The paradox of the two boxes

Rémi Peyre

December 7, 2005

1 A benevolent genius

Let's imagine a benevolent genius says to me : "Here are two boxes A and B. Each of them contains a non-zero amount of money, and one of the boxes contains exactly twice as much money as the other. Choose the box you want; its content will be for you". I am about to take box A, when the genius asks me: "Are you sure?"

First I take the following reasoning: "The two boxes play the same role, so there is no reason why I should modify my choice".

But then I tell myself: "Suppose that the box A is transparent, and that I see £x in it. Then, as I don't know which box is the good one, there is one chance in two that B contains £$\frac{x}{2}$, and one chance in two that it contains £$2x$, which gives a mean £$\frac{5}{4}x > x$ in B, and thus I would always be well advised to change boxes.

How can we explain that paradox?

2 Formal resolution of the paradox

Let's call x the amount of money in the box A, and y the amount of money in B. (x, y) can be considered as a random variable, call P its law. P is supported by the half-lines $\{(x, y) : x = 2y \text{ and } y > 0\} \cup \{(x, y) : y = 2x \text{ and } x > 0\}$. Moreover P is symmetric w.r.t. switching x and y.

First, I claim that it is possible that we have $E[y|x] > x$, which means, $E[y|x] \geq x$ a.e. and $E[y|x] \neq x$. To prove that, just consider the following example : let $x_0 > 0$ and $\alpha \in (1, 2)$. We take

$$P(y = 2x \text{ and } x \in [x + \mathrm{d}x]) = \frac{\alpha}{2{x_0}^{\alpha-1}} \frac{1}{x^\alpha} \mathbf{1}_{x \geq x_0} \mathrm{d}x$$

and thus

$$P(x = 2y \text{ and } y \in [y + \mathrm{d}y]) = \frac{\alpha}{2{x_0}^{\alpha-1}} \frac{1}{y^\alpha} \mathbf{1}_{y \geq x_0} \mathrm{d}y$$

Then, we check easily that

$$E[y|x] = 2x \mathbf{1}_{x_0 \leq x < 2x_0} + \frac{2^{\alpha-2} + 2}{2^{\alpha-1} + 1} x \mathbf{1}_{x \geq 2x_0} > x$$

Notice that $E[x] = +\infty$ in that example. Actually, that is unavoidable, as I now claim :

Theorem 1 *If $E[x] < \infty$ then it is impossible that $E[y|x] > x$.*

Proof. By symmetry $E[y] = E[x]$. Suppose then that $E[x] < \infty$, thus $y - x$ in integrable and $E[y - x] = E[y] - E[x] = 0$. Now if $E[y|x] > x$ that can also be written $E[y - x|x] > 0$ and integrating we get $E[y - x] > 0$, a contradiction. □

3 Heuristic interpretation

Now we are able to explain with words what lead to a paradox in section 1. Actually, there are two possible interpretations :

- Either my expected gain by changing boxes is strictly positive indeed, but in that case my expected gain by choosing box A is already infinite. Thus there is no contradiction: adding something positive to an infinite quantity does not change it!
- Either my expected gain is finite. In that case, it turns that if the amount of money I can see in box A is big enough, then it is *not* in my interest to change boxes. For instance, imagine that the quantity of money in the boxes is bounded by M — in which case $E[x]$ is finite indeed —, then each time I see more than £$\frac{M}{2}$ in box A, I am ensured that it is the good box and thus I would lose money if I changed boxes.