CARL G. WAGNER

MISADVENTURES IN CONDITIONAL EXPECTATION: THE TWO-ENVELOPE PROBLEM

ABSTRACT. Several fallacies of conditionalization are illustrated, using the two-envelope problem as a case in point.

1. PREFERENCE AND EXPECTED VALUE

Most normative decision theories prescribe a preference for one act over another just when the expected utility of the former exceeds that of the latter. Expected utility, not utility *tout court*, because we are typically unsure about the exact circumstances in which our acts will be performed, and thus about their consequences.

The calculation of expected utility is often facilitated by exploiting the notion of conditional expectation, but this practice requires some care. We survey in this paper some cautionary examples. Our case in point involves two sums of money, unspecified except for the fact that one is twice as large as the other, which have been placed at random in red and blue envelopes. You may select one envelope and take the sum therein. Which should you choose?

We begin by verifying formally the *prima facie* intuition that there is no reason to prefer one envelope to the other, and that this is the case regardless of the particulars of your utility function over money. We observe that knowledge of the value of, say, the lesser sum would not disturb the prescription of indifference and that, more generally, this is so regardless of your prior probability distribution (if you have one) over the lesser sum. (An intriguing example of Broome (1995), which might be misconstrued to show otherwise, is discussed in §5.)

We then examine two arguments, the *tyro's argument* and the *expert's argument*, each purporting to show that it is preferable to choose the red envelope. It will be seen that each involves an abuse of conditionalization, the first committing what we call the *instantiation fallacy* and the second, devised by Jeffrey (1995), what he calls the *discharge fallacy.*

Erkenntnis **51:** 233–241, 1999.

Throughout the paper we employ a formulation of decision theory due to Savage (1954), in which acts are characterized by tabulating the numerical utilities of their consequences under each scenario in a probabilized set of possible states of the world. Thus acts are simply the familiar random variables of probability theory, and are ranked according to the magnitude of their expected values. In the next section we offer a brief review of pertinent results from probability theory, including the important notion of conditional expectation.

2. EXPECTATION AND CONDITIONAL EXPECTATION

Let $\mathcal{S}$ be a set (assumed here to be finite) of possible states of the world, equipped with a probability distribution P. A *random variable on* $\mathcal{S}$ is simply a function R which assigns to each state $s \in \mathcal{S}$ a real number $R(s)$. The *range of* R is the set of real numbers $\{R(s) : s \in \mathcal{S}\}$ and the *expected value* $E(R)$ *of* R is defined by

$$(1) \qquad E(R) := \sum_r rP(R = r),$$

where the sum in (1) is taken over all r in the range of R and $R = r$ is an abbreviation for the set $\{s \in \mathcal{S} : R(s) = r\}$.

If H is any subset of $\mathcal{S}$ with nonzero probability, we may define the *conditional expected value of* R, *given* H, denoted $E(R|H)$, by

$$(2) \qquad E(R|H) = \sum_r rP(R = r|H),$$

with the sum again taken over all r in the range of R, and

$$P(R = r|H) := P(R = r \cap H)/P(H).$$

In particular, if B is another random variable on $\mathcal{S}$, then for every b in the range of B such that $P(B = b) \neq 0$, the *conditional expected value of* R, *given that* $B = b$, denoted $E(R|B = b)$, is defined by

$$(3) \qquad E(R|B = b) := \sum_r rP(R = r|B = b).$$

Calculations of expected value are often facilitated by

THEOREM 1 (Total Expectation Theorem). If $\{H_1, H_2, \ldots, H_n\}$ is a family of pairwise disjoint sets, and $H_1 \cup H_2 \cup \ldots \cup H_n = \mathcal{S}$, then

$$(4) \qquad E(R) = \sum_i P(H_i)E(R|H_i),$$

with the sum taken over all i for which $P(H_i) \neq 0$.

Proof.

$$\begin{aligned} E(R) &= \sum_r rP(R=r) = \sum_r r \sum_i P(R = r \cap H_i) \\ &= \sum_r r \sum_i P(H_i)P(R=r|H_i) \\ &= \sum_i P(H_i) \sum_r rP(R=r|H_i) \\ &= \sum_i P(H_i)E(R|H_i). \square \end{aligned}$$

A special case of the above is

THEOREM 2 (Conditional Expectation Theorem). If R and B are random variables on $\mathcal{S}$, then

$$(5) \qquad E(R) = \sum_b P(B=b)E(R|B=b),$$

with the sum taken over all b in the range of B such that $P(B=b) \neq 0$.

An immediate consequence of this theorem is

COROLLARY 1 (Conditional Comparison Corollary). If $E(R|B=b) > b$ for every b in the range of B such that $P(B=b) \neq 0$, then $E(R) > E(B)$.

We are now prepared to pursue a careful analysis of the two-envelope problem.

3. TWO ENVELOPES

Two sums of money, unspecified except for the fact that one is twice as large as the other, have been placed in red and blue envelopes, the lesser sum first having been placed in one of the envelopes based on the toss of a fair coin, and the greater sum then having been placed in the remaining empty envelope. You may choose one of the envelopes and take the sum therein. Which envelope should you choose?

The *prima facie* intuition that there is no reason to prefer one envelope to the other is easily confirmed using any of the standard formulations of decision theory. Following Savage (1954), we would proceed as follows: Represent the sums in question by m and $2m$, and denote the set of possible states of the world by $\mathcal{S}$. Here $\mathcal{S} = \{s_1, s_2\}$, where $s_1 = m$ in the red envelope $\wedge$ $2m$ in the blue envelope, and $s_2 = 2m$ in the red envelope $\wedge$ m in the blue envelope. Under the terms of the problem, $P(s_1) = P(s_2) = 1/2$. Let the random variables R and B record, respectively, the amounts in the red and blue envelopes for each possible state of the world, so that $R(s_1) = B(s_2) = m$ and $R(s_2) = B(s_1) = 2m$. If your subjective utility of a given sum of money is identical with the numerical amount of that sum, then the triple $(\mathcal{S}, P, R)$ incorporates all relevant features of the act of choosing to take the contents of the red envelope,and the triple $(\mathcal{S}, P, B)$ all relevant features of the act of choosing to take the contents of the blue envelope.

A simple computation shows that $E(R) = E(B) = 1.5m$, confirming the *prima facie* intuition that you should be indifferent between the two envelopes. More generally, if your subjective utility over money is specified by the function υ, then it is the expected values of the utility-transformed random variables $\upsilon(R)$ and $\upsilon(B)$ that are relevant. For any real-valued function υ, another simple computation shows that $E(\upsilon(R)) = E(\upsilon(B)) = (\upsilon(m) + \upsilon(2m))/2$. Hence the prescription of indifference is robust, holding regardless of the particulars of your utility function over money.[1]

It is also worth noting that knowing the actual value of m would not alter your indifference between the envelopes. Then the less precise information furnished by a prior probability distribution over the possible values of m would surely leave the prescription of indifference undisturbed. An apparent counter-example to this assertion will be discussed in § 5.

In the next section, we analyze two arguments purporting to show, contrary to what was just demonstrated, that it is preferable to choose the red envelope.

4. TWO ARGUMENTS

An argument purporting to show that it is in fact preferable to choose the red envelope goes as follows:

I. The probability that the red envelope contains half the amount in the blue envelope is equal to $\frac{1}{2}$, and the probability that the red envelope contains twice the amount in the blue envelope is equal to $\frac{1}{2}$.

II. Hence, for each possible amount b in the blue envelope, the expected amount in the red envelope is $\frac{1}{2}(b/2) + \frac{1}{2}(2b) = 1.25b$, and this is greater than b, since each such b is positive.

III. Thus it is preferable to choose the red envelope.

Mutatis mutandis, this argument shows that it is preferable to choose the blue envelope, and so the argument is necessarily fallacious. The problem is to identify where the argument goes astray.

The argument is partly correct (which accounts for its superficial plausibility). In particular, proposition **I**, which which states that

$$(6) \qquad P(R = B/2) = P(R = 2B) = \tfrac{1}{2},$$

is true, since $P(R = B/2) = P(\{s \in \mathcal{S} : R(s) = B(s)/2\}) = P(s_1) = \frac{1}{2}$, and $P(R = 2B) = P(\{s \in \mathcal{S} : R(s) = 2B(s)\} = P(s_2) = \frac{1}{2}$. Moreover, proposition **II**, which states that

$$(7) \qquad E(R|B = b) = 1.25b > b, \text{ for } b \in \{m, 2m\},$$

does imply proposition **III**. For it follows from (7) by the Conditional Comparison Corollary that $E(R) > E(B)$. But **I** does not imply **II**, which is false ($E(R|B = m) = 2m$ and $E(R|B = 2m) = m$, as a simple computation shows). How then could the case be made for the false proposition **II**? It could only be made by confusing **I** with the stronger (but, as we shall see, false) assertion that, whatever amount is in the blue envelopė, the red envelope contains either half or twice this amount, with equal probability, i.e., with the assertion that

$$(8) \qquad P(R = b/2|B = b) = P(R = 2b|B = b) = \tfrac{1}{2}, \quad \text{for } b \in \{m, 2m\}.$$

Of course, if (8) were true, then for each $b \in \{m, 2m\}$ it would be the case that $E(R|B = b) = \frac{1}{2}(b/2) + \frac{1}{2}(2b) = 1.25b$, as asserted in (7). But, in fact, (8) is true neither for $b = m$ nor for $b = 2m$, since $P(R = m/2|B = m) = P(R = 4m|B = 2m) = 0$ and $P(R = 2m|B = m) = P(R = m|B = 2m) = 1$.[2]

To confuse (6) with (8) is to commit what might be called the *instantiation fallacy*, namely the confusion of the assertion

$$(9) \qquad P(R = \phi(B)) = p,$$

where ϕ is some real-valued function and p is some fixed probability, with the assertion that

$$(10) \qquad P(R = \phi(b)|B = b) = p,$$

for every b in the range of B. If (10) is true, then

$$\begin{aligned} P(R = \phi(B)) &= \sum_b P(B = b)P(R = \phi(b)|B = b) \\ &= \sum_b P(B = b)p = p. \end{aligned}$$

But, as the example at hand shows, (9) does not imply (10). It is this fallacy on which the argument for taking the red envelope founders.

That argument, in its pre-analyzed form, assumes minimal knowledge of probability. Call it the *tyro's argument*. One has, after all, only to grasp the simple weighted average in proposition **II** and to make the intuitively natural connection between **II** and the preferability of choosing the red envelope.[3] A different argument, call it the *expert's argument*, aims to entrap the more sophisticated. This argument, due to Jeffrey (1995) replaces proposition **II** by the computation **II(J)**:

$$\begin{aligned} E(R) &= P(R = B/2)E(R|R = B/2) \\ &\quad + P(R = 2B)E(R|R = 2B) \\ &= P(R = B/2)E(B/2) + P(R = 2B)E(2B) \\ &= \tfrac{1}{2}(\tfrac{1}{2}E(B)) + \tfrac{1}{2}(2E(B)) = 1.25E(B) > E(B). \end{aligned}$$

Here **III** follows immediately from **II(J)** and **I** is correctly interpreted as (6), so the problem lies with **II(J)**. The first line of this proposition is justified by the Total Expectation Theorem, and the third line by **I** and the linearity of expectation. As Jeffrey notes, the (subtle) defect occurs in the second line, in the unjustified equating of $E(R|R = B/2)$ with $E(B/2)$ and of $E(R|R = 2B)$ with $E(2B)$. These errors are instances of what Jeffrey calls the *discharge fallacy*, i.e., the mistaken belief that

$$E(\psi(R)|R = \phi(B)) = E(\psi(\phi(B))),^4$$

where ψ and ϕ are real-valued functions. Both the tyro's and the expert's arguments thus founder on fallacies of conditionalization.[5]

5. BROOME'S PARADOX

We noted in § 3 that knowing the actual value of m, the lesser of the two amounts placed in the envelopes, would not have changed the prescription of indifference regarding the choice of an envelope. Then surely the less precise information furnished by a prior probability distribution over the possible values of m would leave this prescription undisturbed. But consider the following intriguing example due to Broome (1995):

The lesser amount is chosen by a chance mechanism, taking the value 2^n with probability $2^n/3^{n+1}$ for each nonnegative integer n. The amount so chosen and twice that amount are then placed at random in the red and blue envelopes. Here the set of possible states of the world is $\mathcal{S} = \{(1,2), (2,1), (2,4), (4,2), \ldots, (2^n, 2^{n+1}), (2^{n+1}, 2^n), \ldots\}$, where (x, y) denotes the state in which the amount x is in the red envelope and the amount y is in the blue envelope, and $P((2^n, 2^{n+1})) = P((2^{n+1}, 2^n)) = 2^{n-1}/3^{n+1}$ for every n. Of course, $R((x, y)) = x$ and $B((x, y)) = y$, and the range of R and the range of B are both equal to the infinite set $I = \{1, 2, 4, \ldots, 2^n, \ldots\}$.

A computation shows that $E(R|B = 1) = 2$ and $E(R|B = b) = 11b/10$ for all $b \in I - \{1\}$, i.e., that

$$(11) \qquad E(R|B = b) > b, \ \text{for all } b \in I,$$

from which it might appear to follow from the Conditional Comparison Corollary that $E(R) > E(B)$. But that corollary does not warrant this inference, being restricted to cases where $\mathcal{S}$ is finite. Of course, there are substantial generalizations of the version of the Conditional Expectation Theorem and its corollary that sufficed for our purposes above.[6] But they warrant the inference from (11) to the inequality $E(R) > E(B)$ only when $E(R)$ and $E(B)$ are finite. Here, as Broome notes, the expected value of the lesser amount, $E(R)$, and $E(B)$ are all infinite.

Indeed, it had better not be the case that (11) entails the inequality $E(R) > E(B)$. For a further computation shows that $E(B|R = 1) = 2$ and $E(B|R = r) = 11r/10$ for all $r \in I - \{1\}$, i.e., that

$$(12) \qquad E(B|R = r) > r, \ \text{for all } r \in I,$$

which would then correspondingly entail the contrary inequality $E(B) > E(R)$. The lesson to be learned here is to keep steadfastly in mind that it is an inequality between unconditional expectations that is our criterion for preferring one act to another. While a certain family of inequalities involving conditional expectation can often serve as a surrogate for that criterion, this is not universally the case.[7]

ACKNOWLEDGEMENTS

Research supported by the National Science Foundation (SBR-9528893). I am grateful to Keith Lehrer for introducing me to the two-envelope problem and to Richard Jeffrey for sharing with me his manuscript *Probabilistic Thinking* (1995), and for bringing to my attention the paper of John Broome (1995).

NOTES

1 A slightly different version of the two-envelope problem, treated by Chihara (1995), admits of a similar analysis. In Chihara's version, all the money in some individual's pocket is placed in a white envelope, which is then placed, based on the toss of a fair coin, in either the red or the blue envelope. Instead of placing twice that amount in the remaining envelope, however, it is decided by another toss of the coin whether twice or half the amount in the white envelope goes in the remaining envelope. Here $\mathcal{S} = \{(\frac{w}{2}, w), (2w, w), (w, \frac{w}{2}), (w, 2w)\}$, where w is the amount in the white envelope, $(x, y) = x$ in the red envelope $\wedge y$ in the blue envelope, and each state has probability $\frac{1}{4}$. A simple computation shows that for any real-valued function v, $E(v(R)) = E(v(B)) = \frac{1}{4}v(\frac{w}{2}) + \frac{1}{2}v(w) + \frac{1}{4}v(2w)$. Chihara is under the mistaken impression that utility must be a linear function of money in order to justify indifference (see p. 9 and endnote 10 on p. 16 of his paper).

2 Condition (8) would hold if the envelopes were filled by a different procedure: A coin weighted so that it comes up 'heads' with probability p is tossed, and the sum m or $2m$ is placed in the blue envelope depending on whether this coin comes up 'heads' or 'tails.' The amount placed in the red envelope is then either half or twice the amount in the blue envelope, depending on whether the toss of a second, *fair* coin comes up 'heads' or 'tails.'

3 We know that the inference from **II** to the inequality $E(R) > E(B)$, and thus to **III**, is warranted by the Conditional Comparison Corollary, but our tyro need not.

4 This formula is true for the special case where $\phi(B) = c$, with c a constant. Then, as Jeffrey notes, $E(\psi(R)|R = c) = E(\psi(c)) = \psi(c)$. Thus the mistakes in **II(J)** might arise from confusing the random variables $B/2$ and $2B$ with constants. In his analysis of what appears to be a less formal version of the expert's argument, Chihara (1995) may be making something like the latter point, in his distinction between quantities "in the dollar sense" and "in the whatever-sense".

5 Both arguments have assumed for simplicity that the utility of a given sum of money is identical with the numerical amount of that sum. But the relevant versions of **II** and **II(J)** go through for any utility function v satisfying the inequality $\frac{1}{2}v(\frac{x}{2}) + \frac{1}{2}v(2x) > v(x)$. In particular, one may take $v(x) = x^r$ for *any* $r > 0$, and so this inequality is compatible with considerable risk aversion.

6 See, e.g., Bickel and Doksum (1977) or Parzen (1960).

7 This homily is directed at hypothetical sinners, not at Broome, who does not infer from (11) that $E(R) > E(B)$, and who clearly characterizes the conjunction of (11) and (12) as a paradox, not a contradiction. Note that if the amount in the blue envelope were revealed to you and you could take it or take whatever was in the red envelope, you would have good

reason for switching no matter what the amount in the blue envelope turned out to be. Yet, *a priori*, you would have no reason to prefer the red envelope(*cf.* Zabell (1988), where it is asserted that this could never be the case). Indeed, no matter what envelope was opened for your inspection, and no matter what the contents turned out to be, you would apparently have good reason to switch. Of course, one could object to the unrealistic assumption of *Heidengeld* underlying the paradox, or point out that it will not entrap the sufficiently risk averse (but see Broome (1995) at p.9 for a variant of the paradox immune to these objections).

REFERENCES

Bickel, P. and K. Doksum: 1977, *Mathematical Statistics*, Holden-Day, San Francisco, p. 6.

Broome, J.: 1995, 'The Two-Envelope Paradox', *Analysis* **55**(1), 6–11.

Chihara, C.: 1995, 'The Mystery of Julius: A Paradox in Decision Theory', *Philosophical Studies* **80**, 1–16.

Jeffrey, R.: 1995, *Probabilistic Thinking*. Princeton University.

Parzen, E.: 1960, *Modern Probability Theory and its Applications*, John Wiley, New York, p. 384.

Savage, L.: 1954, *The Foundations of Statistics*, John Wiley, New York.

Zabell, S.: 1988, 'Loss and Gain: The Exchange Paradox', In: J. Bernardo et al. (eds.), *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting*. Oxford, pp. 233–236.

Mathematics Department
University of Tennessee
121 Ayres Hall
Knoxville TN 37996, 1300
U.S.A.
E-mail: cwagner@utk.edu