The Stata Journal

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

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

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

http://www.stata-journal.com

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

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

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

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

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

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

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

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

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

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

The Stata Journal (2013)
13, Number 3, pp. 640–666

Speaking Stata: Trimming to taste

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

Abstract. Trimmed means are means calculated after setting aside zero or more values in each tail of a sample distribution. Here we focus on trimming equal numbers in each tail. Such trimmed means define a family or function with mean and median as extreme members and are attractive as simple and easily understood summaries of the general level (location, central tendency) of a variable. This article provides a tutorial review of trimmed means, emphasizing the scope for trimming to varying degrees in describing and exploring data. Detailed remarks are included on the idea's history, plotting of results, and confidence interval procedures. Examples are given using astronomical and medical data. The new Stata commands `trimmean` and `trimplot` are also included.

Keywords: st0313, trimmean, trimplot, trimming, means, medians, midmeans, Winsorizing, robust, resistant, graphics

1 Introduction

Trimmed means are means calculated after setting aside some values in one or both tails of a sample distribution. In the simplest and most common case, the same percent or number is set aside in each tail, omitting equal numbers of lowest and highest values.

The idea of trimming has been reinvented repeatedly since the 18th century. Trimmed means have been prominent as one of the simpler methods within the field of robust statistics for over 50 years, since their reintroduction by J. W. Tukey (1960), W. J. Dixon (1960), and others. The idea of trimming binds means and medians together in a wider family: the mean, strictly speaking, is the mean with no values trimmed, while the median is the mean with all values trimmed except the one or two values that define the median. Intermediate degrees of trimming offer varying compromises between the urge to use all the information in the data and any need to discount extreme values that may appear unreliable.

Trimming before averaging is easy to understand and to explain to general scientific audiences. Trimmed means are likely to be useful as a cautious check on means or as an alternative summary when using means seems dubious or even dangerous. Why then are they not more widely used? Lack of detailed explanations and implementations may be one answer, and this article addresses that lack for Stata users. Two earlier user-written programs will be discussed later, but in general, Stata has lagged behind other statistical software in this field: "a trimmed mean" was added to BMDP in 1977 (Hill and Dixon 1982, 378).

Trimmed means may be based on trimming differently in each tail, including the case of trimming in one tail only. Staudte and Sheather (1990), for example, first introduce trimmed means in terms of trimming only in the right tail when estimating the scale of an exponential distribution. Here we focus on trimming symmetrically.

The new commands `trimmean` and `trimplot` are included in this article. Discussion will follow a detailed explanation of the basic statistics, together with historical remarks.

Some comments follow first on the literature. The reviews by Dixon and Yuen (1974) and Rosenberger and Gasko (1983) remain clear and helpful on both specific details and wider context. Both have often been overlooked in later surveys for no obvious good reason. Kafadar (2003) gives an excellent concise review of Tukey's work on robustness. Stigler (2010) gives a light and brisk historical perspective on robust statistics in general.

Introductory or intermediate texts featuring trimmed means include Breiman (1973, 244), Mosteller and Tukey (1977, 34–36), Dixon and Massey (1983, 380–382), Siegel (1988, 66–68), Helsel and Hirsch (1992, 7), Venables and Ripley (2002, 122), van Belle et al. (2004, 276–277), Rice (2007, 397), Sprent and Smeeton (2007, 461–465), Reimann et al. (2008, 43), and Feigelson and Babu (2012, 110).

2 Definitions

2.1 Simplest case

The order statistics of a sample of n values of a variable y are defined by

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)}$$

so that $y_{(1)}$ is the smallest value and $y_{(n)}$ is the largest.

The method for trimmed means at its simplest is to set aside some fraction of the lowest-order statistics and the same fraction of the highest-order statistics and then to calculate the mean of what remains, thus providing some protection against possible stretched tails or outliers in a sample. For example, suppose $n = 100$, and we set aside 5% in each tail, namely, $y_{(1)}, \ldots, y_{(5)}$ and $y_{(96)}, \ldots, y_{(100)}$. We can then take the mean of $y_{(6)}, \ldots, y_{(95)}$.

For such a definition, see Tukey and McLaughlin (1963, 336), Bickel (1965, 848), Huber (1981, 57–58), Lehmann (1983, 360), Rosenberger and Gasko (1983, 307–308), Hampel et al. (1986, 178), Staudte and Sheather (1990, 104), Barnett and Lewis (1994, 79), Miller (1986, 29), David and Nagaraja (2003, 213), Jurečková and Picek (2006, 67), Pearson (2011, 228, 267), or Wilcox (2003, 62–63; 2009, 26; 2012a, 55; 2012b, 25).

The 0% trimmed mean is thus just the usual mean.

By courtesy, or as a limiting case, the 50% trimmed mean is taken to be the median. The small detail here is that trimming exactly half the values in each tail will leave no values at all; hence, the courtesy exercised in leaving the one or two values required to determine the median. (Averaging the two central values whenever n is even to calculate the median is explained to mathematical audiences as a convention and to nonmathematical audiences as a rule.)

One other trimmed mean has often been given a special name. The 25% trimmed mean has been called the "midmean", that is, the mean of the middle half of the data (Tukey [1970a; 1970b, 168], adopting earlier scientific usage, on which see Tukey [1986, 871]); the "interquartile mean" (for example, Tilanus and Rey [1964]; Erickson and Nosanchuk [1977, 40; 1992, 44]); and the "quartile-discard average" (Daniell 1920).

A more general rule is that the lowest value included in the calculation of the p% trimmed mean is $y_{(g+1)}$, where $g = \lfloor n\,p/100 \rfloor$, and the highest value included is thus $y_{(n-g)}$.

The very useful floor notation, $\lfloor\ \rfloor$, here specifies rounding down to the nearest integer. Incidentally, almost all the literature on trimmed means uses [ ] with the same meaning. Despite a lengthy pedigree, that notation needs to be explained repeatedly: many readers might disregard it as merely a standard use of brackets. See Cox (2003) for more discussion and further references on floors and ceilings.

2.2 Weighting

Some authors use a more elaborate definition of trimmed means in which some values may be given fractional weights. See Andrews et al. (1972, 7, 31), Stigler (1977, 1060), Kleiner and Graedel (1980, 706), Huber (1981, 57–58), Rosenberger and Gasko (1983, 311), Barnett and Lewis (1994, 79), Huber and Ronchetti (2009, 57–58), or Wilcox (2012a, 55).

The precise rule is usually that $\lfloor n\,p/100 \rfloor$ values are removed in each tail, and the smallest and largest remaining values are assigned weight $1 + \lfloor n\,p/100 \rfloor - n\,p/100$. So, for example, given $n = 74$ and percent 5/100, their product is 3.7. Rounding down gives 3 and so we work with $y_{(4)}, \dots, y_{(71)}$. However, $y_{(4)}$ and $y_{(71)}$ are assigned weight $4 - 3.7 = 0.3$, and $y_{(5)}, \dots, y_{(70)}$ assigned weight 1. Then a weighted mean is taken.

The idea underlying this alternative definition appears twofold: p% should mean precisely that, and the result of trimming should vary as smoothly as possible with p. Rosenberger and Gasko (1983, 310–311) explain this especially clearly with two helpful diagrams.

The difference is partly a matter of taste. But always using weights that are 1 or 0 is appealingly simple and appears entirely adequate for descriptive and exploratory uses. Moreover, any fine structure that results from the inclusion and exclusion of particular values as trimming proportion varies is likely to be trivial or part of what we are watching for. Either way, there is little loss.

2.3 Number instead of percent trimmed

In some situations, it is more natural to specify trimming in terms of the number of values trimmed rather than the percent. For example, trimming or truncating procedures have been used in combining the scores of a panel of judges in various sports to discourage or discount bias for or against competitors. Here the rules might require, for example, trimming the highest and lowest values.

Focusing on the number trimmed allows a slightly different definition of trimmed means. We can describe the order statistics $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n-1)} \leq y_{(n)}$ using the idea of depth (for example, Tukey [1977]). Depth is defined as 1 for $y_{(1)}$ and $y_{(n)}$, 2 for $y_{(2)}$ and $y_{(n-1)}$, and so forth: it is the smaller number reached by counting inward from either extreme $y_{(1)}$ or extreme $y_{(n)}$ toward any specified value. So the depth of $y_{(i)}$ is the smaller of i and $n - i + 1$.

A trimmed mean may be defined for any particular depth as the mean of all values with that depth or greater. Thus the trimmed mean for depth 1 is the mean of all values. The trimmed mean for depth 2 is the mean of all values except those of depth 1, that is, all values except for the extremes. The trimmed mean for depth 3 is the mean of all values except those of depths 1 and 2, and so forth.

The highest depth observed for a distribution occurs once if n is odd and twice if n is even; either way it labels values whose mean is the median. Thus, again, trimmed means range from the mean to the median.

2.4 Symmetry or asymmetry?

Whatever the precise definition, trimming the same number of order statistics in each tail is arguably based on a symmetry assumption—if not that the distribution of interest is approximately symmetric, then that the chances of contamination are approximately equal in either tail. Certainly, estimation of location (level, central tendency) is easy to think about whenever the underlying distribution is symmetric (and easier still if it is unimodal). Then estimators of location can typically be thought of as aimed at precisely the same target, the middle or center of the distribution.

The opposite argument is that the estimand is whatever the estimator points to. As Tukey (1962, 60) urged, "We must give even more attention to starting with an estimator and discovering what is a reasonable estimand, to discovering what is it reasonable to think of the estimator as estimating". A similar point of view has been elaborated formally in considerable detail by Bickel and Lehmann (1975) and informally with considerable lucidity by Mosteller and Tukey (1977, 32–34). There is also an elementary version. Using, say, sample median or geometric mean to estimate the corresponding population parameter makes sense regardless of whether the underlying distribution is symmetric, and the same goodwill can be extended to sample trimmed means, which are regarded as estimators of their population counterparts.

All that said, symmetric trimmed means are unlikely to be ideal for strongly asymmetric distributions. So-called J-shaped distributions such as the exponential or Pareto are examples. As remarked in the *Introduction*, there has been work on asymmetric trimmed means, but that is not discussed further in this article.

2.5 Confidence intervals

Trimmed means can be given confidence intervals. The approach here follows Tukey and McLaughlin (1963). Note also Dixon and Tukey (1968) as the sequel to that article. For a one-sentence summary, see Huber (1972, 1053–1054). For lucid textbook accounts, see Staudte and Sheather (1990, 98), Miller (1986, 30–31), Huber and Ronchetti (2009, 147–148), or Wilcox (2003, 126–132; 2009, 98–99, 127–128, 150–151; 2010, 153–154; 2012a, 57–61, 111–114; 2012b, 153–159).

Suppose we have n values and trim g in each tail and we seek *level*% confidence intervals (for example, *level* = 95). We need first a Winsorized standard deviation. Winsorizing is replacing values in each tail by the next inward value; that is, $y_{(1)}, \ldots, y_{(g)}$ are each replaced by $y_{(g+1)}$, and $y_{(n-g+1)}, \ldots, y_{(n)}$ are each replaced by $y_{(n-g)}$ before calculation, so long as $g \geq 1$. Let sd_W denote the standard deviation of the Winsorized values. Then intervals are mean $\pm$ (t multiplier $\times$ sd_W) / $\{\sqrt{n}\,(1 - 2g/n)\}$, where the t multiplier in Stata terms is `invttail(`*n* `- 2*g - 1, (100 -` *level*`)/200)`. If the latter expression looks too much like a strange incantation, consider an example such as $n = 100$, $g = 5$, and *level* = 95:

```
. display invttail(100 - 2*5 - 1, (100 - 95)/200)
1.9869787
```

Note that the `trimmean` command uses `summarize` to calculate the standard deviation; thus, as documented in [R] **summarize**, the divisor before rooting is $n - 1$. The fraction of values used in the trimmed mean $1 - 2g/n$ is calculated from the number actually used, not from any percent trimming specified.

This approach does not in the limit as trimming approaches 50% give reasonable confidence intervals for the median, because the number of degrees of freedom in this method approaches 0. `trimmean` declines to cite confidence intervals for the median; otherwise, obtaining intervals for large trimming fractions is left to the judgment of the user.

As another approach to confidence intervals, bootstrapping is quite attractive. Efron and Tibshirani (1993) and Davison and Hinkley (1997) discuss bootstrapping trimmed means. Although all results are returned in a matrix, `trimmean` also saves each trimmed mean separately as a convenience. However, bootstrapping necessarily implies that wild values could be selected repeatedly in a bootstrap sample, so some individual trimmed means could be much less resistant than the mean based on the sample as a whole. The converse is also true.

2.6 Metric trimming

A different definition of trimmed means, often called metric trimming, yields means of values satisfying some constraint on the absolute deviation from the median $|y - \text{median}(y)| =: d$. The name "metric" echoes Bickel (1965), Kim (1992), and Venables and Ripley (2002, 122). None of those cited the earlier work of Short (1763), similar in spirit, except that he worked with $|y - \text{mean}(y)|$.

Two simple merits of this definition deserve mention. Like the usual definition, it defines a family spanning the median and mean as extremes and including intermediate compromises. Specifying allowed deviations on the scale of the variable may make much sense to working scientists accustomed to thinking about their measurements.

Thus $d = 0$ identifies data points equal to the median. (A small detail here is that quite possibly, $d = 0$ identifies no data points at all: that will necessarily happen whenever n is even and the median is calculated as the mean of two different values.)

At the other extreme, such a trimmed mean equals the mean so long as d exceeds the largest possible absolute deviation, the larger of $\text{median}(y) - y_{(1)}$ and $y_{(n)} - \text{median}(y)$.

Lest readers confuse this with other procedures, this is not an iterative calculation. The overall median is calculated just once; there is no cycling such that the median is redefined to be the median of those values $\leq d$.

Metric trimming can be combined with trimming based on order statistics (for example, Hampel [1997, 150] and Olive [2001]), but only the simplest flavor is supported by `trimmean`. See also Huber (1964) for a brief mention and Hampel (1985) for broader discussion.

3 Historical remarks

The idea of a trimmed mean is quite old. For some related history, see Stigler (1973, 1976), Harter (1974a,b), Hampel et al. (1986, 34–36), and Barnett and Lewis (1994, 27–31). The episodes identified here seem best thought of as independent inventions and not as evidence of a continuous thread of thought intermittently made visible.

Throughout several centuries, trimmed means have been one of several practices in science and, typically, only an occasional practice. Pooling all measurements and calculating a single mean has at best been one approach and only slowly came to be regarded as one standard. Choosing the best measurement from several as a matter of judgment was, and remains, an alternative often used both in science and in everyday life. Some scientists and a few statisticians have focused on "rejection of outliers", that is, identification either by judgment or by some formal rule of outlying values not to be trusted. This might be included under trimming in a broad sense. However, it seems best to distinguish clearly: Trimmed means are based on choosing a rule for trimming, whether a percent or number to be trimmed or a maximum allowed deviation. Rejection of outliers is based on looking at the data, deciding which, if any, values need to be rejected, and then averaging what remains.

Five case studies follow from the literature, with absolutely no claim to completeness. Biographical vignettes on the main individuals follow the References.

3.1 James Short and the 1761 transit of Venus

James Short (1763) used a form of what is now called metric trimming in 1763 for estimating the sun's parallax based on observations of the transit of Venus across the face of the sun, namely, taking the mean of values closer than some chosen distance from the mean of all. The parallax here is the angle subtended by the earth's radius, as if viewed and measured from the surface of the sun. The units are seconds of a degree. Note that repeating Short's (1763) calculations points up small errors in his arithmetic. For much more on measuring the transit of Venus in 1761 (and 1769) as a major research program in astronomy, see Woolf (1959) or Wulf (2012). Woolf (1959, 147) comments: "One of the factors that had rendered Short's results so homogeneous had been the rather judicious series of alterations which he had made in the original data concerning longitude and time of contact at various stations". Short's (1763) line, however, was that he was fixing the mistakes of others.

Stigler (1977) included Short's (1763) data in an evaluation of robust estimation methods with real data.

Short (1763) provides 53 measurements on page 310 of his article. These are datasets (1) to (3) in Stigler (1977, 1074). Short (1763) first averages all 53 and gets 8.61; then all 45 within 1 of that mean and gets 8.55; then all 37 within 0.5 of that mean and gets 8.57. Then he takes the mean of all 3 means and gets 8.58. In effect, his final mean is weighted according to deviations from the initial overall mean.

Similarly, Short (1763) provides 63 measurements on page 316 of his article. These are datasets (4) to (6) in Stigler (1977, 1074). Short (1763) first averages all 63 and gets 8.63; then all 49 within 1 of that mean and gets 8.50; then all 37 within 0.5 and gets 8.535. The mean of all 3 means is 8.55.

Short's (1763) data on page 325 of his article are datasets (7) and (8) in Stigler (1977, 1074). The mean of 21 values in the first set, for the Cape of Good Hope, is 8.56. All 29 values are within 0.2 of that. The mean of 21 values in the second set, for Rodrigues, is 8.57; the same mean is obtained for all 13 within 0.2.

3.2 A French custom

An anonymous writer (identified by Stigler [1976] as Joseph Diaz Gergonne, 1771–1859) included an example of trimmed means in a discussion of how to calculate means (Anonymous 1821, 189): "For example, there are certain provinces of France where, to determine the mean yield of a property of land, there is a custom to observe this yield during twenty consecutive years, to remove the strongest and the weakest yield and then to take one eighteenth of the sum of the others" (translation in Huber [1972, 1043]).

3.3 Mendeleev on metrology

Mendeleev (1895) (reference in Harter [1974b, 241]) reported his method "to evaluate the harmony of a series of observations that must give identical numbers, namely, I divide all the numbers into three, if possible equal, groups (if the number of observations is not divisible by three, the greatest number is left in the middle group): those of greatest magnitude, those of medium magnitude, and those of smallest magnitude; the mean of the middle group is considered the most probable ... and if the mean of the remaining groups is close to it ... the observations are considered harmonious". Thus Mendeleev used a 1/3, roughly 33%, trimmed mean.

3.4 Daniell's theoretical treatment

Daniell (1920) gave an elegant and pathbreaking general treatment of statistics that are linear combinations of the order statistics, including various estimators of location and scale. It was apparently inspired by a reading of Poincaré's *Calcul des probabilités* (1912). Daniell derived optimal weighting functions and gave the first mathematical treatment of the trimmed mean. However, his article had essentially no impact before its rediscovery by Stigler (1973). Its placement in a journal rarely read by statisticians cannot have helped.

3.5 Tukey and modern robust statistics

Tukey (1960) surveyed the problem of location estimation when data are likely to come from distributions heavier tailed than the normal (Gaussian) in an outstanding article that was one of the founding documents of modern robust statistics. He combined a literature review with a report on his own work on the subject since the mid-1940s, some published partly as technical memoranda. In particular, Tukey (1960) showed that truncated means (his term in this article) calculated after dropping the same percent of the lowest and highest values offered considerable protection in the face of such distributions. Dixon (1960) also deserves credit for work in this territory.

The term "trimmed mean" was introduced shortly afterward by Tukey (1962). Names in earlier use include "truncated mean" (Tukey [1960], as above) and "discard average" (Daniell 1920). Dixon (1960) discussed using means of a censored sample. Talking of truncation or censoring raises the need to distinguish carefully between truncation or censoring of the data before they arrive and such truncation or censoring used deliberately in data analysis—reason enough for using the term "trimming" instead.

After 1960, trimmed means became an established method within the field of robust statistics, a field repeatedly surveyed and unified by monographs and textbooks and now covering many other statistical problems, including robust regression.

4 How much to trim in practice?

A simple and natural question not raised so far is how much to trim in practice. There is a simple and natural answer: that depends on your dataset, your ideas about generating processes, and your attitude to risk. But let us back up and consider more generally how we might evaluate trimmed means and make choices.

One of the advantages of trimmed means is that their behavior is easy to think about. Trimming p% in each tail of a distribution offers protection against (up to) the same fraction of dubious values in each tail. A more formal treatment would be phrased in terms of the idea of "breakdown point" (Donoho and Huber 1983). Whether these dubious values are called outliers, or something else, is partly a matter of taste and judgment. There is also some taste and judgment in trading off the protection afforded against dubious values against the loss of information incurred by ignoring values that may be genuine.

As with any other method, trimmed means may be studied theoretically, including by simulation from distributions deemed credible as generating processes, or empirically, by studying how trimmed means behave with real data. The broad advantages and disadvantages of each method are clear.

The monograph by Andrews et al. (1972) remains the most impressive compendium of simulation results for trimmed means (and other robust estimators of location). It is striking to learn from Hampel (1997) that only a small part of the project was ever written up. One important omission was what happens with skewed or asymmetric distributions. But the mass of previous simulation results, both in that volume and elsewhere, is increasingly redundant. It is now easier to simulate afresh using whatever underlying distributions and sample sizes appear pertinent to any particular project than to comb through the literature searching for relevant results.

Similarly, studies with other datasets still raise the question of judging which other results are pertinent to the current project. Such studies (for example, Stigler [1977] and Hill and Dixon [1982]) are often more positive than reports of simulation studies, implying in particular that mild trimming (loosely, of the order of 5–10%) may be all that is required in many cases.

In either case, there are severe selection problems. How do you reasonably sample from the space of possible distributions, whether theoretical or empirical? What part of that space is relevant to your project? These are difficult questions.

One tactic that seems underplayed is to use trimming percents across a wide range and see what happens. An obvious aid here is to plot trimmed mean versus number or percent trimmed or allowed deviation. Examples of such plots can be found in Rosenberger and Gasko (1983, 315) and Davison and Hinkley (1997, 122).

Then the possibilities include, but are not limited to, some leading cases:

1. Results are stable, whatever the trimming proportion. We can relax and just use means in any way. This case is like a health check or machine service that found no problems.

2. Results are stable provided that you trim at least a certain proportion.

3. Results vary systematically with trimming proportion, from mean to median. Note that this is expected with most asymmetric or skewed distributions, regardless of whether outliers or heavy tails are present. (It is often forgotten that there are skewed distributions for which mean and median are identical or very close. For example, this is true of some binomial distributions and usually a good approximation for Poisson distributions.) In turn, there will be choices, including living with the fact; realizing that multiple descriptors—say, mean and median and perhaps others as well—may be advisable in reporting data; and considering an appropriate transformation or link function (in the jargon of generalized linear models, see [R] **glm**).

4. Something else that needs consideration. Trimmed means may here indicate a problem, but they do not promise to provide a solution.

`trimmean` is designed to make it easy to produce several trimmed means at once by specifying differing trimming percents, numbers, or allowed deviations from the median. The accompanying program `trimplot` provides graphical display of results. In fact, you might prefer to look at a graph from `trimplot` first. (For the syntax of both commands, see sections 7 and 8.)

5 Applications to real data

5.1 Short (1763) revisited

For a first example, we will revisit Short's (1763) measurements of parallax. His data are provided with the media for this issue in `short.dta`. The variables are `parallax`, measured in decimal seconds of a degree, and `page`, meaning page in the original article. (As already mentioned in section 3.1, there were two datasets on page 325.)

For a first look at data, I often turn to quantile plots as capable of showing both broad features and any unusual details and specifically to the program `qplot`. See Cox (1999, 2005) for general discussion and many references and the files associated with Cox (2012a) for code download. (A `search` for `qplot` in an up-to-date Stata may reveal a later version, depending on when you read this.)

```
. use short
. qplot parallax, by(page)
```

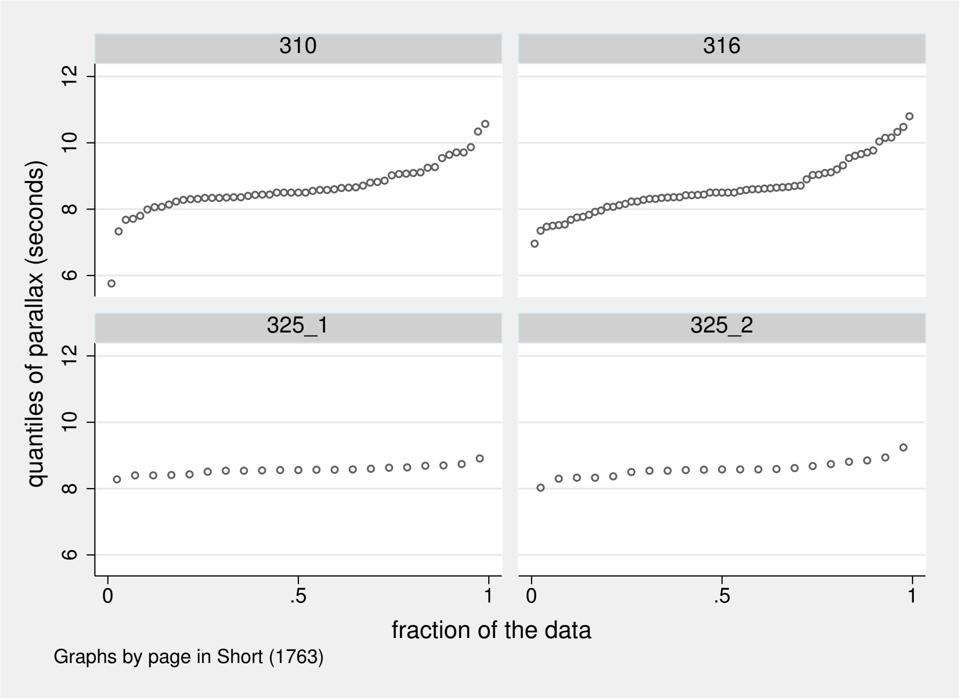

Figure 1. Quantile plots for subsets of parallax measurements in Short (1763)

Given four subsets, there is a choice between showing them separately (juxtaposed) with the `by()` option of `qplot` and showing them together (superposed) with its `over()` option. Both can be useful. We see immediately from figure 1 that the two smaller subsets are distinctly less variable than the two larger ones. No wild outliers are apparent in each case.

`trimplot` shows all possible trimmed means. Specifying `percent` as an option is useful for comparing subsets of differing sizes. Instead of using the `by()` option, we can use the `over` option to show results in a single panel. As with `qplot`, both choices are allowed. There is enough space to put the legend inside the plot region in the top right-hand corner. See figure 2.

```
. trimplot parallax, over(page) percent legend(position(1) ring(0) cols(1))
```

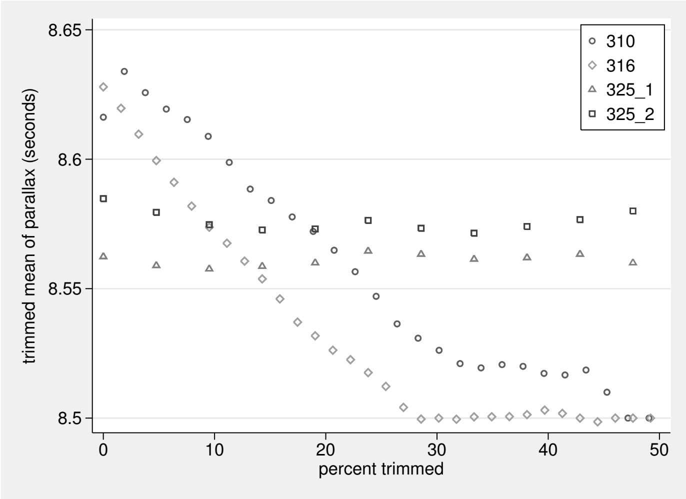

Figure 2. All possible trimmed means for subsets of parallax measurements in Short (1763)

The results show a simple contrast. Trimmed means are stable for the two smaller subsets but drift fairly systematically with trimming proportion for the two larger subsets. The graph quantifies agreement (all results shown are between $8.50''$ and $8.63''$ to 2 decimal places) and also disagreement (there appears to be some systematic difference between the two pairs). Were this a live issue, the results could now be taken back to the astronomical community for reflection and further analysis.

Graphs such as figure 2 show percent trimmed on the x axis. This encourages thinking in terms of what happens as we trim more and more. The opposite interpretation is also possible: what happens as we use more and more of the values in the data? That interpretation would be made easier by reversing the axis, which is just a standard `twoway` option call, `xsc(reverse)`.

To see the numbers, we can fire up `trimmean`. If desired, `trimmean` will show all possible trimmed means through its `number()` option, but a display of values for percents `0(5)50` appears to be adequate detail for many problems. The tabulation is suppressed here to save space.

```
. by page, sort: trimmean parallax, percent(0(5)50) format(%3.2f)
```
(*output omitted*)

A final flourish with Short's (1763) data is to get closer to what Short (1763) actually did and trim metrically by using an idea of maximum allowed deviation. As mentioned already, Short (1763) worked with deviation from the overall mean, whereas we find the idea of working with deviation from the overall median much more appealing. Crucially, means can be pulled way off by any dubious values, and focusing only on values close to the mean may not be enough protection.

Here again a question arises about axis direction. To be consistent with the previous plot, in which trimming amount increases left to right, we reverse this plot's x axis scale. Partly to underline what is possible and partly to use the available space more fully, we use separate scales for each subset. See figure 3.

```
. trimplot para, by(page, xrescale) metric xsc(reverse)
```

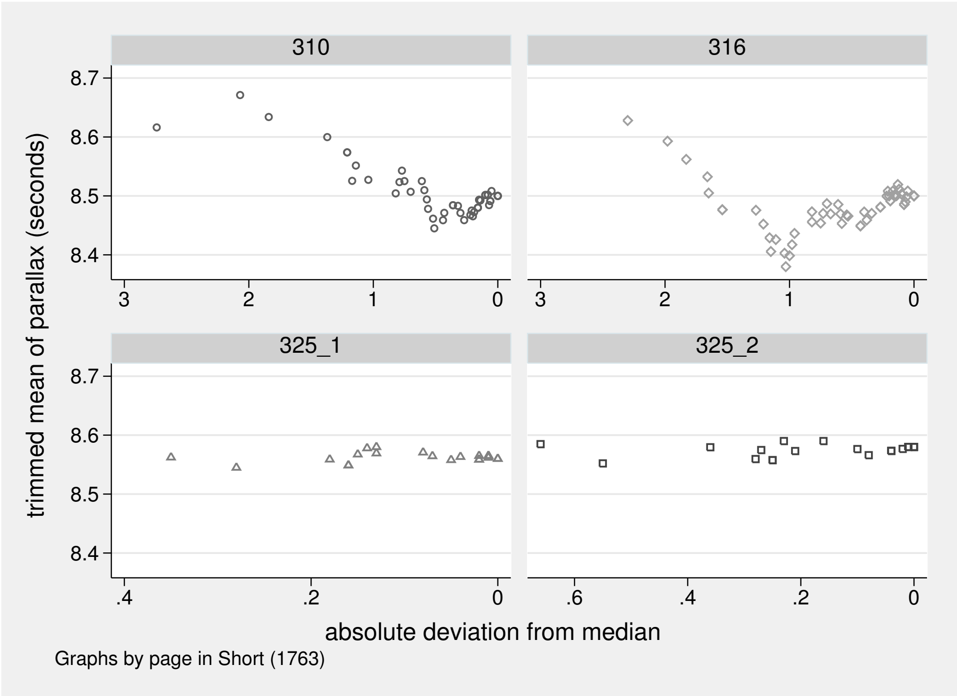

Figure 3. All possible metrically trimmed means for subsets of parallax measurements in Short (1763)

5.2 Chapman data

For a second example, we turn to a dataset on 200 men from Dixon and Massey (1983, 17–20) called the Chapman heart study data. (Details about Chapman were not recorded there.) This appears to be of fairly high quality. We pick four variables with quite different characters and units of measurement. For a basic view, a multiple quantile plot is obtained with `multqplot` (Cox 2012b). See figure 4.

```
. use chapman
. multqplot diastolic cholesterol height weight
```

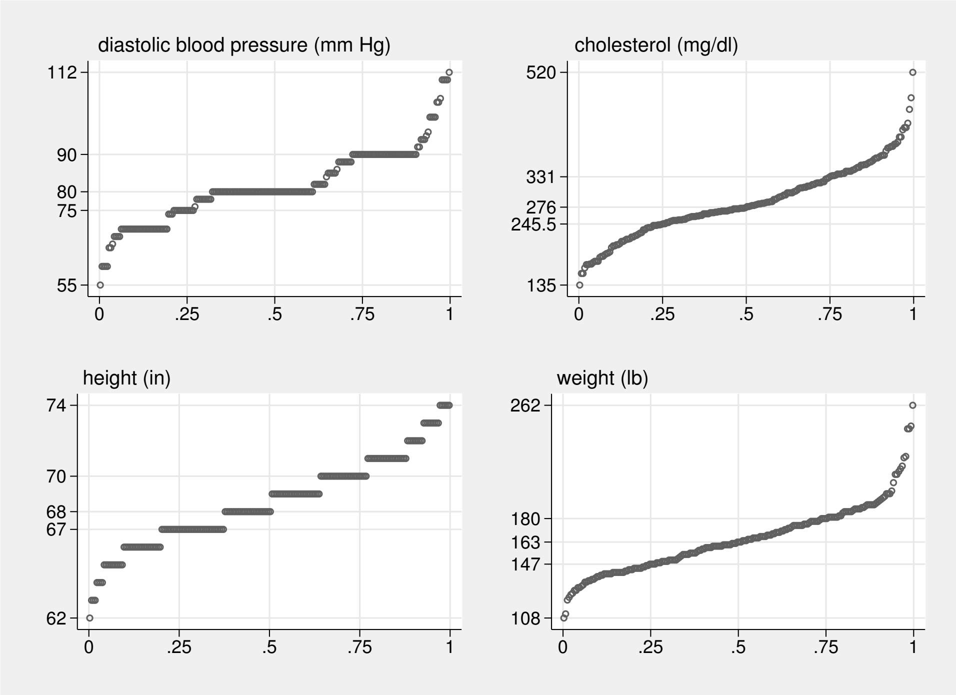

Figure 4. Quantile plots for diastolic blood pressure, cholesterol, height, and weight for 200 men in the Chapman heart study data

By default in this graph, 5 quantiles are labeled on the y axis: minimum, lower quartile, median, upper quartile, and maximum, or the 0, 25, 50, 75, and 100% points in a distribution. These are precisely the quantiles explicit in the most common kind of box plots. The difference from box plots is that quantile plots show all distinct values. They will necessarily, but unproblematically, blur into each other to some extent in many large datasets. Conversely, it is almost a definition that outliers will remain distinct.

Three broad features stand out from the multiple display, not all of which would be strongly evident from corresponding box plots or histograms. First, all variables are approximately symmetric to mildly right skewed. Second, there are no marked outliers. Third, diastolic blood pressure and height stand out as granular in detail with several ties at rounded values. The motive of the measures of avoiding spurious precision may easily be guessed, but the granularity is still strong. Diastolic blood pressure is often reported as 70, 80, or 90 mm Hg. Height is measured to the nearest inch. (Note: 1 inch = 25.4 mm.)

Given the diversity of values and units of measurements, separate `trimplots` are advisable. We loop over the four variables and then use `graph combine`. The result is shown in figure 5.

```
. foreach v in diastolic cholesterol height weight {
  2.          trimplot `v´, percent name(`v´)
> xlabel(0(5)50) xline(0 25 50, lcolor(gs10) lwidth(medium))
  3. }
. graph combine diastolic cholesterol height weight
```

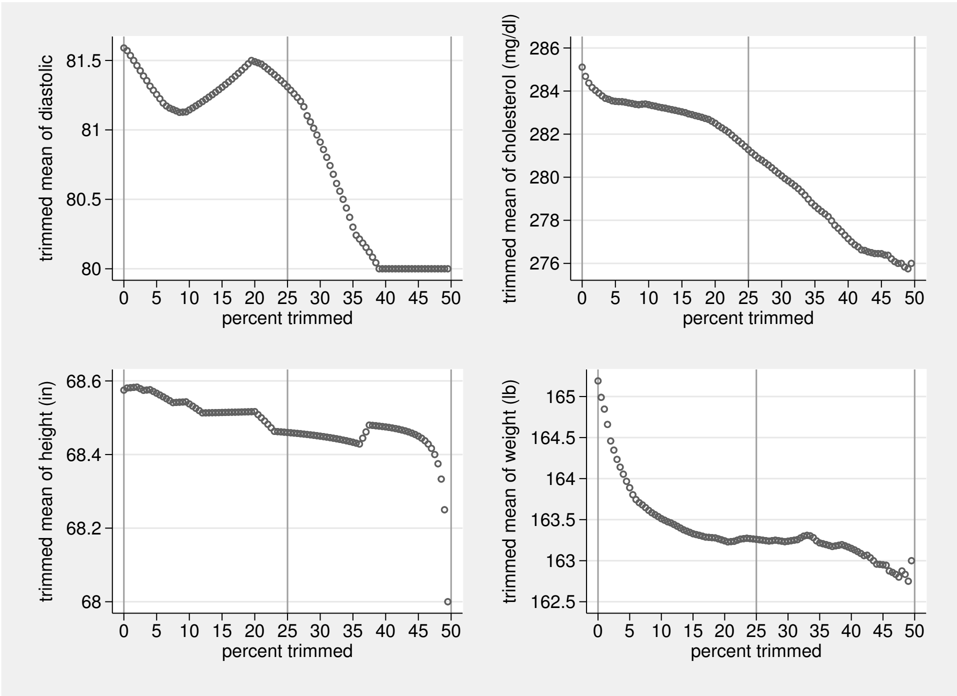

Figure 5. All possible trimmed means for diastolic blood pressure, cholesterol, height, and weight for 200 men in the Chapman heart study data

Grid lines have been added at 0, 25, and 50% trimming to flag the positions of the mean, midmean, and median. In detail, some linear segments of the traces for `diastolic` and `height` are evidently side effects or artifacts of the granularity of each distribution. Relatively smooth, approximately flat, or monotonic patterns are expectable when a distribution is symmetric or skewed and when there are many differing values.

For description or exploration, the main idea is that a plot of trimmed means could help underline that different summaries are close enough that one will be adequate, or that different summaries need to be quoted together, or indeed that the variable should be transformed. The midmean appears unjustly neglected as a simple descriptor. It has the merit that quartiles are widely familiar to users of statistics, so the idea can easily be explained. Having a memorable and evocative name also helps, even if it is pure superstition to suppose that a name is anything but a label.

For inference, how do confidence interval procedures work? As an experiment, I calculated 95% confidence intervals for the midmean using the Tukey–McLaughlin method and three standard `bootstrap` procedures. The script is too long to reproduce here but is available with the media for this issue as `trimming_ci.do`. I used 10,000 replications for bootstrapping. The random number seed is explicit in the do-file. Figure 6 shows good broad agreement, or for the skeptical or cynical, a reminder that different procedures usually give at least slightly different results.

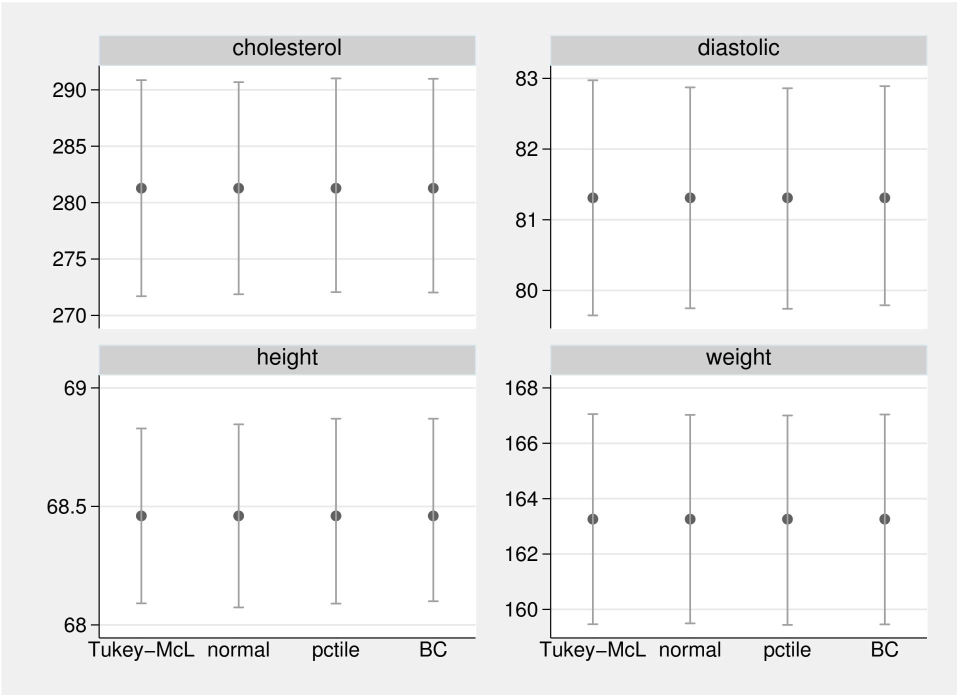

Figure 6. 95% confidence intervals for midmeans of diastolic blood pressure, cholesterol, height, and weight for the Chapman data, obtained by the Tukey–McLaughlin method and three bootstrap procedures

6 A note on other Stata implementations

The user-written program `iqr` by Hamilton (1991) calculates the 10% trimmed mean (only) as a sideline to other aims. His definition is the mean of values greater than the

10% percentile and less than the 90% percentile as calculated by [R] **summarize**, so results may often differ at least slightly from those calculated by `trimmean`.

The user-written program `robmean` by Ender (2009) calculates trimmed means according to the fraction trimmed (this is equivalent to the default of `trimmean` with the `percent()` option), together with some other quantities.

7 The trimmean command

7.1 Syntax

`trimmean` *varname* [*if*] [*in*] ,
{`percent(`*numlist*`)` | `number(`*numlist*`)` | `metric(`*numlist*`)`} [`ceiling` `weighted` `ci` `level(`#`)` `format(`*format*`)` `generate(`*newvar*`)`]

`by ...` : may also be used with `trimmean`; see help on `by`.

7.2 Description

`trimmean` calculates symmetric trimmed means as descriptive or inferential statistics for *varname*.

7.3 Options

`percent(`*numlist*`)` specifies percents of trimming for one or more trimmed means. Percents must be integers between 0 and 50 but otherwise can be specified as a *numlist*. Precisely one of `percent()`, `number()`, or `metric()` is required.

`number(`*numlist*`)` specifies numbers of values to be trimmed for one or more trimmed means. Numbers must be zero or positive integers less than half the number of observations available but otherwise can be specified as a *numlist*. Precisely one of `percent()`, `number()`, or `metric()` is required.

`metric(`*numlist*`)` specifies trimming such that means are of values within a specified absolute deviation of the median of a variable, say, y. Suppose `metric(0 100 200)` is specified. Then the means are means of values satisfying $|y - \text{med}(y)| \leq 0, 100, 200$. Deviations must be zero or positive values but otherwise can be specified as a *numlist*. Precisely one of `percent()`, `number()`, or `metric()` is required.

`ceiling` specifies the use of `ceil()` rather than `floor()` in the calculation of ranks to be included. It is allowed with `number()` or `metric()` but ignored as irrelevant. This variation is occasionally suggested in the literature (for example, Huber in Andrews et al. [1972, 254]).

`weighted` implements a weighted variant explained in detail in the *Definitions* section in the help. It is allowed with `number()` or `metric` but ignored as irrelevant. This option may not be combined with `ci`.

`ci` specifies production of confidence intervals. This option may not be combined with `weighted` or `metric()`. For detailed discussion, see the *Definitions* section in the help.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **20.7 Specifying the width of confidence intervals**.

`format(`*format*`)` specifies a numeric format for displaying trimmed means (and confidence limits when requested). The default is the display format of *varname*.

`generate(`*newvar*`)` specifies that an indicator (a.k.a. dummy) variable be generated with value 1 if an observation was included in the last trimmed mean calculated and 0 otherwise. The trimmed mean with the highest trimming percent or number or allowed deviation is always produced last, regardless of user input.

7.4 Stored results

`trimmean` stores the following in `r()`:

Scalars

`r(tmean#)`	each trimmed mean for percent or number # (for example, `r(tmean5)` for 5%) (with the `metric()` option, labeling is 1 upward, not with deviations specified)

Matrices

`r(results)`	Stata matrix with columns as percents or numbers, number averaged, and trimmed means (and confidence limits when requested)

8 The trimplot command

8.1 Syntax

`trimplot` *varname* [*if*] [*in*] [`,` {`over(`*overvar*`)`|`by(`*byvar* [`,` *by_subopts*]`)`}
`percent metric mad` *scatter_options*]

`trimplot` *varlist* [*if*] [*in*] [`, percent metric mad` *scatter_options*]

8.2 Description

`trimplot` produces plots of trimmed means versus depth or percent trimmed or deviation for one or more numeric variables. Such plots may help specifically in choosing or assessing measures of level and generally in assessing the symmetry or skewness of distributions. They can be used to compare distributions or to assess whether transformations are necessary or effective.

`trimplot` may be used to show trimmed means for one variable, in which case different groups may be distinguished by the `over()` or the `by()` option, or for several variables.

8.3 Options

`over(`*overvar*`)` specifies that calculations be carried out separately for each group defined by *overvar* but plotted in the same panel. `over()` is allowed only with a single variable to be plotted. `over()` and `by()` may not be combined.

`by(`*byvar* [`,` *by_subopts*]`)` specifies that calculations be carried out separately for each group defined by *byvar* and plotted in separate panels. Suboptions may be specified to tune the graphical display; see help on *by_option*. `by()` is allowed only with a single variable to be plotted. `over()` and `by()` may not be combined.

`percent` specifies that depth be scaled and plotted as percent trimmed, which will range from 0 to nearly 50 (a median cannot be based on no observed values, so 50 cannot be attained).

`metric` specifies that trimmed means be defined and plotted in terms of allowed absolute deviation from the median.

`mad` specifies metric trimming as above, but values will be plotted versus absolute deviation from the median / median absolute deviation from the median. The median (absolute) deviation (from the median) can be traced to Gauss (1816).

scatter_options are options of `twoway scatter`.

9 Conclusions

All careful users of statistics worry about how to handle awkward data. One well-documented kind of awkwardness consists of outliers and tails heavier than normal (Gaussian), a source of worry if techniques being used work best whenever some distribution is normal. Even if the modeling assumption is only that some conditional distribution be normal, marked nonnormality can still be awkward in various senses, not least in terms of how best to summarize and report such distributions.

Skepticism about the normal or Gaussian is often presented as a recent phenomenon, but it has deeper roots. Poincaré (1896, 149; 1912, 170–171) quoted a remark by Lippmann (Gabriel Lippmann 1845–1921, Nobel Prize for Physics 1908). This remark has often been misquoted or loosely translated, so here is a close translation from Mazliak (2012, 187): "[This distribution] cannot be obtained by rigorous deductions; many a proof one had wanted to give it is rough, among others the one based on the statement that the probability of the gaps is proportional to the gaps. Everyone believes it, however, as M. Lippmann told me one day, because the experimenters imagine it is a mathematical theorem, and the mathematicians that it is an experimental fact".

The field of robust statistics offers various solutions in this area. Trimmed means are one of the oldest and simplest methods for summarizing sample evidence on location as robustly as you wish. Why then are they not more frequently used? Speculation is easy. Perhaps they fall uncomfortably between teaching material and the research literature: they are too much of a complication or distraction to be included in many courses or texts and too simple or too well known to receive sustained focus in monographs on robust statistics. The larger problem for the latter is, naturally, how to model robustly the relationships between outcomes and predictors. The ever-elusive goal of a robust regression that is easy to understand, always reliable, and suitably fast seems likely to drive research for the indefinite future.

Whatever the precise diagnosis, I have focused here on providing constructive answers by way of a tutorial review and usable programs for trimmed means. The most distinctive emphases are including confidence interval procedures and emphasizing the scope for plotting results. Confidence intervals are often not mentioned in introductory accounts. The importance of plotting both raw data and results for trimmed means is also often understated. A more specific suggestion is that the midmean, the mean of the middle half of the data, seems unfairly neglected.

10 Acknowledgments

Rebecca Pope and Ariel Linden gave useful comments on earlier versions of `trimmean` and `trimplot`. David Hoaglin was very helpful in identifying citations in and around Tukey's work.

11 References

Aldrich, J. 2007. "But you have to remember P. J. Daniell of Sheffield". *Electronic Journal for History of Probability and Statistics* 3(2): 1–58.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. 1972. *Robust Estimates of Location: Survey and Advances.* Princeton, NJ: Princeton University Press.

Anonymous. 1821. Probabilités. Dissertation sur la recherche du milieu le plus probable, entre les résultats de plusieurs observations ou expériences. *Annales de Mathématiques pures et appliquées* 12: 181–204.
http://www.numdam.org/item?id=AMPA_1821-1822__12__181_0.

Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data.* 3rd ed. Chichester, UK: Wiley.

Bickel, P. J. 1965. On some robust estimates of location. *Annals of Mathematical Statistics* 36: 847–858.

Bickel, P. J., and E. L. Lehmann. 1975. Descriptive statistics for nonparametric models II. Location. *Annals of Statistics* 3: 1045–1069.

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

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.

———. 2005. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460.

———. 2012a. Software update: gr42_6: Quantile plots, generalized. *Stata Journal* 12: 167.

———. 2012b. Speaking Stata: Axis practice, or what goes where on a graph. *Stata Journal* 12: 549–561.

Daniell, P. J. 1920. Observations weighted according to order. *American Journal of Mathematics* 42: 222–236.

David, H. A., and H. N. Nagaraja. 2003. *Order Statistics*. 3rd ed. Hoboken, NJ: Wiley.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Dixon, W. J. 1960. Simplified estimation from censored normal samples. *Annals of Mathematical Statistics* 31: 385–391.

Dixon, W. J., and F. J. Massey, Jr. 1983. *Introduction to Statistical Analysis*. 4th ed. New York: McGraw–Hill.

Dixon, W. J., and J. W. Tukey. 1968. Approximate behavior of the distribution of Winsorized t (trimming/Winsorization 2). *Technometrics* 10: 83–98.

Dixon, W. J., and K. K. Yuen. 1974. Trimming and Winsorization: A review. *Statistische Hefte* 15: 157–170.

Donoho, D. L., and P. J. Huber. 1983. The notion of breakdown point. In *A Festschrift For Erich L. Lehmann*, ed. P. J. Bickel, K. Doksum, and J. L. Hodges, Jr., 157–184. Belmont, CA: Wadsworth.

Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Ender, P. B. 2009. robmean: Trimmed, Winsorized means & Huber 1-step estimator. UCLA: Statistical Consulting Group.
http://www.ats.ucla.edu/stat/stata/ado/analysis.

Erickson, B. H., and T. A. Nosanchuk. 1977. *Understanding Data*. Toronto, Canada: McGraw–Hill.

———. 1992. *Understanding Data*. 2nd ed. Toronto, Canada: University of Toronto Press.

Feigelson, E. D., and G. J. Babu. 2012. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge: Cambridge University Press.

Flournoy, N. 1993. A conversation with Wilfrid J. Dixon. *Statistical Science* 8: 458–477.

———. 2010. Wilfrid Joseph Dixon, 1915–2008. *Journal of the Royal Statistical Society, Series A* 173: 455–457.

Gauss, C. F. 1816. Bestimmung der Genauigkeit der Beobachtungen. *Zeitschrift für Astronomie und verwandte Wissenschaften* 1: 187–197.

Gordin, M. D. 2004. *A Well-Ordered Thing: Dmitrii Mendeleev and the Shadow of the Periodic Table*. New York: Basic Books.

Hamilton, L. C. 1991. sed4: Resistant normality check and outlier identification. *Stata Technical Bulletin* 3: 15–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 1, pp. 86–90. College Station, TX: Stata Press.

Hampel, F. R. 1985. The breakdown points of the mean combined with some rejection rules. *Technometrics* 27: 95–107.

———. 1997. Some additional notes on the "Princeton robustness year". In *The Practice of Data Analysis: Essays in Honor of John W. Tukey*, ed. D. R. Brillinger, L. T. Fernholz, and S. Morgenthaler, 133–153. Princeton, NJ: Princeton University Press.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Harter, H. L. 1974a. The method of least squares and some alternatives—Part I. *International Statistical Review* 42: 147–174.

———. 1974b. The method of least squares and some alternatives—Part II. *International Statistical Review* 42: 235–264 and 282.

Helsel, D. R., and R. M. Hirsch. 1992. *Statistical Methods in Water Resources*. Amsterdam: Elsevier. http://pubs.usgs.gov/twri/twri4a3/.

Hill, M., and W. J. Dixon. 1982. Robustness in real life: A study of clinical laboratory data. *Biometrics* 38: 377–396.

Huber, P. J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35: 73–101.

———. 1972. The 1972 Wald lecture: Robust statistics: A review. *Annals of Mathematical Statistics* 43: 1041–1067.

———. 1981. *Robust Statistics*. New York: Wiley.

Huber, P. J., and E. M. Ronchetti. 2009. *Robust Statistics*. 2nd ed. Hoboken, NJ: Wiley.

Jennrich, R. I. 2007. BMDP and some statistical computing history. *Statistical Computing and Graphics* 18: 17–23.

Jurečková, J., and J. Picek. 2006. *Robust Statistical Methods with R*. Boca Raton, FL: Chapman & Hall/CRC.

Kafadar, K. 2003. John Tukey and robustness. *Statistical Science* 18: 319–331.

Kim, S.-J. 1992. The metrically trimmed mean as a robust estimator of location. *Annals of Statistics* 20: 1534–1547.

Kleiner, B., and T. E. Graedel. 1980. Exploratory data analysis in the geophysical sciences. *Reviews of Geophysics and Space Physics* 18: 699–717.

Lehmann, E. L. 1983. *Theory of Point Estimation*. New York: Wiley.

Mazliak, L. 2012. Poincaré's odds. *Séminaire Poincaré* 16: 173–210.

Miller, R. G. 1986. *Beyond ANOVA: Basics of Applied Statistics*. New York: Wiley. Reprinted, London: Chapman & Hall, 1997.

Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison–Wesley.

Olive, D. J. 2001. High breakdown analogs of the trimmed mean. *Statistics & Probability Letters* 51: 87–92.

Pearson, R. K. 2011. *Exploring Data in Engineering, the Sciences, and Medicine*. New York: Oxford University Press.

Poincaré, H. 1896. *Calcul des probabilités*. Paris: Gauthier-Villars.

———. 1912. *Calcul des probabilités*. 2nd ed. Paris: Gauthier-Villars. http://archive.org/details/calculdeprobabil00poinrich.

Reimann, C., P. Filzmoser, R. Garrett, and R. Dutter. 2008. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. Chichester, UK: Wiley.

Rice, J. A. 2007. *Mathematical Statistics and Data Analysis*. 3rd ed. Belmont, CA: Duxbury.

Rosenberger, J. L., and M. Gasko. 1983. Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis*, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, 297–338. New York: Wiley.

Short, J. 1763. Second paper concerning the parallax of the sun determined from the observations of the late transit of Venus, in which this subject is treated of more at length, and the quantity of the parallax more fully ascertained. *Philosophical Transactions of the Royal Society of London* 53: 300–345.

Siegel, A. F. 1988. *Statistics and Data Analysis: An Introduction*. New York: Wiley.

Sprent, P., and N. C. Smeeton. 2007. *Applied Nonparametric Statistical Methods*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

Staudte, R. G., and S. J. Sheather. 1990. *Robust Estimation and Testing*. New York: Wiley.

Stewart, C. A. 1947. P. J. Daniell. *Journal of the London Mathematical Society* 22: 75–80.

Stigler, S. M. 1973. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association* 68: 872–879.

———. 1976. The anonymous Professor Gergonne. *Historia Mathematica* 3: 71–74.

———. 1977. Do robust estimators work with real data? *Annals of Statistics* 5: 1055–1098.

———. 2010. The changing history of robustness. *American Statistician* 64: 277–281.

Tilanus, C. B., and G. Rey. 1964. Input-output volume and value predictions for the Netherlands, 1948–1958. *International Economic Review* 5: 34–45.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, ed. I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, and H. Mann, 448–485. Stanford, CA: Stanford University Press.

———. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33: 1–67.

———. 1970a. *Exploratory Data Analysis*. Limited preliminary ed. Reading, MA: Addison–Wesley.

———. 1970b. Some further inputs. In *Geostatistics: A Colloquium*, ed. D. F. Merriam, 163–174. New York: Plenum.

———. 1977. *Exploratory Data Analysis*. Reading, MA: Addison–Wesley.

———. 1986. Choosing techniques for the analysis of data. In *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis: 1965–1986, Volume IV*, ed. L. V. Jones, 869–874. Monterey, CA: Wadsworth and Brooks/Cole.

Tukey, J. W., and D. H. McLaughlin. 1963. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhyā* 25: 331–352.

Turner, G. L'E. 1969. James Short, F.R.S., and his contribution to the construction of reflecting telescopes. *Notes and Records of the Royal Society of London* 24: 91–108.

van Belle, G., L. D. Fisher, P. J. Heagerty, and T. Lumley. 2004. *Biostatistics: A Methodology For the Health Sciences.* 2nd ed. Hoboken, NJ: Wiley.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* 4th ed. New York: Springer.

Wilcox, R. R. 2003. *Applying Contemporary Statistical Techniques.* San Diego: Academic Press.

———. 2009. *Basic Statistics: Understanding Conventional Methods and Modern Insights.* New York: Oxford University Press.

———. 2010. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy.* 2nd ed. New York: Springer.

———. 2012a. *Introduction to Robust Estimation and Hypothesis Testing.* 3rd ed. Waltham, MA: Elsevier.

———. 2012b. *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction.* Boca Raton, FL: CRC Press.

Woolf, H. 1959. *The Transits of Venus: A Study of Eighteenth-century Science.* Princeton, NJ: Princeton University Press.

Wulf, A. 2012. *Chasing Venus: The Race to Measure the Heavens.* New York: Alfred A. Knopf.

12 Vignettes

Percy John Daniell (1889–1946) was born to British parents in Valparaiso, Chile. He was the last publicly declared Senior Wrangler in Mathematics (top student in his year) at Cambridge in 1909. After brief periods in Liverpool and Göttingen, he taught and researched from 1912 to 1923 at the Rice Institute at Houston, Texas, before returning to Britain as Professor of Mathematics at Sheffield. Daniell's contributions, which span a wide range from pure mathematics to applied mathematics and statistics, were surveyed briefly by Stewart (1947) and Stigler (1973) and in much more detail by Aldrich (2007).

Wilfrid Joseph Dixon (1915–2008) was born in Portland, Oregon. He received degrees in mathematics and statistics from Oregon State College, the University of Wisconsin, and Princeton. He was on the faculty at the University of Oklahoma, the University of Oregon, and University of California, Los Angeles, where he was a leader in biostatistics and biomathematics. Dixon's statistical interests were wide ranging, including robust estimation in the presence of outliers, and he collaborated with medical scientists on many projects. With Frank Jones Massey, Jr. (1919–1995), he wrote a major statistics text for nonmathematicians, which was unusual in including material on trimming and Winsorizing in its third and fourth editions (1969, 1983). Beginning in 1961, he led the development of the package that has morphed over its history from BIMED to BIMD to BMD to BMDP. See Flournoy (1993, 2010) and Jennrich (2007) for more details.

Donald Hatch McLaughlin (1941–) earned degrees in mathematics and psychology from Princeton, the University of Pennsylvania, and Carnegie Mellon and taught psychology at Berkeley for six years. Since 1973, he has worked for the American Institutes for Research in Palo Alto, California, and independently as a senior researcher and consultant on many applied projects in education and several other areas.

Dmitrii Ivanovich Mendeleev (1834–1907) was born near Tobolsk in Siberia. He studied and researched in chemistry in St. Petersburg and Heidelberg, quickly rising to professorial rank and establishing St. Petersburg as a major center in chemical research. Mendeleev is best known for his work developing a periodic table of the elements, distinguished not only for providing a classification but also for allowing the prediction of other elements and correcting errors in the measurement of atomic weights. He was, however, much more than an outstanding chemist: "The same individual who composed the periodic system also helped design the highly protectionist Russian tariff of 1891, battled local Spiritualists, created a smokeless gunpowder, attempted Arctic exploration, consulted on oil development in Baku, investigated iron and coal deposits, published art criticism, flew in balloons, introduced the metric system, and much more" (Gordin 2004, xviii). Numerous different transliterations of his name exist.

James Short (1710–1768) was born in Edinburgh and first educated to become a minister, but with inspiration and support from Colin MacLaurin, he became more interested in mathematics and optics and specifically the construction of telescopes. He used metallic specula and succeeded in giving them true parabolic and elliptic shapes. Short adopted telescope-making as his profession, practicing with great success in Edinburgh and then London. He was elected Fellow of the Royal Society and published many of his observations, including his calculation of solar parallax from the 1761 transit of Venus. See also Turner (1969). Note that Short was Scottish, not English as stated by Stigler (1973, 873).

John Wilder Tukey (1915–2000) was born in New Bedford, Massachusetts. He studied chemistry at Brown and mathematics at Princeton and afterward worked at both Princeton and Bell Labs. He was also involved in a great many government projects, consultancies, and committees. He made outstanding contributions to several areas of statistics, including time series, multiple comparisons, robust statistics, and exploratory data analysis. Tukey was extraordinarily energetic and inventive, not least in his use of terminology: he has been credited with inventing the terms "bit", "analysis of variance", "box plot", "data analysis", "hat matrix", "jackknife", "stem-and-leaf plot", "trimming", and "Winsorizing", among many others. He was awarded the U.S. National Medal of Science in 1973. Tukey's direct and indirect influence marks him as one of the greatest statisticians of all time.

Charles P. Winsor (1895–1951) was educated at Harvard as an engineer and then worked for the New England Telephone and Telegraph Company, but his interests shifted to biological research and biostatistics. After further study at Johns Hopkins and Harvard, he held posts at Iowa State College and Johns Hopkins; in between, in the Second World War, he did government work at Princeton. The term "Winsorize" has been attributed to J. W. Tukey but was first used in publications by Dixon (1960).

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*.