On interpreting and extracting information from the cumulative distribution function curve: A new perspective with applications

Uditha Balasooriya, Jackie Li & Chan Kee Low
Nanyang Technological University, Singapore
<auditha@ntu.edu.sg>
<jackieli@ntu.edu.sg>
<acklow@ntu.edu.sg>

For any density function (or probability function), there always corresponds a *cumulative distribution function* (cdf). It is a well-known mathematical fact that the cdf is more general than the density function, in the sense that for a given distribution the former may exist without the existence of the latter. Nevertheless, while the density function curve is frequently adopted as a graphical device in depicting the main attributes of the distribution it represents, the cdf curve is usually ignored in such practical analysis.

By looking at a density function curve, we instantly obtain a visual decomposition of the shape of the underlying distribution (e.g., whether it is symmetric or skewed, either to the right or left, long-tailed or short-tailed, etc). By inspecting the cdf curve in the usual way, however, we do not receive a comparable amount of impact. In fact, given only the cdf curve, most people would tend to mentally convert it into the corresponding density function curve and then try to visualise its characteristics.

Can the cdf curve be more fruitfully utilised as a graphical device? In this paper, we show that the region above a cdf curve can be interpreted as an aggregate value of the underlying random variable. This perspective would facilitate the graphical display of the information contained in the distribution. We also exploit this approach to give intuition to the derivation of some well-known results.

For certain problems, this approach can be more advantageous than the usual treatment. For example, the Lorenz curve is typically used to illustrate income inequality. As shown later, the cdf curve reveals the same information as in the Lorenz curve, and additionally, it gives a better visual feel for the extent of income inequality.

Apart from its use in practical analysis, this approach of viewing the cdf also has pedagogical value. We introduced it in a few statistics related courses (first and second year of university) to the students who have earlier been exposed to the cdf in the usual manner. From the feedback during tutorial discussions, the students generally appreciate the insights provided by this new perspective and value the additional information that is usually not discussed. This concept may also be useful in teaching some Years 11 and 12 mathematics courses, which introduce a probability distribution with the usual focus on its density function only.

An alternative view of the cdf curve

The main reason of not gaining much information about the distribution via the cdf curve is that people tend to focus on the ordinate of the curve at a given point. We now consider the cdf curve from a different point of view. Suppose we have 100 straight iron rods of uniform shape but different lengths, and we want to portrait the distribution of the lengths. Let us arrange the iron rods in an ascending order of length and pile them up, as in Figure 1. Consider a point x on the horizontal axis. There are 85 iron rods with their lengths less than or equal to x and so we have $\Pr(X \le x) = 0.85$, where X denotes the length as a variable. The curve formed by the lower sides and the right edges of the iron rods is effectively a discretised approximation of the cdf curve. Note that there is an error of approximation in estimating the percentile of X, and this error can be reduced by increasing the number of rods and reducing the width of each rod. For any x, there is always a width such that x corresponds to the length of a rod. The corresponding percentile is then the proportion of rods that have their lengths not greater than x.

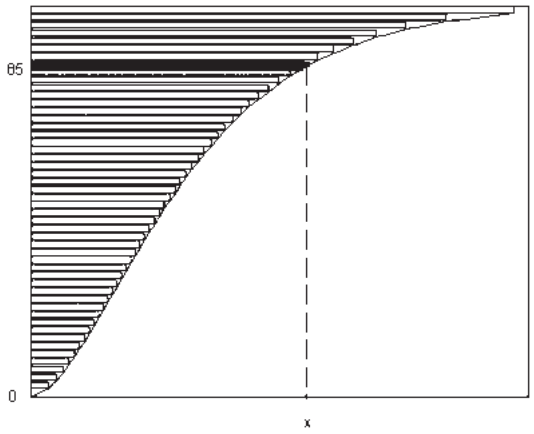

Figure 1. Distribution of 100 iron rods.

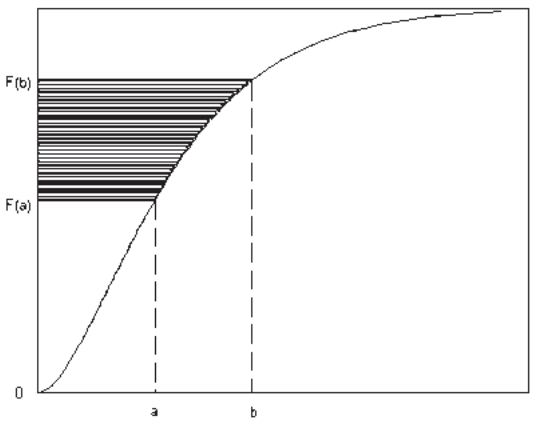

Figure 2. Aggregate value represented as region.

In this view of counting the number of rods with different lengths, any region formed by the horizontal bundles in Figure 1 represents the aggregate value of the corresponding elements in the sample space. As another example, consider a cdf curve of the incomes of a large number of individuals. Then the region formed by a horizontal strip, such as the shaded portion in Figure 2, divided by the total region above the cdf curve, would represent the share of the total income received by those individuals having an income between a and b. In contrast, the usual view would centre on the probability associated with these individuals.

This approach of interpreting the cdf curve is potentially useful in analysing various kinds of economics, population, and insurance data which have positive values, such as income level, future lifetime, and individual claim size. More examples are provided in the later sections. In addition, one may argue that this approach has some correspondence with the relationship $\mathrm{E}(X) = \int_0^\infty \big(1 - F(x)\big)\,dx$, in which the expected value is given by the sum of the areas of the vertical bars covering the region above the cdf curve.

Properties of some distributions

We now attempt to deduce the main features of some distributions using the approach above. First, note that the cdf curve of a degenerate distribution at a point $x = a$ could be interpreted, in the usual view, as a jump function with only one jump at a. In the alternative view, we would say that all the elements have the same value, represented by horizontal lines of the same length all ending at a.

In Figure 3, the usual view would say that the cdf curve is steep around the interval (a, b) and so the density function has a high value there. In the alternative view, we would take a look at how the horizontal lines meet the cdf curve. The lengths of these lines, and hence the values of the corresponding elements in the sample space, change rather slowly around this interval, which means there is a high probability here relative to the unit length of the x-axis.

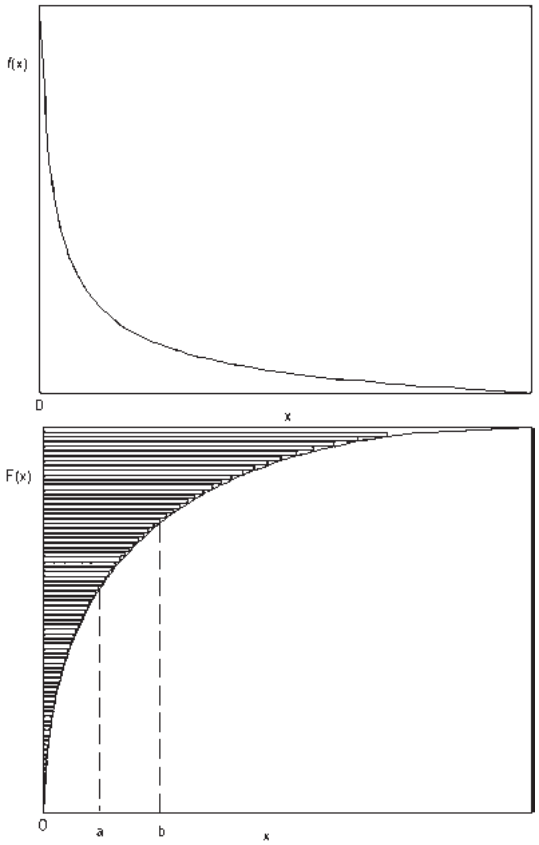

Figure 3. Density and cdf curves of a J-shaped distribution.

Figure 4 shows the density and the corresponding cdf curves of a uniform distribution. The usual view would refer to the constant slope of the cdf curve, which implies the uniformity of the density. The alternative view would regard the individual values, as shown by the horizontal lines, as changing evenly throughout the range.

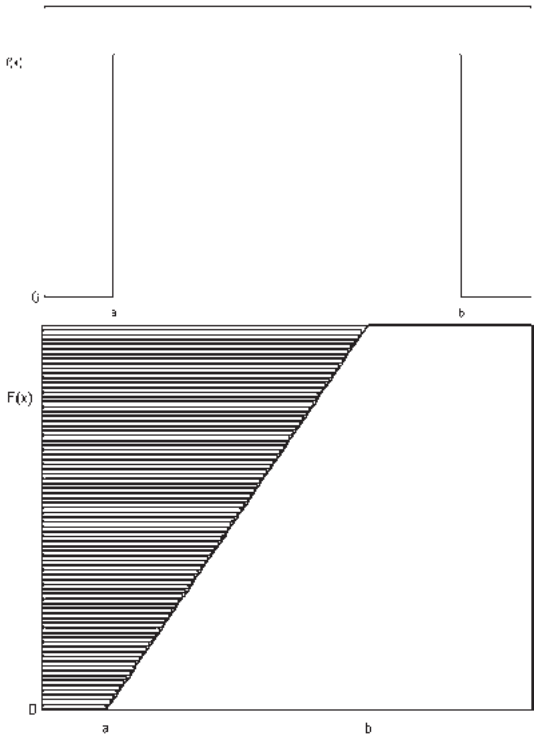

Figure 4. Density and cdf curves of a uniform distribution.

Figure 5 demonstrates a typical right-skewed distribution. The alternative view would suggest that the length of the horizontal lines changes rather quickly at the start, then more slowly in the middle range, and ultimately more quickly (than the start) in the upper end with large values. For a long-tailed distribution as in Figure 6, the length of the horizontal lines changes rapidly at the long tail. The tail behaviour is very important in many applications such as waiting time to failure and long-tailed insurance claims.

In general, a *J*-shaped distribution (with highest frequency initially and decreasing frequency afterwards as in Figure 3) has a cdf curve whose values change first slowly and then more quickly. Likewise a *U*-shaped

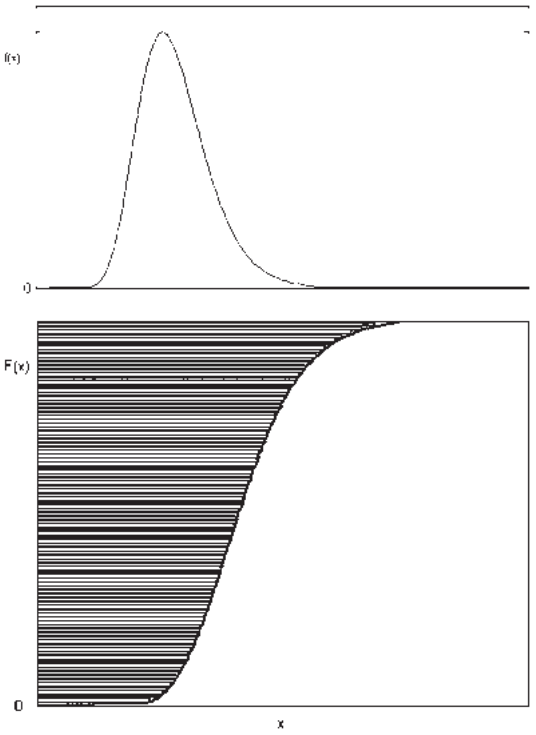

Figure 5. Density and cdf curves of a right-skewed distribution.

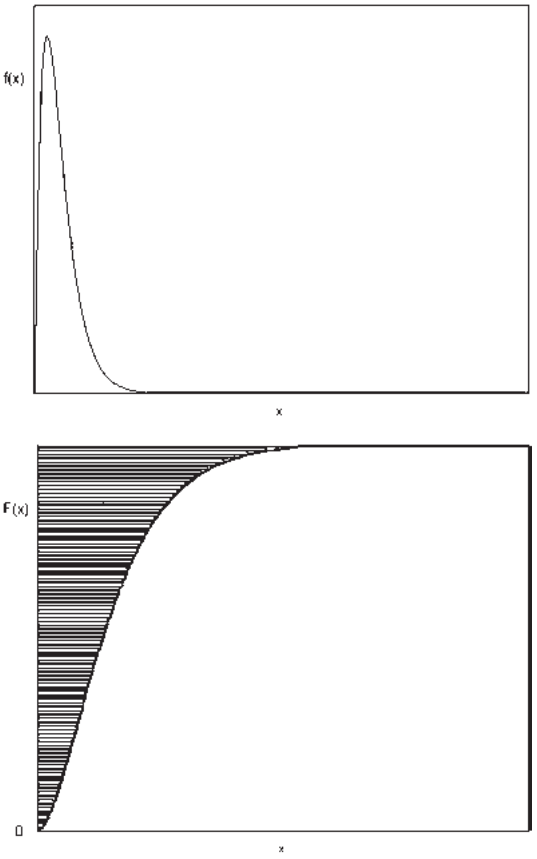

Figure 6. Density and cdf curves of a long-tailed distribution.

distribution (e.g., beta distribution with both parameters smaller than one) has a cdf curve whose values also change slowly at first and more quickly afterwards, but then change slowly again towards the upper end.

Depicting income concentration—an application

Figure 7 shows the distribution of Australian household income in 2005-06, expressed in Australian dollars per week, as reported to the Australian Bureau of Statistics. The aggregate value associated with the lowest 30% income group and that associated with the highest 4.5% income group are represented by the two shaded regions. The equality of the areas of these two regions means that the lowest 30% group earns as much as the highest 4.5% does. Comparisons of other proportions can be made similarly. For example, the broken line at $F(x) = 0.596$ divides the total value, as represented by the entire region northwest of the cdf curve, into two halves. Thus, the total household income is shared equally between the poorest 59.6% and the richest 40.4% of Australian households. This way of inspecting the graph provides us a convenient means of gaining insights into the concentration or inequality of income among Australian households.

The existing method of displaying income concentration graphically is almost exclusively via the celebrated Lorenz curve. See Stuart and Ord (1994) for the definition and properties. The Lorenz curve which corresponds to the distribution in Figure 7 is shown in Figure 8. If both the total frequency and the total value, as represented by the whole region northwest of the cdf curve, are scaled to 1, then a typical point P on the Lorenz curve would have its horizontal coordinate equal to the frequency at S and its vertical coordinate equal to the lower shaded area in Figure 7.

A graph like that in Figure 7 should be a good supplement of, or an alternative to, the Lorenz curve in showing such phenomena as income inequality. It displays directly, in terms of area, the contrast between the individuals having high income and those with low income.

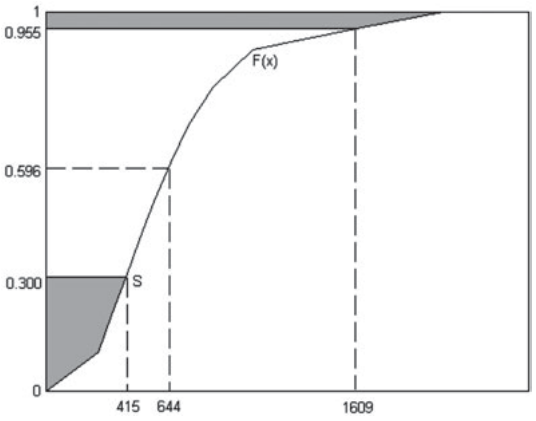

Figure 7. Cdf curve of Australian income.

The relationship between two random variables

The relationship between two random variables, and hence between their distribution functions, are easy to visualise using the same approach as above. For example, Figure 9 shows the cdfs $F(x)$ of a random variable X and $G(x)$ of the random variable $X + a$. One can immediately see via the horizontal lines that $G(x)$ can be obtained from $F(x)$ through extending the $F(x)$ curve horizontally by an amount a. Likewise, Figure 10 shows the cdfs $F(x)$ of X and $H(x)$ of the random variable cX. Here each horizontal element has its value extended, or shrunk, by a multiple c. This interpretation would work well only for monotone transformation of the random variable.

A random variable Y is said to be stochastically greater than another, X, if $\Pr(Y \geq a) > \Pr(X \geq a)$ for any a. Then, in terms of distribution functions, the smaller random variable X would have its cdf $F(.)$ greater than the cdf $G(.)$ of the greater random variable Y at any point, which is not a very convenient relationship to grasp. As shown in Figures 9 and 10, examining the horizontal lines has more intuitive appeal in interpreting such relationships.

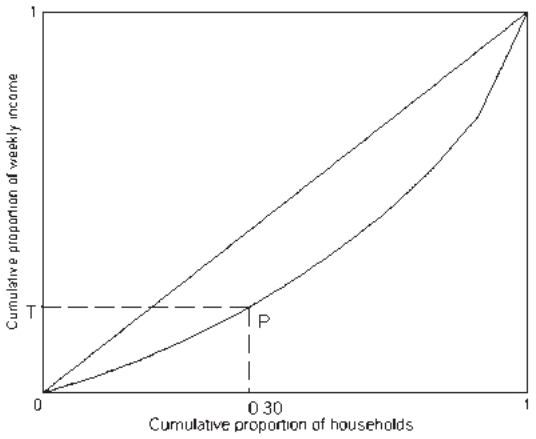

Figure 8. Lorenz curve for Australian income.

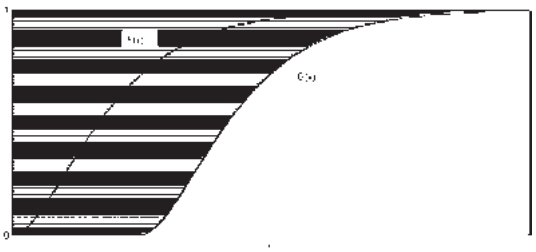
Figure 9. Cdfs F(x) of X and G(x) of X + a.

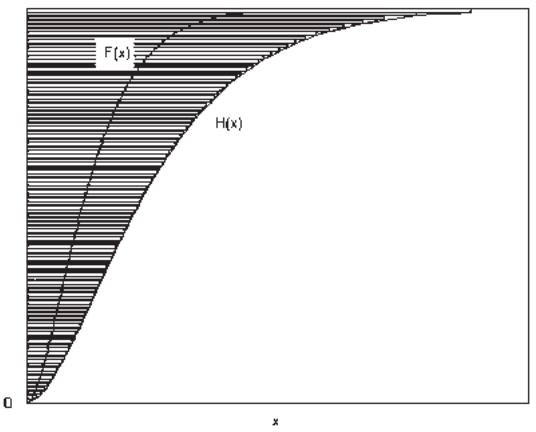

Figure 10. Cdfs F(x) of X and H(x) of cX.

Some expectation formulae

One of the concepts that a student of statistics first encounters is the expected value of a random variable. It is often presented as a measure of the central tendency of a

distribution. In classical mechanics, this is analogous to the centre of mass. If we visualise the dot diagram of a dataset as a uniform horizontal bar on which equal weights are placed at the positions of the data points, the expected value represents the balancing point of the bar. Another graphical representation of the expected value corresponds to the area above the cdf curve.

In Figure 11, the area above the cdf curve can be approximated by the sum of the areas of the n horizontal strips. For the strip corresponding to x_i, the area is given by:

$$x_i \left(F(x_{i+1}) - F(x_i)\right) .$$

Summing over all the strips, we obtain the approximate area above the cdf curve as:

$$\sum_{i=1}^{n} x_i \left(F\left(x_{i+1}\right) - F\left(x_i\right)\right) .$$

As we reduce the width of each strip and increase the number of strips, this sum approaches the exact area above the cdf curve:

$$\lim_{n\to\infty} \sum_{i=1}^{n} x_i \left(F\left(x_{i+1}\right) - F\left(x_i\right)\right) = \int_0^\infty x \; dF\left(x\right) .$$

It is well known that the last expression in the above formula is the expected value of X. Similarly, the area above the cdf curve can also be derived by summing up vertical, instead of horizontal strips. In this case, we obtain the alternative representation of the expected value:

$$\mathrm{E}\left(X\right) = \int_0^\infty \left(1 - F\left(x\right)\right) \, dx .$$

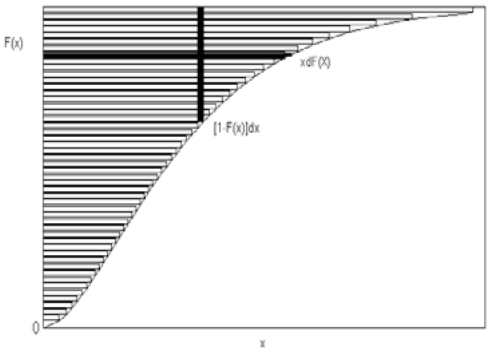

Figure 11. $\int x \, dF(x) = \int (1 - F(x)) \, dx$.

In mortality studies, the complete expectation of life (i.e., expected future lifetime or life expectancy) of a person aged exactly x is calculated as $\int_0^\infty t \; dF_x\left(t\right)$, where $F_x(t)$ is the cdf of the future lifetime of this person. For example, the life expectancy at birth (i.e., $x = 0$) in 2007 for Australia and New Zealand are computed as 81.5 and 80.3 respectively, based on the data from Human Mortality Database and Statistics New Zealand. Accordingly, the region above the cdf curve can be seen as horizontal strips representing the lifetimes of a large group of individuals of the same age, and its area refers to the expected lifetime of these individuals.

Sometimes we encounter a random variable X which is truncated above at a constant c:

$$Y = \begin{cases} X & \text{if } X \le c \\ c & \text{otherwise.} \end{cases}$$

For example, an insurance cover may compensate only up to an amount c for a loss X. Then the expected value of Y (i.e., expected claim size to the insurer) is:

$$\mathrm{E}(Y) = \int_0^c x \; dF(x) + c \; (1 - F(c)) = \int_0^c (1 - F(x)) \; dx$$

This idea is illustrated in Figure 12. The integral in the middle equation above corresponds to the area of the region occupied by horizontal strips and the next expression in the equation corresponds to the shaded area. Thus, the expected value of a truncated random variable can also be linked to a particular region above the cdf curve. This region can also be seen as horizontal strips representing the individual claim sizes, whose values are either below c or capped at c.

The situation is similar when a random variable X is truncated from below at a constant d:

$$Z = \begin{cases} 0 & \text{if } X \le d \\ X - d & \text{otherwise.} \end{cases}$$

In Australia, a motor insurance cover normally has an excess level d, in which the insurer pays only the amount in excess of d for a loss X. Then the expected value of Z (i.e., expected claim size to the insurer) is:

$$\mathrm{E}(Z) = \int_d^\infty (x - d) \; dF(x) = \int_d^\infty x \; dF(x) - d \; (1 - F(d))$$

which can also be interpreted as a particular area or region above the cdf curve.

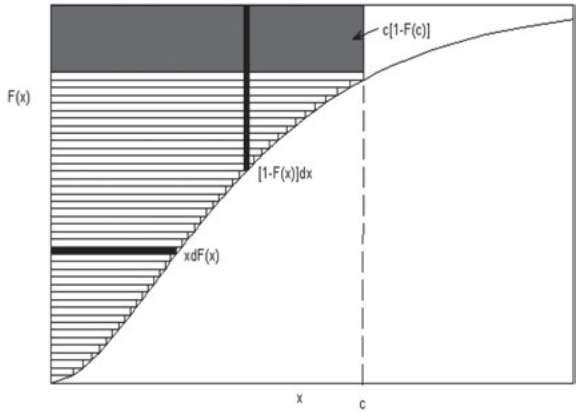

Figure 12. $\int_0^c x \; dF(x) + c \; (1 - F(c)) = \int_0^c (1 - F(x)) \; dx$.

A theorem in decision theory

In statistical estimation theory, many methods have been proposed for estimating the parameters of a distribution. The commonly used methods include the ordinary least squares, maximum likelihood, and method of moments. In the Bayesian approach, estimation is viewed as a decision-making problem. Given an observed random sample, $x_1, x_2, \ldots, x_n$, from some density function, the statistician has to decide what the estimate of the unknown parameter to be. One might then call the value of some estimator a decision and the estimator itself, $t = g(x_1, x_2, \ldots, x_n)$, decision function since it tells us what decision to make. The estimate is likely to be subject to error, so some measure of the severity of the error seems appropriate. In the literature, the word 'loss' is used in place of 'error' and 'loss function' is used as a measure of the 'error'. Given a loss function, the Bayes estimator seeks a decision that minimises the expected loss, where the expectation is taken with respect to a statistical distribution for the unknown parameter, known as the posterior distribution of the parameter. That is, the unknown parameter is treated as a random variable conditional on the data collected. For further details on Bayesian estimation, refer to Mood, Graybill, and Boes (1974).

Consider the Bayes estimator of the unknown parameter θ where the loss function, $l(t;\theta)$, for a given estimator t is:

$$l\ (t\ ;\theta) = \begin{cases} a\ (\theta - t) & \text{if } t \le \theta \\ b\ (t - \theta) & \text{otherwise.} \end{cases}$$

The constants a and b can be defined to permit different amount of penalty for under- and over-estimation of θ. For example, if $a = b$, under- and over-estimation of θ are considered to be equally serious, whereas if $a > b$ ($a < b$), under-estimation is considered to be more (less) serious with a relatively heavier (lighter) penalty imposed on the error.

Suppose the posterior density function is given as $f(\theta)$. The Bayes estimate is the value that minimises the expected loss, that is, it is the t such that $\mathrm{E}(l(t;\theta))$ is a minimum, where the expectation is taken with respect to $f(\theta)$. Using the previous graphical approach, we can provide an alternative illustration that this minimum is attained at the

$$\frac{a}{a+b}$$

quantile of the posterior distribution.

First, following the definition of expectation, the expected loss can be expressed as:

$$\mathrm{E}\left(l\ (t\ ;\theta)\right) = b\int_0^t (t-\theta)\, f(\theta)\ d\theta + a\int_t^\infty (\theta - t)\, f(\theta)\ d\theta$$

The two integrals above can indeed be represented by the areas of the shaded regions below and above the cdf curve in Figure 13, respectively. To understand this view, the first integral is expressed as:

$$\int_0^t (t-\theta)\, f(\theta)\ d\theta = \int_0^t t\, f(\theta)\ d\theta - \int_0^t \theta\, f(\theta)\ d\theta = t\ F(t) - \int_0^t \theta\ dF(\theta)$$

Referring to Figure 13, $F(t)$ is equal to the distance oq, and so $t\,F(t)$ is the area of the rectangle $oqpt$. Next, the second integral $\int_0^t \theta\ dF(\theta)$, by the same argument given in the previous section, is the area of that portion of the rectangle $oqpt$ that lies above the cdf curve. Hence, $\int_0^t (t-\theta) f(\theta)\ d\theta$ is the shaded area opt. In a similar fashion, one can establish that $\int_t^\infty (\theta - t) f(\theta)\ d\theta$ is the upper shaded area above the cdf curve in Figure 13.

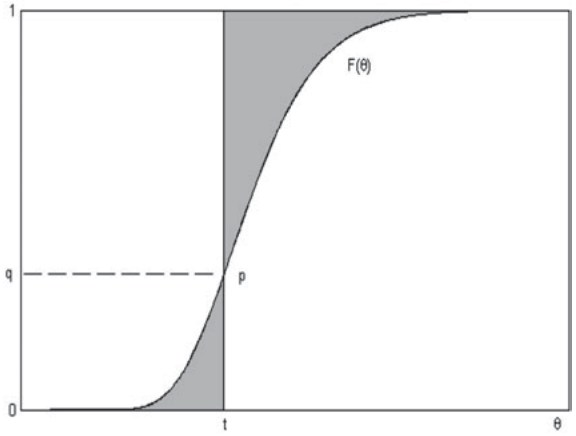

Figure 13. $\int_0^t (t-\theta) f(\theta)\ d\theta$ *and* $\int_t^\infty (\theta - t) f(\theta)\ d\theta$

Suppose $a = 1$ and $b = 2$, so that $\dfrac{a}{a+b} = \dfrac{1}{3}$.

Let t be the (1/3)rd quantile in Figure 14. The expected loss corresponding to the (1/3)rd quantile of the estimate of θ is determined by the two shaded regions. Suppose t is to be increased to $t + h$. Then $E(l(t;\theta))$ is increased by twice the area of A (the marked region below the cdf curve) and decreased by the area of B (the marked region above the curve). The net change in $E(l(t;\theta))$ would be:

$$2\left(\frac{1}{3}h + \quad h\right) - \left(\frac{2}{3}h - \quad h\right) = 3 \quad h$$

which is a non-negative quantity, with Δh being the triangle-like area shown in the graph. A similar argument with the help of the graph shows that, if t is decreased by an amount h, there will again be an increase in $E(l(t;\theta))$. Thus the expected loss is at a minimum when the estimate t is the $a/(a+b)$ quantile.

The same graphical technique has been applied to demonstrate an optimality property of the median, namely that the mean absolute deviation is at a minimum when evaluated from a median. Note that this method is applicable to a posterior density function of any form, including discrete and mixed. Comparatively, the formal mathematical proof is not straightforward. For difficulties arising in the algebraic proof of the optimality property of the median, see Lee (1995).

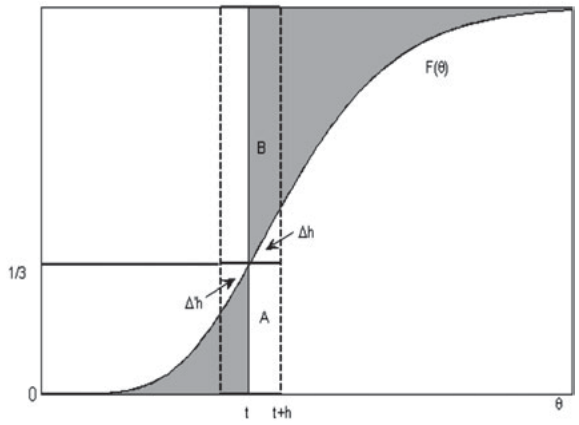

Figure 14. Expected loss is minimum at a specific quantile.

Concluding remarks

In this paper, we demonstrate an alternative approach to extract information from the cdf curve. Instead of the usual convention of converting the cdf curve into the density function curve, we show how the former can be more fruitfully utilised to describe the key features of the underlying distribution. As illustrated in the several examples above, this approach is potentially useful in a number of applications, such as economic analysis, mortality studies, and insurance claims modelling. Regarding teaching of statistics, it can also serve as a supplementary graphical device to help students understand better the rationale of some well-known results by gaining more intuitive insights from the cdf curve.

References

Human Mortality Database. (2010). University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). www.mortality.org.

Lee, Y. S. (1995). Searching for the right sample size. *The American Statistician, 49*(4), 369–372.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (Vol. 2). New York: McGraw Hill.

Statistics New Zealand. (2010). *Infoshare*. www.stats.govt.nz.

Stuart, A. & Ord, J. K. (1994). *Kendall's advanced theory of statistics* (Vol. 1, Chapter 2). London: Hodder Arnold.

Acknowledgments

The authors thank the editors and the anonymous referees for their helpful comments. This work was partially supported by Nanyang Technological University AcRF Grant, Singapore.