Geometry of Transformations of Random Variables

Univariate distributions

We are interested in the problem of finding the distribution of $Y = h(X)$ when the transformation h is one-to-one so that there is a unique $x = h^{-1}(y)$ for each x and y with positive probability or density. In the case of discrete random variables, the transformation is simple.

$$P(Y = y) = P(h(X) = y) = P\left(X = h^{-1}(y)\right)$$

In contrast, for absolutely continuous random variables, the density $f_Y(y)$ is in general not equal to $f_X(h^{-1}(y))$. The reason is that the geometry of the transformation becomes more complex as the dimension increases. For discrete distributions, probability is located at zero-dimensional points, and the transformations do not affect the size of points. For univariate absolutely continuous distributions, however, probability is associated with the integral of a density over a one-dimensional line segment. Transformations can change the lengths of intervals, as shown here where an interval of length $\mathrm{d}x$ is transformed to smaller interval of length $\mathrm{d}y$.

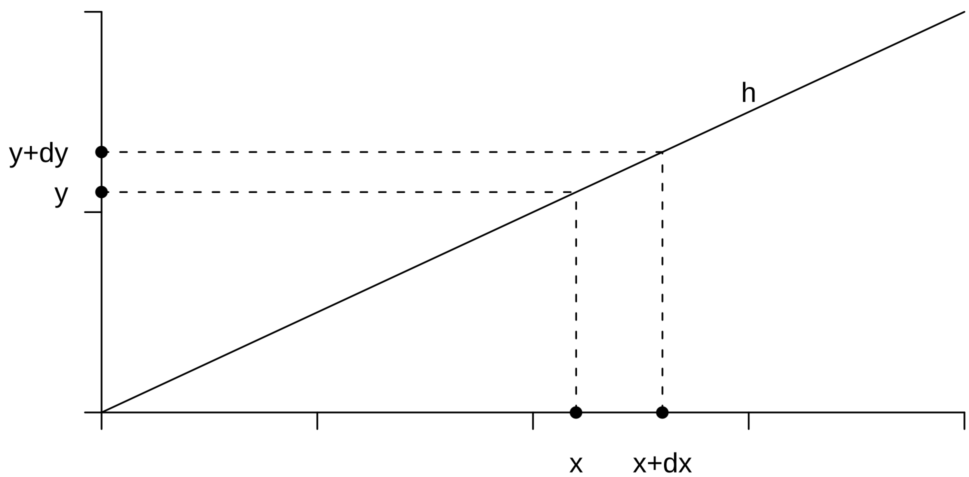

Figure 1: **Transformation** $Y = h(X)$. The figure shows $Y = h(X)$ over a very small interval so that h appears to be essentially linear.

For small $\mathrm{d}x$, the probability in the interval $(x, x + \mathrm{d}x)$ is approximately $f_X(x)\mathrm{d}x$. The density at $y = h(x)$ will be the limit of the ratio of this probability over the length of the interval between $h(x)$ and $h(x+\mathrm{d}x)$ which is $|h(x+\mathrm{d}x) - h(x)|$. (If $h'(x) < 0$, then $h(x+\mathrm{d}x) < h(x)$ so the absolute value is needed.) As h is differentiable, the approximation $h(x + \mathrm{d}x) \approx h(x) + h'(x)\mathrm{d}x$ is accurate for very small $\mathrm{d}x$ and it follows that the transformed interval has approximate length $|h'(x)|\mathrm{d}x$. The density at y is then

$$f_Y(y) = \frac{f_X(x)\mathrm{d}x}{|h'(x)|\mathrm{d}x} = \frac{f_X(h^{-1}(y))}{|h'(h^{-1}(y))|}$$

after applying $x = h^{-1}(y)$.

Multivariate Distributions

We would like to extend this idea to joint densities. If random variables $X = (X_1, \ldots, X_n)$ have joint density f_X, we aim find the joint density f_Y of the random variables $Y = (Y_1, \ldots, Y_n)$ where we write $Y = h(X)$ to mean

$$Y_i = h_i(X_1, \dots, X_n) \quad \text{for } i = 1, \dots, n.$$

We will assume that h is a differentiable bijection which means that all partial derivatives $\partial h_i / \partial x_j$ exist and that the vector equation $(y_1, \dots, y_n) = h(x_1, \dots, x_n)$ has a unique solution (such that $f_X > 0$) with $(x_1, \dots, x_n) = h^{-1}(y_1, \dots, y_n)$.

Bivariate Distributions. We motivate the general answer by examining the bivariate case $(Y_1, Y_2) = h(X_1, X_2)$. The density at $h(x_1, x_2)$ is the limiting ratio of the probability in a rectangle with a corner at (x_1, x_2) with sides of length $\mathrm{d}x_1$ and $\mathrm{d}x_2$ over the area of the rectangle $\mathrm{d}x_1 \mathrm{d}x_2$. The density at $(y_1, y_2) = h(x_1, x_2)$ will depend on the geometry of the transformation of the corners of this rectangle.

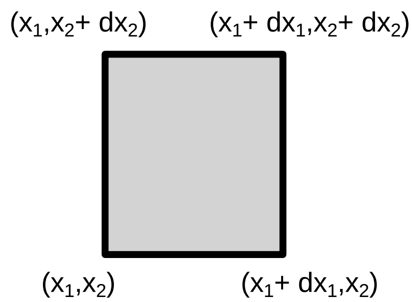

Figure 2: **Rectangle before transformation**.

By the partial differentiability of h in each dimension, the following approximation is true.

$$\begin{aligned} h_1(x_1 + \mathrm{d}x_1, x_2 + \mathrm{d}x_2) &\approx h_1(x_1, x_2) + \frac{\partial h_1}{\partial x_1}\mathrm{d}x_1 + \frac{\partial h_1}{\partial x_2}\mathrm{d}x_2 \\ h_2(x_1 + \mathrm{d}x_1, x_2 + \mathrm{d}x_2) &\approx h_2(x_1, x_2) + \frac{\partial h_2}{\partial x_1}\mathrm{d}x_1 + \frac{\partial h_2}{\partial x_2}\mathrm{d}x_2 \end{aligned}$$

To simplify the expressions, let $y_1 = h_1(x_1, x_2)$, $y_2 = h_2(x_1, x_2)$, $a = \frac{\partial h_1}{\partial x_1}\mathrm{d}x_1$, $b = \frac{\partial h_1}{\partial x_2}\mathrm{d}x_2$, $c = \frac{\partial h_2}{\partial x_1}\mathrm{d}x_1$, and $d = \frac{\partial h_2}{\partial x_2}\mathrm{d}x_2$. With this notation, the four corners of the rectangle are mapped approximately as follows:

$$\begin{aligned} (x_1, x_2) &\mapsto (y_1, y_2) \\ (x_1 + \mathrm{d}x_1, x_2) &\mapsto (y_1 + a, y_2 + b) \\ (x_1 + \mathrm{d}x_1, x_2 + \mathrm{d}x_2) &\mapsto (y_1 + a + c, y_2 + b + d) \\ (x_1, x_2 + \mathrm{d}x_2) &\mapsto (y_1 + c, y_2 + d) \end{aligned}$$

These points will not be arranged as a rectangle in general, but will be a parallelogram. The parallelogram can be understood geometrically as being formed by the two vectors (a, b) and (c, d)

extending from (y_1, y_2) to form two adjacent sides with the other sides then being parallel and equal length to these. The proper scaling of the density $f_Y(y)$ will then depend on the relative area of this parallelogram to the original rectangle.

The following figure shows a parallelogram where lower left corner corresponds to the point (y_1, y_2) and the two adjacent sides are described by the vectors (a, b) and (c, d). In this figure, $a, b, c, d > 0$, which corresponds to all of the partial derivatives $\partial h_i/\partial x_j$ being positive. In addition, $a > c$ and $d > b$ so that $ad > bc$. The following geometric argument relies on these choices, but the result will be true in general.

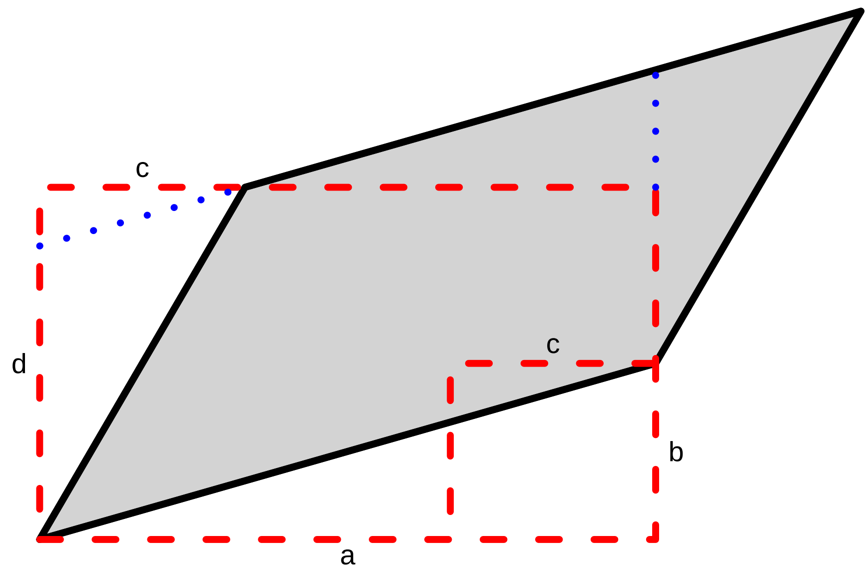

Figure 3: **Parallelogram after transformation.**

There are two rectangles with dashed lines added to the figure. The larger of these rectangles has width a and height d and the smaller one has width c and height b. In addition, there are two small dotted lines added to the figure which create six triangles and two larger polygons. Notice that the six triangles come in pairs which are the same size and orientation. Each pair includes one shaded triangle within the parallelogram and one outside. When the shading of the triangles are reversed, we get the following figure. The total shaded area is the same and is equal to $|ad - bc|$ as it is the difference in the areas of the rectangles.

Thus, the area of the parallelogram depends only on the lengths and orientations of the vectors (a, b) and (c, d). When these vectors are combined to form a matrix, we see that the area is equal to the absolute value of the determinant of this matrix.

$$|ad - bc| = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

In fact, the absolute value of this determinant measures the area of the corresponding parallelogram for any real a, b, c, d which can be shown by working through all cases. If we substitute back in our original expressions, we see that the area of the parallelogram is

$$|ad - bc| = \left(\frac{\partial h_1}{\partial x_1}\mathrm{d}x_1\right)\left(\frac{\partial h_2}{\partial x_2}\mathrm{d}x_2\right) - \left(\frac{\partial h_1}{\partial x_2}\mathrm{d}x_2\right)\left(\frac{\partial h_2}{\partial x_1}\mathrm{d}x_1\right) = |J|\mathrm{d}x_1\mathrm{d}x_2$$

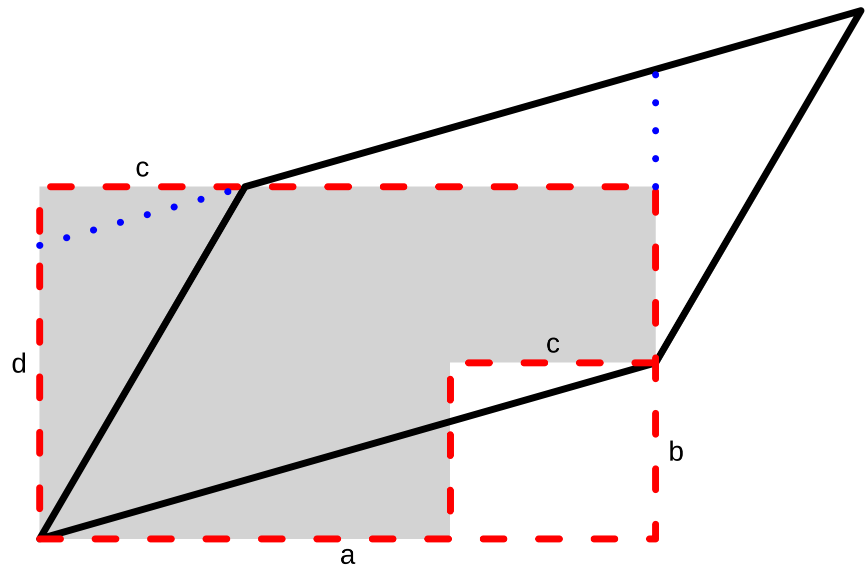

Figure 4: **Equal area.** The area of the parallelogram is equal to the difference in the areas of the rectangles.

where

$$J = \det \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} \end{pmatrix}$$

is called the *Jacobian* or *Jacobian derivative* of the transformation. The ratio of the area of the parallelogram to the area of the original rectangle is J and it follows then that the joint density of the random variables Y_1 and Y_2 is

$$f_Y(y_1, y_2) = \frac{1}{|J(h^{-1}(y_1, y_2))|} f_X(h^{-1}(y_1, y_2)) \ .$$

More than two dimensions. It is natural to then ask how this extends to joint distributions of n random variables. The answer is that the density requires a rescaling which is found by calculating the reciprocal of the absolute value of the Jacobian derivative for this larger transformation which is simply a determinant of a larger matrix of partial derivatives.

The derivation above found the Jacobian deriavative by computing $\partial y_j / \partial x_i$ for each i, j, but it is also possible to take derivatives of the inverse relationships $\partial x_j / \partial y_j$ and find the corresponding Jacobian deriavative. The value of this second derivative is the reciprocal of the first. In order to better distinguish these cases, it is useful to introduce a different notation that make the direction of differentiation clear. We can define the Jacobian derivative as follows.

$$\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} = \det \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

The density of $Y = (Y_1, \ldots, Y_n)$ can then be computed by finding one of two Jacobian derivatives.

$$\begin{aligned} f_Y(y_1, \ldots, y_n) &= \frac{1}{\left|\frac{\partial(y_1,\ldots,y_n)}{\partial(x_1,\ldots,x_n)}\right|} f_X(h^{-1}(x_1, \ldots, x_n)) \\ &= \left|\frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)}\right| f_X(h^{-1}(x_1, \ldots, x_n)) \end{aligned}$$

If you simply memorize the expression

$$f_Y(y_1, \ldots, y_n)\partial(y_1, \ldots, y_n) = f_X(x_1, \ldots, x_n)\partial(x_1, \ldots, x_n)$$

you can rerrange this algebraically to find either Jacobian and then properly use it or its reciprocal to find the desired density after the transformation.