Figured Out: Making frequency distributions tangible

Paul Cuffe,[1]
[1]University College Dublin

Figured Out is a recurring column about how to produce clear and effective data graphics. This installment discusses how to portray sets of numbers in a *concrete* and *intuitive* way.

How to think about sets of numbers

There is an isolated Brazilian tribe, the *Pirahã*, whose language, Mura, seems to lacks even a basic vocabulary for discussing numbers and quantities. While this might seem surprising, we should remember that even the simplest words for discussing *sets* of numbers are quite new to the English language. The well-known linguist Mark Liberman has pointed out this equivalence at the excellent *Language Log* blog, where he noted:

> *"Until about a hundred years ago, our language and culture lacked the words and ideas needed to deal with the evaluation and comparison of sampled properties of groups"*

It seems that discussing numeric data is by no means a natural human ability! Indeed, Liberman notes that, even though a handful of statistical terms have drifted into the mainstream vocabulary over the last century, they have not spread much enlightenment:

> *"most of the population still relies on crude modes of expression like the attribution of numerical properties to prototypes (*'A woman uses about 20,000 words

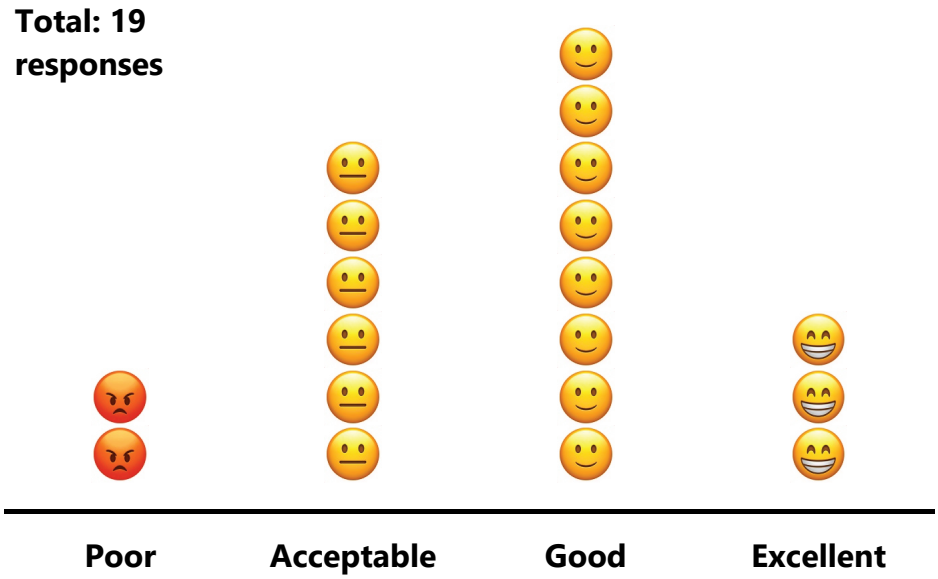

Figure 1: A direct and intuitive way to show nineteen distinct responses to a customer satisfaction question

per day while a man uses about 7,000'*) or the comparison of bare-plural nouns (*'men are happier than women'*).*"

While it is easy to get muddled when talking about distributions, as usual the best advice is: *show, don't tell*. Let yourself see just how happy and/or talkative *every* person in your dataset is, rather than worrying about that phantasmal creature, the "typical woman" (Though however you disseminate your claims, the *absolute most* essential requirement is their basic truth: work published by Mehl et al in *Science* in 2007 found no meaningful difference in daily utterance counts between men and women!)

Showing frequency distributions

The histogram is often the most intuitive way to show a frequency distribution. It is basically just a bar chart, where the vertical axis counts the portion of the data that falls within certain

intervals.

Bin intervals: Depending on how you pick your binning intervals, you might see quite a different shape in your distribution. Ideally, you'll pick intervals that neatly show the shape of the distribution while also providing some sensible context. Unfortunately, software cannot make such decisions for you! So for instance, if you were trying to show the distribution of body mass index in a population, you could bin it according to the medical thresholds for *'underweight', 'normal', 'overweight'* and *'obese'*. Natural thresholds in your data should be exploited when binning: you want to have meaningful categories underpinning a histogram, instead of arbitrary numerical ranges.

Handling the vertical axis: Labelling and interpreting the vertical axis in a histogram can be a bother. Suppose that you're examining the power output of a solar panel, and you have samples taken every fifteen minutes. How should the vertical axis be labelled in such a graph? To be literal, it ought to be: *"Number of fifteen minute intervals having solar power outputs within stated range"*. That's rather a mouthful, of course! Instead, you could agglomerate your data to hourly intervals, to at least give a familiar unit for your vertical axis, or you could multiple by fifteen to get a minute granularity.

Dots and symbols: Often, a frequency distribution can be made more intuitive and tangible by representing it as a *dot plot* rather than a histogram. In this style of graphic, discrete dots are used to explicitly show which bin each data point falls into. The graphic can be made yet more concrete by using appropriate symbols, as in figure 1, which stacks up various emoji to tangibly portray the different responses to a satisfaction survey. The lay reader will often struggle to grasp what histograms mean: symbols can be an effective tool to remove abstraction and boost comprehension. Dot plots also allow specific data points to be labelled, if that is necessary.

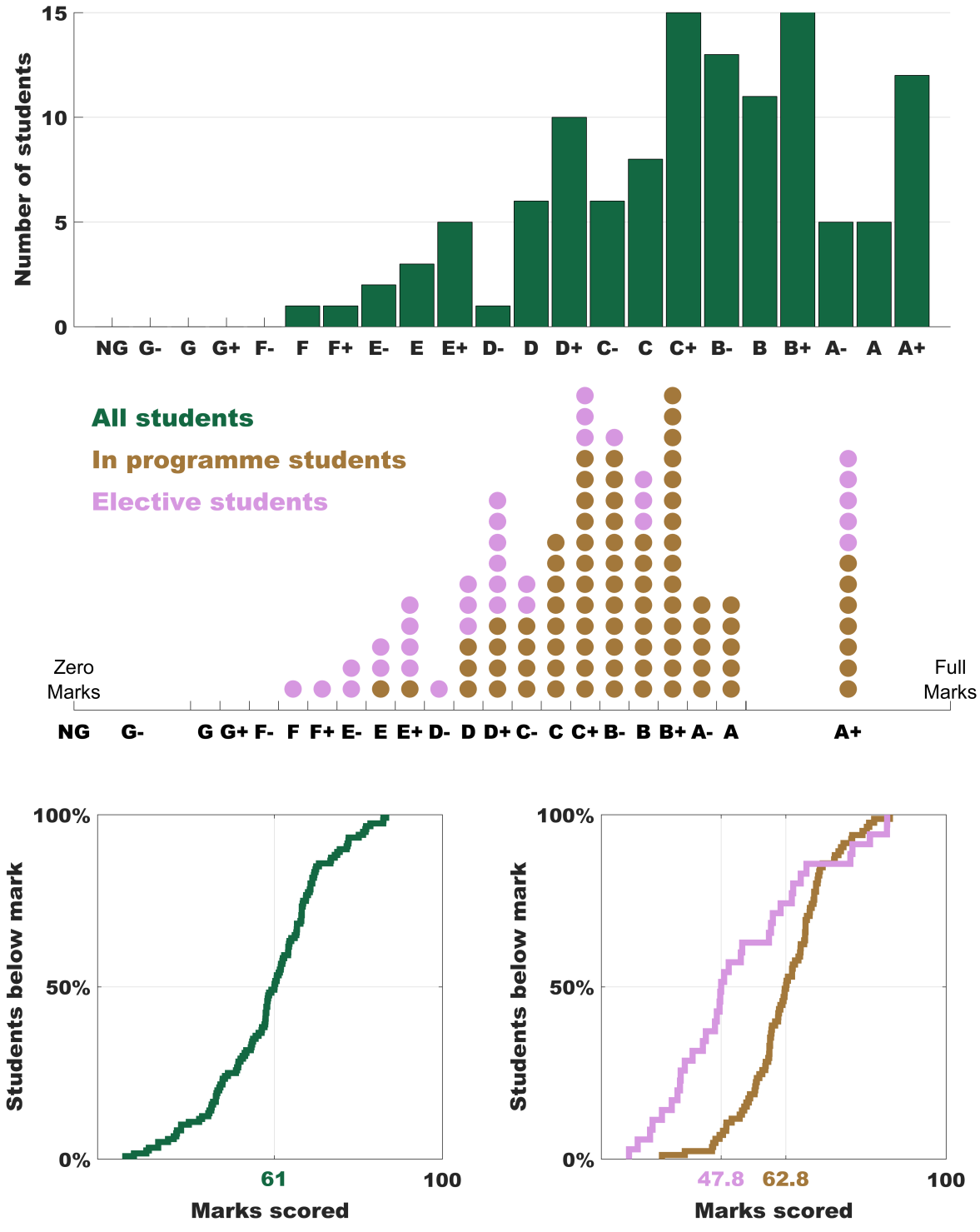

Figure 2: Four different ways of showing the distribution of exam performance between students

Note that making a dot plot like this needn't be too tricky: for instance, you could paste the symbols into the cells of a table in Microsoft Word and build up your graphic that way. It's important to clearly picture the graphic you want in your mind's eye before you sit down to make it: whatever software you use is merely a means to an end, and should never dictate your design decisions.

Empirical cumulative distributions: Plotting the *empirical cumulative distribution function* is often a useful alternative to histograms. The key to this graphic is that your data is rank-ordered: it is a stairstep plot of your ordered dataset, running from the smallest at left to largest at right, with the vertical axis showing the percentage of data points below a threshold. As no binning is required, there is no less of precision. The median values are represented explicitly, and can be projected directly from the vertical axis. As this graph uses a line to show the shape of the distribution, several traces of different colour can be overlaid, to compare the distributions of different datasets.

An academic example

The ensemble of graphs in figure 2 includes four distinct ways of showing the variation in student performance in an engineering exam. The histogram to the top is perhaps the most familiar way of representing this data: it bins students' raw marks by the traditional academic grade letters. A shortcoming of this view is that it tacitly implies that the categories are equivalent, which is actually not the case! The second pane, a dot plot, shows the situation more clearly: the ranges of actual *marks* that correspond to an $A+$ or a $G-$ are actually wider here than for other grades. These wider categories are directly shown on the horizontal axis. Imposing these wider bin thresholds in a traditional histogram would cause the corresponding bars to grow in width and attain a disproportionate area. The use of dots here avoids this effect, and also

helps to underscore that each grade was given to just one specific student. Showing the data in this way also allows the population to be colour coded into those who took the exam as an *'In programme'* or *'Elective'* students.

Two *empirical cumulative distribution functions* for the same data are plotted at the bottom of figure 2. At the left, the aggregate performance is shown, and the median mark, 61, is denoted by a gridline and is explicitly labelled on the horizontal axis. The plot to the right breaks down the same data into the two distinct student populations. The leftward shift of the pink trace for the 'elective students' denotes the generally poor performance of this group compared to the 'in programme students'. A crossover at around 70 marks is also evident, though, meaning that a distinct subset of the elective students secured elite marks. This style of plot facilitates such an in-depth cohort analysis, however it will not be as accessible for a general audience as a traditional histogram. As always, fit the graphic to the task: are you trying to provide a quick take-home message in a short presentation, or are you trying to tease out a subtle and technical point?

A few other little design details to note about figure 2: to maintain clarity, the three categorical groupings are represented by consistent colours throughout. The heaviest colour, the green, is reserved for the most important category, the population of *all* students, and the progressively lighter shades of brown and then pink are used for, respectively, the larger and the smaller subpopulations of interest. This deliberate colour weighting is chosen to subtly guide the reader. To respect the symmetry in the axis ranges of the bottom two graphs, they are drawn with a precisely square plotting area.

Conclusion

It is easy to confuse readers when showing frequency distributions – for many audiences, a histogram is just one notch too abstract to grasp easily. Careful choice of axes and binning

intervals can provide helpful context, though, as can the use of symbols and sensible colours.

About the author

Paul Cuffe (paul.cuffe@ucd.ie) is an Assistant Professor in the School of Electrical & Electronic Engineering at University College Dublin. With his regular *Figured Out* column in IEEE Potentials, he seeks to equip early career engineers with the graphical skills they need to understand complex data and to make an impact in their organisations. Paul would like to acknowledge Chris Dent, Amy Wilson and Harold Kirkham for their help with an earlier version of this article.

References and Notes

1. M. Liberman, "The Pirahã and us," Oct 2007. Available: http://itre.cis.upenn.edu/%7Emyl/languagelog/archives/004992.html.

2. M. R. Mehl, S. Vazire, N. Ramírez-Esparza, R. B. Slatcher, and J. W. Pennebaker, "Are women really more talkative than men?," *Science*, vol. 317, no. 5834, pp. 82–82, 2007.