Economics Letters (2017)
doi: 10.1016/j.econlet.2017.07.027

The Sure-thing Principle and P2

YANG LIU
University of Cambridge

This paper offers a fine analysis of different versions of the well known sure-thing principle. We show that Savage's formal formulation of the principle, i.e., his second postulate (P2), is strictly stronger than what is intended originally.

1. Introduction

Savage's theory of subjective expected utility (1972) is widely celebrated as *the* paradigmatic system of Bayesian decision theory. The cornerstone of Savage's theory is a rationality postulate known as the "sure-thing principle." The principle is formally stated in Savage's system as the second postulate (P2). In this short note, we point out that there is certain discrepancy between P2 and the sure-thing principle in its original form. We show that Savage's P2 is, in fact, a strictly stronger principle.

Recall that a *Savage decision model* is a structure of the form $(S, \mathcal{B}, X, \mathcal{A}, \succcurlyeq)$ where S is an (infinite) set of *states* of the world; $\mathcal{B}$ is a Boolean algebra equipped on S, each element of which is referred to as an *event* in a given decision situation; X is a set of consequences; and a (Savage) *act* is a function f mapping from S to X, the intended interpretation is that $f(s)$ is the consequence of the agent's action f performed when the state of the world is in s. As a primitive notion of the model, $\succcurlyeq$ is a binary relation on the set of all acts, denoted by $\mathcal{A}$. For any $f, g \in \mathcal{A}$, $f \succcurlyeq g$ says that f is *weakly preferred to* g. Say that f is *strictly preferred to* g, written $f \succ g$, if f is weakly preferred to g but not vice versa, and that f is *indifferent to* g, denoted by $f \sim g$, if f is weakly preferred to g and vice versa.

Definition 1.1 (combined acts)**.** For any $f, g \in \mathcal{A}$, define the *combination* of acts f and g with respect to any given event $E \in \mathcal{B}$, written $f|E + g|\overline{E}$, to be such that:

$$\left(f\middle|E + g\middle|\overline{E}\right)(s) =_{\text{df}} \begin{cases} f(s) & \text{if } s \in E \\ g(s) & \text{if } s \in \overline{E}, \end{cases} \tag{1.1}$$

where $\overline{E} = S - E$ is the compliment of E.

In other words, $f|E + g|\overline{E}$ is the act which agrees with f on event E, with g on $\overline{E}$. It is easily seen that $f|E + g|\overline{E} \in \mathcal{A}$. Normative constraints governing the behavior of the agent's preference relation $\succcurlyeq$ over acts were then introduced by Savage. Our focus here is on the "sure-thing principle," as well as its formal renderings.

2. STP and P2

The following is the example used by Savage to motivate this principle.

> A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate were going to win, and again finds that he would do so. Seeing that the would buy in either event, he decides that the should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say. (*ibid.* p. 20)

As illustrated by this example, the principle stems from an intuitive idea of *reasoning by cases* which says that if a decision maker is prepared to take certain course of action given the occurrence of some event and she will behave in the same manner if the event does not occur, then she shall proceed with *that* action. In other words, the implementation of this course of action is a "sure-thing". To state in terms of preferences over acts, the sure-thing principle says that

STP If the decision maker (weakly) prefers one act over another assuming either certain event or its compliment occurs, then her preference over the two acts shall remain unchanged.

The principle is sometimes referred to as the *dominance principle*, which can be generalized as follows: suppose that the state space is partitioned into n-many mutually exclusive events (usually representing n different possible decision scenarios), if one act weakly dominates another in each possible decision scenario, then the act is weakly preferred throughout.

Savage takes the consideration above as fundamental to rational decision making: "I," he says, "know of no other extra-logical principle governing decisions that finds such ready acceptance." (p. 21)

Note that in order to formalize this version of the sure-thing principle one needs to invoke a concept of *conditional preferences* – that is, the concept of one act being (weakly) preferred to another *given* the occurrence of certain event. Savage, however, was unwilling to incorporate this conception directly into his system for the concern that the notion of conditional preference may lead, according to him, to unsought philosophical complications.[1]

Instead, Savage sought to define conditional preferences in a roundabout way using *unconditional* preferences. To this end, he posited the following alternative approximation to **STP** which is officially stated as his second postulate for rational decision making:

1. Savage (1972, p. 22) explains: "The sure-thing principle [i.e., **STP** above] cannot appropriately be accepted as a postulate in the sense that P1 [i.e., the assumption that $\succcurlyeq$ is a complete preorder] is, because it would introduce new undefined technical terms referring to knowledge and possibility that would refer it mathematically useless without still more postulates governing these terms."

Table 1: Illustrations of

(a) Savage's P2

	E	$\overline{E}$
f	a	c
g	b	c
f'	a	d
g'	b	d

(b) conditional preference

	E	$\overline{E}$
f	a	c
g	b	d
f'	a	e
g'	b	e

P2 If the consequences of two acts differ in the states in some event E but otherwise agree with each other, then the preference between these two acts shall be determined by their differences in E.[2]

In symbols, **P2** says that for any acts $f, g, h, h' \in \mathcal{A}$ and for any event $E \in \mathcal{B}$,

$$f|E + h|\overline{E} \succcurlyeq g|E + h|\overline{E} \iff f|E + h'|\overline{E} \succcurlyeq g|E + h'|\overline{E}, \tag{P2}$$

Table 1(A) contains an illustration of this principle where $\{E, \overline{E}\}$ forms a simple partition of S for which $f(s) = a$ for all $s \in E$ and $f(s) = c$ for $s \in \overline{E}$ (other acts are defined similarly). Then (P2) mandates that if f and g agree with one another on $\overline{E}$ and if f' and g' agree, respectively, with f and g on E and with each other on $\overline{E}$, then $f \succcurlyeq g$ if and only if $f' \succcurlyeq g'$.

The next technical construction is what connects **STP** and **P2**.

Definition 2.1 (conditional preference)**.** Let E be some event, then, given acts $f, g \in \mathcal{A}$, f is said to be weakly preferred to g *given* E, written $f \succcurlyeq_E g$, if *for all* pairs of acts $f', g' \in \mathcal{A}$, if f' and g' agree, respectively, with f and g on E and with each other on $\overline{E}$ then $f' \succcurlyeq g'$. In symbols, we have that $f \succcurlyeq_E g$ if

$$\forall f', g' \in \mathcal{A}, \quad \left.\begin{array}{ll} f(s) = f'(s),\ g(s) = g'(s) & \text{if } s \in E \\ f'(s) = g'(s) & \text{if } s \in \overline{E} \end{array}\right\} \implies f' \succcurlyeq g'. \tag{2.1}$$

It is important to note the universal quantifier in (2.1). Table 1(B) provides an illustration of this definition: f is weakly preferred to g given E if $f' \succcurlyeq g'$ *for all* such f's and g's. As seen, (A) differs from (B) in that, in the case of conditional preferences, f and g do *not* need to agree with one another on $\overline{E}$ in order for f to be conditionally preferred to g given E. (This turns out to the precise technical detail that separates **STP** and **P2**, upon which Example 2.3 below is constructed.)

The crucial role **P2** plays in the definition of conditional preference above is thus to ensure that, for any two pairs of acts, say, (f', g') and (f'', g''), if they both satisfy the antecedent of the conditional in (2.1) then, by (P2), the consequent must follow in the

2. Savage (1967, p. 306) describes P2 as follows: "If two acts have the same consequences for some states, the preference between the two acts will not be changed if they are given new common consequences on those states where they are already in agreement and each is left unaltered elsewhere."

same manner (i.e., $f' \succcurlyeq g'$ iff $f'' \succcurlyeq g''$). This is Savage's peculiar way of circumventing conditional preference in his system using unconditional preference and **P2**. Let's put Definition 2.1 in the following concise form:

$$f \succcurlyeq_E g \;=_{\text{df}}\; f|E + h|\overline{E} \succcurlyeq g|E + h|\overline{E}, \;\text{ for all } h \in \mathcal{A}. \tag{CP}$$

Now, under this definition of conditional preference, the **STP** above can be translated straightforwardly into

$$\left[f \succcurlyeq_E g, f \succcurlyeq_{\overline{E}} g\right] \implies f \succcurlyeq g. \tag{STP}$$

Savage considers **P2** to be the *formal* version of the sure-thing principle incorporated in his decision theory, leaving **STP** itself as an *informal* – or, to use Savage's phrase, a "loose" version of the sure-thing principle (p. 22). At this point, it is not difficult to see that **P2** is a special case of **STP** (because on the part where two acts agree with one another, one act trivially weakly dominates the other.), hence, by that token, **P2** is a more restrictive principle than **STP**. In fact, we show that (P2) is strictly stronger than (STP) in Savage's system.

Proposition 2.2. Let $\succcurlyeq$ be a preorder on $\mathcal{A}$, then (P2) implies (STP).

Proof. Assuming (P2), then the antecedent of (STP) can be rewritten, via (CP), as

$$f|E + h|\overline{E} \succcurlyeq g|E + h|\overline{E} \tag{2.2}$$
$$f|\overline{E} + h'|E \succcurlyeq g|\overline{E} + h'|E. \tag{2.3}$$

where h and h' are arbitrary acts in $\mathcal{A}$.

Now, in (2.2), substitute h with $h|E + f|\overline{E}$, then, by (P2), we get

$$f = f|E + \big(h|E + f|\overline{E}\big)\big|\overline{E} \succcurlyeq g|E + \big(h|E + f|\overline{E}\big)\big|\overline{E} = g|E + f|\overline{E}$$

Similarly, Similarly, in (2.3), replace h' with $g|E + h'|\overline{E}$, then

$$f|\overline{E} + g|E = f|\overline{E} + \big(g|E + h'|\overline{E}\big)\big|E \succcurlyeq g|\overline{E} + \big(g|E + h'|\overline{E}\big)\big|E = g.$$

By transitivity of $\succcurlyeq$, we have $f \succcurlyeq g$, i.e., the consequent of (STP). □

Proposition 2.2 shows that **P2** is sufficient in bringing about **STP** in Savage's decision model *given* the definition of conditional preferences in (CP), and the latter is, in turn, regulated by **P2**. The converse of Proposition 2.2, however, does not necessarily hold. We show this by means of a simple example which, as we shall see, satisfies (STP) but not (P2).[3]

3. For relevant discussion see Gaifman (2013), where the author discusses the relationships between versions of STP and P2 in the context of "partial-act" systems (i.e., acts defined in terms of partial, instead of total, functions mapping from S to X). Our focus in this note is rather on Savage' original formulations with "total acts".

Example 2.3. Let $S = E \cup \overline{E}$ for some (non-null) events E and $\overline{E}$, and $X = \{a, b\}$. Consider the following four acts:

	E	$\overline{E}$
f_1	a	a
f_2	b	a
f_3	a	b
f_4	b	b

Suppose that $f_1 \succ f_2 \sim f_3 \prec f_4$. Now, for any pair of acts f_i, f_j ($i, j \in \{1, 2, 3, 4\}$), consider the following two cases

1. If $i = j$ then (STP) holds trivially.

2. If $i \neq j$, then it can be easily verified, by applying the definition of conditional preference in (2.1), that at least one of $f_i \not\succcurlyeq_E f_j$ and $f_i \not\succcurlyeq_{\overline{E}} f_j$ is the case. This means that, in all these cases, the antecedent of the conditional in (STP) is false, hence (STP) holds vacuously for f_i and f_j.

Hence, we have that (STP) holds in this example. But, on the other hand, (P2) is obviously violated. ◁

3. Conclusion

As remarked by Savage, the sure-thing principle *in its original form* (i.e., **STP**) is arguably the single most acceptable principle governing rational decision making, and it has indeed been widely adopted in various decision-theoretic analyses ever since its emergence. Savage's own formulation of the sure-thing principle, as we have seen, opts for a strictly stronger principle (i.e., his **P2**) due to his view towards conditional preferences. This approach, however, renders Savage's system a less general theory than it was intended to be. For future works, instead of using Savage's intertwined notions of (P2), (CP), and (STP), it would be of great interest to reconstruct Savage's theory from **STP** itself.

Acknowledgement.

Thanks are due to professor Haim Gaifman and an anonymous reviewer.

References

Gaifman, H. (2013). The sure thing principle, dilations, and objective probabilities. *Journal of Applied Logic*, 11(4):373–385.

Savage, L. J. (1967). Difficulties in the theory of personal probability. *Philosophy of Science*, 34(4):305–310.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover Publications, Inc., New York, second revised edition.