Ntigrams: special-purpose histograms that solve some pernicious problems

Tim Erickson
Epistemological Engineering
Oakland, California USA

August 18, 2004

Abstract

This paper describes a species of variable-width histogram suitable for use in K–12 and university settings. In these histograms, called *Ntigrams*, we divide the sample into bins of equal frequency rather than bins of equal width. Thus an Ntigram tends to show relevant features in distributions by giving more detail where the density of high and less detail where density is low.

If you're looking at a distribution of continuous data—for example, the ages of 100 people—it's good to use a display that shows all of the data, for example, a one-dimensional scatter plot: a dot plot. But sometimes, the dot plot is impractical. Maybe there are too many cases for the plot to make sense. Or perhaps you're comparing the distribution to another, and because the sample sizes are different, it's hard to make the comparison.

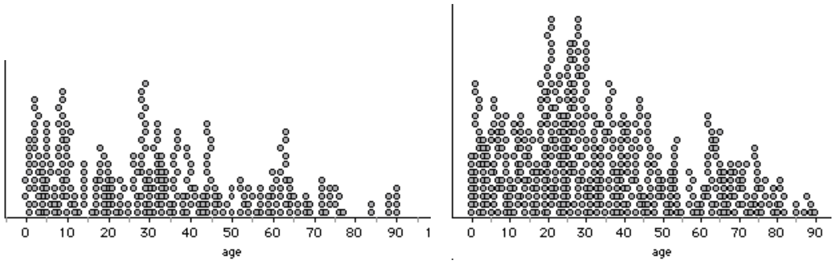

Figure 1: Comparing age distributions. The first graph is a sample of 300 from Miami, Florida. The second is a sample of 500 from New Haven, Connecticut.

Look at Figure 1. Are the distributions very different? It's hard to say. We need to use some other display, one that summarizes the data. We often use histograms or box plots for this purpose, but box plots and traditional histograms have disadvantages. We'll discuss the problems here and propose an alternative we call the "Ntigram" (pronounced *en-ti-gram*). The Ntigram is reminiscent of density-trace plots such as the violin plot (Hintze and Nelson 1998) but has other interesting properties.

1 What's Wrong with Box Plots

Box plots (Tukey 1977) have become very popular in school data curricula starting in the middle years. The box shows the median and the upper and lower quartiles. Depending on the particular species of plot, "whiskers" show the full range, or only part of the range with individual points for the most extreme values. Some box plots show other values as well, such as the mean.

While box plots are useful, especially for comparing more than two groups, and while they fulfill their original mission of being reasonably easy to make by hand, they suffer from an interesting problem: they can confuse students who think that the biggest boxes represent either the most points or the biggest density of points (Bakker 2004; Bakker et al. 2005). They also suggest, by their appearance, that the data in the whiskers might be less important, whereas they are only less central.

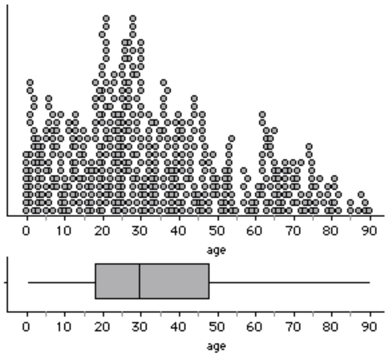

Figure 2: Dot plot and box plot for the New Haven data. Notice how the second quartile, the smallest box, is where the highest density is.

2 What's Wrong with Histograms

The traditional histogram—the version we most often use in school mathematics—has equal-width bars with heights determined by the frequency of cases in the interval. Histograms show us the shapes of distributions.

The problem is deciding on the width of the bins. If the bins are too wide, you can't see important structure in the distribution. But if they're too small—and this is the most dangerous thing—it's easy to see structure where there is none; that is, the counts in the bins get so small that differences with nearby bins are likely to be due to chance. And while experienced statistics students may recognize this, novices may get confused. Figure 3 shows three extreme examples.

To recapitulate, if the bin size is small enough to see the structure in the dense part of the distribution, a histogram may show misleading, "unreal" structure in the sparse regions. Interestingly, this problem occurs even with more sophisticated displays. For example, in computing smooth kernel density estimators, choosing a kernel bandwidth that is too small risks showing spurious details at low densities (Haughton and Phong 2004).

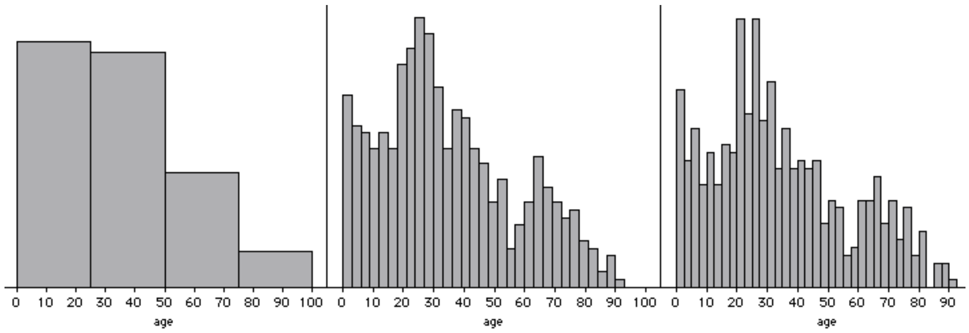

Figure 3: Three histograms of the same 500 ages from New Haven. In the first, the peak near age 25 is invisible. In the second, students may erroneously report a peak at 88 or a dip at 73. The third alternates between tall and short rectangles because bins alternate between including three (integer) ages and only two.

3 A Solution

Why not use small bins when the data is dense and wide bins where the data is sparse?

We can. More general histograms have variable-width bins (Freedman et al. 1997). But to work properly, they must have density, not frequency, on the (usually vertical) axis. Then the area of each box is equal to the number of cases it represents. For an age distribution, for example, the density axis will be in units of people per year. Students seldom see these, and even we practitioners sometimes scratch our heads when we come across one. But they work well to show us the shape of the distribution because of the proportionality between frequency and area. This proportionality also holds for the same-bin-width, traditional school histogram; but it is decidedly not true for the box plot.

What would happen if we made a histogram where the bin boundaries were set at the quartiles? That we see in Figure 4.

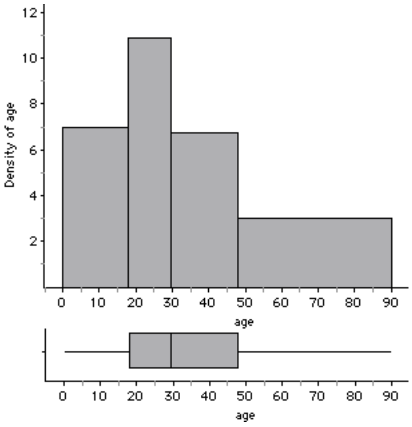

Figure 4: The box plot from Figure 2 beneath a new histogram having four bins of unequal width. Note the density axis, which is in units of people per year. To find the number of people in a rectangle, multiply its width (in years) times the density. In this case, since N=500, there are 125 in each one.

Now the box plot's problems are solved: the densest part of the distribution is highest; the "tails"—where half of the people belong—are not relegated to whisker status; and the eye easily sees the shape of the distribution. What about the problems with the histogram? We have four bars as in

Figure 3, but the main peak is clear—and the little jiggles are invisible.

Suppose we want more detailed information about the distribution? Instead of proceeding to the general histogram—with bin boundaries in arbitrary places—let us keep an important feature of the 4-bin plot: each bin had the same number of people in it. We will limit the histogram in the opposite way from the traditional one. Instead of using bins of equal width but unequal numbers of cases, we will make bins of unequal width but equal number. This has the happy result of widening the bins where the distribution is sparse (and features tend not to be "real") and narrowing them where the distribution is dense (and the higher numbers support finding real features).

With this restriction, we don't have to decide about where to place the bin boundaries, only how many bins we want. The population is thus divided in to quartiles, quintiles, deciles, or whatever. We could call them N-tiles, whence the name Ntigram.

While one would hardly ask students to make Ntigrams by hand (that is what stem-and-leaf plots and box plots were originally designed for, after all), technology makes them accessible. One could use high-end software to construct such plots, but where can K–12 students (and non-programmers) see such things? Tinkerplots (Konold 2004) can create a graph something like what we're talking about. Figure 5 shows the New Haven data with shading that shows the positions of deciles.

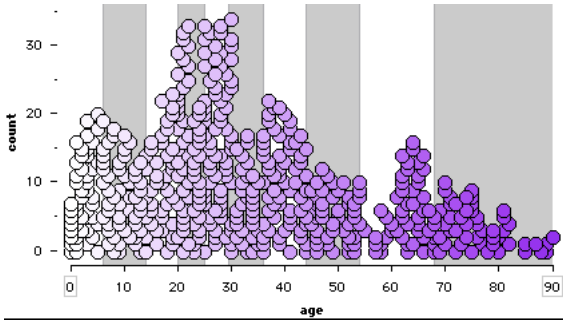

Figure 5: Dot plot from TinkerPlots of ages for the New Haven sample of 500. The stripes in the background show the deciles: the same number of points lie in each stripe.

Still, this is not a histogram. The true Ntigram appears in a data analysis package called Fathom (KCP Technologies 2000). Figure 6 is a 10-bin Ntigram (perhaps a "10tigram") for our New Haven data; you can see most of the features of the distribution.

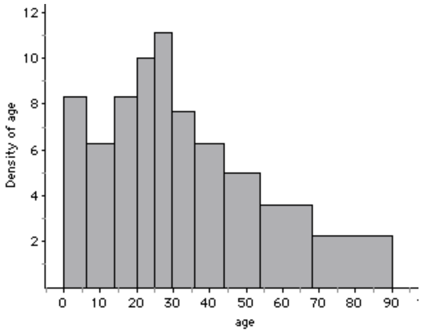

Figure 6: 10-bin Ntigram for New Haven population sample of 500.

The Ntigram has a chunky clarity partway between the traditional histogram and the sleek, hi-tech violin plot. And with an Ntigram, students can (usually) tell which bin each case belongs to; this makes it more concrete. Students can also use the Ntigram to compare n-tiles directly, and thus to compare distributions. Figure 7 shows back-to-back 10-bin Ntigrams for New Haven and Miami. There you can see that there are proportionally fewer children in the New Haven sample, that the “student” peak is younger there, and that Miami’s bulge of older people occurs around age 60, not in the oldest decile.

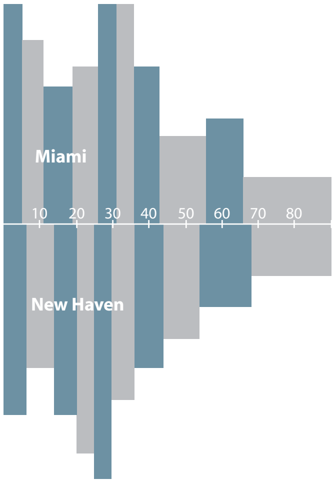

Figure 7: Back-to-back 10-bin Ntigrams of age for Miami and New Haven. As in Figure 1, there are 300 cases from Florida and 500 from Connecticut.

4 Details and Limitations

We will not describe the nitty-gritty of the Ntigram here. There are, indeed, different reasonable ways to construct such a display. There are two main issues. As with the box plot, one has to decide the locations of the bin boundaries, and one solution is to choose the positions of orthodox percentiles. Another issue is whether to split cases between bins so that the bins all have the same frequency—and therefore the same area—even though the frequencies may not be whole numbers. (This is what Fathom does.)

Despite some advantages, Ntigrams are not always the best display. That the vertical axis is density instead of frequency can perplex some learners. And an Ntigram takes up more space than the compact box plot. There are technical limitations as well: you need enough cases that dividing the data into n bins makes sense for your sample size. If there are enough identical values that two n-tiles are the same (i.e., the width of a bin is zero), the density in the bin is infinite and the display does not make as much sense.

References

[Bakker 2004] Bakker, A. (2004). *Design Research in Statistics Education: On symbolizing and computer tools* Dissertation, Utrecht University. Freudenthal Institute.

[Bakker et al. 2005] Bakker, A., Biehler, R., and Konold, C. (2005). “Should Young Students Learn About Box Plots?” *Proceedings of the IASE Roundtable*, Lund, Sweden.

[Freedman et al. 1997] Freedman, D., Pisani, R., and Purves, R. (1997). *Statistics*, third edition. W. W. Norton.

[KCP Technologies 2000] KCP Technologies. (2000). *Fathom^{TM}Dynamic Data software.* Emeryville, CA: Key Curriculum Press.

[Haughton and Phong 2004] Haughton, D. and Phong, N. (2004). "Graphical and Numerical Descriptive Analysis: Exploratory Tools Applied to Vietnamese Data." *Journal of Statistics Education*, **12**. (www.amstat.org/publications/jse/v12n2/haughton.htm)

[Hintze and Nelson 1998] Hintze, J. and Nelson, R.D. (1998). "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician*, **52**, 181–184.

[Konold 2004] Konold, C. (2004). *TinkerPlots.* Emeryville, CA: Key Curriculum Press.

[Tukey 1977] Tukey, J. (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.