depth of 1, where “depth” is just the number of values counted inward from the extremes.

Otherwise put, and without now any compulsion to give novel names or tags to all measures so defined: Calculating the median defines two tails to a distribution, each including values with probability one-half. We look for the medians of each half, defined as the quartiles, which in turn define two quarters to the distribution lying beyond them, one above the upper quartile and one below the lower quartile. Repeating this idea, we halve the tail fractions at each step and look for their medians until we reach the extremes.

It turns out, as does not seem to have been remarked before, that the main idea was used previously by Francis Galton in 1880. People interested in the history of statistical ideas will find further discussion later in the article.

Tukey’s name “letter values” is matched by the “letter value displays” he suggested, supported in Stata’s `lv` (see [R] `lv`) command since Stata 3.0 (1992). Using letter tags (*M*, *F*, *E*, *D*, ...) faces a small difficulty that we could easily run out of possible letters when faced with large samples, unless we complicate the lettering rules somehow. However, only *M*, *F*, *E* and 1 are strongly evocative in any case. Hence, this is no real loss, particularly as the letter tags are not of use or importance for other displays. The older terms for various quantiles (median, quartiles, octiles, etc.) have proved to be much more durable than the letter tag notation.

In what follows, I mix the older names freely with Tukey’s particular algorithm for calculating such quantiles, or approximations thereto. Those wedded to ideas that other calculation methods for quantiles are better, or even correct, will presumably prefer one of the several methods documented by Hyndman and Fan (1996). But while the name “letter values” is no longer especially apt for their best uses, it is also an established term, so there is no compelling need to replace it.

The major contribution of this article is a new command, `lvalues`, for calculation of letter values as new variables, together with examples of its use, particularly for plotting distributions or aspects of distribution level, spread, and shape.

In the next section, I develop the main idea of letter values with slightly more formality and document some minor complications. In later sections, I detail the syntax of the new command and give worked examples.

2 The main idea

2.1 Letter values

Consider a set of n values ordered, smallest first, so that they have ranks 1 to n . The ordered values are often called “order statistics” or (particularly in statistical graphics) the (sample) “quantiles”. In ranking, tied values are here arbitrarily assigned distinct (unique) ranks, so each integer from 1 to n is used just once as a rank.

The depth associated with rank i is the smaller of i and $n-i+1$. Hence, the extremes (minimum and maximum) with ranks 1 and n both have depth 1, the second smallest and second largest values both have depth 2, and so on. Think of depth as giving the number of values counted inward from the extremes. (For more on depth, see [Tukey \[1977\]](#) or the discussion of trimmed means in [Cox \[2013\]](#).)

The conventional rule for calculating a median can be stated in terms of a depth $(1+n)/2$. If n is odd, then the result is an integer; if n is even, then the result is a half-integer. So if $n = 75$, the depth is 38, which means that the median is the single value that has rank 38; if $n = 74$, the depth is 37.5, which is interpreted as the mean of, or midpoint between, the values with ranks 37 and 38. The median may be tagged with the letter M . The median is a “letter value”, in Tukey’s terminology ([Tukey 1977](#)).

Further letter values are calculated by extending this idea to mark successively smaller tail fractions of a sample. Fourths (approximate quartiles) (tagged F , say) both have a depth that is $(1 + \lfloor \text{depth of median} \rfloor)/2$; eighths (approximate first and last octiles) (tagged E , say) have depth $(1 + \lfloor \text{depth of fourths} \rfloor)/2$; and so on. See [Hoaglin \(1983\)](#) for a systematic account. In each case, integer and half-integer depths imply selecting single values and averaging adjacent ordered values, respectively.

As in [Cox \(2013\)](#), the engagingly intuitive floor notation $\lfloor \rfloor$ specifies rounding down to the nearest integer. See [Cox \(2003\)](#) if you seek more discussion and further references on the notation for floors and their siblings, ceilings.

[Tukey \(1970\)](#) discussed medians M , hinges H , eighths E , and, in passing, sixteenths defined in this way. The general idea as discussed here was given concisely by [Tukey \(1975, 529\)](#), but without the terminology of letter values or the tags.

[Tukey \(1977\)](#) used further letter values D (for sixteenths), C , B , A , Z , Y , X , and so on, as needed, stopping when the extremes are reached at depth 1 (each is tagged 1). The labels M , F , E are pleasantly mnemonic, and those and other tags help to simplify tabular displays. However, memorizing the meanings of other tags is harder work. Knowing or using the tags is less important than keeping an eye on the depths, ranks, and plotting positions associated with each letter value. Plotting positions are in effect estimates of the cumulative probabilities associated with each value and at least roughly equivalent to percentile ranks.

The letter values can be shown in letter value displays, produced with relatively little effort from small datasets (for example, [Tukey \[1977\]](#), [Mosteller and Tukey \[1977\]](#), and [Velleman and Hoaglin \[1981\]](#)). `lv` is the standard Stata implementation. Despite the

advent of larger datasets and ubiquitous computing facilities, interest in letter values continues (for example, Hofmann, Kafadar, and Wickham [2011]). In essence, the letter values are interesting and useful as a parsimonious but informative reduction of a sample distribution based on order statistics (quantiles), with detail in the tails. Hence, they are pertinent to data screening and exploratory data analysis, including determination of distribution location, scale, and shape; identification of problematic data points; and consideration of transformations.

Note the possibility of extending the letter values with extra summaries calculated in similar style. Thus Brizzi (2000, 250) calculates values in between the median and the fourths with depth $\{n - \lfloor (n-2)/4 \rfloor\}/2$ and in between the fourths and the eighths with depth $\lfloor \text{depth of } F + \text{depth of } E \rfloor/2$. Such summaries are not further discussed here, but there can be no objection to their being used if found helpful, unless it is a quibble that they disturb the simplicity of the idea.

See also Tukey (1977) and Hoaglin (1985a,b) for more on using letter values in study of distributions; Marsh (1988), Fox (1990), and Mills (1990) for examples of textbook introductions to letter values; Cox (2004) for discussion of related skewness plots; and Brizzi (2000, 2002) for tests of skewness and kurtosis based on letter values.

2.2 The `lvalues` command

The new command introduced with this article, `lvalues`, calculates letter values as defined by Tukey (1975, 1977) and Hoaglin (1983) for each variable in a list of one or more numeric variables. By default, letter values are stored in new variables. Optionally, letter values may be displayed only, without generation of new variables.

By default, `lvalues` calculates new variables as follows. For every variable in *varlist*, there is a new variable containing its letter values for the observations included in the calculation. In addition, variables give ranks, depths, and plotting positions $(i-a)/(n-2a+1)$ for some a . The default variable names for k variables in *varlist* are `_lv1` to `_lvk` and `_rank`, `_depth`, and `_ppos`. If any of those names is in use, and alternatives not in use are not suggested through the `generate()` option, then the command will fail. Unlike `lv`, `lvalues` will not overwrite existing variables.

Because no letter value necessarily corresponds uniquely to any single data value, and because many letter values are means of (midpoints between) data values, the values of any new variables are (contrary to usual Stata practice) not to be considered as aligned with values of other variables in the same observations. However, if the `by()` option is used, values of any new variables will be placed in observations with corresponding values of the *byvarlist* specified. Positively, it is always true that letter value results are aligned with depths, ranks, and plotting positions.

The number of letter values for n values is $1 + 2\lceil \log_2 n \rceil$ (compare Tukey [1975, 529]). Here the ceiling notation $\lceil \cdot \rceil$ specifies rounding up to the nearest integer. For $n = 1$, that is 1, so the single letter value (median) is just the single data value. For $n = 2, 3, 4, 5, 6, 7$, the number of letter values is 3, 5, 5, 7, 7, 7; that is, in some cases,

there are more letter values than data values. For $n \leq 7$, `lvalues` just returns the ordered values. With that small a sample size, looking at all the values is both feasible and sensible.

To see what that implies, let's consider the number of letter values for examples of very different sample sizes: for $n = 1000$, 1 million, and 1 billion, there are 21, 41, and 61 letter values, respectively. Note that `lv` will not display or save more than 21 letter values.

One further detail is usually avoided in discussion, but should be faced by any careful program. If the iteration process selects letter values with depth 2, as it will about half the time, then the next letter value would have depth $(1 + \lfloor 2 \rfloor)/2 = 1.5$, and the last letter value in turn would have depth $(1 + \lfloor 1.5 \rfloor)/2 = 1$. But a letter value for depth 1.5 adds no information to those for depths 2 and 1, being inevitably just the mean of those 2 letter values. Hence, many data analysts might prefer omitting those letter values, and `lvalues` includes an option to do so. [Velleman and Hoaglin \(1981, 45–46\)](#) raised this point. The number of letter values would then be two fewer than given above.

2.3 Historical remarks, especially on Galton's earlier work

Median, quartiles, and other quantiles (as now called) have been in use as descriptive measures for distributions for well over a century. Their current use arguably owes most to Francis Galton (1822–1911), who championed their employment in several of his books and articles. This is well known through biographies surveying his research and general histories of statistics. What seems to have escaped notice to date is that one of his projects contains the essence of letter values (although trivially, not the term that was introduced by Tukey).

Galton ([1880b](#), especially pages 308–309; compare [1880a](#), especially page 313; [1883](#), pages 93–94; [1907](#), pages 64–65) used summaries of data in terms of first suboctile, first octile, first quartile, medium or middlemost, last quartile, last octile, and last suboctile, in practice the 6th, 12th, 25th, 50th or 51st, 75th, 88th, and 94th percentiles. He flagged that “the system admits of indefinite extension” ([1880b](#), 309) and further added the lowest and highest values, but warned of their dependence on sample size and their being liable to be “of an exceptional character” ([1880b](#), 310). His application was to reports of mental imagery, ordered on their vividness: for more on the project, see [Burbidge \(1994\)](#). These are essentially the same summaries as the letter values that would be used in Tukey's method for $n = 9(1)16$ and, together with others, for larger sample sizes.

To be sure, Galton used other schemes both before and after this work. [Galton \(1874\)](#) suggested using the 2, 9, 25, 50, 75, 91, and 98 percentiles (in later terms) as being 0(1)3 probable errors from the mean of a (tacitly approximately normal) distribution. [Galton \(1885\)](#) used the 5 10(10)90 95% percentiles; see also [Galton \(1896\)](#). And similar ideas were in the air both at the time and a little later. For example, [Bowley \(1910, 62; 1952, 73\)](#) recommended using minimum and maximum and 10, 25, 50, 75, and 90% points as a basis for graphical summary.

The unusual name “suboctile” raises the question of terminology for various kinds of quantiles. If interested in this point, see the Appendix at the end of this article.

Also note that the logic behind the plotting position rule $(i - 0.5)/n$ will be found in Galton’s work at this time (for example, Galton [1883, 53–54]). Various later workers independently reasoned their way to the same rule.

3 Syntax

```
lvalues varlist [if] [in] [, a(#) by(byvarlist)
    {generate(newvarlist) | displayonly} list list_options omit1h]
```

3.1 Options

a(#) specifies the constant a in calculating plotting positions. The default is **a(1/3)**, as suggested by Tukey (1975, 528–529) and Hoaglin (1983) in a detailed discussion of letter values and plotting positions. A particular advantage of this choice is that it corresponds closely to the position of the median of the sampling distribution of each order statistic. See also Cox (2016a) on plotting positions in a Stata context.

by(*byvarlist*) specifies one or more variables defining distinct groups for which letter values are to be calculated separately.

generate(*newvarlist*) specifies new variable names as alternatives to the default, up to as many as the number of variables plus three. If fewer variable names are suggested, as many of **_lv1–_lv#**, **_rank**, **_depth**, and **_ppos** are used as needed, but those default names used must still be new in the dataset.

displayonly specifies display of the letter values only, with no generation of new variables. Here “display” implies **list**, as shown below.

list specifies that the letter values be listed. **list** may be specified by itself or together with options of **list** (see [D] **list**). The default options include **sep(0) noobs** or (with the **by()** option) **sepsy(*byvar*) noobs**. Plotting positions are shown to three or more decimal places (but stored as **double** variables).

list_options refers to options of **list** (see [D] **list**).

omit1h omits letter values for depth 1.5 (**1h** in notation used by Tukey [1977]) whenever there are also letter values for depths 2 and 1. The motive would be that such letter values contain no extra information. See Velleman and Hoaglin (1981, 45–46) for discussion. The number of letter values would then be two fewer than given above.

4 Examples

4.1 Basic features

We start with a simple example. Reading in a dataset and specifying a single variable to `lvalues`,

```
. sysuse auto
(1978 Automobile Data)
. lvalues mpg
```

has the result that the letter values, ranks, depths, and plotting positions are stored in new variables with names `_lv1`, `_rank`, `_depth`, and `_ppos`. Other names may be specified using the `generate()` option, say,

```
. lvalues mpg, generate(lv_mpg rank depth ppos)
```

Either method may be combined with separate calculation groupwise, such as

```
. lvalues mpg, generate(lv_mpgf rankf depthf pposf) by(foreign)
```

Further, more variables may also be specified:

```
. lvalues mpg weight, gen(lv_mpg lv_weight rank depth ppos) by(foreign)
```

What is sometimes helpful is just a tabulation, without generation of new variables:

```
. lvalues mpg, displayonly
```

rank	depth	fraction	mpg
1	1	0.009	12
1.5	1.5	0.016	12
2	2	0.022	12
3	3	0.036	14
5.5	5.5	0.070	14
10	10	0.130	15
19	19	0.251	18
37.5	37.5	0.500	20
56	19	0.749	25
65	10	0.870	28
69.5	5.5	0.930	30.5
72	3	0.964	35
73	2	0.978	35
73.5	1.5	0.984	38
74	1	0.991	41

4.2 Skewness plots

Let's now turn to using letter values in graphics. The rationale here is twofold. First, clean and simple but informative plots can be obtained often just by using letter values.

Second, the corresponding graph files are smaller than those based on all the data, which despite cheaper memory in various forms remains a virtue. There can always be a possible downside. In particular, fine structure in the data will often be smoothed away, such as granularity in the form of repeated minor modes. Equally, such fine structure may be immaterial for most purposes.

A small family of plots to examine skewness is based on the so-called midsummaries, namely, the median; the mean of (or midpoint between) the fourths; the mean of the eighths; and so forth, out to the mean of the extremes (itself often called the midrange). For a review of this idea, see [Cox \(2004, sec. 6\)](#) and the references there. Examples in that article were based on all the data, but using just the letter values is often adequate.

In an exactly symmetric distribution, these midsummaries would be identical and would plot as a horizontal constant against any other property. Conversely, such plots can show features such as approximate symmetry in the middle of a distribution combined with marked asymmetry in the far tails. Seemingly odd mixes of symmetry and asymmetry are quite common in practice, and not at all well captured by single scalar measures of skewness, whether moment based or quantile based.

Plausible candidates for what to plot against are the depths, possibly reversed, and the spreads or differences, each of the form (upper letter value – lower letter value). For the median, the natural spread to use is just zero.

Apart from the median, the letter values occur in pairs, and so it is a small step to calculate such quantities.

```
. by _depth (_lv1), sort: generate spread = cond(_N == 1, 0, _lv1[2] - _lv1[1])
(59 missing values generated)
. by _depth: generate mid = cond(_N == 1, _lv1, (_lv1[1] + _lv1[2])/2)
(59 missing values generated)
```

Purists will note that in such code, letter values other than the median all occur in pairs, which typically should not bite.

Example graphs are left to readers' discretion. At this point, you need only a standard graph command, such as `scatter` or `line` or `twoway connected`. Showing letter values for two or more variables or two or more groups is more complicated in detail, but just an extension of the same idea.

4.3 Quantile plots

As a further and final example of the uses of letter values, we show them in quantile plots, capitalizing on `qplot`, which generalizes the official `quantile` command in several respects. For general discussions and further references, see [Cox \(1999, 2005\)](#); for the latest update at the time of writing, see [Cox \(2016b\)](#) or `search qplot`. Here, as is standard in the *Journal*, we use the graph scheme `sj`, but you may well be at liberty to produce graphs that are more colorful.

First, we show letter values being used. Because the letter values are not equally spaced in rank or cumulative probability terms, using the `xvariable()` option is necessary.

```
. sysuse auto, clear
(1978 Automobile Data)
. lvalues mpg, by(foreign) list
(output omitted)
. local common over(foreign) aspect(1) legend(order(2 1))
. qplot _lv1, `common' xvariable(_ppos) name(g1)
. qplot _lv1, `common' xvariable(_ppos) trscale(invnormal(0)) recast(connected)
> name(g2)
. graph combine g1 g2
```

Figure 1 shows two versions of the quantile plot, the plain or vanilla version with cumulative probability (fraction of the data) as horizontal coordinate and a normal quantile plot (also known as normal probability plot) in which the normal or Gaussian distribution is the reference. On that scale, the letter values are closer to being equally spaced.

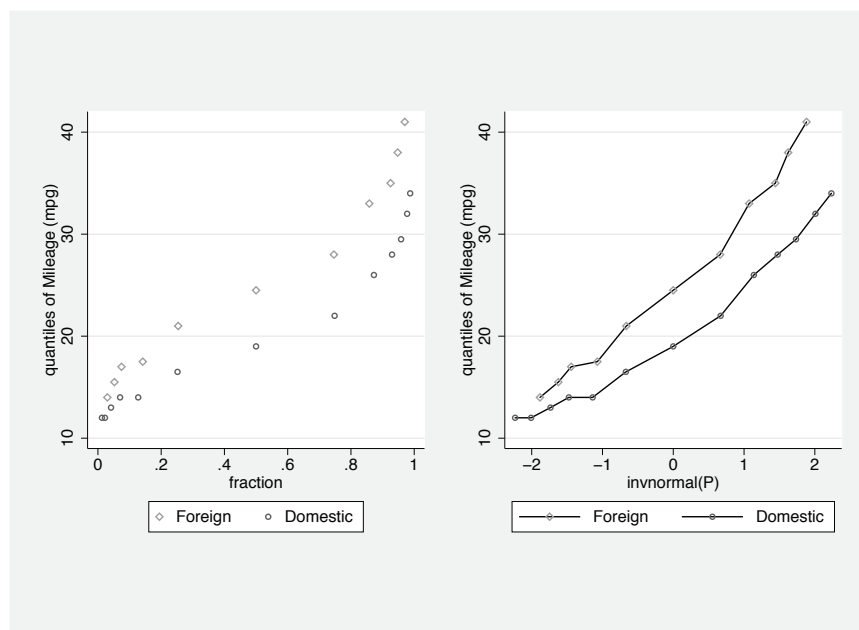


Figure 1. Quantile plots for `mpg` from the `auto` data based on letter values. The version on the left is plotted against cumulative probability (fraction of the data) and the version on the right against quantiles of a standard normal or Gaussian distribution.

We should compare figure 1 with the full quantile plot to see what we missed.

```
. qplot mpg, `common` name(g3)
. qplot mpg, `common` trscale(invnormal(0)) recast.connected name(g4)
. graph combine g3 g4, ycommon
```

Figure 2 shows what we missed, principally, repeated minor modes as a side effect of reporting values as integers. That is worth knowing if not otherwise apparent, but is unlikely to be problematic in any later analysis.

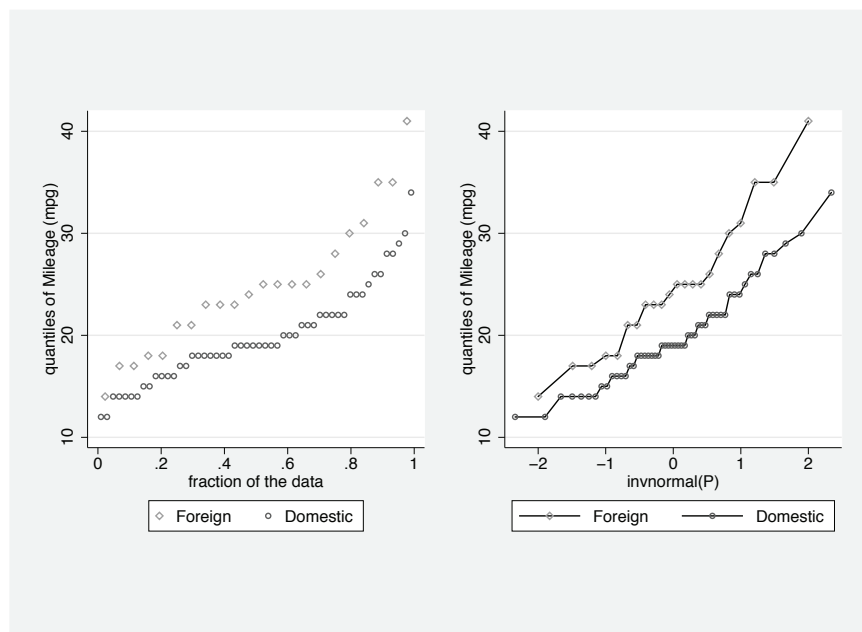


Figure 2. Quantile plots for `mpg` from the `auto` data based on all the data. The version on the left is plotted against cumulative probability, and the version on the right against quantiles of a standard normal or Gaussian distribution.

5 Conclusions

“Data analysis” is one of our standard phrases in statistical science. Literally, or etymologically, “analysis” refers to splitting into components, so that we typically end up with more numbers than we started with, say, parameter estimates describing a model fit, predictions, residuals, figures of merit, and so forth. “Data analysis” is in essence a twentieth-century term, most associated with John Tukey and most evident in his monograph *Exploratory Data Analysis* (1977).

Before data analysis became common parlance, there was often talk, from the nineteenth century at least, of “data reduction”. This also remains central to statistical science and is sometimes used as a title (Ehrenberg 1975). We need to reduce our data to parsimonious and informative summaries. Naturally, Tukey knew all this too. Letter values remain one of his key ideas, simple but serviceable, reductive as well as exploratory. It is also good to note that Galton, anachronistically but appropriately regarded as a great nineteenth-century data analyst, got there earlier.

6 Acknowledgments

David Hoaglin rekindled my interest in letter values by a comment at the Chicago Stata Conference in 2016 and provided helpful encouragement thereafter. Access to Galton’s publications is made immensely easier by <http://galton.org>, an outstanding resource provided by Gavan Tredoux.

7 References

- Aronson, J. K. 2001. Francis Galton and the invention of terms for quantiles. *Journal of Clinical Epidemiology* 54: 1191–1194.
- Bowley, A. L. 1910. *An Elementary Manual of Statistics*. London: Macdonald and Evans.
- . 1952. *An Elementary Manual of Statistics*. 7th ed. London: Macdonald and Evans.
- Brizzi, M. 2000. Detecting skewness and kurtosis by letter values: A new proposal. *Statistica* 60: 243–258.
- . 2002. Testing symmetry by an easy-to-calculate statistic based on letter values. *Metodološki zvezki* 17: 63–74.
- Burbridge, D. 1994. Galton’s 100: An exploration of Francis Galton’s imagery studies. *British Journal for the History of Science* 27: 443–463.
- Cox, N. J. 1999. gr42: Quantile plots, generalized. *Stata Technical Bulletin* 51: 16–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 113–116. College Station, TX: Stata Press.
- . 2003. Stata tip 2: Building with floors and ceilings. *Stata Journal* 3: 446–447.
- . 2004. Speaking Stata: Graphing distributions. *Stata Journal* 4: 66–88.
- . 2005. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460.
- . 2013. Speaking Stata: Trimming to taste. *Stata Journal* 13: 640–666.

- . 2016a. FAQ: How can I calculate percentile ranks? <http://www.stata.com/support/faqs/statistics/percentile-ranks-and-plotting-positions/>.
- . 2016b. Software Updates: Quantile plots, generalized. *Stata Journal* 16: 813–814.
- Ehrenberg, A. S. C. 1975. *Data Reduction: Analysing and Interpreting Statistical Data*. London: Wiley.
- Fisher, R. A., and F. Yates. 1938. *Statistical Tables for Biological, Agricultural and Medical Research*. Edinburgh: Oliver and Boyd.
- Fox, J. 1990. Describing univariate distributions. In *Modern Methods of Data Analysis*, ed. J. Fox and J. S. Long, 58–125. Newbury Park, CA: Sage.
- Galton, F. 1874. On a proposed statistical scale. *Nature* 9: 342–343.
- . 1880a. Mental imagery. *Fortnightly Review* 28: 312–324.
- . 1880b. Statistics of mental imagery. *Mind* 5: 301–318.
- . 1883. *Inquiries into Human Faculty and its Development*. London: Macmillan.
- . 1885. Anthropometric per-centiles. *Nature* 31: 223–225.
- . 1896. Application of the method of percentiles to Mr. Yule's data on the distribution of pauperism. *Journal of the Royal Statistical Society* 59: 392–396.
- . 1907. *Inquiries into Human Faculty and its Development*. 2nd ed. London: J.M. Dent.
- Hoaglin, D. C. 1983. Letter values: A set of selected order statistics. In *Understanding Robust and Exploratory Data Analysis*, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, 33–57. New York: Wiley.
- . 1985a. Summarizing shape numerically: The g-and-h distributions. In *Exploring Data Tables, Trends, and Shapes*, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, 461–513. New York: Wiley.
- . 1985b. 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: Wiley.
- Hofmann, H., K. Kafadar, and H. Wickham. 2011. Letter-value plots: Boxplots for large data. <http://vita.had.co.nz/papers/letter-value-plot.pdf>.
- Hyndman, R. J., and Y. Fan. 1996. Sample quantiles in statistical packages. *American Statistician* 50: 361–365.
- Kendall, M. G. 1940. Note on the distribution of quantiles for large samples. *Supplement to the Journal of the Royal Statistical Society* 7: 83–85.

- Marsh, C. 1988. *Exploring Data: An Introduction to Data Analysis for Social Scientists*. Cambridge: Polity Press.
- Mills, T. C. 1990. *Time Series Techniques for Economists*. Cambridge: Cambridge University Press.
- Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.
- Tukey, J. W. 1970. *Exploratory Data Analysis*. Limited preliminary ed., vol. 1. Reading, MA: Addison–Wesley.
- . 1975. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians: Vancouver, 1974*, ed. R. D. James, 523–531. Vancouver: Canadian Mathematical Congress.
- . 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.
- Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury.

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 was an author of several inserts in the *Stata Technical Bulletin* and is an editor 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).

A Appendix: Further remarks on terminology for quantiles

[Aronson \(2001\)](#) documented first uses of various terms for quantiles. This list adds some earlier dates from searches of the *Oxford English Dictionary* and <http://www.jstor.org> on October 5, 2016. The dates refer to earliest citations of the terms with their statistical meaning and not to other meanings. The general term “quantile” itself is often attributed to [Kendall \(1940\)](#) but can be found in [Fisher and Yates \(1938\)](#).

English ordinal	Statistical term	Earliest citation (Aronson)	2016 additions (Cox)
Third	Tertile	1931	1911
	Tercile	1942	
Fourth	Quartile	1879 (*)	
Fifth	Quintile	1951	1910
Sixth	Sextile	1920	
Seventh	Septile	1993	1981
Eighth	Octile	1879	
Ninth	Nonile	1968	
Tenth	Decile	1881	
Sixteenth	Suboctile	1880	
Twentieth	Vigintile	1936	
Fortieth	Quadragintile	1976	
Hundredth	Percentile	1885	
	Centile	1902	1894
Thousandth	Permille	1904	

(*) [Galton \(1874\)](#) used “quarter points” for the quartiles.