VISUAL REPRESENTATIONS OF EMPIRICAL PROBABILITY DISTRIBUTIONS WHEN USING THE GRANULAR DENSITY METAPHOR

J. Todd Lee[1] and Hollylynne S. Lee[2]
[1]Elon University, North Carolina, USA; [2]NC State University, North Carolina, USA.
tlee@elon.edu

This paper is an expository on searching for intuitive visualizations of empirical probability distributions. The visualizations use the granular density metaphor for probability density functions (Lee & Lee, 2009) to provide a way for students to build from equiprobable intuitions of probability to standard discrete and continuous distributions. The visualizations dynamically link the sampling process with the formation of an empirical distribution, while remaining true to the granular density metaphor for both the empirical and theoretical density functions. We present and discuss several examples along with conceptual and mathematical ideas that support use of these visualizations.

INTRODUCTION

Over the past 15 years, several researchers (e.g., Konold & Kazak, 2008; Lehrer, Kim & Schauble, 2007; Pratt, 2000, 2011; Prodromou, 2012; Stohl & Tarr, 2002) have expanded approaches for teaching and learning probability to include a strong focus on distribution. Understanding probability distributions is at the heart of learning and doing statistics (e.g., Wild, 2006). When applying a probability model to a real world situation, we conceive of that model as including a random process governed by a theoretical density distribution, and that data is sampled from this distribution. This model encapsulates many assumptions made about the real world phenomena, and with these assumptions, the model is hopefully robust in matching the variation seen in the real world. The probability model also gives us that empirical distributions formed from sampled data have bi-directional mathematical connections to the theoretical distribution. Comparing and reforming theoretical and empirical distributions using a variety of representations and data collection tools is a well discussed pedagogical approach for educationally exploring the use of probability models, especially with statistical reasoning as an end goal. Recently, Pfannkuch and Ziedins (2014) aptly summarized such approaches and highlighted the central message that probability education is well served to include opportunities for students to both construct initial "bad" probability models for a real world situation, and then use bidirectional reasoning to build better models.

Introductory probability lessons typically attempt to build from at least one of two primary student intuitions: 1) the concept of equiprobable events, 2) and the law of large numbers (Batanero, Henry, & Parzysz, 2005). Such lessons, however, do not often push beyond the basic combinatorial nature of common chance-makers like dice and spinners. Pratt (2011) argued that current probability curriculum using such common devices does not serve well to educate in the powerful use of probability as one of several modelling tools when drawing information from real world data. As Wild (2006) points out, we often employ pedagogically the notion of a population distribution from which we select individuals in an equiprobable random manner, and gain perfect measurement value(s) from each of these individuals. This allows for the imagining of a truly stable, combinatorial distribution for which we are approximating with theoretical and empirical distributions with varying degrees of accuracy. However, the variations that occur in the real world from measurement, selection, population membership, temporal dependence, and other complexities, suggest the need for a more general conception of distribution for useful probability models. We believe that even using Pratt and Wild's more realistic views of distributions and their use, we can still retain the metaphorical value of equiprobable selections from a fixed population.

Building a working understanding of continuous distributions is a tough transition for all level of students for several reasons. As J. Lee (1999) noted, one issue of particular importance to statistics is developing a successful transition from different representations of discrete event spaces (relative frequency tables, etc.) to those of continuous spaces (density functions, etc.). What follows is an exposition to further describe one particular approach to visualizing probability

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

distributions. If the visualizations are mathematically robust, perhaps they can provide an approach that builds from students' primary intuitions in probability and afford better connections to statistical inference.

SAND DISTRIBUTIONS

As introduced in H. Lee and J. Lee (2009), a granular density paradigm may offer students visualizations for reasoning about probability density function representations of distributions, both theoretical models and empirically derived ones. This paradigm starts with the population metaphor, combined with an approximate infinitesimal approach to area. Marbles drawn from a bag with replacement has been shown to be a powerful intuitive representation for students. However, instead of a bag of marbles (or balls in an urn) we shall consider a large container of sand made up of uniform grains. Be it a quart, a kilogram, or a bucket, we want to think of the whole quantity of sand as one unit; in other words, there is no outside sand that can be added. A unit amount of sand will contain grains, each grain being equally likely to be selected in a trial. The infinitesimal approach comes into play since any sub-portion of sand would still contain a vast number of grains. Every sub-portion of sand represents some fraction of the whole container. Thus, both concepts of equiprobable outcomes and law of large numbers are integrated in the representation of a distribution (see H. Lee & J. Lee, 2009 for further explanation).

The sand in a distribution is magic, in the sense that upon a command, exactly one grain of sand lights up or makes its presence and choice known in some manner. This can be done many times, each time every grain has an equiprobable chance of being chosen. A visualization of this process uses the dichotomy of sand as a continuous, easily divisible substance and sand as the total of a terribly large number of indivisible grains. Figure 1 illustrates using sand to represent a discrete (left) and continuous (right) probability model, and how illuminating grains can indicate the chosen value.

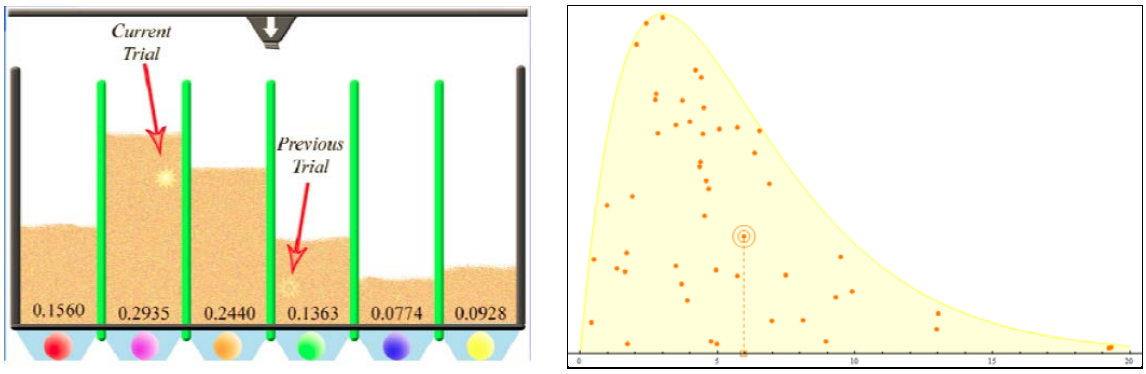

Figure 1. Discrete (left) and continuous (right) sand distributions with sand illuminated.

Recall that a total container of sand is considered as one unit, which leads to a natural correspondence between the "sizable" subsets of sand and fractions from zero to one. We are associating a collection of sand with probability, with the set rules following the probability axioms. The catch here is the same as with getting true infinitesimals to work; subsets containing a relatively small number of grains of sand would have proportions of virtually zero. Most any sample size one would consider would fall in the category of "small number".

For completeness of explanation we are assuming the sand is incompressible and that the density of sand grains is constant. Though a real container of sand is a three-dimensional object, we can imagine pouring the sand onto a plane, pressing it down to a thin, uniform layer. Of course, we still want the layer to be seemingly solid with respect to grain size and the density to be uniform. We want the unit volume of sand to press into a unit area of sand.

The sand represents the area of the probability density function (PDF) of a probability distribution of one variable. Graphs of PDFs are mathematically powerful tools for representing probability measures, it is a very clever use of a single variable function to define values across a vast number of (but not all) subsets over an interval. However, it is extremely tough for students to

move from a focus on the function's $(x,f(x))$ pair to a focus on pairs of intervals and the associated bounded area. This is exacerbated by the fact that many visualizations of empirical distributions are functional representations; histograms where each bin *height* is the frequency or relative frequency. Density histograms are often overlooked in curricula materials. Instead, the transition to continuous PDFs is handled by choosing bin widths in a histogram to be one, so that the relative frequency and density histograms are geometrically equivalent. At this point, textbooks may offer a few obligatory pictures of histograms matched against theoretical PDFs, and then quickly move on to inferential statistics. This is mathematically clever but pedagogically wanting.

EXAMPLE

Suppose we ask a class to go out and take a sample of citizens from a college campus and measure their heights to the nearest inch. If the students consider the measuring process exact, the selection and recording process without error, a measured person's height unchanging, the population well defined and the selection process uniformly random--if all of these things are assumed--the class might think of this being a sample from a discrete collection of a few thousand instances of integers ranging from 40 to 90. A census of this population would determine the underlying theoretical distribution representable by a density histogram with one inch bin widths. In this case, one may be tempted to represent each person as a single grain, but this must be avoided.

In such combinatorial situations, we discussed previously (e.g., H. Lee & J. T. Lee, 2009) the idea of areas of sand being discretized into "bricks". Each brick has the same amount (area) of sand, and a brick "lights" when it contains a grain that lights. In the combinatorial scenario of our height example, we can think of molding each person's sand into a brick and stacking the bricks over corresponding height values. Representing a smooth area distribution that can be discretized into stacks is available using the sampler tool available in *TinkerPlots* 2.0 (Konold &Miller, 2011) that allows for building theoretical models from which data could be sampled. While typical representations such as a mixer (balls in an urn) and spinners can also be used to build models, a smooth curve can be used to produce an area model, which can be re-represented as stacks.

The bricks must be able to be "crumbled" back into sand grains if students question any of the discretizing assumptions. Grains are reserved from being used like the bricks; they are to capture minute, multi-dimensional chunks of the space of varying factors that affect the selection and measuring processes within a real world phenomenon. The leap of faith that must come when moving back and forth from a sand model and the real world situation is that at some tiny scale, the area of possibility, the space of varying factors, can be reasonably divided into equiprobable pieces, each piece associated with a single value in the sample space.

In some ways, a die roll seems an even more natural example of this move from discrete, through continuous, to granular discrete. If we ask a student what the possible influences are when rolling a die, the list will get quite long. Though it is a more subtle observation, if you pick just one influence, it can be seen that not all possibilities of that influence are equally likely. For instance, the height of the toss may be one such influence, but there are some ranges of height for the toss that are far more likely to occur than others. The grain of sand is the unit in which we can zoom in and divide each into tiny, perhaps unevenly sized, subsets of the parameter space for which one value occurs and each grain is equally likely to be "it" for a single experiment.

REPRESENTING SAMPLES

Even when we use a probability model for a real world phenomenon, we often will resort to representations of data that are not oriented towards visualizing an empirical probability density function. For example, results from die tosses are often represented in bar graphs and pie graphs. Data about heights of persons are often represented in dot plots or boxplots. Though these graphs are some of many that can aid in exploratory data analysis, they may not easily afford conceptualizing a probability model. We have been pondering how to represent empirical distributions in a granular manner that may better promote bidirectional thinking between empirical and theoretical probability models. What representations are most accessible for dynamic displays of samples being taken, as opposed to showing a complete sample from the start? Are there representations that help students use empirical data to reason about probability models? What

follows are several possible ways that we have imagined being able to represent the sampling process and empirical probability distribution. Recall that throughout each of our hypothesized visualizations, a main focus is to help students focus on cumulative sand (area) as a representation of likelihood.

One approach could be to use density histograms of uniform bin widths, but rather than represent data points as rectangular bricks, we could use shapes that don't tile so well. As with the ellipses used in Figure 2 (left), there would be irregular spaces between the shapes but none of the shapes would overlap each other. We could have a toggle that allows the sand in the non-tiling bricks to flow into rectangular bricks (Figure 2, right) or straight to sand columns; the level of dynamics could help heighten a focus on the sand/area as the key indicator of probability.

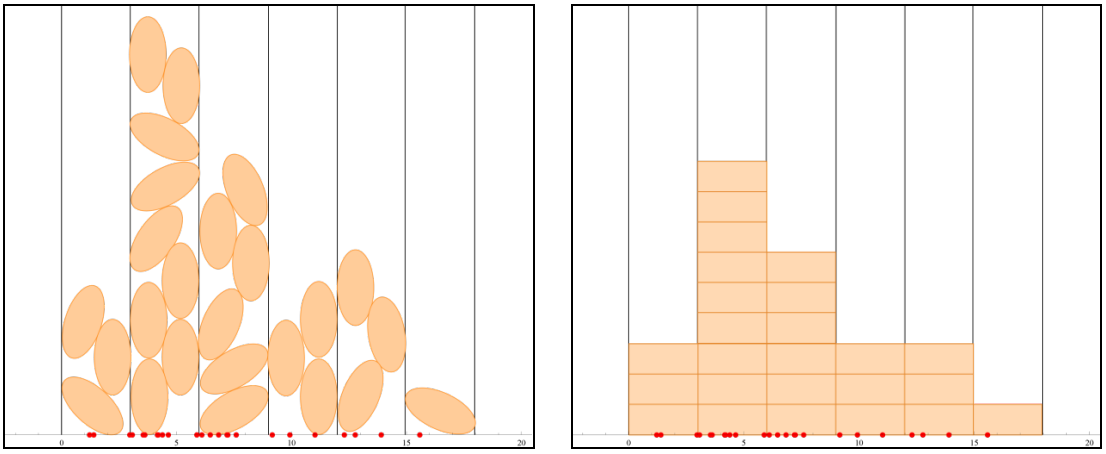

Figure 2. Area-correct elliptical and rectangular bricks used to generate density histograms.

With this approach, we could also dynamically add data points one at a time or all at once, each placed directly or dropped under gravity. And with all of these approaches, we can sample from the generated empirical density histogram using the earlier described process of:

Grain → Containing Brick → Mid-Value of Containing Bin.

These approaches all seem to accentuate the limits of the bin boundaries, which is a good thing. An initial step towards moving from the hard bin walls to smoother distributions could be dynamically removing the bin walls from a column density histogram and allowing the top of the columns to slightly flow and relax (see Figure 3). On the technical side, the tops of the histogram could be re-outlined with splines that maintained unit area while mostly just smoothing out corners. This would not be a statistically robust treatment, but it would allow for lit grains to select specific values, and set up for something truly useful, like a mixed kernel distribution.

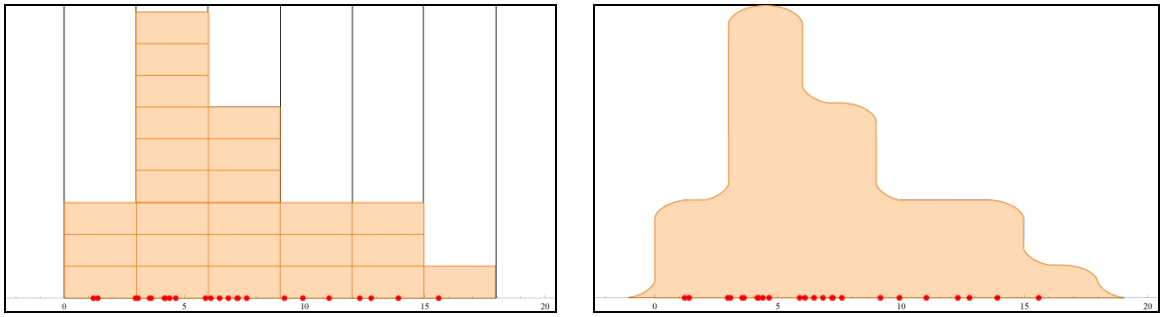

Figure 3. Density histogram (left) and relaxed beyond bin boundaries (right).

Mixed kernel techniques for creating empirical distributions have been around for a long while, and can be calculated and represented quite quickly with current computing power. Thus, consider a possible visualization that could potentially build a strong understanding of PDFs for

upper secondary and college level students. For a given point value in a sample of size *n*, a 1/*n* portion of sand is placed on the sample space axis, centered at the point value and in the shape of the kernel function (say Gaussian). In the left graph of Figure 4, these individual portions are shown, stacked top to bottom in sample order. If these portions are added in this order, then in Figure 4 (right) you can see the end result, with the last added layer centered at sample value 12.3.

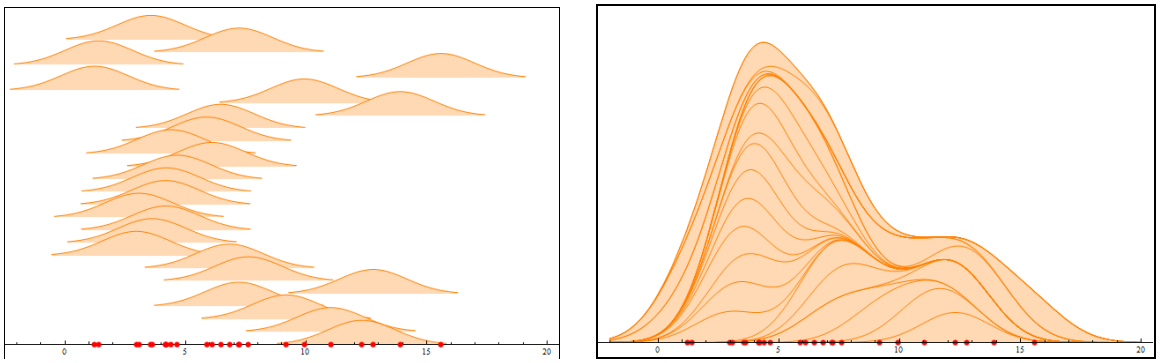

Figure 4. Individual shaped portions of sand for each sample point (left), and the resulting mixed kernel PDF (right).

Seeing an empirical distribution built dynamically, value by value, portion by portion, may be quite informative, especially with the density histogram seen first. Figure 5 illustrates two possible ways that sampling could be visualized from a theoretical probability distribution (in pink). The chosen grains of sand for the sample are shown in red. Then, in the order of the sample, the grain of sand is shown connected to the representation of the data point in the empirical distribution at the top. The brick layering distribution (left) shows that any value within the bin size of the brick results in another brick added to the bin. The numeric value of the data point is still shown as a red dot on the axis. In the mixed kernel layering example, the connection of the sample data point (red grain) in the theoretical probability distribution is done using a cone with gradient coloring. This results in a smooth layer of sand centered at the data value pushing the sand above it upwards. (Actually, the specific current portion is show in bright red at the bottom, but the layer is actually from the top for computational reasons.) The color gradient in the cone and Gaussian shape of the sand layers may help students think about a chosen data value (and the collection of data values) as having both signal and noise. Such attention to center and variation may assist learners in considering the value of interval reasoning when describing distributions in EDA, expressing claims with a degree of uncertainty, and in making inferences about unknown probability distributions and real world phenomena when only given empirical distributions to reason with. The two URLs link to video examples that illustrate this dynamic sampling and building of an empirical distribution.

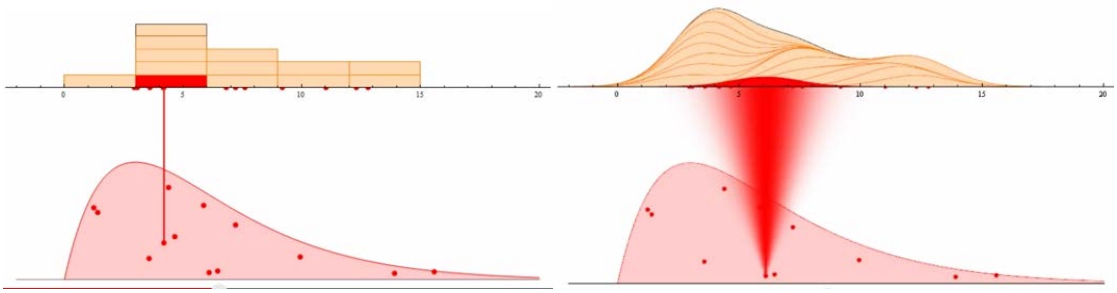

Figure 5. Dynamic sampling with brick layering (http://youtu.be/yGog2USAqXA) and dynamic sampling with mixed kernel layering (http://youtu.be/TLJC1aQrMdA).

REFLECTIVE THOUGHTS

We hypothesize that the granular density metaphor (Lee & Lee, 2009) for representing probability distributions has the potential to provide learning opportunities for students to make strong connections between empirical and theoretical probability. Our intent is to provide a way to reason about and represent probability early in students' experiences with probability models and sampling (say around age 10-12) that can give them stronger foundations for reasoning about about a PDF representation of probability. The difficulties that come with learning how to work with integral areas cannot be overstated. The sand gives physicality to what is being measured by area, and the grains/bricks lend to a blending of classical and epistemic interpretations of probability models. At this point we have used the granular density metaphor and visualizations in our courses, and have used this anecdotal evidence to inform some of our thinking up to this point.

We discussed a few ideas for granular representations for empirical distributions. We also saw how these representations can be dynamically linked with a theoretical distribution. There other possibilities, e.g., since we have deliberately avoided using a vertical axis to accentuate the area over the height, we may go further and show area symmetric about the horizontal axis. However, we have specifically focused on representations that are suggestive of PDFs.

Considering a granular density metaphor for representing probability may also assist others in using area curve models for probability models in their software designs (e.g., see Konold & Miller's sampler in TinkerPlots 2.0, 2011). And since researchers like Pratt, Prodromou, and Lehrer have taken advantage of the ability for students to represent probability distributions with density functions, we have early data that students can reason with area curve models of probability. The next primary goal continues to be to find or construct a suitable software environment that can afford opportunities for us and others to examine carefully how using granular density approaches to modeling probability may impact students' reasoning, in the short run and long run as they learn more statistics and probability.

REFERENCES

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovation in Statistics Education, 2*(1), [Online serial http://www.escholarship.org/uc/uclastat_cts_tise].

Lee, J. T. (1999). It's All in the Area. *Mathematics Teacher*, *92*(8), 670-72.

Lee, H. S., & Lee, J. T. (2009). Reasoning about probabilistic phenomena: Lessons learned and applied in software design. *Technology Innovations in Statistics Education 3*(2).[Online serial http://www.escholarship.org/uc/uclastat_cts_tise].

Lehrer, R., Kim, M., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in modeling and measuring variability. *International Journal of Computers for Mathematics Learning, 12*, 195-216.

Pfannkuch, M., & Ziedins, I. (2014). A modelling perspective on probability. In E. J. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 101-116). Dordrecht, The Netherlands: Springer.

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education, 31*(5), 602-625.

Pratt, D. (2011). Re-connecting probability and reasoning about data in secondary school teaching. In *Proceedings of the 58th World Statistics Conference of the International Statistical Institute (pp. 880-899), Dublin Ireland.* [Online: http://2011.isiproceedings.org/papers/450478.pdf].

Prodromou, T. (2012). Students' construction of meanings about the co-ordination of the two epistemological perspectives on distribution. *International Journal of Statistics and Probability 1*(2), 283-300.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior 21*(3), 319-337.

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal 5*(2), 10-26.