

Chapter 2

Towards Explaining Specific Examples of Suboptimal (and Often Seemingly Irrational) Human Decisions

In this chapter, we show that bounded rationality indeed explains the seemingly irrational human decisions. Specifically, in Sect. 2.1, we briefly recall the traditional utility-based approach to decision making. In Sect. 2.2, we describe our main idea—of optimization under granularity. In Sect. 2.3, we show how this idea can explain the first example of seemingly irrational decision making: the compromise effect. In Sect. 2.4, we provide an explanation for the second example of seemingly irrational decision making: use of bias probability estimates. Finally, in Sect. 2.5, we explain why people use average instead of the sum.

2.1 Traditional Utility-Based Approach to Decision Making: A Brief Reminder

Main assumption behind the traditional decision theory. Traditional approach to decision making is based on an assumption that for each two alternatives A' and A'' , a user can tell:

- whether the first alternative is better for him/her; we will denote this by $A'' < A'$;
- or the second alternative is better; we will denote this by $A' < A''$;
- or the two given alternatives are of equal value to the user; we will denote this by $A' = A''$.

Towards a numerical description of preferences: the notion of utility. Under the above assumption, we can form a natural numerical scale for describing preferences. Namely, let us select a very bad alternative A_0 and a very good alternative A_1 . Then, most other alternatives are better than A_0 but worse than A_1 .

For every probability $p \in [0, 1]$, we can form a lottery $L(p)$ in which we get A_1 with probability p and A_0 with probability $1 - p$.

- When $p = 0$, this lottery coincides with the alternative A_0 : $L(0) = A_0$.
- When $p = 1$, this lottery coincides with the alternative A_1 : $L(1) = A_1$.

For values p between 0 and 1, the lottery is better than A_0 and worse than A_1 . The larger the probability p of the positive outcome increases, the better the result:

$$p' < p'' \text{ implies } L(p') < L(p'').$$

Thus, we have a continuous scale of alternatives $L(p)$ that monotonically goes from $L(0) = A_0$ to $L(1) = A_1$. We will use this scale to gauge the attractiveness of each alternative A .

Due to the above monotonicity, when p increases, we first have $L(p) < A$, then we have $L(p) > A$, and there is a threshold separating values p for which $L(p) < A$ from the values p for which $L(p) > A$. This threshold value is called the *utility* of the alternative A :

$$u(A) \stackrel{\text{def}}{=} \sup\{p : L(p) < A\} = \inf\{p : L(p) > A\}.$$

Then, for every $\varepsilon > 0$, we have

$$L(u(A) - \varepsilon) < A < L(u(A) + \varepsilon).$$

We will describe such (almost) equivalence by \equiv , i.e., we will write that $A \equiv L(u(A))$.

How to elicit the utility from a user: a fast iterative process. Initially, we know the values $\underline{u} = 0$ and $\bar{u} = 1$ such that $A \equiv L(u(A))$ for some $u(A) \in [\underline{u}, \bar{u}]$.

On each stage of this iterative process, once we know values \underline{u} and \bar{u} for which $u(A) \in [\underline{u}, \bar{u}]$, we compute the midpoint u_{mid} of the interval $[\underline{u}, \bar{u}]$ and ask the user to compare A with the lottery $L(u_{\text{mid}})$ corresponding to this midpoint. There are two possible outcomes of this comparison: $A \leq L(u_{\text{mid}})$ and $L(u_{\text{mid}}) \leq A$.

- In the first case, the comparison $A \leq L(u_{\text{mid}})$ means that $u(A) \leq u_{\text{mid}}$, so we can conclude that $u \in [\underline{u}, u_{\text{mid}}]$.
- In the second case, the comparison $L(u_{\text{mid}}) \leq A$ means that $u_{\text{mid}} \leq u(A)$, so we can conclude that $u \in [u_{\text{mid}}, \bar{u}]$.

In both cases, after an iteration, we decrease the width of the interval $[\underline{u}, \bar{u}]$ by half. So, after k iterations, we get an interval of width 2^{-k} which contains $u(A)$ —i.e., we get $u(A)$ with accuracy 2^{-k} .

How to make a decision based on utility values. Suppose that we have found the utilities $u(A')$, $u(A'')$, ..., of the alternatives A' , A'' , ... Which of these alternatives should we choose?

By definition of utility, we have:

- $A \equiv L(u(A))$ for every alternative A , and
- $L(p') < L(p'')$ if and only if $p' < p''$.

We can thus conclude that A' is preferable to A'' if and only if $u(A') > u(A'')$. In other words, we should always select an alternative with the largest possible value of utility. So, to find the best solution, we must solve the corresponding optimization problem.

Before we go further: caution. We are *not* claiming that people estimate probabilities when they make decisions: we know they often don't. Our claim is that when people make *definite* and *consistent* choices, these choices *can* be described by probabilities. (Similarly, a falling rock does not solve equations but follows Newton's equations $ma = m \frac{d^2x}{dt^2} = -mg$.) In practice, decisions are often *not* definite (uncertain) and *not* consistent.

How to estimate utility of an action. For each action, we usually know possible outcomes S_1, \dots, S_n . We can often estimate the probabilities p_1, \dots, p_n of these outcomes.

By definition of utility, each situation S_i is equivalent to a lottery $L(u(S_i))$ in which we get:

- A_1 with probability $u(S_i)$ and
- A_0 with the remaining probability $1 - u(S_i)$.

Thus, the original action is equivalent to a complex lottery in which:

- first, we select one of the situations S_i with probability p_i : $P(S_i) = p_i$;
- then, depending on S_i , we get A_1 with probability $P(A_1 | S_i) = u(S_i)$ and A_0 with probability $1 - u(S_i)$.

The probability of getting A_1 in this complex lottery is:

$$P(A_1) = \sum_{i=1}^n P(A_1 | S_i) \cdot P(S_i) = \sum_{i=1}^n u(S_i) \cdot p_i.$$

In this complex lottery, we get:

- A_1 with probability $u = \sum_{i=1}^n p_i \cdot u(S_i)$, and
- A_0 with probability $1 - u$.

So, the utility of the complex action is equal to the sum u .

From the mathematical viewpoint, the sum defining u coincides with the expected value of the utility of an outcome. Thus, selecting the action with the largest utility means that we should select the action with the largest value of expected utility $u = \sum p_i \cdot u(S_i)$.

How uniquely determined is utility. The above definition of utility u depends on the selection of two fixed alternatives A_0 and A_1 . What if we use different alternatives A'_0 and A'_1 ? How will the new utility u' be related to the original utility u ?

By definition of utility, every alternative A is equivalent to a lottery $L(u(A))$ in which we get A_1 with probability $u(A)$ and A_0 with probability $1 - u(A)$. For

simplicity, let us assume that $A'_0 < A_0 < A_1 < A'_1$. Then, for the utility u' , we get $A_0 \equiv L'(u'(A_0))$ and $A_1 \equiv L'(u'(A_1))$. So, the alternative A is equivalent to a complex lottery in which:

- we select A_1 with probability $u(A)$ and A_0 with probability $1 - u(A)$;
- depending on which of the two alternatives A_i we get, we get A'_1 with probability $u'(A_i)$ and A'_0 with probability $1 - u'(A_i)$.

In this complex lottery, we get A'_1 with probability

$$u'(A) = u(A) \cdot (u'(A_1) - u'(A_0)) + u'(A_0).$$

Thus, the utility $u'(A)$ is related with the utility $u(A)$ by a linear transformation $u' = a \cdot u + b$, with $a > 0$. In other words, utility is defined modulo a linear transformation.

Traditional approach summarized. We assume that

- we know possible actions, and
- we know the exact consequences of each action.

Then, we should select an action with the largest value of expected utility.

2.2 Our Main Idea: Optimization Under Granularity

When we do not have enough time to take all the information into account, a natural idea is to use partial information. For example, when a man sees an animal in the jungle, it could be a predator, so an immediate decision needs to be made on whether to run away or not. Ideally, we should take into account all the details of an animal image, but there is no time for that, a reasonable reaction is to run way if an animal is sufficiently large.

So, instead of considering each data set separately, we, in effect, combine these data sets into “granules” corresponding to the partial information that is actually used in decision making; see, e.g., [126]. In the above example, instead of using the animal’s size, we only take into account whether this size is greater than a certain threshold s_0 or not. In effect, this means that we divide the set of all possible values of size into two granules:

- a granule consisting of small animals, whose size is smaller than s_0 , and
- a granule consisting of large (and thus, potentially dangerous) animals, whose size is larger than or equal to s_0 .

In this chapter, we show is that in many cases, if we take into account only algorithms that process such granular information, then the observed human decision making can be shown to be *optimal* among such granular algorithms—although, of course, if we could take into account all available information, we would be able to make a better decision.

2.3 Explaining the First Example of Seemingly Irrational Human Decision Making: Granularity Explains The Compromise Effect

In this section, we show that granularity explains the first example of seemingly irrational human decision making: the compromise effect. The results from this section first appeared in [85, 86, 90].

Compromise effect: reminder. We have three alternative a , a' and a'' :

- the alternative a is the cheapest—and is, correspondingly, of the lowest quality among the given three alternatives;
- the alternative a' is intermediate in terms of price—and is, correspondingly, intermediate in terms of quality;
- finally, the alternative a'' is the most expensive—and is, correspondingly, of the highest quality among the given three alternatives.

We do not know the exact prices, we just know the order between them; similarly, we do not know the exact values of quality, we just know the order between them. In this situation, most people select an alternative a' .

Let us describe the corresponding granularity. The “utility” of each alternative comes from two factors:

- the first factor comes from the quality: the higher the quality, the better—i.e., larger the corresponding component u_1 of the utility;
- the second factor comes from price: the lower the price, the better for the user—i.e., the larger the corresponding component u_2 of the utility.

The fact that we do not know the exact value of the price means, in effect, that we consider three possible levels of price and thus, three possible levels of the utility u_1 :

- low price, corresponding to high price-related utility;
- medium price, corresponding to medium price-related utility; and
- high price, corresponding to low price-related utility.

In the following text, we will denote “low” by L , “medium” by M , and “high” by H . In these terms, the above description of each alternative by the corresponding pair of utility values takes the following form:

- the alternative a is characterized by the pair (L, H) ;
- the alternative a' is characterized by the pair (M, M) ; and
- the alternative a'' is characterized by the pair (H, L) .

Natural symmetries. We do not know a priori which of the two utility components is more important. As a result, it is reasonable to treat both components equally. In other words, the selection should be the same if we simply swap the two utility components—i.e., we should select the same of three alternatives before and after swap:

- if we are selecting an alternative based on the pairs (L, H) , (M, M) , and (H, L) ,
- then we should select the exact same alternative if the pairs were swapped, i.e., if:
 - the alternative a was characterized by the pair (H, L) ;
 - the alternative a' was characterized by the pair (M, M) ; and
 - the alternative a'' was characterized by the pair (L, H) .

Similarly, there is no reason to a priori prefer one alternative or the other. So, the selection should not depend on which of the alternatives we mark as a , which we mark as a' , and which we mark as a'' . In other words, any permutation of the three alternatives is a reasonable symmetry transformation. For example, if, in our case, we select an alternative a which is characterized by the pair (L, H) , then, after we swap a and a'' and get the choice of the following three alternatives:

- the alternative a which is characterized by the pair (H, L) ;
- the alternative a' is characterized by the pair (M, M) ; and
- the alternative a'' is characterized by the pair (L, H) ,

then we should select the same alternative—which is now denoted by a'' .

General comment: symmetries have been helpful in dealing with uncertainty. It should be mentioned that in situations with major uncertainty, symmetries are often helpful. The main idea behind using symmetries is that if the situation is invariant with respect to some natural symmetries, then it is reasonable to select an action which is also invariant with respect to all these symmetries.

There have been many applications of this idea, starting from the pioneering work of N. Wiener on Cybernetics; see, e.g., [161]. It has been shown that for many empirically successful techniques related to neural fuzzy logic, networks, and interval computations, their empirical success can be explained by the fact that these techniques can be deduced from the appropriate symmetries; see, e.g., [115]. In particular, this explains the use of a sigmoid activation function $s(z) = \frac{1}{1 + \exp(-z)}$ in neural networks, the use of the most efficient “and”-operations (t-norms) and “or”-operations (t-conorms) in fuzzy logic, etc. [115].

Back to our situation: what we can conclude based on the corresponding symmetries. One can observe the following: that if we *both* swap u_1 and u_2 *and* swap a and a'' , then you get the exact same characterization of all alternatives:

- the alternative a is still characterized by the pair (L, H) ;
- the alternative a' is still characterized by the pair (M, M) ; and
- the alternative a'' is still characterized by the pair (H, L) .

The only difference is that:

- now, a indicates an alternative which was previously denoted by a'' , and
- a'' now denotes the alternative which was previously denoted by a .

As we have mentioned, it is reasonable to conclude that:

- if in the original triple selection, we select the alternative a ,

- then in the new selection—which is based on the exact same pairs of utility values—we should also select an alternative denoted by a .

But this “new” alternative a is nothing else but the old a'' . So, we conclude that:

- if we selected a ,
- then we should have selected a different alternative a'' in the original problem.

This is clearly a contradiction:

- we started by assuming that, to the user, a was better than a'' (because otherwise a would not have been selected in the first place), and
- we ended up concluding that to the same user, the original alternative a'' is better than a .

This contradiction shows that, under the symmetry approach, we cannot prefer a .

Similarly:

- if in the original problem, we preferred an alternative a'' ,
- then this would mean that in the new problem, we should still select an alternative which marked by a'' .

But this “new” a'' is nothing else but the old a . So, this means that:

- if we originally selected a'' ,
- then we should have selected a different alternative a in the original problem.

This is also a contradiction:

- we started by assuming that, to the user a'' was better than a (because otherwise a'' would not have been selected in the first place), and
- we ended up concluding that to the same user, the original alternative a is better than a'' . This contradiction shows that, under the symmetry approach, we cannot prefer a'' .

We thus conclude that out of the three alternatives a , a' , and a'' :

- we cannot select a , and
- we cannot select a'' .

This leaves us only once choice: to select the intermediate alternative a' .

This is exactly the compromise effect that we planned to explain.

Conclusion. Experiments show when people are presented with three choices $a < a' < a''$ of increasing price and increasing quality, and they do not have detailed information about these choices, then in the overwhelming majority of cases, they select the intermediate alternative a' .

This “compromise effect” is, at first glance, irrational: selecting a' means that, to the user, a' is better than a'' , but in a similar situation when the user is presented with $a' < a'' < a'''$, the same principle would indicate that the user will select a'' —meaning that a'' is better than a' .

Somewhat surprisingly, a natural granularity approach explains this seemingly irrational decision making.

2.4 Explaining the Second Example of Seemingly Irrational Human Decision Making: Granularity Explains Why Our Probability Estimates Are Biased

In this section, we show that granularity explains the second example of seemingly irrational human decision making: that in our decisions, we use biased estimates of probabilities.

The results from this section appeared in [82, 86, 90].

Main idea. Probability of an event is estimated, from observations, as the frequency with which this event occurs. For example, if out of 100 days of observation, rain occurred in 40 of these days, then we estimate the probability of rain as 40%. In general, if out of n observations, the event was observed in k of them, we estimate the probability as the ratio $\frac{k}{n}$.

This ratio is, in general, different from the actual (unknown) probability. For example, if we take a fair coin, for which the probability of head is exactly 50%, and flip it 100 times, we may get 50 heads, but we may also get 47 heads, 52 heads, etc. Similarly, if we have the coin fall heads 50 times out of 100, the actual probability could be 50%, could be 47% and could be 52%. In other words, instead of the exact value of the probability, we get a *granule* of possible values. (In statistics, this granule is known as a *confidence interval*; see, e.g., [144].)

In other words:

- first, we estimate a probability based on the observations; as a result, instead of the exact value, we get a granule which contains the actual (unknown) probability; this granule is all we know about the actual probability;
- then, when a need comes to estimate the probability, we produce an estimate based on the granule.

Let us analyze these two procedures one by one.

Probability granules: analysis of the first procedure and the resulting formulas.

It is known (see, e.g., [144]), that the expected value of the frequency is equal to p , and that the standard deviation of this frequency is equal to

$$\sigma = \sqrt{\frac{p \cdot (1 - p)}{n}}.$$

It is also known that, due to the Central Limit Theorem, for large n , the distribution of frequency is very close to the normal distribution (with the corresponding mean p and standard deviation σ).

For normal distribution, we know that with a high certainty all the values are located within 2–3 standard deviations from the mean, i.e., in our case, within the interval $(p - k_0 \cdot \sigma, p + k_0 \cdot \sigma)$, where $k_0 = 2$ or $k_0 = 3$: for example, for $k_0 = 3$, this is true with confidence 99.9%. We can thus say that the two values of probability p and p' are (definitely) distinguishable if the corresponding intervals of possible

values of frequency do not intersect—and thus, we can distinguish between these two probabilities just by observing the corresponding frequencies.

In precise terms, the probabilities $p < p'$ are distinguishable if

$$(p - k_0 \cdot \sigma, p + k_0 \cdot \sigma) \cap (p' - k_0 \cdot \sigma', p' + k_0 \cdot \sigma') = \emptyset,$$

where

$$\sigma' \stackrel{\text{def}}{=} \sqrt{\frac{p' \cdot (1 - p')}{n}},$$

i.e., if $p' - k_0 \cdot \sigma' \geq p + k_0 \cdot \sigma$. The smaller p' , the smaller the difference $p' - k_0 \cdot \sigma'$. Thus, for a given probability p , the next distinguishable value p' is the one for which

$$p' - k_0 \cdot \sigma' = p + k_0 \cdot \sigma.$$

When n is large, these value p and p' are close to each other; therefore, $\sigma' \approx \sigma$. Substituting an approximate value σ instead of σ' into the above equality, we conclude that

$$p' \approx p + 2k_0 \cdot \sigma = p + 2k_0 \cdot \sqrt{\frac{p \cdot (1 - p)}{n}}.$$

If the value p corresponds to the i -th level, then the next value p' corresponds to the $(i + 1)$ -st level. Let us denote the value corresponding to the i -th level by $p(i)$. In these terms, the above formula takes the form

$$p(i + 1) - p(i) = 2k_0 \cdot \sqrt{\frac{p \cdot (1 - p)}{n}}.$$

The above notation defines the value $p(i)$ for non-negative integers i . We can extrapolate this dependence so that it will be defined for all non-negative real values i .

When n is large, the values $p(i + 1)$ and $p(i)$ are close, the difference

$$p(i + 1) - p(i)$$

is small, and therefore, we can expand the expression $p(i + 1)$ in Taylor series and keep only linear terms in this expansion:

$$p(i + 1) - p(i) \approx \frac{dp}{di}.$$

Substituting the above expression for $p(i + 1) - p(i)$ into this formula, we conclude that

$$\frac{dp}{di} = \text{const} \cdot \sqrt{p \cdot (1 - p)}.$$

Moving all the terms containing p into the left-hand side and all the terms containing i into the right-hand side, we get

$$\frac{dp}{\sqrt{p \cdot (1-p)}} = \text{const} \cdot di.$$

Integrating this expression and taking into account that $p = 0$ corresponds to the lowest 0-th level—i.e., that $i(0) = 0$ —we conclude that

$$i(p) = \text{const} \cdot \int_0^p \frac{dq}{\sqrt{q \cdot (1-q)}}.$$

This integral can be easily computed if introduce a new variable t for which $q = \sin^2(t)$. In this case,

$$dq = 2 \cdot \sin(t) \cdot \cos(t) \cdot dt,$$

$1 - p = 1 - \sin^2(t) = \cos^2(t)$ and therefore,

$$\sqrt{p \cdot (1-p)} = \sqrt{\sin^2(t) \cdot \cos^2(t)} = \sin(t) \cdot \cos(t).$$

The lower bound $q = 0$ corresponds to $t = 0$ and the upper bound $q = p$ corresponds to the value t_0 for which $\sin^2(t_0) = p$ —i.e., $\sin(t_0) = \sqrt{p}$ and $t_0 = \arcsin(\sqrt{p})$. Therefore,

$$\begin{aligned} i(p) &= \text{const} \cdot \int_0^p \frac{dq}{\sqrt{q \cdot (1-q)}} = \text{const} \cdot \int_0^{t_0} \frac{2 \cdot \sin(t) \cdot \cos(t) \cdot dt}{\sin(t) \cdot \cos(t)} = \\ &= \int_0^{t_0} 2 \cdot dt = 2 \cdot \text{const} \cdot t_0. \end{aligned}$$

We know how t_0 depends on p , so we get

$$i(p) = 2 \cdot \text{const} \cdot \arcsin(\sqrt{p}).$$

We can determine the constant from the condition that the largest possible probability value $p = 1$ should correspond to the largest level $i = m$. From the condition that $i(1) = m$, taking into account that

$$\arcsin(\sqrt{1}) = \arcsin(1) = \frac{\pi}{2},$$

we conclude that

$$i(p) = \frac{2m}{\pi} \cdot \arcsin(\sqrt{p}).$$

Thus,

$$\arcsin(\sqrt{p}) = \frac{\pi \cdot i}{2m},$$

hence

$$\sqrt{p} = \sin\left(\frac{\pi \cdot i}{2m}\right)$$

and thus,

$$p(i) = \sin^2\left(\frac{\pi \cdot i}{2m}\right).$$

Thus, probability granules are formed by intervals $[p(i), p(i+1)]$. Each empirical probability is represented by the granule i to which it belongs.

From granules to probability estimates: analysis of the second procedure. As we have mentioned, instead of the actual probabilities, we have probability *labels* corresponding to m different granules:

- the first label corresponds to the smallest certainty,
- the second label corresponds to the second smallest certainty,
- etc.,
- until we reach the last label which corresponds to the largest certainty.

We need to produce some estimates of the probability based on the granule. In other words, for each i from 1 to m , we need to assign, to each i -th label, a value p_i in such a way that labels corresponding to higher certainty should get larger numbers: $p_1 < p_2 < \dots < p_m$.

Before we analyze how to do it, let us recall that one of the main objectives of assigning numerical values is that we want computers to help us solve the corresponding decision problems, and computers are not very good in dealing with granules; their natural language is the language of numbers. From this viewpoint, it makes sense to consider not all theoretically possible exact real numbers, but only computer-representable real numbers.

In a computer, real numbers from the interval $[0, 1]$ are usually represented by the first d digits of their binary expansion. Thus, computer-representable numbers are $0, h \stackrel{\text{def}}{=} 2^{-d}, 2h, 3h, \dots$, until we reach the value $2^d \cdot h = 1$.

In our analysis, we will assume that the “machine unit” $h > 0$ is fixed, and we will thus assume that only multiples of this machine units are possible values of all n probabilities p_i .

For example, when $h = 0.1$, each probability p_i takes 11 possible values: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

In the modern computers, the value h is extremely small; thus, whenever necessary, we can assume that $h \approx 0$ —i.e., use limit case of $h \rightarrow 0$ instead of the actual small “machine unit” h .

For each h , we consider all possible combinations of probabilities $p_1 < \dots < p_m$ in which all the numbers p_i are proportional to the selected step h , i.e., all possible combinations of values $(k_1 \cdot h, \dots, k_m \cdot h)$ with $k_1 < \dots < k_m$.

For example, when $m = 2$ and $h = 0.1$, we consider all possible combinations of values $(k_1 \cdot h, k_2 \cdot h)$ with $k_1 < k_2$:

- For $k_1 = 0$ and $p_1 = 0$, we have 10 possible combinations $(0, 0.1), (0, 0.2), \dots, (0, 1)$.
- For $k_1 = 1$ and $p_1 = 0.1$, we have 9 possible combinations $(0.1, 0.2), (0.1, 0.3), \dots, (0.1, 1)$.
- ...
- Finally, for $k_1 = 9$ and $p_1 = 0.9$, we have only one possible combination $(0.9, 1)$.

For each i , for different possible combinations (p_1, \dots, p_m) , we get, in general, different value of the probability p_i . According to the complete probability formula, we can obtain the actual (desired) probability P_i if we combine all these value p_i with the weights proportional to the probabilities of corresponding combinations:

$$P_i = \sum_{p_1 < \dots < p_m} p_i \cdot \text{Prob}(p_1, \dots, p_m).$$

Since we have no reason to believe that some combinations (p_1, \dots, p_m) are more probable and some are less probable, it is thus reasonable to assume that all these combinations are equally probable. Hence, P_i is equal to the arithmetic average of the values p_i corresponding to all possible combinations (p_1, \dots, p_m) .

For example, for $m = 2$ and $h = 0.1$, we thus estimate P_1 by taking an arithmetic average of the values p_1 corresponding to all possible pairs. Specifically, we average:

- ten values $p_1 = 0$ corresponding to ten pairs $(0, 0.1), \dots, (0, 1)$;
- nine values $p_1 = 0.1$ corresponding to nine pairs $(0.1, 0.2), \dots, (0.1, 1)$;
- ...
- and a single value $p_1 = 0.9$ corresponding to the single pair $(0.9, 1)$.

As a result, we get the value

$$P_1 = \frac{10 \cdot 0.0 + 0 \cdot 0.1 + \dots + 1 \cdot 0.9}{10 + 9 + \dots + 1} = \frac{16.5}{55} = 0.3.$$

Similarly, to get the value p_2 , we average:

- a single value $p_2 = 0.1$ corresponding to the single pair $(0, 0.1)$;
- two values $p_2 = 0.2$ corresponding to two pairs $(0, 0.2)$ and $(0.1, 0.2)$;
- ...
- ten values $p_2 = 1.0$ corresponding to ten pairs $(0, 1), \dots, (0.9, 1)$.

As a result, we get the value

$$P_2 = \frac{1 \cdot 0.1 + 2 \cdot 0.2 + \dots + 10 \cdot 1.0}{1 + 2 + \dots + 10} = \frac{37.5}{55} = 0.7.$$

The probability p_i of each label can take any of the equidistant values $0, h, 2h, 3h, \dots$, with equal probability. In the limit $h \rightarrow 0$, the resulting probability distribution tends to the uniform distribution on the interval $[0, 1]$.

In this limit $h \rightarrow 0$, we get the following problem:

- we start with m independent random variables v_1, \dots, v_m which are uniformly distributed on the interval $[0, 1]$;
- we then need to find, for each i , the conditional expected value

$$E[v_i \mid v_1 < \dots < v_m]$$

of each variable v_i under the condition that the values v_i are sorted in increasing order.

Conditional expected values are usually more difficult to compute than unconditional ones. So, to solve our problem, let us reduce our problem to the problem of computing the usual (unconditional) expectation.

Let us consider m independent random variables each of which is uniformly distributed on the interval $[0, 1]$. One can easily check that for any two such variables v_i and v_j , the probability that they are equal to each other is 0. Thus, without losing generality, we can safely assume that all m random values are different. Therefore, the whole range $[0, 1]^m$ is divided into $m!$ sub-ranges corresponding to different orders between v_i . Each sub-range can be reduced to the sub-range corresponding to $v_1 < \dots < v_m$ by an appropriate permutation in which v_1 is swapped with the smallest $v_{(1)}$ of m values, v_2 is swapped with the second smallest $v_{(2)}$, etc.

Thus, the conditional expected value of v_i is equal to the (unconditional) expected value of the i -th value $v_{(i)}$ in the increasing order. This value $v_{(i)}$ is known as an *order statistic*, and for uniform distributions, the expected values of all order statistics are

known (see, e.g., [1, 4, 25]): $P_i = \frac{i}{m+1}$.

So, if all we know is that our degree of certainty is expressed by i -th label on an m -label scale of granules, then it is reasonable to assign, to this case, the probability

$$P_i = \frac{i}{m+1}.$$

Let us now combine the two procedures. In the first procedure, based on the empirical frequency p , we find a label i for which

$$p \approx \sin^2 \left(\frac{\pi \cdot i}{2m} \right).$$

Based on this label, we then estimate the probability as $P_i = \frac{i}{m+1}$. For large m , we have $P \approx \frac{i}{m}$. Substituting P instead of $\frac{i}{m}$ into the formula for p , we conclude that

$$p \approx \sin^2 \left(\frac{\pi}{2} \cdot P \right).$$

Based on this formula, we can express the estimate P in terms of the actual probability p , as

$$P \approx \frac{1}{\pi} \cdot \arcsin(\sqrt{p}).$$

Comparing our estimates P with empirical probability estimates \tilde{p}_i : first try. Let us compare the probabilities p_i , Kahneman’s empirical estimates \tilde{p}_i , and the estimates $P_i = \frac{1}{\pi} \cdot \arcsin(\sqrt{p_i})$ computed by using the above formula:

p_i	0	1	2	5	10	20	50	80	90	95	98	99	100
\tilde{p}_i	0	5.5	8.1	13.2	18.6	26.1	42.1	60.1	71.2	79.3	87.1	91.2	100
P_i	0	6.4	9.0	14.4	20.5	29.5	50.0	70.5	79.5	85.6	91.0	93.6	100

For most probabilities p_i , the difference between the values P'_i and the empirical probability estimates \tilde{p}_i is so small that it is below the accuracy with which the empirical weights can be obtained from the experiment.

Thus, granularity ideas indeed explain Kahneman and Tversky’s observation of biased empirical probability estimates.

Summary. Kahneman and Tversky showed that when people make decisions, then instead of—as should be rational—weighting outcomes with weights proportional to probabilities of different outcomes—they use *biased* weights, overestimating the importance of low-probability events and underestimating the importance of high-probability events. In this section, we show that this observable bias can be explained if we take into account granularity—imposed by our limited rationality (i.e., our limited ability to process information).

2.5 Explaining the Third Example of Seemingly Irrational Human Decision Making: Using Average Instead of the Sum

In this section, we show how to explain the third example of seemingly irrational behavior: using average instead of the sum. The result of this section first appeared in [76].

Using average instead of the sum: reminder. According to [61], when in a hurry, people often use an arithmetic average as a substitute for the sum. This substitution leads to a seemingly irrational behavior. For example, when pricing two large dinnerware sets,

- one consisting of 24 pieces in perfect condition, and
- the other consisting of the same 24 pieces plus 16 additional ones, 9 of which are broken,

most people value the second set lower—possibly because in the second set, the average value of a piece is lower. This selection is irrational, since after buying the second set, we can simply throw away the broken pieces, and actually end up with more pieces.

Why do people use the average instead of the sum: a possible explanation. Our explanation for the use of arithmetic averages is that the arithmetic average is much easier to compute than, e.g., the sum.

This may sound somewhat counter-intuitive, because, at first glance, the formula for the arithmetic average $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ looks somewhat more complex to compute than the formulas for the sum $s = x_1 + \dots + x_n$: to compute the average, we need to perform all the additions needed for the sum plus one additional division.

This is indeed the case if we talk about *exact* computations: to compute the exact sum or the exact average, one needs to process each of n numbers at least once—if we do not process one of the numbers, we cannot get the exact value of sum or average. Since each elementary arithmetic operation takes at most two numerical inputs, this means that in both cases, we need at least $n/2$ operations, and $n/2$ is $O(n)$.

However, if we take into account that the values x_i are only known approximately and that, as a result, we only need approximate values of sum and average, then the computational complexity changes. For the sum, we still need to count the intervals, but to compute the approximate values of the average, we can use Monte-Carlo techniques: namely, we can select a random sample of values and take the arithmetic average of this sample.

According to the Large Numbers Theorem, when the sample size is large, this random-sample-based arithmetic average provides a good approximation to the desired exact average—and the larger the sample, the more accurate this approximation; see, e.g., [144].

The required sample size—and thus, the corresponding computational complexity of estimating the average this way—depends only on the desired accuracy of estimating the average, and does not depend on the number n of original values. Thus, for a fixed accuracy, the computational complexity of this algorithm does not grow with n at all, it is $O(1)$, while the complexity of computing the sum still grows with n as $O(n)$. Since for large n , $O(1) \ll O(n)$, this explains why people use an average as a substitute for the sum.

Bounded Rationality in Decision Making Under
Uncertainty: Towards Optimal Granularity

Lorkowski, J.; Kreinovich, V.

2018, IX, 164 p., Hardcover

ISBN: 978-3-319-62213-2