The PDF of a Function of a Random Variable: Teaching its Structure by Transforming Formalism into Intuition

James A. HANLEY and Dana TELTSCH

In many instances, the probability density function (pdf) of a function of a random variable is obtained from the pdf of the random variable, the inverse function and the derivative of this inverse. The formula tends to be memorized rather than fully understood. This article describes how we teach the structure of the new pdf by (a) treating the problem as a change in scale, rather than a "change in variable"; (b) appealing to the concept of "conservation of probabilities"; (c) using the physical analogy of "pouring" probability mass from one set of "containers" to another; and (d) dealing fully with linear transformations before considering nonlinear ones.

KEY WORDS: Graphics; Physical representation; Probability mass; Scales; Transformations.

1. INTRODUCTION

The increasing availability of digital tools is an opportunity for teachers and authors to re-examine how statistical concepts are best presented. We consider the topic found under the rubric "transformation" or "change of variable." The goal is to understand the form of the probability density function (pdf) of a function h (taken, for now, to be monotonic) of a continuous random variable. The way the topic is presented today suggests that no matter whether teachers or authors favor the traditional or the more technological, the teaching could profit from a more intuitive and explanatory approach.

Most texts begin with the "Method of Distribution Functions," or the "Direct Method." If the cumulative distribution function (cdf) of the original random variable has a closed form, the method leads directly to the cdf of the new random variable and, from there, via differentiation, to its pdf. For some distributions, such as the Gaussian one, the cdf does not have a closed form, but the pdf does. In such cases, one may obtain the pdf via the original pdf, the inverse function h^{-1} and the derivative of h^{-1}. Textbooks usually give an algebraic proof, without an intuitive explanation for the formula's structure. As a result, many students merely memorize it for exams, but subsequently are unable to rapidly and confidently reconstruct it.

In an article in this journal a "statistical-generation" ago, Watts (1973, p. 22) described "three highly useful devices for teaching transformations, all centered, quite naturally and obviously, around an overhead transparency projector." Sadly, none of the textbooks we examined have taken up his ideas, and the topic of transformations is often just as "badly received" today as it was when Watts wrote.

We describe ways to make the structure of the pdf formula more intuitive. A commonly used example is given in Section 3, and a practical example in Section 4, but we begin with two simpler ones to show the concepts involved. We deal with the bivariate situation in Section 5.

2. SIMPLE, LOCAL EXAMPLES WITH LINEAR TRANSFORMATIONS

Suppose our focus is on day-to-day variations in temperature (T) in Montreal during a certain time of the year. Even though the early temperatures in the series were originally measured in degrees F, Canada switched to the metric system in 1975, and now the summaries of the entire series from the past 100 years are reported in degrees C. For the sake of illustration, suppose that on this Celsius scale, the distribution is well approximated by $T_{\mathrm{C}} \sim N[\mu = 5, \sigma = 4]$. The pdf of this random variable is shown on the left panel of Figure 1, using the temperature scale -5C to +15C, shown in plain typeface, and with the corresponding pdf scale shown on the left vertical axis, running from 0 to 0.1.

Those not familiar with the Celsius scale will wish to change the temperatures to Fahrenheit, using the conversion

$$T_{\mathrm{F}} = 32 + (9/5)T_{\mathrm{C}}.$$

One might ask students whether, in order to plot the pdf for T_{F}, we must to make an entirely new graph, or whether we could simply keep the existing curve for the pdf and just add (superimpose) the F scale to create the new temperature scale shown in bold. Alert students will notice that the area under the curve would now be greater than 1. Thus, if we *enlarge* the horizontal scale by a factor of 9/5, we must *shrink* the height of the pdf accordingly, by multiplying it by 5/9, so that the area under the new pdf (i.e., the total probability mass) remains unchanged at 1. Without redrawing the curve, this shrinkage in the height of pdf$[T_{\mathrm{F}}]$ can indeed be accomplished *in situ* by adding a new vertical scale, shown in bold. The right-hand panel in Figure 1 shows the two distributions on a single horizontal axis, labeled "degrees," and makes the migration and associated changes in the vertical (pdf) scale more explicit.

The use of a superimposed second horizontal scale helps to illustrate the principle that the probability mass associated with a particular range of T_{C} (say 5C to 10C) must be transferred "intact" to the corresponding range of T_{F} (in this case, 41F to 50F). The same "conservation of probability mass" principle can be applied to the masses associated with ever-smaller T_{C} intervals, say (5.0, 5.5], (5.0, 5.1], and so on. To accommodate these masses on the F scale, the new "containers" located atop (41.0, 41.9], (41.0, 41.18], and so on, will only be 5/9ths as

James A. Hanley is Professor, and Dana Teltsch is a PhD Student, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, H3A 1A2, Canada (E-mail: James.Hanley@McGill.CA *and* dana.teltsch@mail.mcgill.ca). This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, and le Fonds Québécois de la recherche sur la nature et les technologies. The authors thank Lawrence Joseph for comments on earlier versions.

DOI: 10.1198/000313006X89947

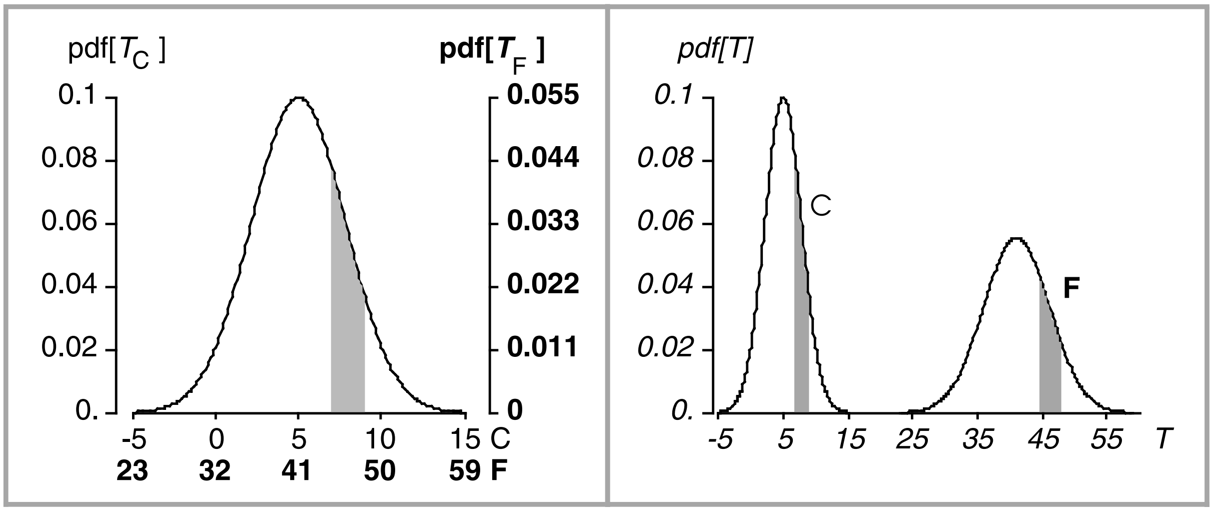

Figure 1. Day-to-day variations in ambient temperature at a certain time of the year. Left panel: distribution plotted on C scale (plain typeface). F scale added underneath in bold typeface, and pdf[T_F] scale, also in bold, obtained by multiplying pdf[T_C] scale by a factor of 5/9 to constrain the total probability mass to be unity. Right panel: the same two distributions plotted on an unspecified ("degree") scale, shown in italics. The probability mass of 0.15 in the 2C wide interval 7C–9C represents the same days in which the temperature falls within the 3.6F wide interval 44.6F–48.2F.

tall, because their bases are wider than those of their original containers by a factor of 9/5.

With ever-narrower T intervals, so that the shape of the probability mass in each bin is approximately rectangular, we can write the redistribution in terms of the bases and heights of the pair of original and new rectangles

$$\text{new width} \times \text{new height} = \text{original width} \times \text{original height},$$

or

$$\text{new height} = \text{original height} \times \frac{\text{original width}}{\text{new width}}.$$

Thus, with the intervals in the new scale made infinitesimal, this expression for the pdf becomes

$$\begin{aligned}&\text{pdf}_{\text{NEW}}[\text{new}]\\ &\qquad = \text{pdf}_{\text{ORIGINAL}}[\text{equivalent of new }] \times \text{scale factor}.\end{aligned}$$

In our example, the scale factor is the ratio of the increment on the C scale corresponding to a 1 unit increment on the new F scale. More generally,

$$\text{scale factor} = \frac{\text{corresponding increment on original scale}}{\text{1 unit increment on new scale}}.$$

This example shows that it is not so much that we created a new random variable as it is that we *changed the scale* on which the variable was measured. Conceptually, there is only one random variable, temperature (T). We measure it either as T_C or T_F.

In this first example, the distribution of the original random variable had a particular parametric form. However, the formula linking the original and the new pdf applies regardless of distribution. Consider, as a new example, the years of publication of the books in the McGill libraries; for the sake of this illustration, we limit our attention to the books published after 1900. These follow the distribution shown in Figure 2, a peculiar pattern shaped by human events and by technological changes. Suppose we now measure how old the books are in decades by converting the year published using

$$\begin{aligned}\text{Age}_{\text{decades}} &= h[\text{year published}]\\ &= 200.5 - 0.1 \times (\text{year published}).\end{aligned}$$

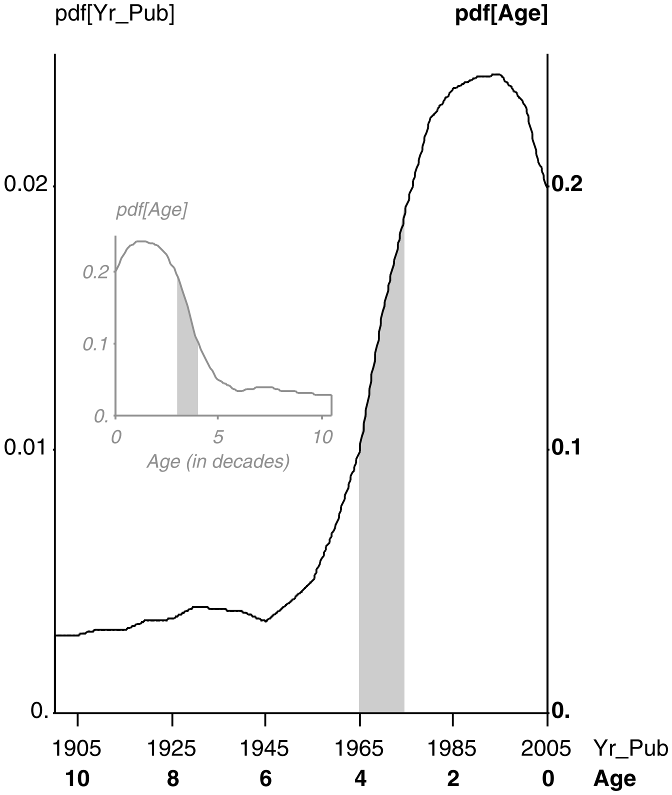

Figure 2. pdf for the variable "year of publication" [Yr_pub] of library holdings, with horizontal (1900–2005) and vertical (0–0.025) scales shown in plain typeface. Scale for derived variable ("age of holding, in decades," range 0 to 10.5) added underneath in reverse in bold typeface, and pdf scale for this new variable obtained by multiplying original pdf scale by a factor of 10 to constrain the total probability mass to be unity, shown on right. Inset: pdf of age distribution, with conventional polarity. For illustrative purposes, range restricted to works published in the years 1900–2005. Works now three to four decades old (i.e., published between 1965 and 1975) are tracked in gray.

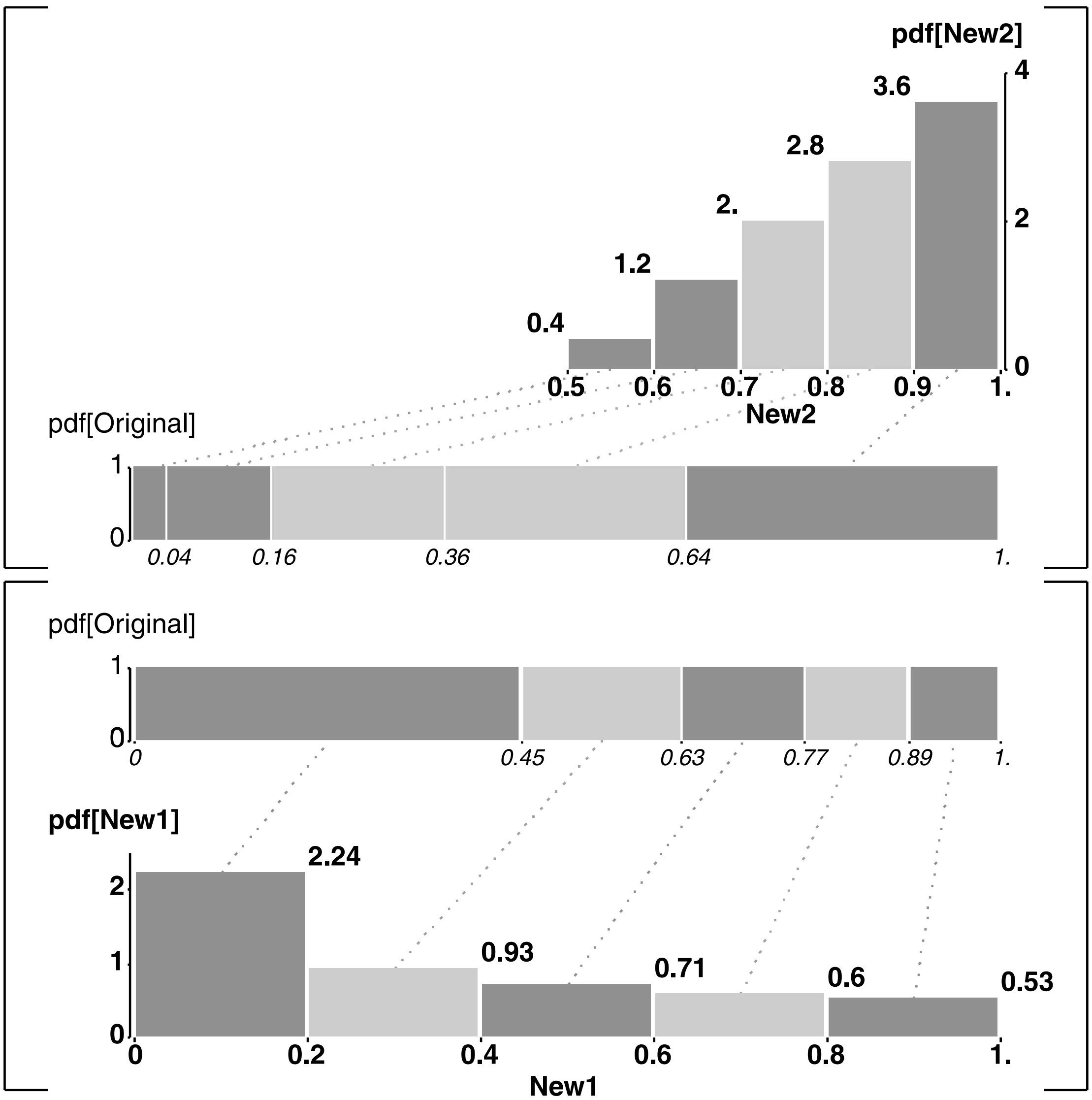

Figure 3. Distribution of original variable Original ∼ Uniform[0,1] and new variables New1 = Original^{2} in the lower panel, and New2 = (Original$^{1/2}$ + 1)/2 in the upper panel. In the lower panel, probability mass is transferred to new containers, each with a width of 0.2 (shown in bold, along bottom) from their corresponding original containers. For example, the probability mass of 0.12 is transferred to the interval 0.6 to 0.8 (new, bold) from the interval 0.77 to 0.89 (original, italic). The heights of the new containers (2.24, 0.93), . . .) are shown in bold. In the limit, with ever-narrower destination containers, this sequence of heights approaches the function pdf[new] = 1 × (1/2) new$^{-(1/2)}$. The upper panel illustrates the same principle in an example where the range of the new variable is different from that of the original.

Because the scale is reversed, the new pdf has the mirror-image shape. Its vertical scale is 10 times larger than the original to compensate for the fact that its horizontal scale is 10 times smaller.

In these two examples, we were able to create the new pdf *in situ* simply by superimposing a new scale below the original one, and changing the vertical scale accordingly. The fundamental shape of the new pdf did not change. However, when we change from the original to a new scale by a nonlinear transformation, this simple relabeling of axes is no longer sufficient.

3. NONLINEAR SCALE CHANGES: A TEXTBOOK EXAMPLE

A theoretical example, commonly used, is one in which the original variable $O \sim U(0, 1)$ and the new variable is $N = h(O) = O^2$. We have changed their names from the traditional X and Y (or Y and X!) to more helpful ones. Although this $\{$distribution, transformation$\}$ pair does not bring out all the issues—the uniform distribution makes one of the two calculations too easy, and the transformation maps the values into the same range—it does illustrate the two important complexities created by a nonlinear transformation.

As shown schematically in the lower panel of Figure 3, one first divides the range of the *new* variable into equal-sized intervals, then determines what intervals these correspond to in the range of the original variable. Since the probability mass in each "destination" rectangle must match that in its corresponding "source" rectangle, the new and the original height continue to be linked via a scale factor, but with one important difference: the scale factor depends on *where* on the new range the focus is. For example, the rectangles at the left end of the new range are taller than their original counterparts, while those at the right end are shorter. The same pattern is seen if we make the width

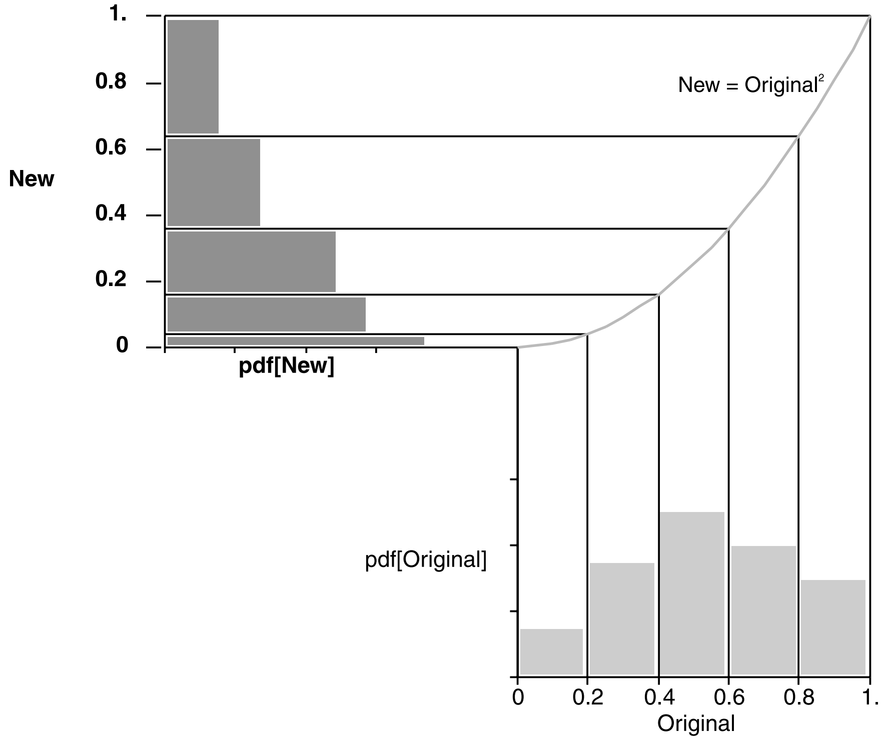

Figure 4. “Pouring probability,” after Watts (1973). Arbitrary original distribution, and transformation New = h[Original] = Original^2. By tipping the device, the probability mass in the original distribution (lower right) is poured into the new probability distribution (on its side, upper left).

of each new container smaller, say δ units wide. At a given value in the new range, the scale factor can be calculated by inverting the transformation $N = O^2$ to obtain $O = N^{1/2}$. By the same principle as above, at the limit the scale factor at a particular value n in the new scale is

$$\begin{aligned}
&\text{scale factor}\\
&\quad= \lim_{\delta\to 0} \frac{\text{corresponding increment on original scale}}{\text{an increment of } \delta \text{ on new scale}}\\
&\quad= \frac{do}{dn}\\
&\quad= \frac{dn^{1/2}}{dn}\\
&\quad= (1/2) \times n^{-1/2}.
\end{aligned}$$

Therefore, in the limit, the pdf of the new random variable is

$$\begin{aligned}
\text{pdf}_N[n] &= \text{pdf}_O[o - \text{equivalent of } n] \times \text{scale factor}[n]\\
&= 1 \times (1/2) \times n^{-1/2}.
\end{aligned}$$

The upper panel of Figure 3 shows that the same reasoning applies to situations where the range of the new variable is different from that of the original. One can also use the pair of panels to show that one can carry out a sequence of transformations, just as children do when playing with water.

After we had completed an earlier version of this note, in which we had used the imagery of “pouring” probability mass from one set of containers to another, a colleague pointed us to the article by Watts (1973). His device for univariate transformations consisted of “a sandwich of 1/8-inch plastic sheets separated by 1/8-inch plastic rods. The probability was fine sand” (taken from an unguarded sand-filled ashtray in the University of Wisconsin Statistics Department). As illustrated in Figure 4, by tipping his device, he could “pour” probability distributions from one scale to another.

4. A NONLINEAR SCALE-CHANGE: PRACTICAL EXAMPLE

The top half of Figure 5 shows the distribution of “fuel-economy,” measured as miles per gallon (MPG, or “M”). To focus on fuel rather than distance, we instead consider how many gallons are required to travel say 100 miles, (G.P.100M, or “G”), a scale that is more logical when investigating factors that influence fuel consumption. In this scale, which reverses the polarity, a low value indicates an economical vehicle, and a high number an un-economical one. The new (G) and original (M) values are linked by the relationship

$$G = h[M] = 100/M,$$

and its inverse

$$M = h^{-1}[G] = 100/G.$$

The figure shows that the new pdf is not a mirror image of the original; one cannot simply reverse the original, and relabel the horizontal and vertical axes. The height of the pdf at a value on the new scale (e.g., five gallons) is obtained by locating its counterpart (20 mpg) on the original scale, measuring the height of the original pdf at this value (0.025 approximately) and multiplying it by the absolute value of the derivative of $100/G$ with respect to G, evaluated at the value in question, that is,

$$\text{pdf}_G[5] = \text{pdf}_M[20] \times (100/G^2)|_{G=5} = 0.025 \times 4 = 0.100.$$

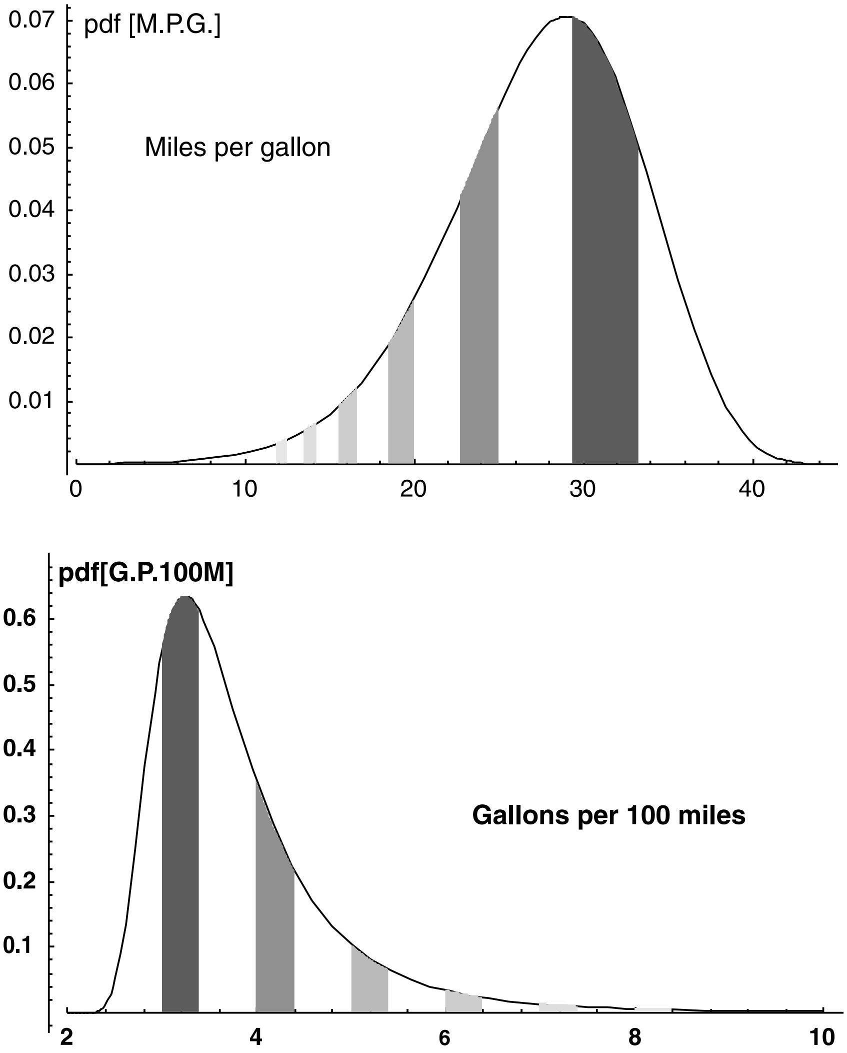

Figure 5. Distribution of variable "fuel economy" of vehicles, measured first in miles per gallon (top) and re-expressed in G.P. 100M = gallons per 100 miles (bottom). These are abbreviated to M and G in the text. The probability mass (fractions of all cars) in several (0.4-gallon wide) intervals is tracked with shaded areas.

5. A BIVARIATE EXAMPLE: THE ROLE OF THE JACOBIAN

Having introduced the idea of transferring probability masses, it becomes easier to consider transformations of one bivariate random variable to another. In his Figure 2, Watts transformed the original sample space in X, Y coordinates into a new one in polar or R, Θ coordinates, using the transformations $r = (x^2+y^2)^{1/2}$ and $\theta = \arctan(y/x)$. The reverse transformations $x = r\cos\theta$ and $y = r\sin\theta$ will be of interest later on. A portion of the original space is shown in the upper level of Figure 6. Again, we begin with "destination" bins, with bases in the shape of rectangles, in the transformed space on the lower level. Each of the two example bins, labeled D1 and D2, has a base with an area of $\Delta R \times \Delta\Theta = (1) \times (\pi/10)$. We have labeled the corresponding bins in the original or "source" space as S1 and S2. Their bases have different nonrectangular shapes and different areas. Visually, we estimate that the base of S1 has an area $\Delta x \times \Delta y$ that is just less than 1/2 that of the unit square. From theory, the exact area of this base is $(1.5) \times (\pi/10)$. Thus, if the probability mass in the larger-base S1 is "poured" into the smaller-base D1, the filled height of D1 becomes 1.5 times higher than what it was in S1. In contrast, when the probability mass in S2 is poured into D2, the filled height becomes half what it was in S2, since the base of S2 is only half that of D2.

Students might be asked to verify that the ratio of the area of the base of the source bin to that of the destination bin reaches a limit as the latter is progressively reduced. This limiting ratio is none other than the Jacobian of the transformation from X, Y to R, Θ.

$$\begin{aligned}\text{scale factor} &= \lim_{\text{Areas}\to 0} \frac{\text{Area}_{\text{source}}}{\text{Area}_{\text{dest.}}} \\ &= \begin{vmatrix} \delta x/\delta r & \delta x/\delta\theta \\ \delta y/\delta r & \delta y/\delta\theta \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.\end{aligned}$$

This fits with our earlier calculations where $\text{Base[S1]}/\text{Base[D1]} = r_{\text{average}} = 1.5$ and $\text{Base[S2]}/\text{Base[D2]} = r_{\text{average}} = 0.5$

Thus far, we have deliberately avoided any mention of a specific probability distribution over X, Y. We did this to emphasize that the *scales*, and the *probability distributions* over them, are two separate components. One example of this second compo-

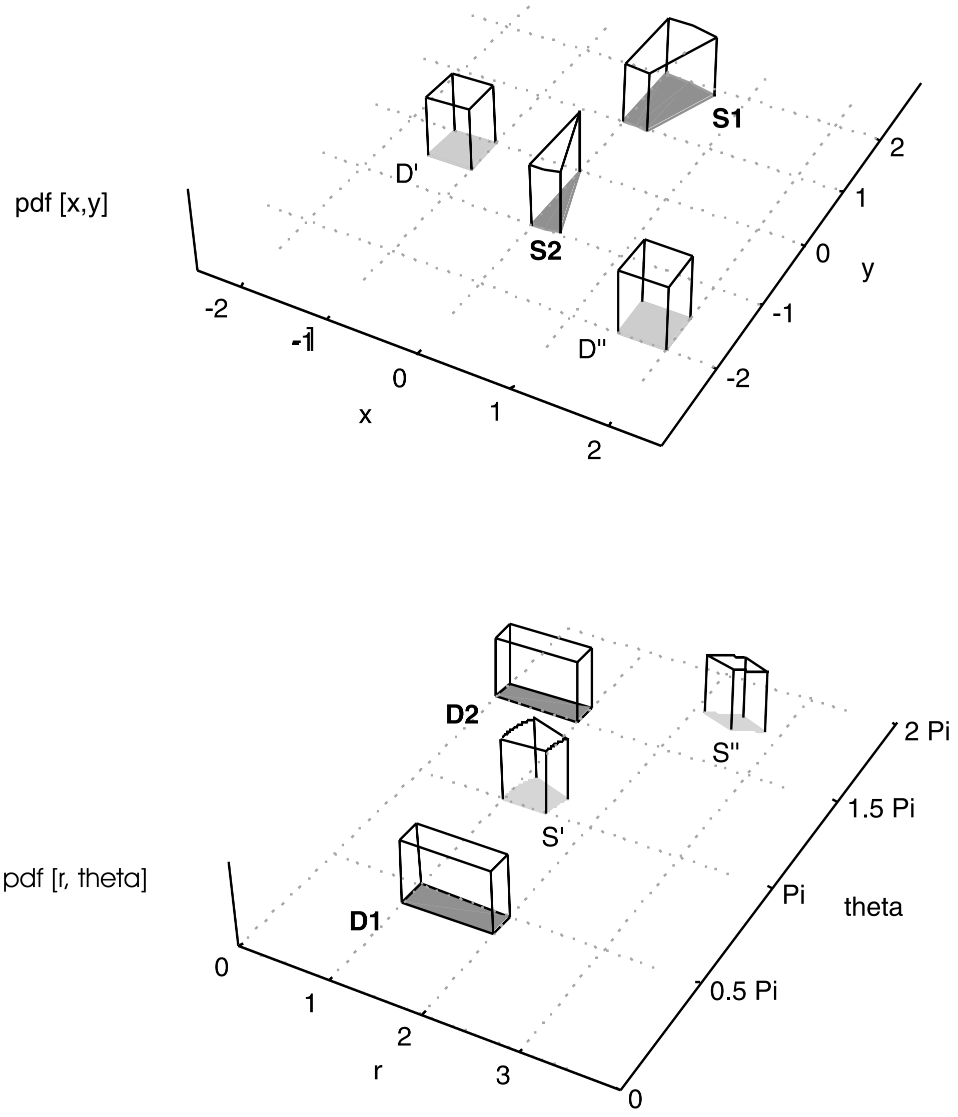

Figure 6. Sample space elements in original X, Y coordinates (upper) and polar R, Θ coordinates (lower). "Destination" bins, **D1** *and* **D2**, *with bases in the shape of rectangles, are shown in the transformed space on the lower level. Their corresponding bins* **S1** *and* **S2** *are shown in the original or "source" space. In the limit, the area of the base of* **S1** *to that of* **D1** *is 1.5. In general, the scale factor is calculated as the Jacobian of the source axes with respect to the destination axes. Thus, when the probability mass in the larger base* **S1** *is poured into the smaller base* **D1**, *the filled height of the new* **D1** *will be 1.5 times what it was in the original* **S1**. *The two sets of bins labeled D′ and D″, and S′ and S″ show the opposite transformation—from the R, Θ sample space to the X, Y one. After Watts (1973).*

nent is the textbook one where $X, Y \overset{\text{iid}}{\sim} N(0, \sigma^2)$. If $\sigma = 1$,

$$\text{pdf}_{X,Y}[x, y] = (1/2\pi)e^{-(1/2)(x^2+y^2)} = (1/2\pi)e^{-(1/2)r^2}.$$

If X and Y were independent horizontal and vertical deviations from a target, there might be a natural focus on the induced (marginal) distribution of R, which reflects the total deviation from the target.

Because the (reverse) $R, \Theta \to X, Y$ transformation is 1:1,

$$\text{pdf}_{R,\Theta}[r, \theta] = \text{pdf}_{X,Y}[x_{[r,\theta]}, y_{[r,\theta]}] \times \text{scale factor},$$

where $x_{[r,\theta]}$ stands for $r \cos\theta$, the "x-equivalent of" r, θ, and similarly for $y_{[r,\theta]}$, or $r \sin\theta$. Thus, in our specific example,

$$\text{pdf}_{R,\Theta}[r, \theta] = (1/2\pi)e^{-(1/2)r^2} \times r,$$

leading to the marginal density

$$\text{pdf}_R[r] = \int_0^{2\pi} \text{pdf}_{R,\Theta}[r, \theta]d\theta = re^{-(1/2)r^2}.$$

This distribution is called the Rayleigh distribution with parameter $\sigma = 1$. Because the Rayleigh random variable for general σ is simply σ times that for the case with $\sigma = 1$, one can obtain its pdf using the results for the linear transform given in Section 2.

The other two sets of bins in Figure 6, labeled D′ and D″, and S′ and S″, are included for completeness, to show the opposite transformation—the original sample space is R, Θ, the transformed one is X, Y, and the reverse transformations are as given above.

6. DISCUSSION

Many textbooks continue to teach mathematical statistics more as mathematics than statistics, with few real examples, with limited use of graphics, and with little appeal to intuition.

We argue that more can be done to make the teaching more relevant without sacrificing mathematical rigor.

Curiously, many of these texts proceed immediately to nonlinear transformations, when the first principles (such as the basic idea of a change of scale) can be more convincingly illustrated with linear ones. For an even simpler linear transformation than our Celsius to Fahrenheit conversion, one might consider converting the heights or weights of students to (from) metric (imperial) units, or—to compare fuel economy of U.S. and European cars—converting gallons per 100 miles to liters per 100 Kilometers. The age-of-books (linear) and fuel economy (nonlinear) examples show the pedagogical advantage of using familiar measurements, but reversing or inverting the scale on which we usually look at them.

Typical theoretical derivations depend on algebraic manipulation, and use of the chain rule. Concrete examples can show the procedure as one where we transfer probability masses from one histogram to another, by adjusting the heights of the new bins or containers to conserve the masses being transferred. In all applications we rely on rescaling to increase/decrease the height of the original pdf's: in the linear case, there is one universal scaling factor, whereas in the nonlinear case, point-specific scaling factors are required, with the derivative at each value providing the localized scaling factor. The Jacobian is seen as a way to calculate the scaling factor for higher dimensional settings. The above approach can also be applied to nonmonotonic h's by dividing the inverse of h into monotonic sections and "transferring" each section separately.

The use of superimposed scales, and the focus on units of measurement, show that is not so much a new variable, as a new scale for the same variable. Often, when the real units are on one scale, we use another one for convenience, such as when we report (or study the distribution of) the strengths of earthquakes on the Richter (log) rather than the "regular" scale.

We found that some textbooks teach only the Direct Method, avoiding examples where the CDF of the original variable does not have a closed form. This is unfortunate. By substituting $h^{-1}(n)$ *directly* into $\mathrm{CDF}_O()$ to arrive at $\mathrm{CDF}_N(n)$, and then simply differentiating this with respect to n, students do not get to see the two-part structure, involving the original pdf and the scaling function.

The strategies we have presented are directed primarily at texts and courses in mathematical statistics. Their purpose is to deepen the understanding of the topic of transformations without sacrificing rigor. They may also allow this topic to be included in texts aimed at mathematically less sophisticated audiences, along with the formulas for the mean and variance of a function of a random variable.

One of us (JH) used these strategies in 2001 in a course on probability theory for undergraduate students in science and engineering who had had differential and integral calculus. The course was based on the first seven chapters of Wackerly, Mendenhall, and Scheaffer (1996); this text provides a large number of exercises under the topic of "Functions of Random Variables." In this accelerated summer course, some students were content to simply calculate the derivative and use it in the formula. Those who were more inquisitive and had fewer other commitments appreciated the "why" and wondered why the text did not use a more graphical approach.

Sadly, none of the textbooks we examined have taken up the ideas put forth by Watts. We did however find the graphical approach illustrated in two Web sites. The first (Glanz 2004) uses a random number generator to produce input/output pdf/CDFs for any writeable mixed pdf inputs and transformation function. The sample pdf of the original random variable, the transformation function, and the sample pdf of the new variable are plotted in three panels in an arrangement similar to that in Figure 4. The menu helps to emphasize that the original random variable is a separate entity from the transformation; the purpose of the latter is to change the scale over which the probability is distributed. Kingsbury (2003) begins by algebraically deriving the new pdf of the new random variable in the usual way, but then illustrates the procedure geometrically using a triangular distribution and a transform which is monotonically increasing. In order to maintain the conventional polarity, he constructed his diagram so that the probability "conduits" cross each other. The noncrossing conduits in our Figure 4 result in a reversed polarity for the new scale, but since Watts had a transparent device, he could flip it so that the new values increase from left to right.

For good pedagogical reasons, lecture rooms are still equipped with overhead transparency projectors, but a teacher would need to be quite agile to use all of Watts's strategies for bivariate transformations: he overlaid then moved a duplicate of an original shape to show translation and rotation; he put the duplicate on a second projector/screen to show scaling, and he curled the screen to show nonlinear transforms. However, in 2005 and beyond, many of these classrooms will also be equipped with computer projectors. Although computer-generated images lack the credibility and realism of the physical objects whose shape Watts and his projector was able to rearrange, computer technology could replace his bivariate transparencies. For example, one could show the (O_1, O_2) space on the left half of the screen and the (N_1, N_2) space on the right. With a computer mouse, or a stylus on a graphics tablet, the teacher or student could delineate rectangular (and other!) regions in the (N_1, N_2) space; the interactive computer program could show their equivalents in the original space. Figure 2 in Watts illustrates how he accomplished this with colored marker pens, while our static Figure 6 uses shading. Two-dimensional points could then be randomly generated from the source pdf and shown as dots in the sample space. At the same time their transformed locations would be placed in the new sample space. Updated reports would show the densities of dots within the region(s) of interest. Using dots is also a helpful way to remind us that the "d" in pdf stands for "density." If one wished to create a 3-D effect, the dots might be cumulated vertically, to create "3-D" skyscrapers. Our Web site (http://www.epi.mcgill.ca/hanley/jacobian) includes a crude beta-version of the $X, Y \leftrightarrows R, \Theta$ version that can be visualized using Macromedia Flash Player. For now, it includes just one bivariate distribution.

However, devices for bivariate transformations are relevant only after students have fully understood the concepts involved in the univariate case. Fortunately, these can be effectively taught with nothing more than chalk, or a pen, or ready-made diagrams such as those in Figures 1 to 5.

[Received March 2005. Revised September 2005.]

REFERENCES

Glanz, F. (2004), "Transformation of a Random Variable Demo," available online, under "Statistics and Probability," at http://www.mathworks.com/matlabcentral/fileexchange/. Accessed October 10, 2004.

Kingsbury, N. (2003), "Functions of Random Variables," connexions module: m11066 Version 2.5: 2003/03/31, http://cnx.rice.edu/content/m11066/latest/.

Wackerly D. D., Mendenhall, W., and Scheaffer, R.L. (1996), *Mathematical Statistics with Applications* (5th ed.), Belmont, CA: Duxbury Press.

Watts, D. G. (1973), "Transformations Made Transparently Easy, or, so That's What a Jacobian Is!" *The American Statistician*, 27, 22–25.