The Stata Journal (2021)
**21**, Number 3, pp. 838–846 DOI: 10.1177/1536867X211045583

# Stata tip 141: Adding marginal spike histograms to quantile and cumulative distribution plots

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

## 1 Marginal spike histograms

Some intriguing examples given by Harrell (2015) prompt experiment with adding marginal spike histograms to quantile and (empirical) cumulative distribution (function) plots. This tip explains why this might be helpful and gives sample code using the official command `quantile`.

The point is at least twofold.

- *Pedagogic.* Such additions help clarify what (for example) a quantile plot means and how it can be explained as stacking values in terms of their associated cumulative probabilities. Every kind of graph becomes comfortable only with increased familiarity. I have encountered students and colleagues evidently long past their first histogram who struggle a little on their first experience with quantile plots. A little help does no harm.

- *Practical.* In principle, marginal histograms add no more information to a quantile plot. In practice, they offer a complementary view of each distribution, affording another way to think about distribution level, spread, and shape—and indeed fine structure too.

Discussion here is phrased entirely in terms of official Stata commands, with the `quantile` command (see [R] **Diagnostic plots**) the main workhorse, but the idea extends easily to community-contributed (user-written) commands such as `qplot` or `distplot` from the *Stata Journal* (Cox [1999a, 1999b, 2004, 2005]; use `search` to find recent updates).

As usual, the idea has longer roots. See Galton (1889, 38) for (in modern terms) a quantile plot and a histogram sharing the same vertical magnitude axis.

## 2 Enhancing quantile plots

The name "quantile plot" goes back at least to Wilk and Gnanadesikan (1968), but the device is much older, being used several times in the 19th century and early 20th century. The main idea is just to plot ordered values, often but not invariably on the $y$ axis, against their associated cumulative probabilities on the other axis. More generally,

 

ordered values are plotted against corresponding quantiles of some relevant distribution. Small print aside, associated cumulative probabilities are just quantiles of a uniform (rectangular, flat) distribution on the interval from 0 to 1. Cumulative distribution plots are—again, often but not invariably—plotted with those axes reversed.

There are many small variations in both terminology and format. Quantile–quantile plot and empirical cumulative distribution function plot are some other terms. As if in apology or compensation for such long-windedness, concise but more cryptic labels such as q–q plots and ECDF plots can be found. As a matter of curious convention rather than compelling logic, quantile plots commonly use markers or point symbols for each distinct value, while cumulative distribution plots commonly use connected lines. That is a distinction without a difference if any distribution is such that points merge into lines, as may well be true with large sample sizes.

Let's start with a simple quantile plot. The official Stata command `quantile` adds a reference line that allows comparison with a uniform (rectangular, flat) distribution with the same range as the observed data. I almost never want that, so I usually make it invisible by changing its color. I also often vary the marker symbol and vertical axis label from the default. Figure 1 shows such a plot.

```
. sysuse auto
(1978 automobile data)
. set scheme sj
. quantile mpg, rlopts(lc(none)) ms(oh) yla(, ang(h))
```

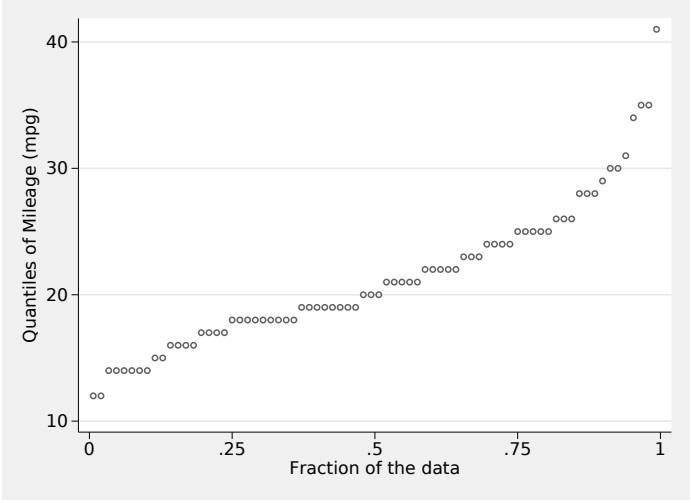


Figure 1. Quantile plot of miles per gallon from the auto data

As implemented in `quantile`, the idea is to plot values in rank order versus cumulative probability, specifically (unique rank − 0.5) / sample size. As explained in, say, Cox (2021), a rule rank / sample size does not treat lower and upper quantiles symmetrically and would cause other problems.

Because one axis is a probability scale, that is compatible in principle with a marginal histogram showing probabilities in each distinct bin. There are various fairly simple ways to get that directly. Here's one.

```
. count if mpg < .
  74
. egen prob = total(1/`r(N)'), by(mpg)
. egen tag = tag(mpg)
```

Counting the sample size explicitly does no harm and is needed if there are missing values (not the case in this example but often true) or if you want to look at a subset of your data (again, not the case here but also often true).

A first attempt just adds spikes to the vertical axis, as in figure 2.

```
. quantile mpg, rlopts(lc(none)) ms(oh) yla(, ang(h))
> addplot(spike prob mpg if tag, horizontal) legend(off)
```

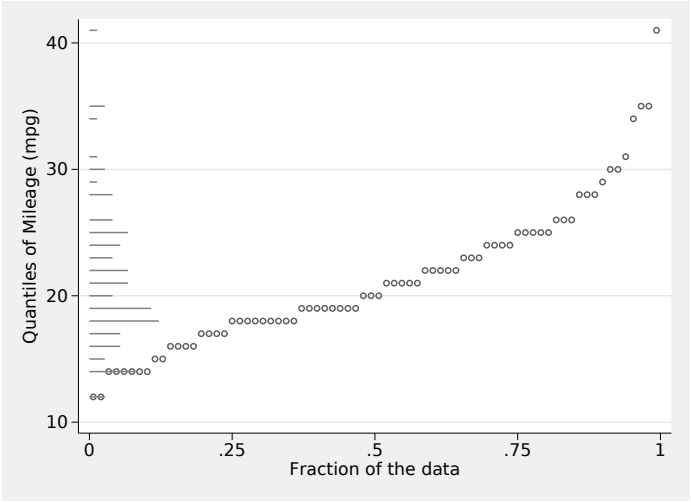


Figure 2. Quantile plot of miles per gallon from the auto data with marginal spike histogram

That is a good start, but we would be better off flipping the histogram so that it lies clear of the quantile plot. Negation is easy enough. See figure 3.

```
. generate nprob = -prob
. quantile mpg, rlopts(lc(none)) ms(oh) yla(, ang(h))
> addplot(spike nprob mpg if tag, horizontal) legend(off)
```

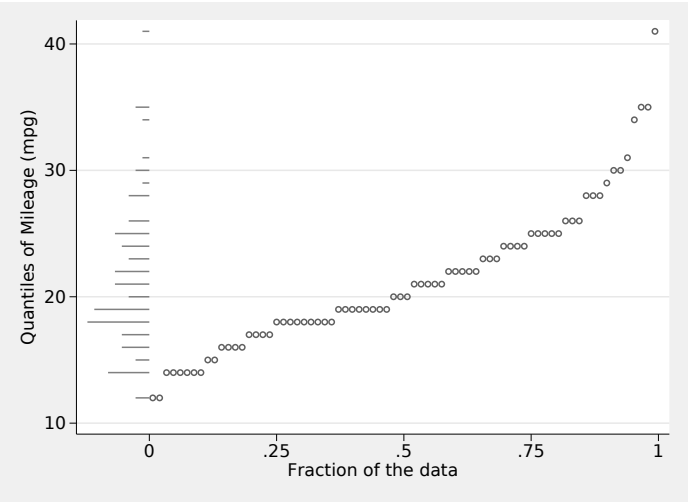


Figure 3. Quantile plot of miles per gallon from the auto data with marginal spike histogram, now flipped over or reflected in the vertical axis

The point about adding `if tag` is partly one of efficiency (the same spikes need not, and should not, be plotted repeatedly) and partly one of avoiding monitor artifacts.

In this example, we were lucky: There are 21 distinct values, so the mean probability in each bin is about 0.05. The maximum probability may naturally be much higher, but here taking probabilities literally (which means numerically) seems to work fine.

In other examples, we might have to do more work and make some decisions to get an extra histogram that is informative yet restrained. Let's look at a larger dataset. Figure 4 shows a distribution of wages on a natural logarithm scale.

```
. webuse nlswork, clear
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)
. egen tag = tag(ln_wage)
. count if ln_wage < .
  28,534
. egen prob = total(1/`r(N)'), by(ln_wage)
. generate nprob = -prob
```

```
. quantile ln_wage, rlopts(lc(none)) ms(oh) legend(off)
> addplot(spike nprob ln_wage if tag, horizontal) yla(, ang(h))
```

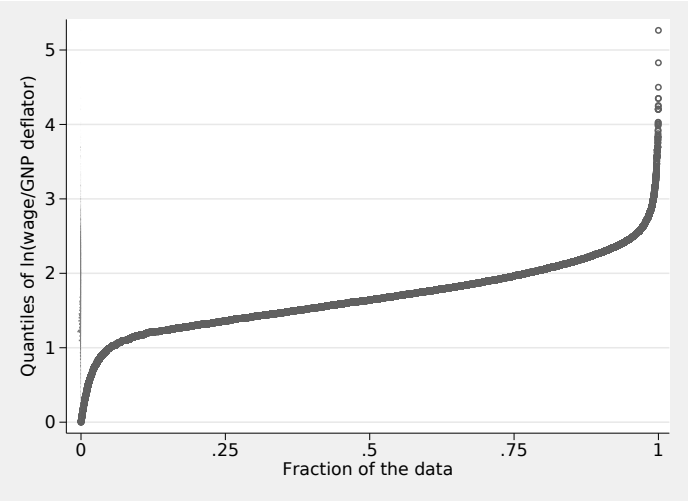


Figure 4. Quantile plot of logarithm of wage from the `nlswork` dataset. The marginal spike histogram is in effect a rug plot given the typically small bin probabilities.

Here there are 8,173 distinct values, and so the mean probability in each bin is about 0.0001. The maximum probability again can be much higher, but here taking probabilities literally (which means numerically) means that the histogram degenerates to a rug plot. If you want a rug plot, you can always get one (for example, Cox [2004]), but that is not the goal here.

There are various ways to move forward. One is to decide how much space to give the histogram and to scale accordingly. Then the probability scale for the quantile plot and that for the quantile plot are different, which needs to be explained somewhere. Figure 5 is an example. Another is to use a square-root scale for probabilities, which often works well to make structure clearer. Figure 6 is an example. The two ideas can be combined.

```
. summarize prob, meanonly
. generate nprob_scaled = nprob * (0.1 / r(max))
. generate prob_root = -sqrt(prob)
. quantile ln_wage, rlopts(lc(none)) ms(oh) legend(off)
> addplot(spike nprob_scaled ln_wage if tag, horizontal) yla(, ang(h))
. quantile ln_wage, rlopts(lc(none)) ms(oh) legend(off)
> addplot(spike prob_root ln_wage if tag, horizontal) yla(, ang(h))
```

Figure 5 scales the histogram to cover 10% of the horizontal extent of the quantile plot.

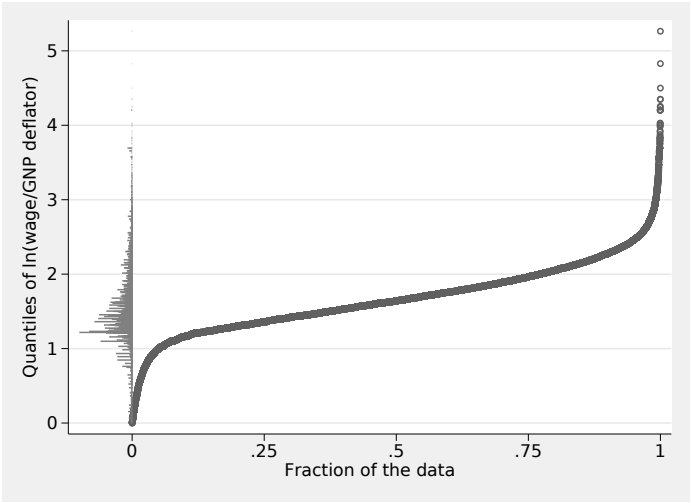


Figure 5. Quantile plot of logarithm of wage from the `nlswork` dataset. The spike histogram shows the same values but with a different probability scale.

Figure 6 shows a square-root scale.

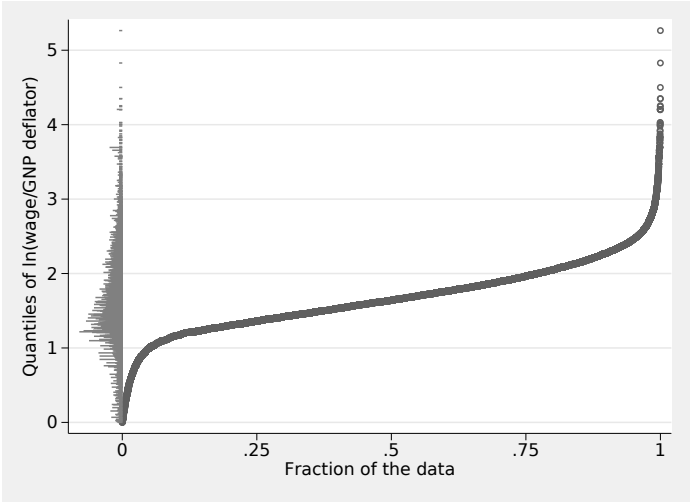


Figure 6. Quantile plot of logarithm of wage from the `nlswork` dataset. The spike histogram shows the same values but with a square-root probability scale.

## 3 Square-root scales for probability or frequency

If the idea of using a square-root scale is unfamiliar, let's spell that out. Square-root scales stretch back at least a century. See discussion and references in Cox (2012) and especially also Perrin (1913, 198; 1916, 131).

Histograms using square-root scales—rootograms!—go back at least to John W. Tukey circa 1965. The idea is to show not bin frequencies but their square roots. Frequencies, as counted variables, tend to have variability that is stabilized by a root transformation, at least approximately. It is now harder to see bins with higher frequency and much easier to see bins with lower frequency. Bins that are empty remain no problem because the square root of zero is zero. Note also that the square root of a normal or Gaussian density is a multiple of another normal or Gaussian density. Hence, if the normal is a reference distribution, we are looking for the same shape on a rootogram, and experience in assessing histograms for approximate normality can be applied directly in assessing rootograms. However, taking the root is only the first step in Tukey's procedure, and we do not implement his hanging or suspended rootograms. See Tukey (1986, 1972, 1977, chap. 17), Tukey and Wilk (1965), or Velleman and Hoaglin (1981).

## 4 A bimodal example

As suggested by Frank Harrell (personal communication), we close with an example that shows bimodal structure. We use the famous iris dataset of Edgar Anderson, as publicized by Fisher (1936) and often used as a sandbox in multivariate statistics and machine learning exercises. Stebbins (1978) gave an appreciation of Anderson, a distinguished and idiosyncratic botanist, and comments on the scientific background to distinguishing three species of the genus *Iris*. Kleinman (2002) surveys Anderson's graphical contributions with statistical flavor.

A Stata-readable copy of the iris data is bundled with the media for this issue. This dataset includes 150 measurements of the width and length of petals and sepals for 50 flowers each of the species *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Using the same recipe as above, figure 7 shows the data for petal length.

The bimodal structure is, I suggest, clearly shown. It is no discovery and easily related to which species is being measured.

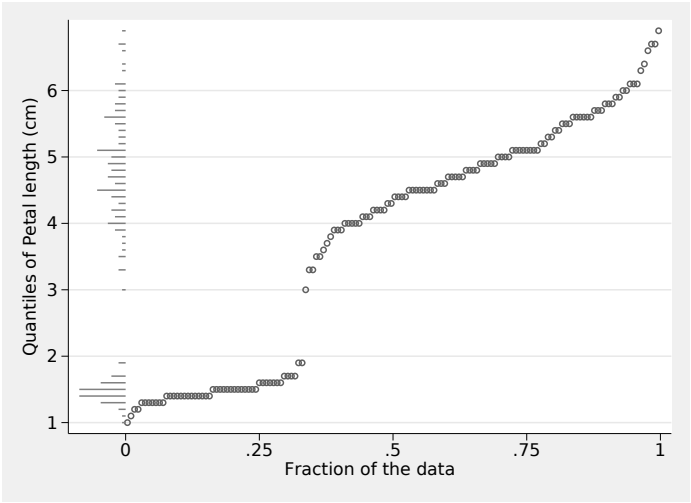


Figure 7. Quantile plot and marginal spike histogram for petal length from the iris data. Note the bimodal structure.

## 5 Graphics choices

In graphics, choices small and large abound, and taste and circumstance should determine your own choices. The device in this tip is for use whenever it is helpful. Other possibilities include offsetting the histogram slightly by introducing some space between it and the display of quantiles, using a marginal dot or strip plot instead, and smoothing the quantiles slightly because sometimes fine structure is just noise.

## 6 Acknowledgments

I thank Frank Harrell and Antony Unwin for encouraging and helpful correspondence.

## References

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

———. 1999b. 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.

———. 2004. Speaking Stata: Graphing distributions. *Stata Journal* 4: 66–88. https://doi.org/10.1177/1536867X0100400106.

———. 2005. Speaking Stata: The protean quantile plot. *Stata Journal* 5: 442–460. https://doi.org/10.1177/1536867X0500500312.

———. 2012. Speaking Stata: Transforming the time axis. *Stata Journal* 12: 332–341. https://doi.org/10.1177/1536867X1201200210.

———. 2021. Speaking Stata: Front-and-back plots to ease spaghetti and paella problems. *Stata Journal* 21: 539–554. https://doi.org/10.1177/1536867X211025838.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Galton, F. 1889. *Natural Inheritance*. London: Macmillan.

Harrell, F. E., Jr. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham, Switzerland: Springer.

Kleinman, K. 2002. How graphical innovations assisted Edgar Anderson's discoveries in evolutionary biology. *Chance* 15(3): 17–21. https://doi.org/10.1080/09332480.2002.10554806.

Perrin, J. 1913. *Les Atomes*. Paris: Félix Alcan.

———. 1916. *Atoms*. New York: Van Nostrand.

Stebbins, G. L. 1978. *Edgar Anderson 1897–1969. Biographical Memoir.* Washington, DC: National Academy of Sciences. http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/anderson-edgar.pdf.

Tukey, J. W. 1972. Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft and S. A. Brown, 293–316. Ames, IA: Iowa State University Press.

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

———. 1986. The future of processes of data analysis. In *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis: 1965–1986, Volume IV*, ed. L. V. Jones, 517–547. Monterey, CA: Wadsworth and Brooks/Cole.

Tukey, J. W., and M. B. Wilk. 1965. Data analysis and statistics: Principles and practice. In *The Collected Works of John W. Tukey: Graphics: 1965–1985, Volume V*, ed. W. S. Cleveland, 23–29. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury.

Wilk, M. B., and R. Gnanadesikan. 1968. Probability plotting methods for the analysis of data. *Biometrika* 55: 1–17. https://doi.org/10.2307/2334448.