Lecture Notes for *Probability & Rational Choice*

J. Dmitri Gallow [1]

Draft from Spring, 2026
Taught at the University of Southern California

[1]These notes are indebted to Jonathan Weisberg's online textbook *Odds and Ends*, Michael Resnik's *Choices: An Introduction to Decision Theory* (1987, University of Minnesota Press: Minneapolis), and Martin Peterson's *An Introduction to Decision Theory* (2013, Cambridge University Press: Cambridge).

Contents

Part I

Probability

1 | Probability Trees

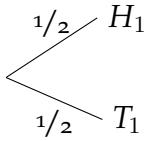
Goal: learn how to build probability trees and use them to calculate probabilities.

Puzzle. *John has two children. Mary asks him*: *"Do you have a boy*?*" John answers "yes". How likely is it that John has two boys*?

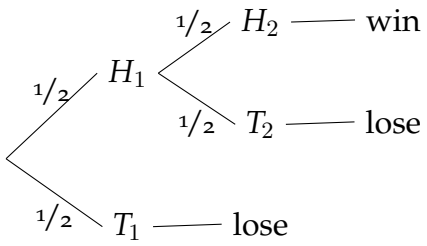
A second (more challenging) puzzle: Suppose Mary had asked John "Do you have a boy born on a Monday?", and John answers "yes". *Then*, what's the probability that John has two boys?

Example 1. *We will flip a fair coin. If it lands heads twice in a row, then you win. Otherwise, you lose. What's the probability that you win*?

There are two ways the coin could land on the first flip: it could land heads (H_1), or it could land tails (T_1), each with equal probability $1/2$.



If the coin lands tails on the first flip, then you've already lost. If, however, it lands heads on the first flip, then there are two more sub-possibilities: it could be that the coin lands heads on the second flip (H_2) or it could be that the coin lands tails on the second flip (T_2), each with equal probability.



This is a *probability tree*. Each path through the tree describes a way things could be. It could be that the coin lands heads, then heads, and you win; or it could land heads, then tails, and you lose; or it could be that it lands tails straightaway and you lose.

These trees consist of *nodes*, *branches*, and *leaves*. A sequence of branches (like the one

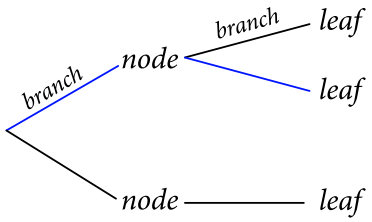
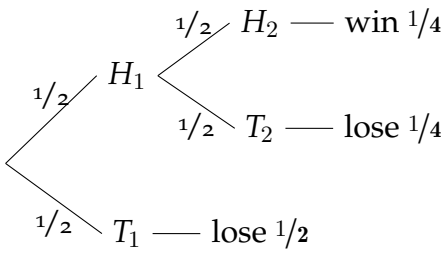


Figure 1.1

in blue in figure 1.1) is a *path* through the tree. Each time a probability tree *branches* (i.e., splits in two or more), there is an associated *number* attached to that branch. This number is the probability that the next node proposition is true, given that the previous ones along that path were true.

To find the probability that you end up at any given leaf, you multiply together all of the probabilities lying along the branches that take you to that leaf. For instance:



To find the probability that you win, you simply add together all of the probabilities for each leaf in which you win. And to find the probability that you lose, you simply add together all of the probabilities for each leaf in which you lose. Thus,

$$\Pr(\textit{win}) = 1/4 \qquad \text{and} \qquad \Pr(\textit{lose}) = 1/4 + 1/2 = 3/4$$

When we build a tree like this, it is important that the propositions branching off from each node are *mutually exclusive* and *jointly exhaustive*. Propositions are *mutually exclusive* if no two can be true at once. And propositions are *jointly exhaustive* if at least one of them must be true—if it's impossible for them all to be false. If a collection of propositions are mutually exclusive and jointly exhaustive, then they are called *a partition*. Exactly one of the propositions in a partition must be true.

Rule #1 When building a tree, each node must branch into a partition.

Rule #2 The numbers you attach to the branches emerging from a node must be positive, and they must add up to 1.

Rule #3 The probability of a leaf is found by multiplying together the probabilities found on each branch leading to that leaf.

Rule #4 To find the probability of a proposition, add up the probabilities from each leaf where the proposition is true.

Coda: Representing, Adding, and Multiplying Probabilities

Probabilities are numbers between zero and one. We can represent these probabilities in four different ways: as a *decimal* number, as a *percentage*, as a *fraction*, or in *odds form*.

Decimal	Percentage	Fraction	Odds
0.5	50%	$1/2$	1 : 1
0.33...	33.33...%	$1/3$	1 : 2
0.05	5%	$1/20$	1 : 19
0.75	75%	$3/4$	3 : 1

The number above the line in a fraction is *the numerator*. The number below the line is *the denominator*. A fraction will have the same value if you multiply both the numerator and the denominator by the same value. So all of the following fractions are the same:

$$1/2 = 2/4 = 4/8 = 12/24 = 60/120 = 175/350$$

If the numerator and denominator have no factors in common, then the fraction is written in its *simplest* form.

To add two fractions, you first write them so that they have *the same* denominator, and then you add together the numerators. So

$$1/4 + 1/2 = 1/4 + 2/4 = 3/4 \qquad \text{and} \qquad 5/8 + 11/16 = 10/16 + 11/16 = 21/16$$

In general, a recipe for adding together two fractions is this (where 'a', 'b', 'c', and 'd' are any positive whole numbers):

$$a/b + c/d = ad/bd + cb/bd = ad+cb/bd$$

To multiply together two fractions, you multiply together their numerators and you multiply together their denominators. For instance,

$$2/3 \cdot 3/4 = 2{\cdot}3/3{\cdot}4 = 6/12 = 1/2$$

In general, if 'a', 'b', 'c', and 'd' are any positive whole numbers, then

$$\boxed{a/b \cdot c/d = ac/bd}$$

One denominator is so common that we've adopted special notation for it. If you have a fraction $a/100$, this gets written 'a%'. So, for instance, 50% is just another way of writing $50/100$, which is just $1/2$ (which we get by dividing both numerator and denominator of $50/100$ by 50).

Odds are perhaps the least common way of representing probabilities, but they turn out to be a very helpful way of keeping track of probabilities. When we write that the odds of a proposition, H, are $a : b$, what we mean is that the probability of H is $a/a+b$.

> If the odds in favor of H are $a : b$ (a-to-b in favor of H), then the probability of H is $a/a+b$.
>
> If the probability of H is a/b, then the odds in favor of H are $a : b-a$.

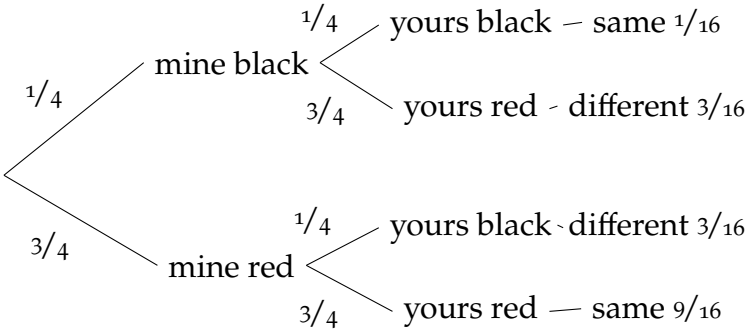
Just like with fractions, if you multiply both the left and the right hand side of the odds $a : b$ by the same number, you'll have the same odds. So all of the following odds are the same:

$$3 : 1 \;=\; 6 : 2 \;=\; 9 : 3 \;=\; 12 : 4 \;=\; 15 : 5$$

If the number on the left and the right have no factors in common, then the odds are written in their *simplest* form. (On problem sets, you should make life easy for your grader by writing final answers which are fractions or odds in their simplest forms.)

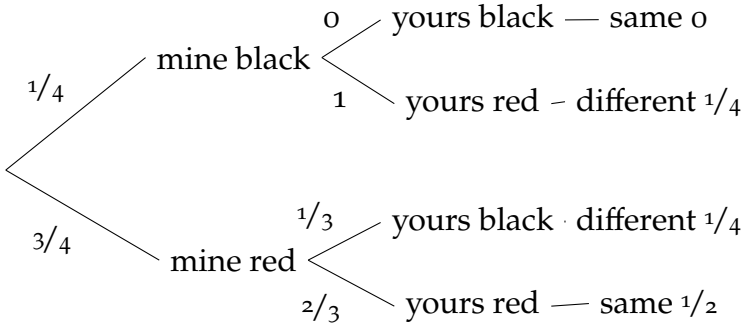
Back to Probability Trees

Example 2. *There is an urn with* 4 *balls inside of it.* 3 *of the balls are red, and* 1 *of the balls is black. First, I draw a ball from the urn and place it back. Next, you draw a ball from the urn. What is the probability that your ball and my ball have the same color*?



So the probability that our balls have the same color is $1/16 + 9/16 = 10/16 = 5/8$.

Example 3. *There is an urn with* 4 *balls inside of it.* 3 *of the balls are red, and* 1 *of the balls is black. First, I draw a ball from the urn and* do not *place it back. Next, you draw a ball from the urn. What is the probability that your ball and my ball have the same color*?



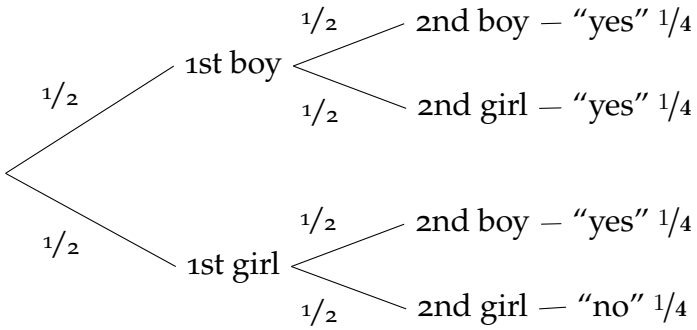
So the probability that our balls are the same color is $1/2 + 0 = 1/2$.

As example 3 illustrates, what happens in the first branch can make a difference to the probability of what happens at later branches. The numbers we place along the branches in our tree diagram are not the *unconditional* probabilities that you draw black or red. Instead, they are the *conditional* probabilities of you drawing black *given* what I have drawn.

Rule #5 The numbers you place on a branch emerging from a node should be the *conditional* probability of the proposition at the end of that branch, *given that* all of the previous propositions leading up to that node are true.

Incorporating New Information

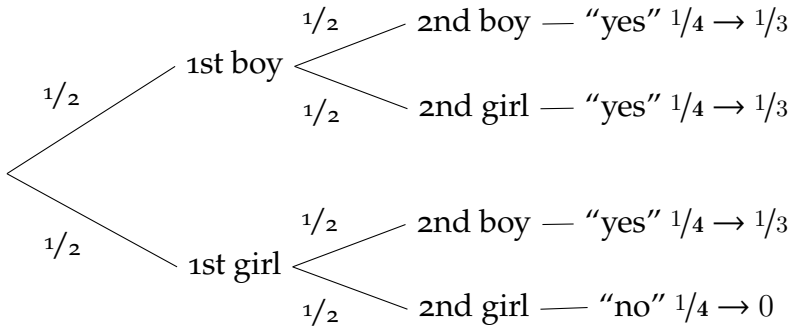
Return to the puzzle: John has two children. Suppose each child had a 50% chance of being a boy and a 50% chance of being a girl (suppose you know that John's children are not intersex).



Mary asks John "Do you have a boy?", and John answers "yes". This is the answer John would give in the first three leafs. He only gives the answer "no" on the final leaf (girl

girl). So when John gives the answer "yes", we learn that we are on one of the first three leafs. This new information *changes* the probabilities. But it does not change the *relative probabilities* of boy boy, boy girl, or girl boy. Each of these three possibilities were compatible with John having at least one boy, and each of these three possibilities were equally likely before. So each of these three possibilities will be equally likely after we incorporate the new information that John has at least one boy.

This means that, when we incorporate the new information that John has at least one boy, we should arrive at the following updated probabilities:

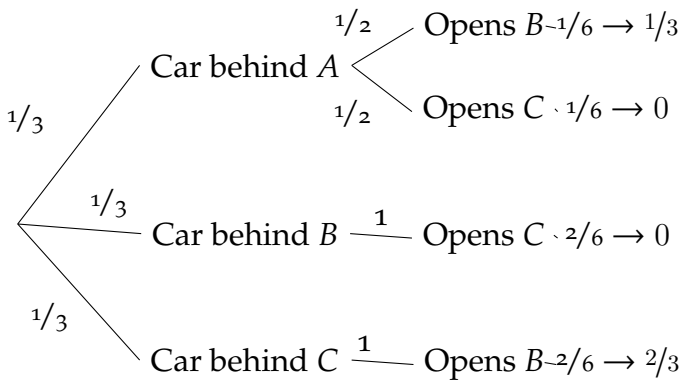


In general, if you gain the information that some leafs are false, you should incorporate that new information by setting the probability of those leafs to zero, taking the sum of the probabilities of the remaining leafs, and then dividing their old probabilities by this sum. This is called *renormalizing*. In our example, we learned that one of the leafs "boy boy", "boy girl", or "girl boy" obtained. The sum of their original probabilities was $1/4 + 1/4 + 1/4 = 3/4$. Dividing through by this sum gives us the new probabilities:

$$\frac{1/4}{3/4} = \frac{1}{3}$$

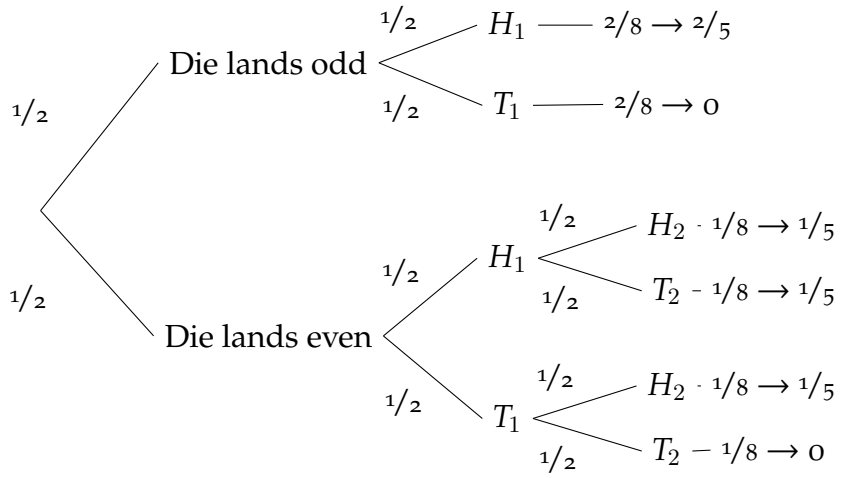
Rule #6 To incorporate the new information that some leafs are false, change the probabilities for these leafs to zero, add up the probabilities for the remaining leafs, and then divide each of the remaining probabilities by this sum.

Example 4. *On the Monty Hall game show, you are asked to choose one of three doors (A, B, and C); you will win whatever is behind your selected door. Behind one of the doors is a car, and behind the other two doors are goats. After you make your choice, Monty always opens one of the unselected doors. He never reveals the car; but otherwise he shows no preference about which door he opens. On this occasion, after you choose A, Monty reveals that there's a goat behind door B. What's the probability that the car is behind C?*



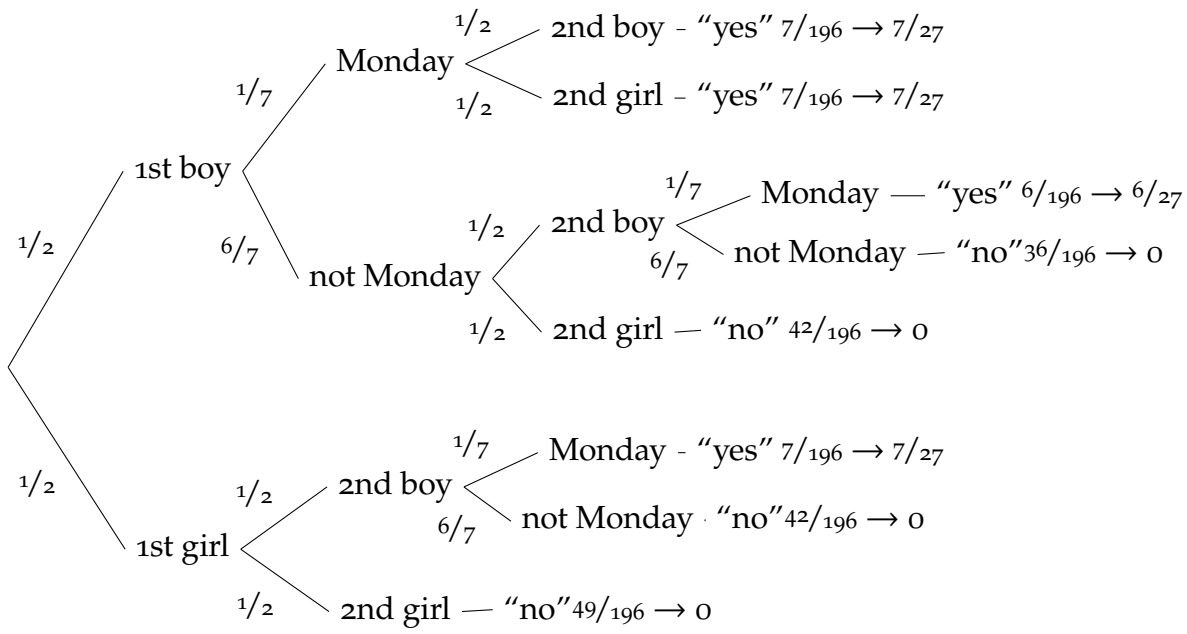
So the probability that the car is behind C, after we've incorporated the new information that Monty has opened door B, is $2/3$.

Example 5. *We roll a fair 6-sided die. If the die lands on an odd number, we will flip a coin once. If the die lands on an even number, we will flip a coin twice. You learn that the coin landed heads at least once. What's the probability that the die landed on odd*?



So the probability that the die landed on odd, after incorporating the new information that the coin landed heads at least once, is $2/5$.

Example 6. *John has two children. Mary asks him*: *"Do you have a boy born on a Monday?" John answers "yes". How likely is it that John has two boys*?



So the probability that John has two boys, given his answer to Mary, is $7/27 + 6/27 = 13/27 \approx 48.15\%$.

2 | Logic

Goal: learn when some propositions give us *conclusive* reason to accept another proposition.

Puzzle. *Alfred and Betty will each flip a coin. They will see the results of their own coin flip, but not the result of the other's. They will then guess how the other's coin landed. Can they come up with a strategy which will guarantee that at least one of them guesses correctly*? (*They are allowed to talk* before *the coins are flipped, but not* afterwards.)

Propositions

A *proposition* is any claim which can be true or false. A test: if "It's true that A" makes sense, then 'A' is a proposition. If 'It's true that A" doesn't make sense, then 'A' is not a proposition.

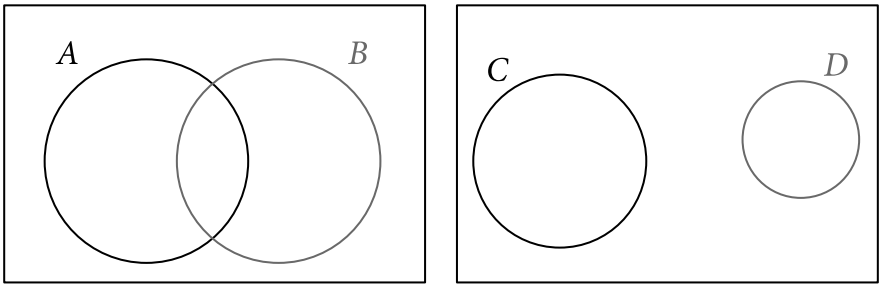
- Propositions: 'I ate my car keys', 'Nobody knows the trouble I've seen', Chocolate is tasty'
- Non-propositions: 'Try jiggling the handle', 'Who ate the car keys?', 'Ouch!'

We can visualize propositions with *Euler diagrams*. This is a very abstract representation of a proposition. We will represent the proposition in terms of the *possible situations in which it is true*. To start, we'll imagine an empty square, each point in which represents a possible way the world could be. We will then enclose some of these points within a circle, labeled with a name for a proposition. The interpretation is that all the points inside the circle are possible worlds in which the proposition is true, and all the points outside the circle are possible worlds in which the proposition is false.

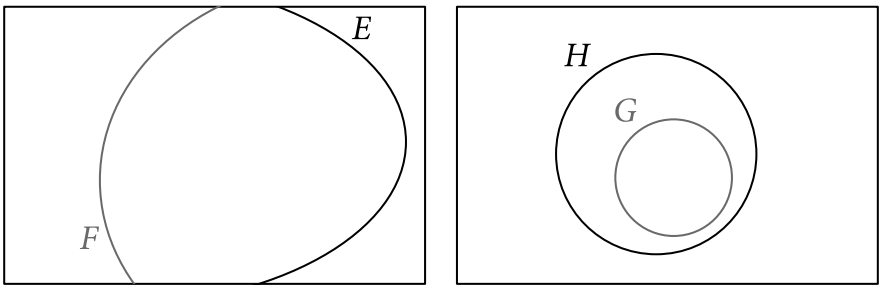
We can use Euler diagrams to illustrate some important logical notions.

Propositions are *compatible* when it is possible for them to all be true at once.

For instance, A and B are *compatible*.



(a) A and B are compatible (b) C and D are mutually exclusive



(c) E and F are jointly exhaustive (d) G entails H

Figure 2.1: Euler diagrams

A: Atlanta is colder than Boston

B: Boston is warmer than Los Angeles

Since it could be that the order of the cities, in terms of warmth, is: Boston > Atlanta > Los Angeles.

> Propositions are *mutually exclusive* when it is not possible for any two of them to be true at once.

For instance, C and D are *mutually exclusive*

C: Home prices are going down.

D: Home prices are going up.

Since there's no way for home prices to *both* be going up *and* going down.

> Propositions are *jointly exhaustive* when it is not possible for all of them to be false.

For instance, E and F are *jointly exhaustive*:

E: John is taller than 5 feet.

F: John is shorter than 6 feet.

Since no matter how tall John is, at least one of E and F will be true. (If he's shorter than 5 feet, then F is true. If he's between 5 and 6 feet, then both E and F are true. And if he's taller than 6 feet, then F is true.)

> Some propositions, $P_1, P_2, \ldots, P_n$ *entail* another proposition, C, when it is not possible for $P_1, P_2, \ldots$, and P_n to all be true while C is false.

For instance, G entails H:

G: Susan is a feminist bank teller

H: Susan is a bank teller

Since there's no way for Susan to be a feminist bank teller without being a bank teller.

Arguments

An *argument* gives you some reasons to accept a proposition. We use arguments to attempt to *persuade* one another

- The proposition we're trying to persuade someone to believe—the thing that the argument is arguing *for*—is called the conclusion of the argument
- The reasons which are presented in the conclusion's favor are called the *premises* of the argument

For our purposes in this class, we want to evaluate whether arguments are any good—whether they give us good reason to accept the conclusion or not. So we won't want our definition to prejudge this question. So we'll adopt a slightly artificial definition of what an argument is:

> An *argument* is a collection of propositions, at most one of which is designated as the conclusion, the others of which are designated as the premises

On this definition, the following will count as an argument, even though the premises don't intuitively give you *any* reason to accept its conclusion:

Bacon isn't meat

Samuel Huntington is spry

Summer will never come

$\therefore$ Elmer Fudd isn't fictional

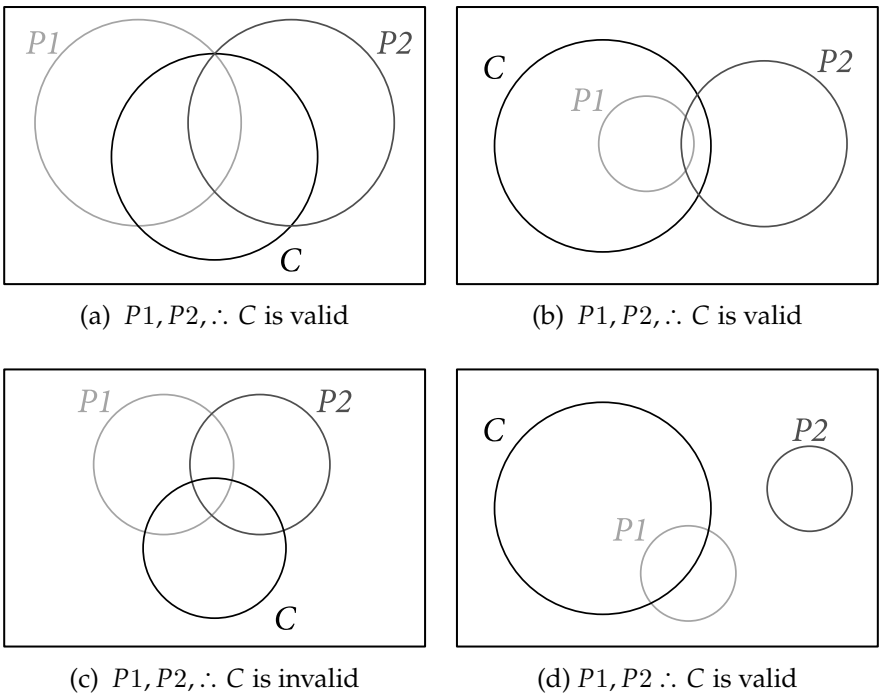
(The final proposition is the conclusion; we indicate this with the symbol "$\therefore$".)

Validity

We'll say that an argument is *valid* if and only if it is not possible for the premises of the argument to be true while the conclusion of the argument is false.

> An argument is *valid* if and only if it is not possible for its premises to be true and its conclusion false.
>
> An argument is *invalid* if and only if it is possible for its premises to all be true and its conclusion false.

In other words, an argument is *valid* iff its premises *entail* its conclusion.



(a) $P1, P2, \therefore C$ is valid

(b) $P1, P2, \therefore C$ is valid

(c) $P1, P2, \therefore C$ is invalid

(d) $P1, P2 \therefore C$ is valid

Figure 2.2: Euler diagrams showing valid and invalid arguments.

For instance, the following arguments are *valid*:

Elvis is alive and Paul McCartney is dead.

∴ Paul McCartney is dead

Either Trump won or Picasso painted *The Scream*.

Picasso didn't paint *The Scream*.

∴ Trump won.

whereas this argument is *invalid*:

All dogs are animals.

Some animals are pets.

∴ Some dogs are pets.

Notice that ***a valid argument can have false premises***, ***a valid argument can have a false conclusion***, and ***an invalid argument can have true premises and a true conclusion***. When it comes to validity, it doesn't matter whether the premises and conclusion are actually true or false. The only thing that matters is whether it's *possible* for the premises to all be true while the conclusion is false.

Notice the following special cases of validity:

- If it is impossible for an argument's premises to all be true, then it is automatically impossible for all of the argument's premises to be true while its conclusion is false—no matter what its conclusion is. So, if it's impossible for an argument's premises to all be true, then it will automatically be valid—no matter what its conclusion is.
- Also, if it is impossible for an argument's conclusion to be false, then it is automatically impossible for all of the argument's premises to be true while its conclusion is false—no matter what its premises are. So, if it's impossible for an argument's conclusion to be false, then it will automatically be valid—no matter what its premises are.

If the premises of a valid argument are all true, then we say that the argument is *sound*.

> An argument is *sound* if and only if it is valid and it has all true premises.

Strength

Some arguments are *invalid* but are nonetheless represent perfectly good ways of reasoning. For instance:

> No one in the clinical trial had an adverse reaction to the drug.
>
> ∴ I will not have an adverse reaction to the drug.

> There is almost always traffic on the 10 East on weekday mornings.
>
> ∴ This weekday morning, there will be traffic on the 10 East.

These arguments are not *valid*, but they are nonetheless *strong*. In general, an argument is *strong* if its premises make its conclusion *highly probable*.

> An argument $P_1, P_2 \therefore C$ is *strong* if and only if the probability of C, given that both P_1 and P_2 are true, is high.
>
> $$\Pr(C \mid P_1 \& P_2) \text{ is high}$$

When an argument is strong, and you know that the premises of the argument are true, then you should be highly confident of the conclusion.

Strength is a more general notion than validity—any valid argument will be strong, but many strong arguments are not valid. One goal of the first part of the course is to come to a better understanding of which arguments are strong; but we will start off by getting clear on when arguments are *valid*.

Formal Logic

Consider these arguments:

Either it rained or John went to the store.	Either Tabitha will hurry or she'll be tale	Either I'm mistaken or the test is Friday
It didn't rain	Tabitha won't hurry	I'm not mistaken
∴ John went to the store	∴ Tabitha will be late	∴ The test is Friday

Each of these arguments is valid; and they appear to be valid for the same reason. Notice: they all share the following *form*:

Either A or B

It is not the case that A

$\therefore$ B

Moreover, it seems like recognizing the form of these arguments is all that it takes to see that they are valid.

With formal logic, we can show that any arguments with this form will be valid. We do so by, firstly, recognizing certain very common ways of building up larger sentences from smaller ones, and secondly, thinking about how the truth of the large sentence is determined by the truth of the smaller ones out of which they are built.

Negation. Whenever 'A' is a proposition, we have another proposition, 'It is not the case that A', which we'll call *the negation of* A, and which we'll abbreviate '$\sim A$'.

Whether '$\sim A$' is true or false seems to be completely determined by whether 'A' is true or false. Whenever 'A' is true, '$\sim A$' is false. And whenever 'A' is false, '$\sim A$' is true. We can summarize this with a *truth-table*, or with an Euler diagram:

A	$\sim A$
T	F
F	T

Table 2.1: Truth-table for '$\sim$'. '$\sim A$' is true whenever 'A' is false, and false whenever 'A' is true.

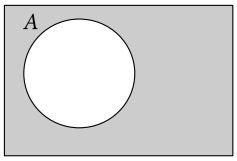


Figure 2.3: 'A' is true at all possibilities inside the circle. '$\sim A$' is true at all possibilities outside the circle (in grey).

Disjunction. Whenever 'A' and 'B' are propositions, we have another proposition, 'Either A or B', which we'll call the *disjunction* of A and B, and we'll abbreviate '$(A \vee B)$'. (Pay attention to the parenthases; they will be important later on.)

A	B	$(A \vee B)$
T	T	T
T	F	T
F	T	T
F	F	F

Table 2.2: Truth-table for '$\vee$'. '$(A \vee B)$' is true whenever either 'A' or 'B' is true.

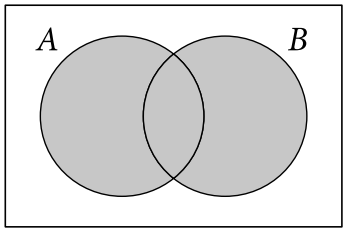


Figure 2.4: '$(A \vee B)$' is true at all possibilities inside either the A circle or the B circle (in grey).

Whether '$(A \vee B)$' is true or false seems to be completely determined by whether 'A' and 'B' are true or false. Whenever at least one of 'A' and 'B' are true, '$(A \vee B)$' is true.

And whenever both 'A' and 'B' are false, '$(A \vee B)$' is false. We can summarize this with the truth-table in table **??**, or with the Euler diagram in figure **??**.

You might expect that 'Either A or B' will be *false* if both A and B are true. That is called the *exclusive* 'or' (sometimes written 'xor'). The 'or' we're writting '$\vee$', on the other hand, is called the *inclusive* 'or'. An exclusive 'or' means 'A or B, but not both', whereas an inclusive 'or' means 'A or B, or perhaps both'. In this class, whenever I say 'or', I mean the *inclusive* 'or'.

Conjunction. Whenever 'A' and 'B' are propositions, we have another proposition, 'Both A and B', which I'll abbreviate '$(A\&B)$'. (Notice the parentheses, these are important.) This sentence is called the *conjunction* of A and B. It is true if and only if both 'A' and 'B' are true.

A	B	$A\&B$
T	T	T
T	F	F
F	T	F
F	F	F

Table 2.3: Truth-table for '&'. '$(A\&B)$' is true when both 'A' and 'B' is true.

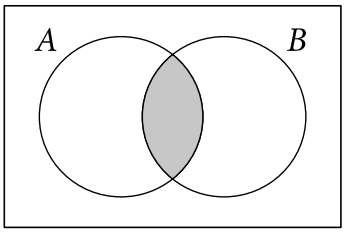
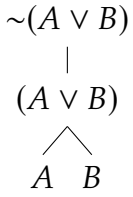
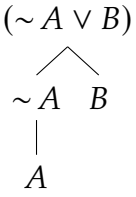


Figure 2.5: '$(A\&B)$' is true at all possibilities inside both the A circle and the B circle (in grey).

More Complicated Propositions. By combining these operations, we can build up more complicated propositions. For instance, consider the proposition '$\sim(A \vee B)$'. This is built up out of the disjunction '$(A \vee B)$', and negation '$\sim$'. Notice first that the parentheses here are important. '$\sim(A \vee B)$' is a different proposition than '$(\sim A \vee B)$'. They are distinguished by the order in which they are 'built up' out of the basic propositions 'A' and 'B'.

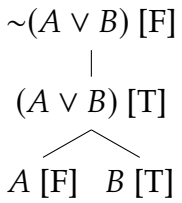


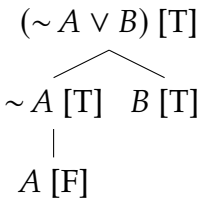


That is, to 'build up' '$\sim(A \vee B)$', we would first introduce a '$\vee$' and *next* introduce a '$\sim$'. But to 'build up' '$(\sim A \vee B)$', we would *first* introduce a '$\sim$' (to get '$\sim A$'), and *next* introduce a '$\vee$'. The final operator we introduce when building a complex proposition up is called the *main* operator of that proposition. For instance, the main operator of '$\sim(A \vee B)$' is $\sim$, and the main operator of '$(\sim A \vee B)$' is '$\vee$'.

We can use truth-tables to determine when more complicated propositions like these are true and false. We do so by determining whether their more basic parts are true or

false, and then we use the truth-tables for the operators $\sim$, $\vee$, and $\&$ to determine whether their more complicated parts are true or false. For instance, suppose that A is false and B is true. Then, we can work out whether '$\sim(A \vee B)$' and '$(\sim A \vee B)$' are true or false by 'working our way up' from the bottom to the top:





We can likewise do this for every possibility for whether A and B are true or false. Here's how: first, write out a truth-table with all the possible combinations of truth and falsehood for A and B in the rows, and transfer over the truth-values from the left-hand columns so that they are sitting beneath A and B, like this:

A	B	$(\sim$	A	$\vee$	$B)$
T	T		T		T
T	F		T		F
F	T		F		T
F	F		F		F

Next, we need to work out whether '$\sim A$' is true or false in each row of the table. Since '$\sim$' is the main operator of '$\sim A$', I'll write the truth-values for '$\sim A$' underneath '$\sim$', like so:

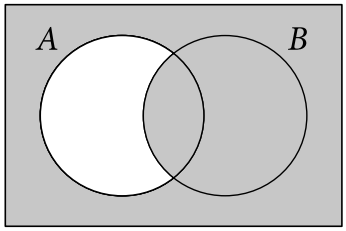
A	B	$(\sim$	A	$\vee$	$B)$
T	T	F	T		T
T	F	F	T		F
F	T	T	F		T
F	F	T	F		F

Finally, we can work out whether '$(\sim A \vee B)$' is true or false in each row in the truth-table. Since '$\vee$' is the main operator of '$(\sim A \vee B)$', I'll write the truth-value for this whole claim underneath '$\vee$', like so:

A	B	$(\sim$	A	$\vee$	$B)$
T	T	F	T	T	T
T	F	F	T	F	F
F	T	T	F	T	T
F	F	T	F	T	F

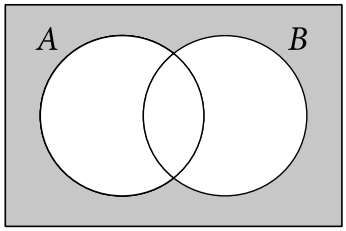
To help remind ourselves which column of truth-values is associated with the main operator of the proposition, we can place a box around that column. Note that we can view the rows of this truth-table as corresponding to regions in an Euler diagram.

A	B	$(\sim$	A	$\vee$	$B)$
T	T	F	T	T	T
T	F	F	T	F	F
F	T	T	F	T	T
F	F	T	F	T	F



We can do the same thing for a different proposition like $\sim(A \vee B)$:

A	B	$\sim$	$(A$	$\vee$	$B)$
T	T	F	T	T	T
T	F	F	T	T	F
F	T	F	F	T	T
F	F	T	F	F	F



The parentheses are important for disambiguating propositions like '$\sim(A \vee B)$' and '$(\sim A \vee B)$'. But, if the proposition *begins* and *ends* with parentheses, then we can safely ignore these starting and ending parentheses. So, going forward, I won't bother writing parentheses like these. So I'll write '$(A\&(B \vee C))$', for instance, as '$A\&(B \vee C)$'. But we should remember that, if we negate this proposition, we will get '$\sim(A\&(B \vee C))$', and *not* '$\sim A\&(B \vee C)$'.

Formal Validity. We can now show that any argument with the following *form* will be valid:

$$A \vee B$$
$$\sim A$$
$$\therefore\ B$$

Consider the truth-table below.

A	B	A	$\vee$	B	$\sim$	A	$\therefore$	B
T	T	T	T	T	F	T		T
T	F	T	T	F	F	T		F
F	T	F	T	T	T	T		T
F	F	F	F	F	T	T		F

Here, after the initial columns of truth-values for 'A' and 'B', I've written out the premises and the conclusion of the argument form. The question to ask ourselves is this: are there any rows of the table in which the premises are both true, yet the conclusion is false? Well, both premises are only true in row 3 of the table (the row in which A is false and B is true. And, in that row, the conclusion is true. So we can conclude that there's no way for the premises to be true while the conclusion is false. So we know that the argument is valid.

An *argument form* is valid if and only if there's no row of the truth-table in which the premises are true and the conclusion is false.

An argument is *formally valid* if and only if it has a form which is valid.

Formal validity is not the same thing as validity. If an argument is formally valid, then it is valid, also. But it is possible for an argument to be valid without being formally valid. For instance, consider this argument:

John is taller than himself

$\therefore$ Snow is white

This argument is valid—there's no possibility in which its premises is true while its conclusion is false (because there's no possibility in which its premise is true). But it is not formally valid. The only form this argument has is $A \therefore B$, which is not a valid form.

However, if every row of the truth-table represents a genuine possibility, then formal validity and regular validity will come to the same thing.

Exercise 1. *Use truth-tables to determine whether the argument form* $\sim(A \vee B), \therefore \sim A$ *is valid.*

A	B		$\sim$	$(A$	$\vee$	$B)$		$\therefore$	$\sim$	A
T	T									
T	F									
F	T									
F	F									

Formal Tautologies and Contradictions.

A proposition is a *formal tautology* if and only if it is true in *every* row of the truth-table.

Exercise 2. *Use a truth-table to determine whether* $C \vee \sim(C \& D)$ *is a formal tautology.*

C	D		C	$\vee$	$\sim$	$(C$	$\&$	$D)$

Exercise 3. *Use a truth-table to determine whether* $\sim B \vee (\sim A \vee (\sim C \vee B))$ *is a formal tautology.*

A	B	C	$\sim$	B	$\vee$	$(\sim$	A	$\vee$	$(\sim$	C	$\vee$	$B))$

A proposition is a *formal contradiction* if and only if it is false in every row of the truth-table.

Exercise 4. *Use a truth-table to determine whether* $(J\&K)\&(\sim J \vee \sim K)$ *is a formal contradiction.*

J	K	$(J$	$\&$	$K)$	$\&$	$(\sim$	J	$\vee$	$\sim$	$K)$

Formal Equivalence.

Two propositions are *formally equivalent* if and only if there is no row of the truth-table in which one of the propositions is true and the other is false.

That is: propositions are formally equivalent if and only if there is no row of the truth-table in which they have different truth-values.

Some formal equivalences which it will be useful to know about: (a) 'A' and '$\sim\sim A$' (b) '$A\&B$' and '$B\&A$' (c) '$A \vee B$' and '$B \vee A$' (d) '$A\&(B\&C)$' and '$(A\&B)\&C$' (e) '$A \vee (B \vee C)$' and '$(A \vee B) \vee C$' (f) '$\sim(A \vee B)$' and '$\sim A\& \sim B$' (g) '$\sim(A\&B)$' and '$\sim A \vee \sim B$' (h) '$A\&(B \vee C)$' and '$(A\&B) \vee (A\&C)$' (i) '$A \vee (B\&C)$' and '$(A \vee B)\&(A \vee C)$'.

Exercise 5. *Use a truth-table to determine whether* $\sim(A\&B)$ *and* $\sim A \vee \sim B$ *are formally equivalent.*

A	B	$\sim$	$(A$	$\&$	$B)$	$\sim$	A	$\vee$	$\sim$	B

Mutual Exclusivity. We can also use truth-tables to test for mutually exclusivity. If there's no row of the truth-table in which both propositions are true, then those propositions are mutually exclusive.

Exercise 6. *Use a truth-table to show that '$\sim(A \vee B)$' and 'A' are mutually exclusive.*

A	B	$\sim$	$(A$	$\vee$	$B)$	A

Jointly Exhaustive. We can use truth-tables to show that propositions are jointly exhaustive. If we have a collection of propositions and there's no row of the truth-table in which every one of those propositions is false, then the propositions are jointly exhaustive.

Exercise 7. *Use truth-tables to show that '$(A\&B) \vee (\sim A\& \sim B)$' and '$(A\& \sim B) \vee (\sim A\&B)$' are jointly exhaustive.*

A	B	$(A$	$\&$	$B)$	$\vee$	$(\sim$	A	$\&$	$\sim$	$B)$	$(A$	$\&$	$\sim$	$B)$	$\vee$	$(\sim$	A	$\&$	$B)$

Careful consideration of exercise 7 will help you answer the puzzle raised at the start of this chapter.

1. Fill in the missing values in the following table (write all fractions and odds in the simplest terms). [48pts]

Decimal	Percentage	Fraction	Odds
0.14			
0.99			
0.2			
0.08			
	67%		
	75%		
	40%		
	18%		
		7/10	
		17/25	
		4/5	
		47/50	
			1 : 1
			2 : 3
			1 : 9
			19 : 1

2. For each of the following collections of propositions, say whether the propositions in the collection are mutually exclusive or not. Then, say whether they are jointly exhaustive or not.

   (a) The die will land on a number less than 5. The die will land on a number greater than 2. (Assume we have rolled a standard six-sided die.) [20pts]

   (b) The Republican's plan would stop inflation. The Democrats' plan would stop inflation. Neither plan would stop inflation. (Assume that both Democrats and Republicans have a plan to stop inflation.) [20pts]

   (c) More people died in the Falklands war than died in the battle of Gettysberg. The Russian revolution was in the 19th century. (Assume that the Falklands war, the battle of Gettysberg, and the Russian revoluation all occurred.) [20pts]

   (d) The coin will land heads on the first flip. The coin will land tails on the first flip and heads on the second flip. The coin will land tails on both flips. (Assume that we will flip a coin twice, and that the coin will either land heads or tails on each flip.) [20pts]

(e) Al is taller than Betty. Betty is taller than Carol. Carol is taller than Al. (Assume that Al, Betty, and Carol all exist, and all have heights.) [30pts]

3. Use probability trees to answer the following questions.

   (a) i. We will flip a fair coin three times. Jack will win if the coin lands heads on the first flip, tails on the second flip, and heads on the third flip. Mary will win if the coin lands tails on all three flips. How probable is it that either Jack wins or Mary wins? [30pts]

   ii. Suppose you learn that Mary did not win. Incorporating this new information, how probable is it that Jack won? [30pts]

   (b) i. We will flip a fair coin, and then roll a fair 3-sided die. You will win if the coin lands heads or if the coin lands tails and the die comes up 1 or 2. What is the probability that you win? [30pts]

   ii. Suppose you learn that the die did not land on 2. Incorporating this new information, what is the probability that you won? [30pts]

   (c) I randomly draw a card from a standard deck (with 52 cards in total, 13 of each suit) and place it back. Then, you randomly draw a card. What is the probability that our cards have the same suit? [30pts]

   (d) I randomly draw a card from a standard deck (with 52 cards in total, 13 of each suit) and *do not* place it back. Then, you randomly draw a card from the deck. What is the probability that our cards have the same suit? [30pts]

   (e) John has three children. Mary asks him "Do you have a boy?" John answers "yes". Incorporating this new information, how probable is it that John has exactly two boys? [30pts]

   (f) There are three chests. The first chest contains two gold coins. The second chest contains two silver coins. The third chest contains one gold coin and one silver coin. The chests are indistinguishable on the outside. One of the three chests is selected randomly and then one coin is selected randomly from that chest. It is a gold coin. Incorporating this new information, what is the probability that the remaining coin is also gold? [30pts]

4. Say whether the following inferences are valid or invalid. If they are invalid, describe a possibility in which the premises are true but the conclusion is false.

   (a) Every raven I've ever seen has been black. Therefore, all ravens are black. [20pts]

   (b) Aristotle was the first philosopher. Plato died before Aristotle was born. Therefore, Plato was not a philosopher. (Assume that 'philosopher' means the same thing each time it is used.) [20pts]

(c) Either it always rains in Los Angeles or it never rains in Los Angeles. On Monday, January 5th, 2026, it rained in Los Angeles. Therefore, it always rains in Los Angeles. [20pts]

(d) Trump is the president. Trump is not the president. Therefore, Harris is the president. (Assume that "is the president" means exactly the same thing each time it is used.) [20pts]

(e) Mt. Everest is not taller than itself. Therefore, no mountain is taller than itself. [20pts]

3 | Independence

Goal: understand when some propositions *irrelevant* to other propositions, in the sense that they give us no information whatsoever about those other propositions.

Probabilistic Truth-Tables and Probabilistic Euler Diagrams

We've seen two different ways to represent the possibilities in which propositions are true and false: *truth-tables* and *Euler diagrams*. And one very helpful way of representing a probability function uses these same tools. For instance, suppose that we're going to flip a fair coin twice. We can think about this situation with two propositions: H_1 (which says that the coin lands heads on the first flip) and H_2 (which says that the coin lands heads on the second flip). And we can represent a probability distribution over these propositions like so:

H_1	H_2	Pr
T	T	$1/4$
T	F	$1/4$
F	T	$1/4$
F	F	$1/4$

Table 3.1: A probabilistic truth-table

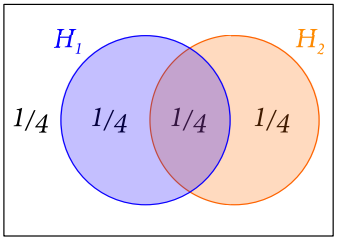


Figure 3.1: A probabilistic Euler diagram

Both of these contain exactly the same information. They both tell us that

$$\Pr(H_1 \& H_2) = \Pr(H_1 \& \sim H_2) = \Pr(\sim H_1 \& H_2) = \Pr(\sim H_1 \& \sim H_2) = 1/4$$

This gives us a probability distribution over four *basic* propositions. But from this, we can work out the probabilities for many more propositions. To do so, we just have to look at which rows of the truth-table the more complicated proposition is true in. The proposition's probability will be equal to the sum of the probabilities given to each row in which it is true.

> The probability of a proposition is equal to the sum of the probabilities given to each row of the truth-table in which the proposition is true.

For instance, given the probabilistic truth-table above, we can determine that the probability of at least one heads ($H_1 \vee H_2$)

H_1	H_2	Pr	H_1	$\vee$	H_2
T	T	$1/4$	T	T	T
T	F	$1/4$	T	T	F
F	T	$1/4$	F	T	T
F	F	$1/4$	F	F	F

Since this proposition is true in rows 1, 2, and 3, its probability must be equal to the sum of the probability given to these rows:

$$\Pr(H_1 \vee H_2) = 1/4 + 1/4 + 1/4 = 3/4$$

Probabilistic Euler diagrams work similarly: We can think about laying an Euler diagram out on a table, and then slopping a bunch of mud on top of it. Regions of the Euler diagram which are more probable get more mud on top of them. Regions of the diagram which are less probable get less mud on top of them. The probabilistic Euler diagram in figure 1 tells us that one fourth of the total mud is sitting in each of the four regions—so the mud is evenly spread. If you want to know the probability of a proposition, you just ask *what proportion* of the total mud is sitting on top of regions in which that proposition is true.

> The probability of a proposition is equal to the sum of the probabilities given to each region of the Euler diagram in which the proposition is true.

(By 'region', I mean a part of the diagram which is enclosed by lines, and which no other lines cross.)

Probabilistic Independence and Conditional Probability

Here is a definition:

> **Probabilistic Independence** Two propositions, A and B, are *probabilistically independent* if and only if
>
> $$\Pr(A\&B) = \Pr(A) \cdot \Pr(B)$$

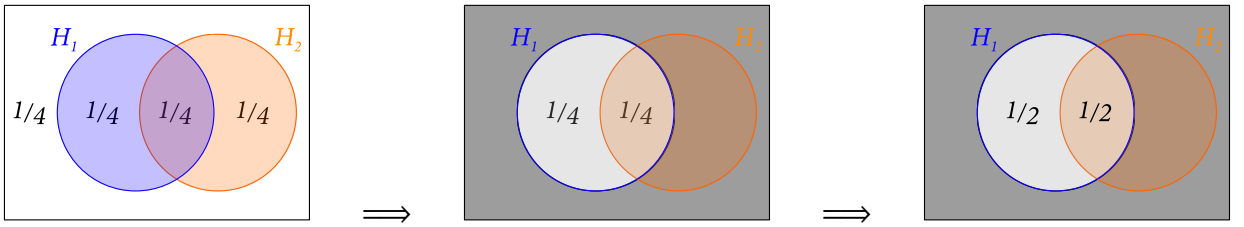
For instance, given the probabilities from table 1, the propositions H_1 and H_2 are probabilistically independent. That's because

$$\Pr(H_1 \& H_2) = 1/4 = \Pr(H_1) \cdot \Pr(H_2) = 1/2 \cdot 1/2$$

To appreciate this definition, let me introduce another:

> **Conditional Probability** Given any two propositions, A and B, if $\Pr(B) > 0$, then the *conditional* probability of A, given B, $\Pr(A \mid B)$, is defined to be
>
> $$\Pr(A \mid B) \stackrel{\text{def}}{=} \frac{\Pr(A \& B)}{\Pr(B)}$$

A *conditional* probability $\Pr(A \mid B)$ tells you how probable A is *on the supposition* that B is true. Probabilistic Euler diagrams give us a helpful way of thinking about what a conditional probability is telling us. To illustrate, let's ask about the conditional probability of the coin landing heads on the second flip, H_2, *given that* it lands heads on the first flip, H_1. To calculate this, we can imagine first sweeping away all of the mud which sits on top of regions in which H_1 is false. After this is done, we ask how much of the *remaining* mud lies on top of regions in which H_1 is true. (See figure 8.4b.)



(a) original probability

(b) proportion of the original probability in H_1

(c) proportion of *remaining* probability, after $\sim H_1$ probability is removed

Figure 3.2: Calculating a conditional probability $\Pr(H_2 \mid H_1)$. *First, remove all probability from regions incompatible with* H_1. *Then, ask yourself: how much of the* remaining *probability lies in regions in which* H_2 *is true?*

These operations are encoded in the definition of conditional probability. To calculate $\Pr(A \mid B)$, we first ask: how much of the original probability sits on top of regions in which both A and B are true? (This is tantamount to asking: how much of the original probability will be left, after we take away any probability incompatible with B?) Then, we ask: of the *remaining* probability, what proportion will be left sitting on top of regions in which A is true? To calculate this, we have to divide by the total amount of probability remaining. And this is just $\Pr(B)$—since this is all the probability that we didn't take away. Dividing

through by $\Pr(B)$ is known as *renormalizing*. Renormalizing allows the total amount of probability left behind to still add up to one.

The conditional probability of A, *given* B, is the probability that you would give to A, if you were to learn that B is true.

Conditional probabilities allow us to give another way of understanding probabilistic independence:

> **Probabilistic Independence (v2)** For any two propositions, A and B, if $\Pr(B) > 0$, then A and B are probabilistically independent if and only if
>
> $$\Pr(A \mid B) = \Pr(A)$$

To appreciate how this follows from our two earlier definitions, note that, for independent A and B,

$$\Pr(A \mid B) = \frac{\Pr(A\&B)}{\Pr(B)} = \frac{\Pr(A) \cdot \Pr(B)}{\Pr(B)} = \Pr(A)$$

So probabilistic independence corresponds to *zero information*. If the probability of A wouldn't change, were you to learn B, then B does not tell you anything at all about whether A. Probabilistic independence is a powerful tool in the theory of probability: it tells you when information can be safely *ignored*.

Properties of Independence.

1. Probabilistic independence is *symmetric*—if A is probabilistically independent of B, then B is probabilistically independent of A. This can be easy to see with our first defintion of probabilistic independence, since it follows straightfowardly from the symmetry of multiplication. It can be more difficult to see with our second definition, since it might seem that we could have $\Pr(A \mid B) = \Pr(A)$ even while $\Pr(B \mid A) \neq \Pr(B)$. However, this is impossible so long as both $\Pr(A)$ and $\Pr(B)$ are greater than zero:

$$\begin{aligned}
\Pr(A \mid B) &= \Pr(A) \\
\frac{\Pr(A\&B)}{\Pr(B)} &= \Pr(A) \\
\frac{\Pr(A\&B)}{\Pr(A)} &= \Pr(B) \\
\frac{\Pr(B\&A)}{\Pr(A)} &= \Pr(B) \\
\Pr(B \mid A) &= \Pr(B)
\end{aligned}$$

2. Probabilistic independence is *not* transitive. It is possible for A to be probabilistically independent of B, for B to be probabilistically independent of C, but for A to not be probabilistically independent of C. For instance, consider the probabilistic truth-table shown

in table 2 (or the probabilistic Euler diagram shown in figure 3). There, the probability of A is one third, which is also the probability of A, given B. And the probability of C is one third, which is also the probability of C, given B. So A and B are independent. And B and C are independent. But A and C are *not* independent. A gives us information about C, and C gives us information about A. The probability of A, given C, is *zero*, which is different than the probability of A ($1/3$).

A	B	C	Pr
T	T	T	0
T	T	F	$1/6$
T	F	T	0
T	F	F	$1/6$
F	T	T	$1/6$
F	T	F	$1/6$
F	F	T	$1/6$
F	F	F	$1/6$

Table 3.2: A probabilistic truth-table showing that probabilistic independence is not transitive

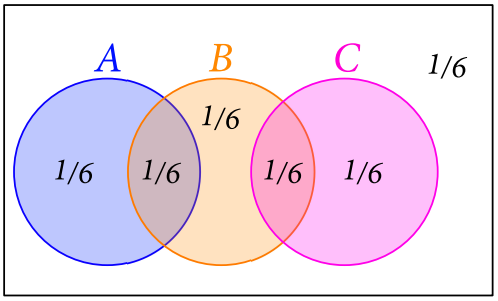


Figure 3.3: A probabilistic Euler diagram showing that probabilistic independence is not transitive

3. If A and B are probabilistically independent, then $\sim A$ and $\sim B$ will *also* be probabilistically independent—though we'll have to wait until we have more probability rules on the table to give a careful proof of this fact.

4. We should clearly distinguish between *pairwise* independence and *mutual* independence. It is possible for us to have three propositions, A, B, and C, which are *pairwise* independent, even though the probability of $A\&B\&C$ is not the product of the probabilities of A, B, and C. For instance, consider the probability shown below:

A	B	C	Pr
T	T	T	$1/4$
T	T	F	0
T	F	T	0
T	F	F	$1/4$
F	T	T	0
F	T	F	$1/4$
F	F	T	$1/4$
F	F	F	0

Table 3.3: A probabilistic truth-table showing that pairwise independence doesn't imply mutual independence

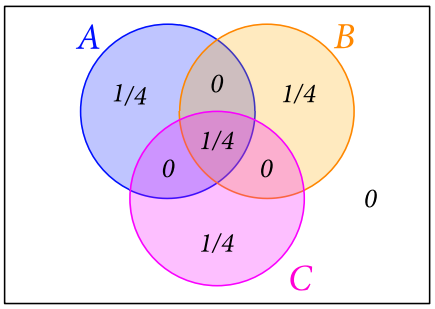


Figure 3.4: A probabilistic Euler diagram showing that pairwise independence doesn't imply mutual independence

Mutual Independence Three propositions, A, B, and C, are *mutually independent* if and only if:

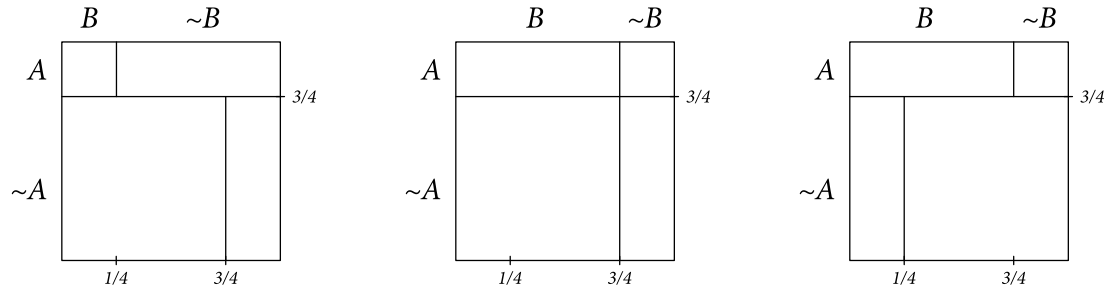
$$\Pr(A\&B) = \Pr(A) \cdot \Pr(B)$$
$$\Pr(B\&C) = \Pr(B) \cdot \Pr(C)$$
$$\Pr(A\&C) = \Pr(A) \cdot \Pr(C)$$
$$\text{and} \quad \Pr(A\&B\&C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$$

More generally, N propositions, $A_1, A_2, \ldots, A_N$ are mutually independent if and only if, for any subset of these propositions, the conjunction of the propositions in that subset is equal to the product of the probability of the propositions in that subset. For instance, if we know that A, B, C, and D are mutually independent, then we know that $\Pr(A\&B\&D) = \Pr(A) \cdot \Pr(B) \cdot \Pr(D)$, and we know that $\Pr(B\&C) = \Pr(B) \cdot \Pr(C)$, and likewise for every other conjunction of the propositions A, B, C, and D.

Eikosograms

If we're only interested in two propositions, then we can visually represent a probability distribution with an *eikosogram*. An eikosogram represents probability with *area*. It gives us a 1×1 square, and then sub-divides this square into rectangles which represent the propositions $A\&B$, $\sim A\&B$, $A\&\sim B$, and $\sim A\&\sim B$. For instance, consider the eikosgram from figure 5a. Here, the probability of $A\&B$ is the $1/4 \times 1/4$ square in the upper left corner. The area of this square is $1/16$, so that's the probability of $A\&B$. The probability of $\sim A\&B$ is the larger $3/4 \times 3/4$ square in the lower left. The area of this square is $9/16$, so that's the probability of $\sim A\&B$. Likewise, the probability of $A\&\sim B$ is the area of the $1/4 \times 3/4$ rectangle: $3/16$.

Looking at Eikosograms allows us to visually connect the two definitions of probabilistic independence: if A and B are probabilistically independent, then the area of the $A\&B$ rectangle in the upper left hand corner is just the product of the probability of A (the length of the rectangle along the y-axis) and the probability of B (the width of the



(a) A lowers the probability of B, $\Pr(B \mid A) < \Pr(B)$

(b) A and B are independent, $\Pr(A\&B) = \Pr(A) \cdot \Pr(B)$

(c) A raises the probability of B, $\Pr(B \mid A) > \Pr(B)$

Figure 3.5: Eikosograms

rectangle along the x-axis). And, if this is the case, then learning that you're in the upper A-rectangle won't change the probability of B at all. For instance, in figure 5b, B takes up $3/4$ of the total area of the square, and it also takes up $3/4$ of the upper A-rectangle at the top of the eikosogram.

Independent and Identically Distributed Propositions

Example 7. *Each time a slot machine lever is pulled, there is a tiny probability that it pays out, and whether it pays out on any given pull is probabilistically independent of its previous history of paying out. You have a choice between two slot machines*: *one that has just paid out, and another which hasn't paid out in months. Should you choose one machine over the other*? *Or does it make no difference*?

A very common thought is that you should go with the second slot machine—the one that hasn't paid out in months. This common thought is known as *the gambler's fallacy*. It is a fallacy because both slots machines have exactly the same probability of paying out on the next pull. Knowing how they have paid out in the past does not affect this probability at all.

Each pull of the slot machine is probabilistically independent of the others, and each pull has exactly the same probability of paying out. So each pull of the slot machine is *independent* and has an *identical* probability distributions. This kind of phenomenon is common enough that it has been given an acronym:

IID A collection of propositions $A_1, A_2, \ldots, A_N$ are *independent* and *identically distributed* (IID) if and only if:

1. The propositions are mutually independent; and
2. They all have the same probability, $\Pr(A_1) = \Pr(A_2) = \cdots = \Pr(A_N)$.

Example 8. *I will draw a card from a shuffled standard* (52 *card*) *deck of playing cards, and then return it to the deck and shuffle. I will then draw another card, return it to the deck, shuffle, and so on. What is the probability that I draw a heart five times in a row? What is the probability that I draw a heart on the fifth draw, given that I drew a heart on the first four draws*?

In this case, the propositions $\heartsuit_1$ (my first draw is a heart), $\heartsuit_2$ (my second draw is a heart), $\heartsuit_3$ (my third draw is a heart), and so on are IID. So we know that

$$\begin{aligned}\Pr(\heartsuit_1 \& \heartsuit_2 \& \heartsuit_3 \& \heartsuit_4 \& \heartsuit_5) &= \Pr(\heartsuit_1) \cdot \Pr(\heartsuit_2) \cdot \Pr(\heartsuit_3) \cdot \Pr(\heartsuit_4) \cdot \Pr(\heartsuit_5) \\ &= 1/4 \cdot 1/4 \cdot 1/4 \cdot 1/4 \cdot 1/4 \\ &= 1/4^5 = 1/1024\end{aligned}$$

(Notice the importance of our assumption that the draws were *mutually* independent, and not just that they were *pairwise* independent.) Because the draws are IID, we also know that

$$\Pr(\heartsuit_1 \& \heartsuit_2 \& \heartsuit_3 \& \heartsuit_4) = 1/4 \cdot 1/4 \cdot 1/4 \cdot 1/4 = 1/4^4$$

So:

$$\Pr(\heartsuit_5 \mid \heartsuit_1 \& \heartsuit_2 \& \heartsuit_3 \& \heartsuit_4) = \frac{\Pr(\heartsuit_1 \& \heartsuit_2 \& \heartsuit_3 \& \heartsuit_4 \& \heartsuit_5)}{\Pr(\heartsuit_1 \& \heartsuit_2 \& \heartsuit_3 \& \heartsuit_4)} = \frac{1/4^5}{1/4^4} = 1/4$$

Many nice probabilistic processes are IID, but not all are.

Example 9. *I will draw cards from a shuffled deck one at a time. What is the probability that I draw a heart twice in a row*?

In this case, the propositions $\heartsuit_1$ and $\heartsuit_2$ are *not* independent and identically distributed. That's because they are not independent. Getting a heart on the first draw makes it less likely that you'll get a heart on later draws. It would be a *mistake* to calculate this probability by just multiplying together $\Pr(\heartsuit_1)$ and $\Pr(\heartsuit_2)$. Next time, we'll learn more about how to calculate probabilities like these.

Example 10. *I have two dice*: *one is a D*6 (*containing six different sides*) *and the other is a D*12 (*containing twelve different sides*). *Both dice are fair. I will roll the D*6 *and then the D*12. *What is the probability that I get snake eyes* (*double ones*)?

In this case, the propositions $\boxdot_1$ (the D6 lands on one) and $\boxdot_2$ (the D12 lands on one) are *not* IID. They are independent, but they are not identically distributed. So we need to distinguish the probability of the first roll landing on one from the probability of the second roll landing on one:

$$\begin{aligned}\Pr(\boxdot_1 \& \boxdot_2) &= \Pr(\boxdot_1) \cdot \Pr(\boxdot_2) \\ &= 1/6 \cdot 1/12 \\ &= 1/72\end{aligned}$$

A Warning About Independence

It's easy to make unwarranted assumptions about IID processes. For instance,

Example 11. *We will flip a fair coin twice. You learn that it lands heads at least once. Incorporating this new information, what's the probability that it lands heads both times*?

You *might* try to reason about this example in the following way: well, the outcome of the two flips are independent of each other, so learning that it landed heads on one of the flips shouldn't affect the probability that it lands heads on the *other* flip. By independence, the probability that it lands heads on both flips should be $1/2$. But this would be a mistake.

H_1 *is* independent of H_2. But neither H_1, H_2, nor $H_1 \& H_2$ is independent of $H_1 \vee H_2$. Using the definition of conditional probability,

$$\Pr(H_1 \& H_2 \,|\, H_1 \vee H_2) = \frac{\Pr((H_1 \& H_2) \& (H_1 \vee H_2))}{\Pr(H_1 \vee H_2)}$$

And using truth-tables,

H_1	H_2	Pr	$(H_1$	&	$H_2)$	&	$(H_1$	$\vee$	$H_2)$	H_1	$\vee$	H_2
T	T	$1/4$	T	T	T	T	T	T	T	T	T	T
T	F	$1/4$	T	F	F	F	T	T	F	T	T	F
F	T	$1/4$	F	F	T	F	F	T	T	F	T	T
F	F	$1/4$	F	F	F	F	F	F	F	F	F	F

we can see that the probability of $(H_1 \& H_2) \& (H_1 \vee H_2)$ is $1/4$, while the probability of $H_1 \vee H_2$ is $3/4$. So

$$\Pr(H_1 \& H_2 \,|\, H_1 \vee H_2) = \frac{1/4}{3/4} = \frac{1}{3}$$

Example 12. *You have a choice between two sequences*: HHH *and* THH*. We will flip a fair coin until one of these sequences of outcomes appears. You will win a prize if your chosen sequence appears before the other one. Should you prefer one of these sequences to the other, or does it not make any difference which one you choose*?

It's natural to think: the outcomes of the flips are probabilistically independent of each other, so it won't make any difference which sequence I choose. But this is incorrect. Notice the following: if the first coin lands tails, it is impossible for HHH to win. Once the first flip lands tails, the first time the sequence HHH appears, it will have to be immediately preceded by a T—just for illustration, perhaps the coin will land like this:

$$TTHTTTTHTHTHHH$$

On flips 12–14, we finally get three heads landings in a row. But because this was the first time we got three heads in a row, the sequence HHH was immediately preceded by a T. So we got the sequence THH *before* we got the sequence HHH. So THH wins. And this is going to happen no matter what, so long as the first coin lands tails.

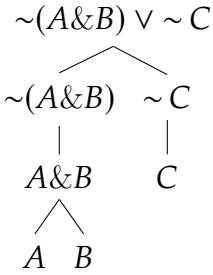
In fact, all the same reasoning goes through so long as the coin *ever* lands tails. Even if it lands tails on the second or the third flip, there will be no way for HHH to win, by exactly the same reasoning. So the only way for HHH to win is for the first three flips to land heads right out of the gate. And the probability of this happening is $\Pr(H_1 \& H_2 \& H_3) = \Pr(H_1) \cdot \Pr(H_2) \cdot \Pr(H_3) = 1/2 \cdot 1/2 \cdot 1/2 = 1/8$.

So the odds in favor of THH winning are $7 : 1$, and you should definitely prefer THH to HHH.

1. For each of the propositions below, show how they would be 'built up' from their parts, and say what their main operator is. (For clarity, the first problem is done for you.)

   (a) $\sim(A\&B) \vee \sim C$

   Answer: It is built up like this:

   Since the final operator added when building it up was '$\vee$', this is the main operator.

   (b) $A\&(\sim B \vee C)$ [20pts]

   (c) $D \vee \sim(E\&F)$ [20pts]

   (d) $(H\&I) \vee \sim\sim(J \vee \sim K)$ [20pts]

   (e) $(L\&(M\&N)) \vee ((\sim L\& \sim M)\& \sim N)$ [20pts]

2. Use truth-tables to determine whether the following arguments are formally valid or not. (When you make your truth-table, be sure to put a box around the column corresponding to the *main operator* of each proposition. And be sure to say whether the argument is formally valid or not.)

   (a) $(H\&V) \vee (\sim H\& \sim V), H \therefore V$ [30pts]

   (b) $A\& \sim A \therefore B$ [30pts]

   (c) $X \vee (Y\&Z) \therefore (X \vee Y)$ [40pts]

3. Use truth-tables to determine whether the following propositions are formal tautologies, formal contradictions, or neither.

   (a) $\sim A\&(A\&B)$ [30pts]

   (b) $\sim(C\& \sim C)$ [30pts]

   (c) $\sim D \vee (\sim E \vee D)$ [30pts]

4. Write out a truth-table for the following pairs of propositions and then say (i) whether the propositions are mutually exclusive, (ii) whether they are jointly exhaustive, and (iii) whether they are formally equivalent.

(a) $A \vee \sim A$ and $B \vee \sim B$ [30pts]

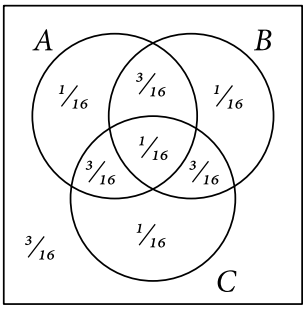
(b) $\sim(C \vee D)$ and $\sim C \& \sim D$ [30pts]

(c) $E \vee F$ and $\sim E \& \sim F$ [30pts]

(d) $A \vee B$ and $(A \& \sim B) \vee ((A \& B) \vee (\sim A \& B))$ [30pts]

5. Consider the probabilistic truth-table below (the corresponding probabilistic Euler diagram is shown next to it).

A	B	C	Pr
T	T	T	1/16
T	T	F	3/16
T	F	T	3/16
T	F	F	1/16
F	T	T	3/16
F	T	F	1/16
F	F	T	1/16
F	F	F	3/16



(a) Using this table, work out the following probabilities. [50pts]

i. $\Pr(A)$

ii. $\Pr(B)$

iii. $\Pr(C)$

iv. $\Pr(A \& B)$

v. $\Pr(B \& C)$

vi. $\Pr(A \& C)$

vii. $\Pr(A \& (B \& C))$

viii. $\Pr(A \vee (B \& \sim C))$

ix. $\Pr(A \& (B \vee C))$

x. $\Pr(\sim A \& (B \vee C))$

(b) Are A and B probabilistically independent? [10pts]

(c) Are B and C probabilistically independent? [10pts]

(d) Are A and C probabilistically independent? [10pts]

(e) Are A and $B \& C$ probabilistically independent? [10pts]

(f) Are B and $A \& C$ probabilistically independent? [10pts]

(g) Are C and $A \& B$ probabilistically independent? [10pts]

(h) Is there anything surprising about your answers to (b) through (g)? [10pts]

4 | Probability Rules

Goal: learn rules we can use to work out the probabilities of complex propositions.

The Tautology and Contradiction Rules

The first rule tells us that certain propositions must always have the same probability. If a proposition is *guaranteed* to be true, then its probability must be 100%.

For any proposition T, if T is a tautology, then

$$\Pr(T) = 1$$

Example 13. *We will flip a coin twice. What is the probability that it either lands heads once or lands tails twice,* $\Pr((H_1 \vee H_2) \vee (\sim H_1 \& \sim H_2))$?

Since $(H_1 \vee H_2) \vee (\sim H_1 \& \sim H_2)$ is a formal tautology,

H_1	H_2	$(H_1$	$\vee$	$H_2)$	$\vee$	$(\sim$	H_1	$\&$	$\sim$	$H_2)$
T	T	T	T	T	T	F	T	F	F	T
T	F	T	T	F	T	F	T	F	T	F
F	T	F	T	T	T	T	F	F	F	T
F	F	F	F	F	T	T	F	T	T	F

we know that it is a tautology. And since it is a tautology, it must have a probability of 1.

$$\Pr((H_1 \vee H_2) \vee (\sim H_1 \& \sim H_2)) = 1$$

On the other hand, if a propositions is *guaranteed* to be false, then its probability must be 0%.

For any proposition C, if C is a contradiction, then

$$\Pr(C) = 0$$

Example 14. *Alfred and Betty both flip a coin. Alfred guesses that Betty's coin landed the same way as his, and Betty guesses that Alfred's coin landed differently than hers. What's the probability that neither's guess is correct*?

It is impossible that neither's guess is correct. If their coins landed on the same side, then Alfred's guess is correct. But if their coins landed differently, then Betty's guess is correct. So there's no way that both of their guesses could be wrong. So the probability that both of their guesses are wrong must be zero.

The Equivalence Rule

Our next rule tells us that equivalent propositions have the same probability—if there's no way for A to be true while B is false, and no way for B to be true while A is false, then the probabilities of A and B must be equal.

> For any propositions A and B, if A and B are equivalent, then
>
> $$\Pr(A) = \Pr(B)$$

Example 15. *We will flip a fair coin twice in a row. If 'H_1' is the proposition that the coin lands heads on the first flip, and 'H_2' is the proposition that the coin lands heads on the second flip, then what is the probability that neither flip lands heads,* $\Pr(\sim(H_1 \vee H_2))$?

To work this out, we can make use of the fact that '$\sim(H_1 \vee H_2)$' is logically equivalent to '$\sim H_1 \& \sim H_2$', together with the fact that the propositions $\sim H_1$ and $\sim H_2$ are independent, with a shared probability of $1/2$.

$$\begin{aligned}
\Pr(\sim(H_1 \vee H_2)) &= \Pr(\sim H_1 \& \sim H_2) && \text{[by the equivalence rule]} \\
&= \Pr(\sim H_1) \cdot \Pr(\sim H_2) && \text{[by independence]} \\
&= 1/2 \cdot 1/2 \\
&= 1/4
\end{aligned}$$

The Disjunction Rule

The next rule tells us how to relate the probability of $A \vee B$ to the probabilities of A and B—at least in the special case where A and B are mutually exclusive.

> For any propositions A and B, if A and B are mutually exclusive, then
>
> $$\Pr(A \vee B) = \Pr(A) + \Pr(B)$$

Example 16. *There is an urn containing* 30 *red balls,* 30 *green balls, and* 40 *yellow balls. The urn is shook and a ball is drawn at random. What is the probability that the drawn ball is either red or yellow?*

Since the drawn ball cannot be both red and yellow, R and Y are mutually exclusive. So

$$\begin{aligned} \Pr(R \vee Y) &= \Pr(R) + \Pr(Y) && \text{[by the disjunction rule]} \\ &= {}^{3}/_{10} + {}^{4}/_{10} \\ &= {}^{7}/_{10} \end{aligned}$$

What about if the disjuncts are not mutually exclusive?

Example 17. *There is an urn containing* 30 *red balls,* 30 *green balls, and* 40 *yellow balls. The urn is shook and a ball is drawn at random. We then return that ball to the urn and draw another. What is the probability that either the first ball is red or the second ball is yellow?*

In this case, the propositions R_1 (the first ball drawn is red) and Y_2 (the second ball drawn is yellow) are not mutually exclusive. It is possible that both are true—if the first ball drawn was red and the second ball drawn was yellow. So we cannot just use the disjunction rule. Think about what would happen if we did: if we added together $\Pr(R_1)$ and $\Pr(Y_2)$, then we'd add up all the probability sitting on top of R_1 with all of the probability sitting on top of Y_2. But this would include the probability in $R_1 \& Y_2$ *twice*. So we'd be double counting. But we could decide to just subtract off the probability in the overlap,

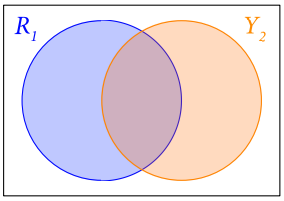


Figure 4.1: An Euler diagram for R_1 (the first ball drawn is red) and Y_2 (the second ball drawn is yellow).

$$\begin{aligned} \Pr(R_1 \vee Y_2) &= \Pr(R_1) + \Pr(Y_2) - \Pr(R_1 \& Y_2) \\ &= \Pr(R_1) + \Pr(Y_2) - \Pr(R_1) \cdot \Pr(Y_2) && \text{[by independence]} \\ &= {}^{3}/_{10} + {}^{4}/_{10} - {}^{3}/_{10} \cdot {}^{4}/_{10} \\ &= {}^{7}/_{10} - {}^{12}/_{100} \\ &= {}^{58}/_{100} = {}^{29}/_{50} = 58\% \end{aligned}$$

And the same trick will work in general: we can always add together the probability of A with the probability of B, and then subtract the probability from the 'overlapping' region $A \& B$ that we've counted twice, and this will give us the probability of $A \vee B$:

For any propositions A and B,

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \& B)$$

The Negation Rule

The next rule tells us how to relate the probability of $\sim A$ and the probability of A, for any proposition A.

> For any proposition A,
>
> $$\Pr(\sim A) = 1 - \Pr(A)$$

Notice that it follows from this rule that $\Pr(A) = 1 - \Pr(\sim A)$.

Example 18. *Every hiring company uses the same job search algorithm. The algorithm assigns a score, s, to your application, which represents the algorithm's evaluation of your quality as an applicant. For each job, it then passes your application along to a human with a probability equal to the score s. You know that your score is $1/3$, and you have applied to three jobs. Each new job, you have a new chance of getting your application passed along to a human; and the propositions "A human sees your application at job i" (for $i = 1, 2, 3$) are IID. What is the probability that a human ever sees your job application*?

In this example, we know something about the propositions H_1, H_2, and H_3, where H_i says that, at job i, a human looks at your application. In particular, we know that these propositions are IID, with a shared probability of $1/3$. We want to work out the probability that *someone* looks at your job application. This is the proposition

$$H_1 \vee H_2 \vee H_3$$

We *could* calculate this using the disjunction rule, but it quickly gets messy:

$$\begin{aligned}
&\Pr(H_1 \vee (H_2 \vee H_3)) \\
&= \Pr(H_1) + \Pr(H_2 \vee H_3) - \Pr(H_1 \& (H_2 \vee H_3)) \\
&= \Pr(H_1) + \Pr(H_2 \vee H_3) - \Pr((H_1 \& H_2) \vee (H_1 \& H_3)) \\
&= \Pr(H_1) + [\Pr(H_2) + \Pr(H_3) - \Pr(H_2 \& H_3)] - \Pr((H_1 \& H_2) \vee (H_1 \& H_3)) \\
&= \Pr(H_1) + \Pr(H_2) + \Pr(H_3) - \Pr(H_2 \& H_3) - [\Pr(H_1 \& H_2) + \Pr(H_1 \& H_3) - \Pr(H_1 \& H_2 \& H_3)] \\
&= \Pr(H_1) + \Pr(H_2) + \Pr(H_3) - \Pr(H_2 \& H_3) - \Pr(H_1 \& H_2) - \Pr(H_1 \& H_3) + \Pr(H_1 \& H_2 \& H_3) \\
&= \Pr(H_1) + \Pr(H_2) + \Pr(H_3) - \Pr(H_2) \cdot \Pr(H_3) - \Pr(H_1) \cdot \Pr(H_2) - \Pr(H_1) \cdot \Pr(H_3) \\
&\qquad + \Pr(H_1) \cdot \Pr(H_2) \cdot \Pr(H_3) \\
&= 1/3 + 1/3 + 1/3 - 1/3 \cdot 1/3 - 1/3 \cdot 1/3 - 1/3 \cdot 1/3 + 1/3 \cdot 1/3 \cdot 1/3 \\
&= 1 - 1/9 - 1/9 - 1/9 + 1/27 \\
&= 1 - 1/3 + 1/27 \\
&= 19/27 \approx 70\%
\end{aligned}$$

Things are *significantly* easier if we instead calculate the probability that *no one* looks at your job application:

$$\begin{aligned}
\Pr(\textit{no one sees}) &= \Pr(\sim(H_1 \vee H_2 \vee H_3)) \\
&= \Pr(\sim H_1 \& \sim H_2 \& \sim H_3) && \text{[by equivalence]} \\
&= \Pr(\sim H_1) \cdot \Pr(\sim H_2) \cdot \Pr(\sim H_3) && \text{[by independence]} \\
&= [1 - \Pr(H_1)] \cdot [1 - \Pr(H_2)] \cdot [1 - \Pr(H_3)] && \text{[by negation rule]} \\
&= [1 - 1/3] \cdot [1 - 1/3] \cdot [1 - 1/3] \\
&= 2/3 \cdot 2/3 \cdot 2/3 \\
&= 8/27
\end{aligned}$$

So $8/27$ is the probability that *no one* looks at your application. Applying the negation rule, we get that the probability that *someone* looks at your application must be:

$$\Pr(H_1 \vee H_2 \vee H_3) = 1 - \Pr(\sim(H_1 \vee H_2 \vee H_3)) = 1 - 8/27 = 19/27$$

Example 19. *As in example* 18*, except that now your score is* 1/10 *and you have applied to ten jobs. What's the probability that a human ever sees your application*?

Here, using the disjunction rule would be prohibitively difficult. But the negation rule makes the problem easy:

$$\begin{aligned}
\Pr(\textit{someone sees}) &= 1 - \Pr(\textit{no one sees}) && \text{[by the negation rule]} \\
&= 1 - \Pr(\sim H_1) \cdot \Pr(\sim H_2) \cdot \Pr(\sim H_3) \cdot \cdots \cdot \Pr(\sim H_{10}) && \text{[by independence]} \\
&= 1 - (9/10)^{10} \\
&\approx 65\%
\end{aligned}$$

Independence and Negation. When we were talking about probabilistic independence, I informed you of the following fact: if A and B are independent, then $\sim A$ and $\sim B$ will be independent, as well. At the time, we weren't in a position to see a *proof* of this fact. But with the equivalence, disjunction, and negation rules under our belts, we are in a position to appreciate why this is true. Start off by assuming that A and B are independent—that is, that $\Pr(A \& B) = \Pr(A) \cdot \Pr(B)$. Then, we will show that it must be that $\Pr(\sim A \& \sim B) = \Pr(\sim A) \cdot \Pr(\sim B)$:

$$\begin{aligned}
\Pr(\sim A \& \sim B) &= \Pr(\sim(A \vee B)) && \text{[by the equivalence rule]} \\
&= 1 - \Pr(A \vee B) && \text{[by the negation rule]} \\
&= 1 - \Pr(A) - \Pr(B) + \Pr(A \& B) && \text{[by the disjunction rule]} \\
&= 1 - \Pr(A) - \Pr(B) + \Pr(A) \cdot \Pr(B) && \text{[by independence]} \\
&= [1 - \Pr(A)] - \Pr(B) \cdot [1 - \Pr(A)]
\end{aligned}$$

$$
\begin{aligned}
&= \Pr(\sim A) - \Pr(B) \cdot \Pr(\sim A) && \text{[by the negation rule]} \\
&= \Pr(\sim A) \cdot [1 - \Pr(B)] \\
&= \Pr(\sim A) \cdot \Pr(\sim B) && \text{[by the negation rule]}
\end{aligned}
$$

The Conjunction Rule

How do you work out the probability of a conjunction $A\&B$? We already know how to do this when A and B are *independent*. In that case, the probability of the conjunction is just the product of the probabilities of the conjuncts: $\Pr(A\&B) = \Pr(A) \cdot \Pr(B)$. But what if A and B are *not* independent?

Example 20. *I will draw cards from a shuffled deck one at a time* (*I will* not *put the first card back after it is drawn*). *What is the probability that I draw a heart twice in a row*?

Here, it would be a *mistake* to try to calculate $\Pr(\heartsuit_1\&\heartsuit_2) = \Pr(\heartsuit_1) \cdot \Pr(\heartsuit_2)$. The two propositions $\heartsuit_1$ (the first card was a heart) and $\heartsuit_2$ (the second card was a heart) are not probabilistically independent. Once a heart has been taken from the deck, the probability of drawing another heart drops from $13/52$ (or $1/4$) to $12/52$. However, we can appeal to our definition of conditional probability. Notice that, by definition,

$$\Pr(\heartsuit_2 \mid \heartsuit_1) = \frac{\Pr(\heartsuit_2\&\heartsuit_1)}{\Pr(\heartsuit_1)}$$

which implies that $\Pr(\heartsuit_1\&\heartsuit_2) = \Pr(\heartsuit_1) \cdot \Pr(\heartsuit_2 \mid \heartsuit_1)$. And we already know both of these probabilities!

$$
\begin{aligned}
\Pr(\heartsuit_1\&\heartsuit_2) &= \Pr(\heartsuit_1) \cdot \Pr(\heartsuit_2 \mid \heartsuit_1) \\
&= 1/4 \cdot 12/52 \\
&= 12/208 = 3/52
\end{aligned}
$$

And we can use the same trick in general:

> For any propositions A and B:
>
> $$\Pr(A\&B) = \Pr(A \mid B) \cdot \Pr(B)$$

Notice that the equivalence rule and the conjunction rule teach us that

$$\Pr(A \mid B) \cdot \Pr(B) = \Pr(B \mid A) \cdot \Pr(A)$$

After all, by the conjunction rule, the left hand side is equal to $\Pr(A\&B)$ and the right hand side is equal to $\Pr(B\&A)$. But by the equivalence rule, these two propositions must be given the same probability.

Example 21. *There are* 50 *people in this class. Assuming that birthdays are IID, with a* $1/365$ *probability that anyone is born on any day, what is the probability that two people in the class share a birthday*?

Here, it is best to use the negation rule. To calculate the probability that at least two people share a birthday, we should calculate the probability that *no one* shares a birthday, and then subtract from 1.

In solving this problem, choosing the right propositions is key. Let's suppose that we have some way of listing out everyone in the class (for instance, alphabetically). Then, let D_n say that the person in nth place does not share a birthday with any of the people in places 1 through $n-1$. And the proposition that *no one* shares a birthday is the conjunction

$$\textit{no shared birthdays} = D_1 \& D_2 \& \ldots \& D_{49} \& D_{50}$$

So:

$$\begin{aligned}
&\Pr(\textit{some shared birthdays}) \\
&= 1 - \Pr(\textit{no shared birthdays}) \\
&= 1 - \Pr(D_1 \& D_2 \& \ldots \& D_{50}) \\
&= 1 - \Pr(D_1) \cdot \Pr(D_2 \mid D_1) \cdot \Pr(D_3 \mid D_1 \& D_2) \cdot \cdots \cdot \Pr(D_{50} \mid D_1 \& D_2 \& \cdots \& D_{49}) \\
&= 1 - \left(\frac{365}{365}\right) \cdot \left(\frac{364}{365}\right) \cdot \left(\frac{363}{365}\right) \cdot \cdots \cdot \left(\frac{316}{365}\right) \\
&= 1 - \left(\frac{365-0}{365}\right) \cdot \left(\frac{365-1}{365}\right) \cdot \left(\frac{365-2}{365}\right) \cdot \cdots \cdot \left(\frac{365-49}{365}\right) \\
&= 1 - \frac{365!/(365-50)!}{365^{50}} \\
&\approx 97\%
\end{aligned}$$

de Méré's Paradox

In the 1560's, the French author Chevalier de Méré was swindling people out of money with a game of dice. He would roll a six-sided die four times, and offer even odds that the die would land on one on at least one of the four rolls. He found that he made money in the long run playing this way. But then, he tried swindling people with a new game: he would roll *two* dice twenty four times, and offer even odds that the dice would land on snake eyes (double ones) at least once. He reasoned about this case like this: getting snakes eyes with two dice should be $1/6$th as probable as getting a one with a single dice. So, in order to keep the probability of winning unchanged, you should increase the total number of rolls by a factor of six. Since the original game was favorable to him, and since $4 \times 6 = 24$, the new game should also be favorable to him. But when de Méré played this game, he found that he ended up losing money after many plays.

Nowadays, we can recognize that de Méré's reasoning was incorrect—de Méré's mistake was thinking that the probability of winning increased with the number of rolls *multiplicatively*, so that, if the probability of winning on any one roll is p, then the probability of winning on n rolls was $n \cdot p$. (Really, on reflection, this *couldn't possibly* be correct—think about what would happen with *seven* rolls of a single die.) In fact, when the probability of winning on any one roll is p, the probability of winning on n rolls is the more complicated function $1 - (1 - p)^n$ (comprehension check: *why?*). So the probability of winning the first game is $1 - (5/6)^4 \approx 51.8\%$, whereas the probability of winning the second game is $1 - (35/36)^{24} \approx 49.1\%$. This is why de Méré found himself winning money playing the first game, but losing with the second.

Figure 4.2: The Chevalier de Méré

Tautology For any proposition T, if T is a tautology, then

$$\Pr(T) = 1$$

Contradiction For any proposition C, if C is a contradiction, then

$$\Pr(C) = 0$$

Equivalence For any propositions A and B, if A and B are equivalent, then

$$\Pr(A) = \Pr(B)$$

Disjunction For any propositions A and B, if A and B are mutually exclusive, then

$$\Pr(A \vee B) = \Pr(A) + \Pr(B)$$

General Disjunction For any propositions A and B,

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A\&B)$$

Negation For any proposition A,

$$\Pr(\sim A) = 1 - \Pr(A)$$

Conjunction For any propositions A and B:

$$\Pr(A\&B) = \Pr(A \mid B) \cdot \Pr(B)$$

Exercise 8. *Assume that* $\Pr(A\&B) = 1/3$ *and* $\Pr(A\& \sim B) = 1/5$. *What is the probability of* A?

Exercise 9. *Suppose* A *entails* B, $\Pr(A) = 1/3$, *and* $\Pr(B) = 1/2$. *Then, what is the probability of* $B\& \sim A$?

Exercise 10. *Suppose that* $\Pr(A) = 1/3$, $\Pr(B) = 1/4$, *and that* A *and* B *are independent. What is the probability of* $A \vee B$? *What is the probability of* $\sim A\& \sim B$?

[1]Many of this problems are lifted or adapted from Weisberg, *Odds and Ends*, chapter 5

Exercise 11. *Suppose that* $\Pr(A) = \Pr(B)$*. Does it follow that* A *and* B *are equivalent? If so, why? If not, provide a probability distribution in which* $\Pr(A) = \Pr(B)$ *even though* A *and* B *are not equivalent.*

Exercise 12. *There are* 3 *empty buckets lined up. Someone takes apples and places each one in a bucket. The placement of each apple is random and independent of the others. What is the probability that the first two buckets end up with no apples?*

Exercise 13. *Each time you go to a restaurant, there is a* $\frac{1}{3}$ *probability that you'll enjoy your meal. And whether you enjoy your meal at one restaurant is independent of whether you enjoy your meal at the other restaurants (I mean* mutual *independence, not* pairwise *independence). Next week, you will go to seven restaurants. What is the probability that you will enjoy at least one of your meals?*

Exercise 14. *Suppose we will roll a fair, six-sided die four times. What is the probability of it landing on the same number each time? What's the probability of it landing on a different number each time?*

Exercise 15. *There are two urns: urn* X *and urn* Y*. Urn* X *contains* 3 *black marbles and* 1 *red marble, whereas urn* Y *contains* 3 *red marbles and* 1 *black one. We will flip a fair coin to decide which urn to draw from, and then we will draw a marble from the selected urn. What is the probability that we either draw a marble from urn* X *or we draw a red marble,* $X \vee R$*?*

1. Consider the Euler diagram in figure 4.3. (To avoid any confusion: the A circle is the one covering the left side of the box, and the B circle is the one covering the right side of the box.)

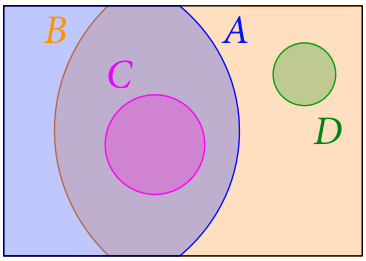


Figure 4.3: An Euler diagram

Then, say whether the following claims are true or false. [80pts]

(a) C and D are compatible. [10pts]

(b) A and D are mutually exclusive. [10pts]

(c) C, D, and B are mutually exclusive. [10pts]

(d) A, B, and D are jointly exhaustive. [10pts]

(e) C entails $A\&B$. [10pts]

(f) D entails $A\&B$. [10pts].

(g) A and B entail C. [10pts]

(h) $A \vee B$ is a tautology. [10pts]

2. Consider the eikosograms below. [70pts]

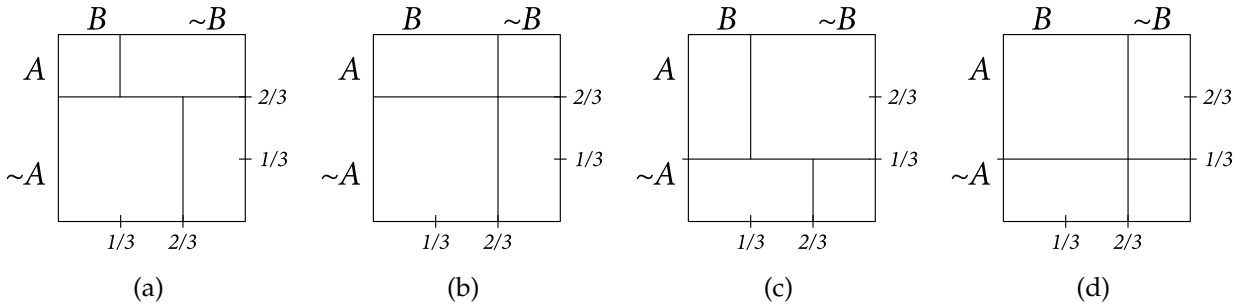


Figure 4.4: Eikosograms

(a) For each of the eikosograms, give the probability of A and the probability of B. [20pts]

(b) For each of the eikosograms, give the probability of A, given B, and the probability of B, given A. [20pts]

(c) In which of the eikosograms are A and B independent? [10pts]

(d) In which of the eikosograms are A and B identically distributed? [10pts]

(e) In which of the eikosograms are A and B independent and identically distributed? [10pts]

3. Say whether the following claims are true or false, and why. For these questions, you should refer back to the definitions from the handouts and think about what they tell you. ('True' means that the statement must be true, and 'false' means that the statement could be false.) [80pts]

(a) If A and B are mutually exclusive, then A and B are probabilistically independent. [10pts]

(b) If C is a contradiction, then C and C are probabilistically independent (that is, C is probabilistically independent of itself). [10pts]

(c) If T is a tautology, then T and T are probabilistically independent (that is, T is probabilistically independent of itself). [10pts]

(d) If A and B have positive probability and are probabilistically independent, then A and B are compatible. [10pts]

(e) If A, B, and C are IID, then $\Pr(A\&B\&C) = \Pr(A) \cdot \Pr(B) \cdot \Pr(C)$. [10pts]

(f) If A, B, and C are IID, then $\Pr(A) = \Pr(B\&C)$. [10pts]

(g) If A, B, and C are jointly exhaustive, then $\Pr(A) + \Pr(B) + \Pr(C) \geqslant 1$. [10pts]

(h) If A, B, and C are a partition (mutually exclusive and jointly exhaustive), then $\Pr(A) + \Pr(B) + \Pr(C) = 1$. [10pts]

4. Al, Betty, and Carol have each purchased a ticket in the lottery. There are 100 tickets overall, and one will be drawn randomly to determine the winner. [60pts]

(a) What is the probability that either Al, Betty, or Carol wins the lottery? [20pts]

(b) What is the probability that Al wins the lottery, given that either Al, Betty, or Carol wins? [20pts]

(c) What is the probability that Al wins the lottery, given that Betty and Carol do not? [20pts]

5. Each time you go on a date, there is a $1/100$ probability that you and your date are compatible, and whether you are compatible with any one date is independent of whether you are compatible with any other. You have decided to go on 50 dates. What is the probability that you meet *somebody* you're compatible with? [40pts]

6. We will shuffle a standard deck of cards and draw two of them from the deck, one after the other.[2] [75pts]

   (a) What is the probability that both of the cards are diamonds? [25pts]

   (b) What is the probability that at least one of the cards is a diamond? [25pts]

   (c) What is the probability that at least one of the cards is the ace of spades? [25pts]

7. We will flip a fair coin ten times in a row.

   (a) What is the probability that the coin lands heads at least once? [20pts]

   (b) Suppose that flips 1 through 9 all land tails. Given this information, what is the probability that the coin lands heads on the tenth flip? [20pts]

8. Here is another rule of probability:

   The Entailment Rule For any propositions A and B, if A entails B, then

   $$\Pr(B) \geqslant \Pr(A)$$

   Show that the entailment rule must be true by using the equivalence rule and the disjunction rule.[3] [40pts]

9. Using the entailment rule and the conjunction rule, show that the following claim must be true. [35pts]

   For any propositions A, B, and C, if A and B entail C, then

   $$\Pr(C) \geqslant \Pr(B \mid A) \cdot \Pr(A)$$

[2] Just a reminder: in a standard deck of playing cards, there are four suits: hearts, diamonds, clubs, and spades. Each suit contains 13 cards. So there are $4 \times 13 = 52$ cards in total.

[3] Hint: remind yourself of the definition of entailment, and then draw out the Euler diagram for A and B, assuming that A entails B. Looking at the Euler diagram, find an expression involving both A and B which is equivalent to B.

5 | The Law of Total Probability

Goal: Learn about a very helpful probability rule: the law of total probability

Let's start with an observation: A and $(A\&B) \vee (A\& \sim B)$ are logically equivalent. We could observe this by considering an Euler diagram like the one below.

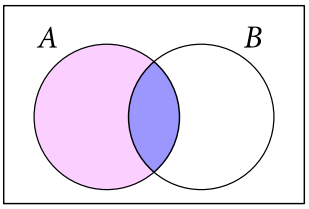


Figure 5.1: A and $(A\&B) \vee (A\& \sim B)$ are logically equivalent

Or we could fill out a truth-table like the one below:

A	B	$(A$	$\&$	$B)$	$\vee$	$(A$	$\&$	$\sim$	$B)$
T	T	T	T	T	T	T	F	F	T
T	F	T	F	F	T	T	T	T	T
F	T	F	F	T	F	F	F	F	T
F	F	F	F	F	F	F	F	T	T

Since A and $(A\&B)\vee(A\& \sim B)$ are true in the same rows of the truth-table, they are logically equivalent.

By the equivalence rule, then, we can conclude that these two propositions must have the same probability.

$$\Pr(A) = \Pr((A\&B) \vee (A\& \sim B))$$

Moreover, $A\&B$ and $A\& \sim B$ are mutually exclusive. If $A\&B$ is true, then B must be true, so $A\& \sim B$ cannot be true. So there's no way for both $A\&B$ and $A\& \sim B$ to be true at once. So they are mutually exclusive. By the disjunction rule,

$$\Pr(A) = \Pr(A\&B) + \Pr(A\& \sim B)$$

Now, the conjunction rule tells us that $\Pr(A\&B) = \Pr(A \mid B) \cdot \Pr(B)$. So we have:

$$\Pr(A) = \Pr(A \mid B) \cdot \Pr(B) + \Pr(A\& \sim B)$$

Likewise, the conjunction rule tells us that $\Pr(A\& \sim B) = \Pr(A \mid \sim B) \cdot \Pr(\sim B)$. So we have:

$$\Pr(A) = \Pr(A \mid B) \cdot \Pr(B) + \Pr(A \mid \sim B) \cdot \Pr(\sim B)$$

This identity is known as the *law of total probability*:

Law of Total Probability For any propositions A and B,[a]

$$\Pr(A) = \Pr(A \mid B) \cdot \Pr(B) + \Pr(A \mid \sim B) \cdot \Pr(\sim B)$$

[a] So long as $\Pr(B) > 0$ and $\Pr(\sim B) > 0$.

Example 22. *We will flip a fair coin. If it lands heads, then we will roll a D*6 (*a fair six sided die*). *If it doesn't land heads, then we will roll a D*12 (*a fair twelve sided die*). *What is the probability that we roll a one*?

Let H say that the coin lands heads, and let $\boxdot$ say that we roll a one. Then:

$$\begin{aligned}
\Pr(\boxdot) &= \Pr(\boxdot \mid H) \cdot \Pr(H) + \Pr(\boxdot \mid \sim H) \cdot \Pr(\sim H) \\
&= 1/6 \cdot 1/2 + 1/12 \cdot 1/2 \\
&= 1/12 + 1/24 \\
&= 2/24 + 1/24 \\
&= 3/24 = 1/8
\end{aligned}$$

Exercise 16. *There are only two candidates*: *Xander and Yurlady. There's a* 60% *chance that Xander wins and a* 40% *chance that Yurlady wins. Given that Xander wins, the odds are* 1 : 3 *that he'll repeal the unpopular housing ordinance. Given that Yurlady wins, odds are* 3 : 1 *that she'll repeal the ordinance. What are the odds that the housing ordinance gets repealed*?

Exercise 17. *There are two urns, which we'll call 'A' and 'B'. Within A, there are four white marbles and one red marble. Within B, there are four red marbles and one white marble. We will start by drawing a marble from A. If it is white, then we will draw another marble from A* (*without placing the first marble back*). *However, if it is red, then we will draw a marble from B. What is the probability that the second marble we draw is white*?

The same reasoning we used to derive the law of total probability can be used to derive a more general rule. Consider the *partition* of propositions B_1, B_2, B_3. (Recall: a *partition* is a collection of mutually exclusive and jointly exhaustive propositions—so exactly one of the propositions in the set must be true.) Then, A and $(A\&B_1) \vee (A\&B_2) \vee (A\&B_3)$ are logically equivalent.

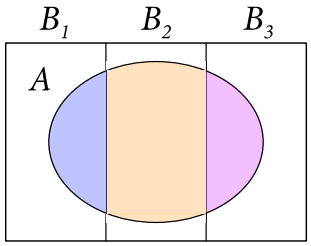


Figure 5.2: If B_1, B_2, and B_3 are a partition, then A and $(A\&B_1)\vee(A\&B_2)\vee(A\&B_3)$ are logically equivalent

Sp, by the rule of equivalence, we must have that

$$\Pr(A) = \Pr((A\&B_1) \vee (A\&B_2) \vee (A\&B_3))$$

Moreover, $A\&B_1$, $A\&B_2$, and $A\&B_3$ are mutually exclusive—since no two of B_1, B_2, and B_3 can be true at once, no two of $A\&B_1$, $A\&B_2$, and $A\&B_3$ can be true at once, either. So, by the disjunction rule,

$$\Pr(A) = \Pr(A\&B_1) + \Pr(A\&B_2) + \Pr(A\&B_3)$$

Now, the conjunction rule tells us that $\Pr(A\&B_1) = \Pr(A \mid B_1) \cdot \Pr(B_1)$. And likewise $\Pr(A\&B_2) = \Pr(A \mid B_2) \cdot \Pr(B_2)$ and $\Pr(A\&B_3) = \Pr(A \mid B_3) \cdot \Pr(B_3)$. So:

$$\Pr(A) = \Pr(A \mid B_1) \cdot \Pr(B_1) + \Pr(A \mid B_2) \cdot \Pr(B_2) + \Pr(A \mid B_3) \cdot \Pr(B_3)$$

And we can go through similar reasoning given *any* (finite) partition of propositions $B_1, B_2, \ldots, B_N$. So we have a more general version of the law of total probability:

General Law of Total Probability For any proposition A and any partition $B_1, B_2, \ldots, B_N$,[a]

$$\Pr(A) = \Pr(A \mid B_1) \cdot \Pr(B_1) + \Pr(A \mid B_2) \cdot \Pr(B_2) + \cdots + \Pr(A \mid B_N) \cdot \Pr(B_N)$$

[a] So long as $\Pr(B_i) > 0$ for each i.

There's another, more concise way of writing the same thing. Instead of writing a long sum $a_1 + a_2 + \cdots + a_N$, we can instead write '$\sum_{i=1}^{N} a_i$'. The letter 'i' is here being used as a variable which is ranging over the subscripts on the summands a_i. The way to read '$\sum_{i=1}^{N}$'

is 'the sum, from i equals 1 to N'. Using this notation, we can write the general law of total probability this way: for any proposition A and any partition $B_1, B_2, \ldots, B_N$,

$$\Pr(A) = \sum_{i=1}^{N} \Pr(A \mid B_i) \cdot \Pr(B_i)$$

Example 23. *James either took the A train, the C train, or the E train, depending upon which one came first. There's a* 25% *chance that the A train came first, a* 25% *chance that the C train came first, and a* 50% *chance that the E train came first. James will only arrive on time if the train he took is running express. The probability that any given A train is running express is* 10%*. The probability that any given C train is running express is* 20%*. And the probability that any given E train is running express is* 100%*. What is the probability that James arrives on time*?

Let's use 'T' for 'James arrives on time', 'A' for 'James took the A train, 'C' for 'James took the C train', and 'E' for 'James took the E train'. Then,

$$\begin{aligned}
\Pr(T) &= \Pr(T \mid A) \cdot \Pr(A) + \Pr(T \mid C) \cdot \Pr(C) + \Pr(T \mid E) \cdot \Pr(E) \\
&= 1/10 \cdot 1/4 + 1/5 \cdot 1/4 + 1 \cdot 1/2 \\
&= 1/40 + 1/20 + 1/2 \\
&= 1/40 + 2/40 + 20/40 \\
&= 23/40 = 57.5\%
\end{aligned}$$

Exercise 18. *There is an urn containing* 9 *marbles;* 3 *of them are black,* 3 *of them are red, and* 3 *of them are yellow. We will draw a marble from the urn at random. If it is black, then we will flip a coin which is biased* 3/4 *towards heads twice. If the drawn marble is red, then will be flip a fair coin twice. If the drawn marble is yellow, then we will flip a coin which is biased* 1/4 *towards heads twice. What is the probability that the coin lands heads on the first flip*?

What is the probability that it lands heads on both flips?

What is the probability that it lands heads on the second flip, given that it lands heads on the first flip?

6 | Bayes' Theorem

Goal: Understand the relationship between $\Pr(H \mid E)$ and $\Pr(E \mid H)$, and learn another way of calculating $\Pr(H \mid E)$ using $\Pr(E \mid H)$.

Puzzle. *You return from a week in the Amazon jungle. About* 10% *of people who return from this vacation end up with malaria, so you decide to get tested. The test you take has a* 10% *rate of false positives/false negatives, meaning that, if you have malaria, then the probability that the test comes back positive is* 90%*, and, if you do not have malaria, then the probability that the test comes back negative is* 90%*. Your test comes back positive. Given this information, what's the probability that you have malaria*?

A *very* natural first thought is: the probability that you now have malaria is now 90%. This is what a significant number of doctors say. But it is the wrong answer. The probability that you have malaria, given that the test came back positive, is actually 50%.

To think about the case, let's imagine 100 people all coming back from the Amazon. Since we know that 10% of these people get malaria, imagine that 10 of the 100 have malaria, and the other 90 do not. Everyone then takes the test. Of the 10 that have malaria, we expect the test will correctly diagnose 9 of them, but give a *false negative* for 1 (since the rate of false negatives is 10%). And of the 90 that do not have malaria, we expect the test to correctly diagnose 81 of them, but give a *false positive* for 9 (since the rate of false positives is 10%). So here's the breakdown we should expect:

	Positive	Negative
Malaria	9	1
No Malaria	9	81

Table 6.1: How many of 100 people returning from the Amazon we should expect to have or not have malaria, and how many of each of them we should expect to have a positive or negative test result.

When you learn that the test came back positive, what you learn is that you are in the left-hand column of table 6.1. But *half* of the people in that column do not have malaria! So the probability that you have malaria, given that the test came back positive, is 50%.

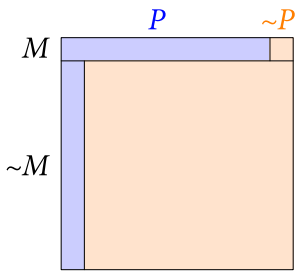


Figure 6.1: Eikosogram for you having malaria (M) and you having a positive test result (P)

We can see the same thing in the eikosogram in figure 6.1: when you learn that you have a positive test result, what you learn that is that you are in one of the two thin rectangles (the $M\&P$ rectangle and the $\sim M\&P$ rectangle). But these two rectangles have exactly the same area ($0.9 \times 0.1 = 0.09$). So you're equally likely to be in either. And so the probability that you have malaria is 50%.

The error that doctors often make is known as *the base rate fallacy*, because it *ignores the base rate of the disease*. The important point is that *it matters* how many people emerging from the Amazon jungle have malaria. When you learn that you have a positive test result, there are two things that could explain the positive test result: you having malaria and the test giving a false positive. In order to know which explanation is most likely, we need to know *both* the probability of a false positive *and* the probability of you having malaria (the base rate).

Bayes' Theorem

In the medical test case, we should carefully distinguish between two different conditional probabilities:

- The probability that you have malaria, given the positive test result, $\Pr(M \mid P) = 1/2$
- The probability that you get a positive test result, given that you have malaria, $\Pr(P \mid M) = 9/10$

The order of the propositions before and after the '|' makes a difference. Nonetheless, there is a relationship between these two probabilities. That relationship is known as *Bayes' Theorem*.

When we're talking about Bayes' Theorem, I'm going to use the proposition letters 'H' and 'E', and I invite you to think about these propositions as standing for some *hypothesis* and some piece of *evidence* which may be relevant to that hypothesis. For instance, in the medical test, we are interested in how likely the hypothesis that you have malaria is made by the evidence that your test came back positive.

Some terminology: let's call $\Pr(H \mid E)$ the *posterior* probability of H, *given* E. It is the probability that H will have, after you have learned E. And let's call $\Pr(H)$ the *prior* probability of H. Then, what Bayes' Theorem gives us is a relationship between the *prior* probability of a hypothesis and the *posterior* probability of that hypothesis, given some piece of evidence, E.

To appreciate the theorem, let's start off with the *definition* of conditional probability.

$$\Pr(H \mid E) \stackrel{\text{def}}{=} \frac{\Pr(H \& E)}{\Pr(E)}$$

The conjunction rule tells us that, for any propositions A and B, $\Pr(A \& B) = \Pr(B \mid A) \cdot \Pr(A)$. So, in particular, $\Pr(H \& E) = \Pr(E \mid H) \cdot \Pr(H)$. So:

Bayes' Theorem (v1) For any propositions H and E,[a]

$$\Pr(H \mid E) = \frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E)}$$

[a] So long as $\Pr(E) > 0$ and $\Pr(H) > 0$.

Moreover, the law of total probability tells us that, for any propositions A and B, $\Pr(A) = \Pr(A \mid B) \cdot \Pr(B) + \Pr(A \mid \sim B) \cdot \Pr(\sim B)$. In particular, $\Pr(E) = \Pr(E \mid H) \cdot \Pr(H) + \Pr(E \mid \sim H) \cdot \Pr(\sim H)$. So:

Bayes' Theorem (v2) For any propositions H and E,[a]

$$\Pr(H \mid E) = \frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E \mid H) \cdot \Pr(H) + \Pr(E \mid \sim H) \cdot \Pr(\sim H)}$$

[a] So long as $\Pr(E) > 0$, $\Pr(H) > 0$, and $\Pr(\sim H) > 0$.

Applying the second version of Bayes' theorem to our puzzle, we have:

$$\Pr(M \mid P) = \frac{\Pr(P \mid M) \cdot \Pr(M)}{\Pr(P \mid M) \cdot \Pr(M) + \Pr(P \mid \sim M) \cdot \Pr(\sim M)}$$

By the description of the case, we know that $\Pr(M) = 1/10$, so $\Pr(\sim M) = 9/10$. And we know that $\Pr(P \mid M) = 9/10$ and $\Pr(P \mid \sim M) = 1/10$. So

$$\begin{aligned} \Pr(M \mid P) &= \frac{9/10 \cdot 1/10}{9/10 \cdot 1/10 + 1/10 \cdot 9/10} \\ &= \frac{9/100}{9/100 + 9/100} \\ &= \frac{9/100}{18/100} \end{aligned}$$

$$= 9/18$$
$$= 1/2$$

Exercise 19. *95% of people have HPV, so you decide to get tested. The test has a 10% false positive rate and a 10% false negative rate, meaning that the probability that you get a positive result, given that you have HPV, is 90%, and the probability that you get a negative result, given that you don't have HPV, is 90%. The test comes back negative. What's the probability that you have HPV?*

Exercise 20. *There are two coins. Coin A has bias of ¼ towards heads, and coin B has a bias of ¾ towards heads. Not knowing which is which, I select one of the coins at random (with equal probability), and give the chosen coin a flip. The coin lands heads. Given this new information, what is the probability that I chose coin A?*

7 | More on Bayes' Theorem

Goal: Use odds to simply Bayes' Theorem, and come to a better understanding of how evidence impacts the probability of a hypothesis.

Odds Ratios

Recall: if the probability of A is x, then the *odds* of A are $x : 1-x$. For instance, if $\Pr(A) = 1/4$, then the odds of A are $1 : 3$. (Recall that odds stay the same when both sides of the ':' are multiplied by the same factor. So the odds $1/4 : 3/4$ are the same as the odds $1 : 3$. The latter way of writing the odds is just simpler.)

Corresponding to the odds of A is a certain fraction, known as the *odds ratio*. If the probability of A is $1/4$, then the odds ratio for A is $1/3$. And, in general, if the probability of A is x, then the odds ratio for A is $x/1-x$. I'll use '$\mathbb{O}(A)$' for the odds ratio for A.

$$\mathbb{O}(A) \stackrel{\text{def}}{=} \frac{\Pr(A)}{\Pr(\sim A)}$$

Similarly, we can write '$\mathbb{O}(A \mid B)$' for the *conditional* odds ratio of A, given B. This is just the ratio corresponding to the odds that A will have, after you've learned that B. For instance, if $\Pr(A \mid B) = 2/5$, then $\mathbb{O}(A \mid B) = 2/3$ And, in general,

$$\mathbb{O}(A \mid B) \stackrel{\text{def}}{=} \frac{\Pr(A \mid B)}{\Pr(\sim A \mid B)}$$

Odds Ratio version of Bayes' Theorem

We can use odds ratios to develop a very simple and powerful version of Bayes' theorem. To do so, let's start with the *posterior* odds ratio for H, given E, $\mathbb{O}(H \mid E)$:

$$\frac{\Pr(H \mid E)}{\Pr(\sim H \mid E)}$$

We can determine what this fraction is by applying our first version of Bayes' theorem to both the numerator and the denominator:

$$\frac{\Pr(H \mid E)}{\Pr(\sim H \mid E)} = \frac{\frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E)}}{\frac{\Pr(E \mid \sim H) \cdot \Pr(\sim H)}{\Pr(E)}}$$

On the right-hand-side, if we multiply both the numerator and the denominator by $\Pr(E)$, then we will get:

$$\frac{\Pr(H \mid E)}{\Pr(\sim H \mid E)} = \frac{\Pr(E \mid H) \cdot \Pr(H)}{\Pr(E \mid \sim H) \cdot \Pr(\sim H)}$$

Or, writing the same thing in a different way:

$$\underbrace{\frac{\Pr(H \mid E)}{\Pr(\sim H \mid E)}}_{\textit{posterior odds ratio}} = \underbrace{\frac{\Pr(E \mid H)}{\Pr(E \mid \sim H)}}_{\textit{strength of evidence}} \cdot \underbrace{\frac{\Pr(H)}{\Pr(\sim H)}}_{\textit{prior odds ratio}}$$

On the left, we have the posterior odds ratio. On the right, we have the prior odds ratio, multiplied by something that we will call the *strength of the evidence E for H*.

Strength of Evidence The *strength of the evidence E for H* is the fraction which tells us how much more likely H makes E than $\sim H$ makes E:[1]

$$\textit{strength of } E \textit{ for } H = \frac{\Pr(E \mid H)}{\Pr(E \mid \sim H)}$$

If the evidential strength is less than 1, then H makes E more likely than $\sim H$ does, and E is evidence *for* H. If it is less than 1, then H makes E *less* likely than $\sim H$ does, and E is evidence *against* H. If it equals 1, then H makes E exactly as likely as $\sim H$ does, and E is neither evidence for nor against H.

In odds form, Bayes' theorem tells us that we can get the posterior odds ratio for a hypothesis from the prior odds ratio by just multiplying the prior odds by the strength of the evidence.

Bayes' Theorem (Odds Form) For any propositions H and E,[a]

$$\mathbb{O}(H \mid E) = \textit{strength of } E \textit{ for } H \cdot \mathbb{O}(H)$$

[a] So long as $\Pr(E) > 0$ and $0 < \Pr(H) < 1$

[1] If $\Pr(H) = 0$ or $\Pr(\sim H) = 0$, then the strength of the evidence E for H is undefined.

Example 24. 10% *of people returning from the Amazon jungle with your symptoms have malaria, so you decide to get tested. The tests have a* 10% *false positive/false negative rate, meaning that the probability of a positive result, given that you have the disease, is* 90%*, and the probability of a positive result, given that you don't have the disease, is* 10%*. Your test comes back positive. Given this evidence, what are the odds that you have malaria*?

The hypothesis we are interested in is the hypothesis that you have malaria (M). The *prior odds* of this hypothesis are $1 : 9$, so $\mathbb{O}(M) = 1/9$. Our evidence is that the test result came back positive (P). The *strength* of this evidence is

$$\textit{strength of } P \textit{ for } M = \frac{\Pr(P \mid M)}{\Pr(P \mid \sim M)} = \frac{9/10}{1/10} = 9$$

So we can calculate the *posterior odds* that you have malaria, given the positive test result, with

$$\begin{aligned}\mathbb{O}(M \mid P) &= \textit{strength of } P \textit{ for } M \cdot \mathbb{O}(M) \\ &= 9 \cdot (1/9) \\ &= 9/9 \\ &= 1/1\end{aligned}$$

So the posterior odds that you have malaria are $1 : 1$. That is, the posterior odds are even. So the probability of malaria, given the positive test result, is $1/2$.

Exercise 21. *You are inspecting a printer which has been delivered to your company. You know that one out of every* 5 *printers has a defect which causes it to misprint. Those printers with a defect misprint one fifth of pages—and whether a defective printer misprints one page is probabilistically independent of whether it misprints any other pages. You print one page and see that it has no error. Given this information, how likely is it that the printer is defective*?

Example 25. *Suppose you print another page and see that* it *has no error, either. Given this information, how likely is it* now *that the printer is defective*?

Use D for the hypothesis that the printer is defective, and N_1 for the evidence that the first page has no error. Then, we have from the previous exercise that $\mathbb{O}(D \mid N_1) = 1/5$ (that is, the probability that it is defective, given that there's no error on the first page, is $1/6$. We have now gained the additional information N_2, that the second page has been printed without any error. Since errors on successive pages are independent of each other, the strength of the evidence N_2 for D will be the same as the strength of the evidence N_1 for D,

$$\textit{strength of } N_2 \textit{ for } D = \frac{\Pr(N_2 \mid D \& N_1)}{\Pr(N_2 \mid \sim D \& N_1)} = \frac{\Pr(N_2 \mid D)}{\Pr(N_2 \mid \sim D)} = \frac{4/5}{1} = 4/5$$

And we get that

$$\begin{aligned} \mathbb{O}(D \mid N_1 \& N_2) &= \textit{strength of } N_2 \textit{ for } D \cdot \mathbb{O}(D \mid N_1) \\ &= 4/5 \cdot 1/5 \\ &= 4/25 \end{aligned}$$

So the new odds that the printer is defective are even lower—$4 : 25$. So the probability that the printer is defective is now $4/29 \approx 13.8\%$.

Exercise 22. *There is a very rare disease that* 1 *in every* 1,000,001 *people have. There are two tests for the disease. Both tests have a false positive/false negative rate of* 1% (*meaning that the probability that you get a positive result, given that you don't have the disease, is* 1%, *and the probability that you get a negative result, given that you do have the disease, is* 1%). *The outcomes of the two tests are independent, conditional on you having the disease*; *and they are also independent, conditional on you* not *having the disease. The first test comes back positive. Given this new information, what are the odds that you have the disease*?

The second test also comes back positive. Given this new information, what are the odds that you have the disease?

1. Complete the following sentences

   (a) A, B, and C are mutually exclusive if and only if _____

   (b) A, B, C, and D are jointly exhaustive if and only if _____

   (c) E, F, and G are a partition if and only if _____

   (d) H and E are probabilistically independent if and only if _____

   (e) The conditional probability of A, given B, is defined to be _____

2. (a) Terry will ask either Joann or Geraldine to the dance (and he won't ask both). If he asks Joann, there's a 50% probability that she says 'yes'. If he asks Geraldine, she will certainly say 'yes'. He's 75% likely to ask Joann, and only 25% likely to ask Geraldine. What's the probability that Terry ends up with a date for the dance?

   (b) Suppose you learn that Terry ends up with a date, but you don't learn who that date is. Given your new information, what's the probability that Terry is taking Joann to the dance?

3. There are two social clubs on campus: the Achaeans and the Parisians. The Achaeans have 10 members, 9 of which are cool, and 1 of which is uncool. The Parisians have 18 members, 16 of which are uncool, and 2 of which are cool. Hubert plans to visit the Achaeans and meet with a random member. If that random member is cool, then he'll stay to meet with another (different) Achaean. If, however, they are uncool, then he'll go over to the Parisians where he'll meet a random member of *their* social club. If the second person he meets is cool, he'll join whichever social club that second person belongs to.

   (a) What's the probability that Hubert ends up joining the Achaeans?

   (b) What's the probability that the second person Hubert meets with is cool?

   (c) Suppose you learn that the second person Hubert met with was cool. Given this new information, what's the probability that he has joined the Achaeans?

4. (a) You are a chicken inspector testing chickens leaving the factory. You know from experience that one out of every twenty five chickens is contaminated with salmonella.[2] Your test has a 5% false positive rate and a 15% false negative rate. That is: the probability of a negative result, given that the chicken doesn't have salmonella, is 95%, and the probability of a positive result, given that the chicken *does* have salmonella, is 85%. What is the probability that your test comes back positive?

[2] This is the actual rate of salmonella contamination. Source: https://www.cdc.gov/foodsafety/foods/chicken.html.

(b) Suppose you learn that the test has come back positive. Given this new information, what's the probability that the chicken is contaminated with salmonella?

5. According to one study, 76% of people with COVID-19 report no symptoms, while about 2% of people without COVID-19 reported symptoms.[3] One model estimated that as of January 20,2021, 3% of residents of Los Angeles County were infected with COVID-19.[4]

   (a) If someone in Los Angeles County on January 20, 2021 reports symptoms, what are the odds that they have COVID-19?

   (b) The same model estimated that at the same time 0.7% of residents of King County, Washington were infected. If someone in King County reports symptoms, what are the odds that they have COVID?

   (c) A COVID-19 nasopharyngeal swab test has a 7% false negative rate and close to 0% false positive rate. (That is,93% of people with COVID test positive, and close to 100% of people without COVID test negative).[5] If someone in Los Angeles County in January, 2021 has symptoms and tests negative, what are the odds that they have COVID? (Assume that the false positive and false negative rate of the test is the same, whether they have symptoms or not.)

[3]I owe this problem set question to Jeff Russell. Source: https://www.dovepress.com/three-quarters-of-people-with-sars-cov-2-infection-are-asymptomatic-an-peer-reviewed-article-CLEP

[4]Source: https://covid19-projections.com/infections/counties/ca/

[5]Source: https://www.news-medical.net/whitepaper/20201019/Sensitive-and-Accurate-Diagnostic-Tests-of-COVID-19-with-High-Quality-SARS-CoV-2-Proteins.aspx.

8 | Probability & Induction

Inductive Inference

Deductive inferences are ones in which the conclusion follows necessarily from the premises—ones in which it is impossible for the premises to be true while the conclusion is false. *Inductive* inferences, on the other hand, are ones in which the conclusion does *not* follow necessarily from the premises. Let's focus on one type of inductive inference (Weisberg discusses others):

Enumerative Induction In enumerative induction, you begin with the premises that some collection of F things each have a certain property, and you conclude that *all* F things have that property.

The first raven is black	The first emerald is green	The first F is G
The second raven is black	The second emerald is green	The second F is G
⋮	⋮	⋮
The nth raven is black	The nth emerald is green	The nth F is G
∴ All ravens are black	∴ All emeralds are green	∴ All Fs are G

The Search for an Inductive Logic

The theory of *deductive* logic is ancient; though there has been considerable progress in the study of deductive logic over the past century and a half. In contrast, at the start of the 20th century, there was nothing comparable to be said about *inductive* logic. Philosophers had absolutely no theory of *when* an inductive inference was a good one.

One reason philosophers were especially concerned to have a theory of inductive inference was Hume's *problem of induction*. While we won't have the time to go into it here, Hume had argued that any *justification* of induction was going to be *circular*. A first step in providing a justification of induction would be to figure out what the rules of induction even *are*.

So, in the middle of the twentieth century, philosophers set out to codify the rules of induction and develop a theory of *formal inductive logic* to sit alongside our theory of *formal deductive logic*. The ambition at the time was that we would have a theory which is both:

1. *Formal*, in the same sense that formal logic is formal. In order to know whether $A \vee B, \sim A \therefore B$ is valid, we don't have to know anything about the *meaning* of 'A' and 'B'. And the hope was that, in exactly the same way, we would be able to say that the argument 'all observed Fs are Gs, $\therefore$ all Fs are Gs' was an inductively strong inference, *without* knowing anything about the meaning of 'F' and 'G'.

2. *Intersubjective*, in the sense that any two reasonable people will be able to agree about whether some inference is inductively strong.

Some terminology: if the inference $E \therefore H$ is strong, then we'll say that the evidence E *supports* the hypothesis H. (In just the same way that we say that A *entails* B when the inference $A \therefore B$ is valid.)

You Can't Always Get What You Want

Unfortunately, they quickly realized that we simply *could not have* a theory like this: a theory of inductive strength which was both formal and intersubjective. In his classic paper *Studies in the Logic of Confirmation*, Carl Hempel considered two plausible principles which you might expect any reasonable theory of inductive strength to satisfy:

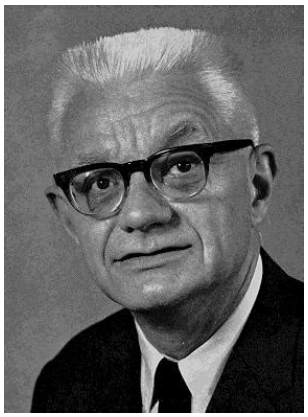


Figure 8.1: Carl Hempel

Entailment Condition If H entails E, then E supports H

Transmission Condition If E supports H, and H entails H^*, then E supports H^*.

In favor of the principle (EC): when we test a theory like Newton's theory of graviation, we figure out what the theory *entails* (e.g., in a vacuum, a dropped hammer and a dropped feather will touch the ground at the same time), and if we go on to *observe* those entailments, we take that to be a reason to accept the theory.

In favor of the principle (TC): when we gather evidence in favor of Newton's theory of gravitation, we *also* take this evidence to support Newton's celestial mechanics (the theory describing the orbit of the planets), precisely because his predictions about celestial mechanics are *logically entailed by* his theory of gravitation.

However, Hempel points out that these two principles (EC) and (CC) together would tell us that $A \therefore B$ is inductively strong—*no matter what A and B are*. Take any two propositions, A and B. Then,

1. $A\&B$ entails A. So, by the Entailment Condition, A supports $A\&B$
2. A supports $A\&B$ (from above), and $A\&B$ entails B. So, by the Transmission Condition, A supports B.

This is a disaster—we cannot have the proposition that Daniel is tall supporting the theory of quantum mechanics. So Hempel considers *weakening* the two principles (EC) and (TC), like so:

Laws are Supported by Their Instances A law statement of the form 'All Fs are Gs' is supported by a G F.

Equivalence Condition If E supports H, then E supports anything logically equivalent to H

The Paradox of the Ravens

Unfortunately, even these weak principles are far too strong. Hempel points out that accepting these two principles would mean saying that *almost all* evidence supports a law statement. For a toy example, take the law statement "All ravens are black".

1. "All ravens are black" is logically equivalent to "All non-black things are non-ravens".
2. By **Laws are Supported by Their Instances**, the evidence that a leaf is green supports the hypothesis that all non-black things are non-ravens.
3. By (1), (2), and **Equivalence Condition**, the evidence that a leaf is green supports the hypothesis that all ravens are black.

As Hempel quips, this result opens up the possibility of *indoor ornithology*; but surely we cannot learn about the color of ravens by gathering evidence about the colors of leaves.

The Grue Paradox

Nelson Goodman twisted the knife. He points out that, in order to know whether "All Fs are G" is confirmed by a G F, we must know something about what F and G *mean*. So we cannot have a purely formal theory of inductive support.

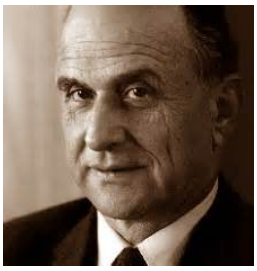


Figure 8.2: Nelson Goodman

1. Say that a thing is grue if and only if it has been observed before 2027 and is green, or else it has *not* been observed before 2027 and is blue.
2. Then, there is no *syntactic, formal* difference between these two enumerative inductions:

The first observed emerald is green

The second observed emerald is green

⋮

The nth observed emerald is green

$\therefore$ All emeralds are green

The first observed emerald is grue

The second observed emerald is grue

⋮

The nth observed emerald is grue

$\therefore$ All emeralds are grue

But *surely* our observations support the hypothesis that all emeralds are *green* and they do *not* support the hypothesis that all emeralds are *grue*. In 2027, we should expect the first new emerald we observe to be *green*. We should not expect it to be *blue*.

Goodman's conclusion: any theory of enumerative induction will have to pay attention to *content* as well as *form*.

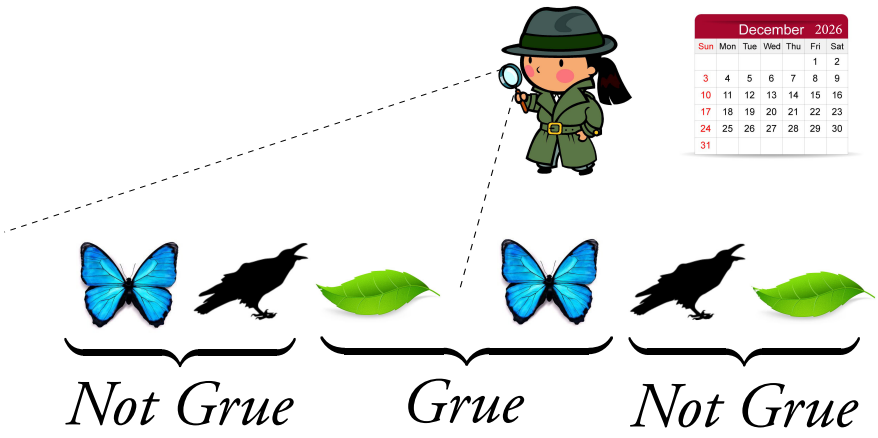


Figure 8.3: An object is grue *iff it is first observed before 2027 and is green, or else it is first not first observed before 2027 and is blue*

The Bayesian Theory of Induction

The *Bayesian* interprets probability claims as entirely *subjective*—they are about your (or anybody's) *degrees of belief* or *degrees of confidence* or *credence*. And the Bayesian endorses two normative claims about these degrees of belief:

Probabilism Your degrees of belief ought to be probabilities (that is: they ought to obey the rules of probability that we've learnt in this course)

Conditionalization If your (prior) subjective probability in H is $\Pr(H)$, and you learn about some new evidence, E, then your new (posterior) subjective probabilities ought to be $\Pr(H \mid E)$.

The Bayesian then says that some piece of evidence, E, *supports* a hypothesis, H, if and only if learning E should raise your probability in H. That is:

> **Bayesian Theory of Evidential Support** The evidence E *supports* a hypothesis H if and only if
>
> $$\Pr(H \mid E) > \Pr(H)$$

Notice that it follows from this that, according to the Bayesian, E supports H if and only if $\Pr(E \mid H) > \Pr(E)$. So we could have equivalently said:

> **Bayesian Theory of Evidential Support (v2)** The evidence E *supports* a hypothesis H if and only if
>
> $$\Pr(E \mid H) > \Pr(E)$$

The Bayesian's theory of inductive support is not formal—since not every hypothesis of the form 'All Fs are G' will receive the same probability. Nor is it *intersubjective*, insofar as the Bayesian allows different people to have different prior probabilities.

Why the Bayesian Thinks You Can't Always Get What You Want

The Ravens Paradox

The Bayesian thinks that the principle **Laws are Supported By Their Instances** is false. Here's a toy model: suppose you are certain that there are four things in existence: two ravens, and two non-ravens. And you only entertain two hypotheses about their colors: *All* and ~*All*:

All	Black	Non-Black
Raven	2	0
Non-Raven	1	1

~ *All*	Black	Non-Black
Raven	1	1
Non-Raven	1	1

According to the hypothesis *All*, both of the ravens are black. And according to the hypothesis ~*All*, one of the ravens are black and the other is non-black.

Suppose you are going to randomly select some thing, and you find that it is a non-black non-raven. Call this evidence 'E'. Notice that

$$\Pr(E \mid All) = \Pr(E \mid \sim All) = 1/4$$

Since, whether all ravens are black or not, the probability that a randomly selected thing would be a non-black non-raven is $1/4$. So, according to the Bayesian, in this toy model, a non-black non-raven does not support the hypothesis that all ravens are black. From the Bayesian perspective, the principle **Laws are Supported By Their Instances** was a hasty generalization; it doesn't hold in general.

On the other hand, suppose that you randomly select some thing, and you find that it is a black raven. Call this evidence 'E^*'. Notice,

$$\Pr(E^* \mid All) = 1/2 \qquad \text{and} \qquad \Pr(E^* \mid \sim All) = 1/4$$

Since, if all ravens are black, half of all things are black ravens; whereas, if not all ravens are black, then one fourth of all things are black ravens. So the observation of a black raven supports the hypothesis that all ravens are black.

The Grue Paradox

Let *Green* be the hypothesis that all emeralds are green, and let *Grue* be the hypothesis that all emeralds are grue. Let E be the evidence that all observed emeralds have been green (and, therefore, grue). Notice that both *Green* and *Grue entail* E. So

$$\Pr(E \mid Green) = \Pr(E \mid Grue) = 1$$

So long as we thought there was *some* prior probability that not all observed emeralds would be green, we will say that *both Green and Grue* are supported by the observation of many emeralds, all of which are green/grue.

But perhaps one of the hypotheses is supported *more*? Not if we measure support in terms of *factor* by which their probability is raised. Note that

$$\frac{\Pr(Green \mid E)}{\Pr(Grue \mid E)} = \frac{\frac{\Pr(E \mid Green) \cdot \Pr(Green)}{\Pr(E)}}{\frac{\Pr(E \mid Grue) \cdot \Pr(Grue)}{\Pr(E)}} = \frac{\Pr(E \mid Green) \cdot \Pr(Green)}{\Pr(E \mid Grue) \cdot \Pr(Grue)} = \frac{1 \cdot \Pr(Green)}{1 \cdot \Pr(Grue)} = \frac{\Pr(Green)}{\Pr(Grue)}$$

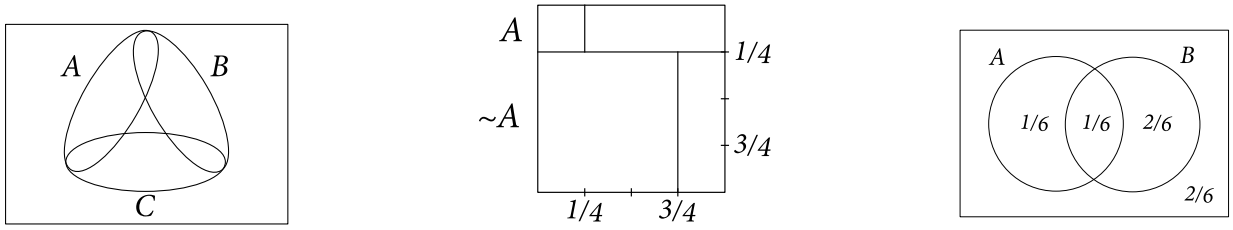
So the ratio between your posterior probability for the *Green* hypothesis and your posterior probability for the *Grue* hypothesis must be equal to the ratio between your *prior* probability for the *Green* hypothesis and your *prior* probability for the *Grue* hypothesis. So both hypotheses have their probabilities raised by the same factor.

So the Bayesian treatment of the Grue paradox is just this: you started out with an *inductive bias* towards the property of greenness. The Bayesian *can* say that you should be

more confident in the *Green* hypothesis than you should be in the *Grue* hypothesis, but they can only say that by saying that you should have *started out* more confident in the *Green* hypothesis than the *Grue* hypothesis. The difference wasn't made by the evidence. It was made by the prior.

A. TRUE/FALSE. If a statement below *must* be true, write 'T' in the provided space. If it *could* be false, write 'F'. (Write legibly. If I cannot tell whether you have written 'T' or 'F', then you will get the question wrong. You may write '1' for true and '0' for false, if you prefer.) [15%]

1. _____ If A is a tautology, and B is a tautology, then A and B are logically equivalent.
2. _____ If A is a contradiction, then $A \therefore B$ is a valid argument.
3. _____ If A entails B, then $A \vee B$ is logically equivalent to B.
4. _____ In the Euler diagram in figure 8.4a, A, B, and C are compatible.
5. _____ If C is a contradiction, then A and C are probabilistically independent.
6. _____ If A and B are probabilistically independent, B and C are probabilistically independent, and A and C are probabilistically independent, then A and $B\&C$ are probabilistically independent.
7. _____ If A, B, and C are a partition, then $\Pr(A) + \Pr(B) + \Pr(C) = 1$.
8. _____ In the eikosogram from figure 8.4b, A and B have the same probability.
9. _____ In the eikosogram from figure 8.4b, A and B are probabilistically independent.
10. _____ In the probabilistic Euler diagram from figure 8.4c, the probability of B, given A, is equal to the probability of A, given B.



(a) (b) (c)

Figure 8.4

B. PROBABILITY TREES. Draw a probability tree to answer the following question. (Write legibly. If I cannot read the labels on your tree, you will not receive credit.) [10%]

> *There are two urns, labeled 'A' and 'B'. In A, there are seven red marbles and three yellow marbles. In B, there are seven yellow marbles and three red marbles. We will flip a fair coin. If it lands heads, then we will draw a random marble from urn A. If it lands tails, then we will draw a random marble from urn B. What is the probability that we draw a yellow marble?*

C. TRUTH-TABLES. Use a truth-table to determine whether the following propositions are formally equivalent or not. Be sure to put a box around the main operator of each proposition, and to explicitly say whether they are formally equivalent.[1] [10%]

$$\sim A\&(C \vee A) \qquad \text{and} \qquad \sim(A \vee \sim C)$$

A	C	$\sim A$	$\&$	$(C$	$\vee$	$A)$	$\sim$	$(A$	$\vee$	$\sim$	$C)$	Pr

D. PROBABILISTIC TRUTH-TABLES. Suppose that $\Pr(A\&C) = {}^{1}/_{12}$, $\Pr(A\& \sim C) = {}^{2}/_{12}$, $\Pr(\sim A\&C) = {}^{3}/_{12}$, and $\Pr(\sim A\& \sim C) = {}^{6}/_{12}$. Then, create a probabilistic truth-table by filling out the final column of the table above, and use it to determine the values below. [5%]

1. $\Pr(\sim A\&(C \vee A))$
2. $\Pr(\sim(A \vee \sim C))$

[1] Note: on the actual midterm, you may be asked to use truth-tables to determine whether arguments are formally valid, whether propositions are formal tautologies/contradictions, or whether propositions are mutually exclusive or jointly exhaustive.

E. PROBABILITY RULES, PART 1. Use the probability rules to answer the following question. (Show your work. You must clearly identify which proposition any capital letters you use stand for, and you must first write down a correct formula only involving a probability function and these propositions—*without* any numbers—in order to receive credit.) [20%]

> *We will draw two cards from a well-shuffled deck. What is the probability that* either *we do* not *get a diamond on the first draw* or *we do* not *get a spade on the second draw?*

F. PROBABILITY RULES, PART 2. Use the probability rules to answer the following question. (Show your work. You must clearly identify which proposition any capital letters you use stand for, and you must first write down a correct formula only involving a probability function and these propositions—*without* any numbers—in order to receive credit.) [20%]

> *You will meet ten people at a speed dating event. The probability that you are compatible with any particular person is ¼, and whether you are compatible with one person is probabilistically independent of whether you are compatible with any other people. What is the probability that you meet somebody you are compatible with?*

G. UPDATING PROBABILITIES. Use some version of Bayes' Theorem to answer the following question. (Again, you must show your work. That is, you must clearly identify which propositions any capital letters stand for, and you must first write down a correct formula involving only a probability function and these propositions—without any numbers—in order to receive credit.) [20%]

> *You are testing sprockets leaving the factory to see whether they are defective. You know from past experience that one in one hundred sprockets are defective. Your test has a 20% false positive/false negative rate. (Meaning: the probability that the test says a sprocket is defective, given that it is not defective, is 20%, and the probability that a test says a sprocket is not defective, given that it is, is also 20%.) Your test says that the sprocket is defective. Given this information, what is the probability that the sprocket is defective?*

Part II

Decision Theory

9 | The Decision Matrix

Goal: Learn how to formally represent a decision with a 'decision matrix'.

Suppose that your friends ask you whether you would rather eat dinner at Sichuan Impression or at Jitlada Thai. If you say 'Sichuan', then you'll order mapo tofu at Sichuan Impression; and, if you say 'Jitlada', then you'll order pad see ew at Jitlada. In this decision,[1]

1. You have available two different *acts*: saying 'Jitlada' and saying 'Sichuan'.
2. There are two possible *outcomes*: you will either eat mapo tofu or pad see ew.

The relationship between these two components of your decision can be formally represented like this:

You say 'Sichuan'	You eat mapo tofu
You say 'Jitlada'	You eat pad see ew

Along the rows, we have listed the available acts; inside the boxes, we have listed which outcome will result, if that act is selected. For this decision, you should first figure out which meal you would most prefer, and then make the choice that will get you that meal. If you prefer mapo tofu to pad see ew, then you should say 'Sichuan'. If, however, you prefer pad see ew to mapo tofu, then you should say 'Jitlada'. (If you could go either way, and don't prefer either to the other, then you can make either choice.)

Consider now a second decision: your friends ask you whether you would rather eat dinner at Sichuan Impression or Jitlada Thai—however, you know that there's construction on the 405, so if you and your friends try to go to Sichuan Impression, you'll get there after they close. So, in this decision, if you say 'Sichuan Impression', then you'll sit in traffic for an hour and miss dinner. If, however, you say 'Jitlada Thai', then you'll eat pad see ew. In *this* decision,

[1] A note on terminology: English allows us to use both 'choice' and 'decision' to refer to both 1) the situation in which you must select an act; and 2) the act you end up selecting. I will (stipulatively) reserve 'decision' for the former and 'choice' for the latter. Thus, as we will use the terminology in this class, you *face* a decision and you *make* a choice.

1. You have available the same two acts: saying 'Jitlada' and saying 'Sichuan'.
2. There are two *different* possible outcomes: you will either eat pad see ew or sit in traffic for an hour and miss dinner.

Again, we can represent the relationship between these two components of your decision formally like this:

You say 'Sichuan'	You sit in traffic
You say 'Jitlada'	You eat pad see ew

Assuming that you prefer pad see ew to sitting in traffic, you should say 'Jitlada'.

In the first two decisions, there is no uncertainty; you know precisely which outcomes will result from each available act. Let's consider a third decision in which there is uncertainty. Suppose you do not know whether or not there is construction on the 405. If there is, then you'll sit in traffic for an hour and you won't be able to get to Sichuan Impression before it closes. If there is not, then you will be able to get there on time, and you will eat mapo tofu. In this decision,

1. You have available the same two acts: saying 'Jitlada' and saying 'Sichuan'.
2. There are *three* possible outcomes: you will either sit in traffic, eat mapo tofu, or eat pad see ew.
3. There are two relevant *states of the world*: one in which there is construction on the 405, and one in which there is no construction on the 405.

We can represent the relationship between these *three* components of your decision in the following table:

	There is construction	There is not construction
You say 'Sichuan'	You sit in traffic	You eat mapo tofu
You say 'Jitlada'	You eat pad see ew	You eat pad see ew

If you prefer pad see ew to mapo tofu, then this choice is easy—no matter whether there's construction or not, you'll get a better outcome by saying 'Jitlada'. However, if you prefer mapo tofu to pad see ew, then the decision is harder—should you risk getting stuck in traffic, or play it safe?

Our goal in this section of the course is to come up with a theory—known as a *decision theory*—which will tell us what to do in circumstances like these. We will provide this theory with an abstract formal representation of the decision you are facing, and the theory will tell us which actions are the rational ones for you to choose. The goal today is to learn about how to construct an important part of that formal representation.

In general, our formal representation of the decision will include a collection of available *acts*, a collection of possible *outcomes*, and a collection of possible *states* that the world could be in. We will display this information in a *decision matrix* like this:

	State 1	State 2
Act 1	$\text{outcome}_{1,1}$	$\text{outcome}_{1,2}$
Act 2	$\text{outcome}_{2,1}$	$\text{outcome}_{2,2}$

(Of course, there may be more states, more acts, and more outcomes than this.) Let's take each aspect of the decision matrix in turn, starting with the outcomes.

Outcomes

An *outcome* is a proposition—it is something which can be true or false, which is expressed in English with a complete sentence. Intuitively, the outcome proposition says how things turn out, after (and as a result of) the choice you make in the decision.

Outcomes include everything (relevant) that you care about

The outcome proposition should settle everything that you care about. If we include less information in our specification of the outcomes, we won't have enough information to say what it is you should do in the decision.

Example 26. *We will flip a fair coin. You face a decision between two bets: A and B. Bet A pays out $1 if the coin lands heads, and loses you $1 if the coin lands tails. Bet B pays out $100 if the coin lands heads, and loses you $1 if the coin lands tails.*

Suppose that we model this decision with two outcomes: *you win money* and *you lose money*, and we create the following decision matrix:

	The coin lands heads	The coin lands tails
You take bet A	You win money	You lose money
You take bet B	You win money	You lose money

Then, we wouldn't be able to distinguish the choices of taking bet A and taking bet B—using these outcomes, both acts lead to the same outcome in each state of the world. But there's a clear difference between these choices. It would be irrational to take bet A. Bet B will certainly get you more money if the coin lands heads, and will certainly not get you any less money. So our specification of the possible outcomes in this decision should be: *you win $1*, *you win $100*, and *you lose $1*. And in general, if there's a difference that you care about, then that difference needs to show up in the possible outcomes.

But wait—if the outcomes have to say something about *everything* that you care about, then won't they *also* have to include information about, e.g., the age at which you die, whether the Dodgers win the world series in 2036, the quality of the champagne at your wedding, and so on? If so, then it looks like modeling any particular decision is going to get prohibitively difficult. While it's presumably possible *in principle* to create such a

decision matrix, practically, we won't be able to do so for any of the decisions we face in life.

A decision in which we've included a different outcome for literally *everything* that you care about is called a *grand world* decision. And a decision in which we've only included outcomes for *relevant* things you care about is called a *small world* decision. What do we mean by 'relevant things'? We will talk more about this later on, when we discuss problems of act-state dependence. But, for now, let's just say that something is *relevant* to your decision if how you act here and now makes a difference to that thing.

> If there's something that you care about, and how you choose in this decision makes a difference to that thing, then it must be included in your specification of the possible outcomes.

Since you care not just about *whether* you win money, but also *how much* money you win, and since how you choose in example 26 makes a difference to how much money you win, you must include a specification of how much money you win in the outcomes in example 26.

You have preferences between outcomes

In order to determine whether to go for Sichuan or Thai food, one thing you must do is figure out which you prefer. And, in general, in order to make a decision, you must determine which outcomes you prefer to which others.

In decision theory, we will typically use *numbers* to represent these preferences between outcomes. That is, we will assign higher numbers to outcomes which are preferred, and lower numbers to outcomes which are dispreferred.

This can be formalized with the notion of a *value function*, V, which maps an outcome to a real number. And we will say that you prefer the outcome o to the outcome o' if $V(o) > V(o')$. If, on the other hand, $V(o) < V(o')$, then we will say that you prefer o' to o. If $V(o) = V(o')$, then we will say that you are *indifferent* between o and o'—you don't prefer either to the other.

In the decision theory we will learn about in this course, these numbers are all we will need to know about the outcomes in order to say which choices are rational and which are irrational. So at times, we will just fill the decision matrix with the numbers which represent your preferences between the outcomes. For instance,

Example 27. *We will roll a fair die. You are offered a bet which will pay out* \$10 *if the die lands on a* 1*, and which will lose you* \$1 *if the die lands on* 2–6*. You can either accept the bet or reject it.*

The relevant outcomes are: *win \$10* (which will happen if you accept the bet and win), *lose \$1* (which will happen if you accept the bet and lose), and *break even* (which will happen if you reject the bet). Suppose that we can represent your preferences between these

	Die lands 1	Die lands 2–6
Accept bet	win \$10	lose \$1
Reject bet	break even	break even

(a)

	Die lands 1	Die lands 2–6
Accept the bet	10	−1
Reject the bet	0	0

(b)

Table 9.1: In 9.1a, a decision matrix with outcomes described in words. In 9.1b, a decision matrix with your preferences *between outcomes described with numbers.*

outcomes with the following numbers:

$$V(\textit{win \$10}) = 10$$
$$V(\textit{lose \$1}) = -1$$
$$V(\textit{break even}) = 0$$

Then, we could choose to represent your decision with the matrix in table 9.1a, or we could instead decide to represent it with the matrix in table 9.1b, since (after all) it's only the numbers associated with an outcome that are going to be needed for our decision theory.

Acts

When you start formalizing a decision, you will need to figure out what the available *acts* are. Like outcomes, the things we're going to call *acts* are also propositions—things that can be true or false, and which are expressed by complete English sentences. While outcomes are propositions describing the things you care about, *acts* are propositions which describe the choices you might make in the decision. (Even though, officially, we are taking acts to be propositions describing the potential choices you might make, I will at times refer to the acts with nominalizations like 'your accepting the bet', rather than the proposition 'you accept the bet'.)

Acts are certain to be entirely within your control

The act propositions need to be ones that you know for sure you are able to make true at will. That is, you must know for sure that which act you select is entirely under your control.

Example 28. *The coin is flipped, and you are asked to call how it lands while it is in the air. If your call is correct, your team will go first. If your call is incorrect, the other team will go first.*

It would be a mistake to model this decision with the following matrix:

	Coin lands heads	Coin lands tails
You call correctly	go first	go first
You call incorrectly	go second	go second

Looking at this matrix, you might conclude that it would be irrational to call the coin incorrectly. But there's nothing irrational about getting the call wrong. It's unfortunate, of course, but it's just a matter of bad luck, not bad choices. The issue with this choice of act propositions is that whether you call the coin correctly or incorrectly is not entirely under your control.

Example 29. *You are recovering from a bicycle accident in which you injured your hands, and you're not sure whether your hands are healed enough for you to open the pickle jar. If you try to open it and you are not healed, you will experience pain and delay the healing. If you ask for help opening it, you won't experience pain, but you will be mocked by your friends. In fact—unbeknownst to you—your hands have healed enough, and if you were to try to open the pickle jar, you'd succeed and feel no pain.*

We should *not* model this decision with the acts: *you open the pickle jar* and *you ask for help*. Even though, in fact, it is in your power to make the proposition *you open the pickle jar* true, this is something about which you are uncertain. The correct way to model this decision is like this:

	Hands are healed	Hands are not healed
You try to open the jar	You get a pickle	You experience pain
You ask for help	You get a pickle and mockery	You get a pickle and mockery

Acts form a partition

Or is it? You might think that I've left out a relevant alternative act: the act of neither opening the jar nor asking for help. Surely this is an available act, too? Indeed it is. So really, if we're going to model example 29 correctly, we'll need to include an additional act for this option.

In general, the acts between which you are choosing need to be mutually exclusive and jointly exhaustive. That is, using the terminology we introduced back in lecture 1, the available acts must form a *partition*—it must be impossible for more than one of them to be true at once, and it must be impossible for all of them to be false at once.

Example 30. *There are two horses racing today*: *Secretariat and Mudskipper. The betting house is offering* $1:3$ *odds on Secretariat winning* (*that is*: *you lose* $\$1$ *if Secretariat loses and you gain* $\$3$ *if Secretariat wins*), *and they are also offering* $1:3$ *odds on Mudskipper winning. You are at the betting house with* $\$2$ *and face a decision of which bets to make.*

You might try modeling example 30 with the following decision matrix:

	Secretariat wins	Mudskipper wins
You bet on Secretariat	You win $3	You lose $1
You bet on Mudskipper	You lose $1	You win $3

But notice that these acts are not jointly exhaustive. We have left out, of course, the option of not making *either* bet. But we have *also* left out the option of making *both* bets. Moreover, this might be a better option. Notice that, if you make both bets, then you are *guaranteed* to make a $2 profit! (You'll definitely lose $1 on one of the bets, but definitely win $3 on the other.)

	Secretariat wins	Mudskipper wins
You bet on Secretariat	You win $3	You lose $1
You bet on Mudskipper	You lose $1	You win $3
You bet on both	You win $2	You win $2
You bet on neither	You break even	You break even

If we didn't stop to make sure that the acts we were using were jointly exhaustive, we would have missed this important alternative.

Example 31. *There are three cups on the table before you, labeled A, B, and C. One of them hides a ball. If you can correctly guess which cup hides the ball, then you will win a prize. But if you guess incorrectly, then you will win nothing.*

It would be a mistake to model this decision with the acts *you guess A* and *you guess C*. These acts are not jointly exhaustive—if you choose B, then they are both false. Similarly, it would be a mistake to model this decision with two acts: *you don't guess A* and *you don't guess C*. These act propositions are jointly exhaustive, but they are not mutually exclusive. If you choose B, then they are both true. In general, the act propositions you use to represent your decision should be both mutually exclusive and jointly exhaustive. Or, using the terminology we introduced in lecture 1: they should form a *partition*.

But wait—won't there be other possible acts, too? Acts like *you cluck like a chicken and dance around the three cups*. Do all of these acts have to be explicitly considered? Strictly speaking, yes, though they don't necessarily have to be considered *individually*. You can always include a 'none of the above' option as the final act. And you should think about what other alternatives might be available to you to choose instead. In example 31, it seems plausible that making *some* guess is going to get you a better outcome than any other thing you can do right then and there—but it's still worth thinking through which other acts might be available that you haven't yet considered. Having a good sense of when you've enumerated all of the individual acts worth considering is more a matter of art than science. But in general, when facing a serious decision, you should probably dedicate some time to thinking about which other acts are available that you're failing to explicitly consider.

Acts and States Determine Outcomes

Why does it matter that we don't use the acts *you don't guess A* and *you don't guess C* to model example 31? In part because your choice of act, together with the state of the world, should determine which outcome you get. If they did not, then we would not be able to fill out the decision matrix. For instance, think about which outcomes we could put in this decision matrix:

	Ball under A	Ball under B	Ball under C
You don't guess A	You win nothing	???	???
You don't guess C	???	???	You win nothing

Of course, if the ball is under A and you guess A, then you won't win anything. And if the ball is under C and you don't guess C, you won't win anything. But what outcome will you get if the ball is under B and you don't guess A? It's not clear, since it's not clear whether you would guess B or C.

Notice also that acts can be *finer* or *coarser-grained*. That is, we can include *more* or *less* in our description of the act. We could have the act proposition *You try to open the pickle jar*, or we could instead have the act propositions *You try to open the pickle jar with your left hand* and *You try to open the pickle jar with your right hand*. If it doesn't matter whether you open the pickle jar with your left or right hand, then we could use either—since it will make no difference which we use. Of course, we might prefer to use the single act proposition in the interest of simplicity. However, if it *matters* whether you use your left or right hand—perhaps your left hand is more likely to be healed than your right hand—then we should distinguish those two act propositions. After all, we want acts and states to together determine outcomes, and if it matters which hand you use, then the act proposition *You try to open the pickle jar* won't tell us enough to determine which outcome you get.

States

As you are figuring out what the available acts and possible outcomes are, you must also be figuring out which *states of the world* are needed to formally model your decision. Like acts and outcomes, the things we're going to call states are propositions—things that can be true or false, and which are expressed by complete English sentences. While *acts* specify your choice, *states* fill in all of the additional information needed in order to determine which outcome a given choice will lead to.

States form a partition

Just like *acts* should form a partition, so too should *states* form a partition.

Example 32. *We will roll a fair die. You are offered a bet which will pay out* $10 *if the die lands on a* 1 *and lose you* $1,000,000 *if the die lands on a* 6*.*

Suppose you modeled this decision with the states *the die lands on a 1*, *the die lands on a 2*, *the die lands on a 3*, *the die lands on a 4*, and *the die lands on a 5*.

	Die lands 1	Die lands 2	Die lands 3	Die lands 4	Die lands 5
You accept the bet	win $10	break even	break even	break even	break even
You reject the bet	break even	break even	break even	break even	break even

Then, you might reason: well, accepting the bet will never get me a *worse* outcome than rejecting the bet will; and in one state, accepting the bet gets me a *better* outcome! This would be a fallacious bit of reasoning. The matrix above leaves out the important possibility in which you lose $1,000,000. It leaves this possibility out because its states are not jointly exhaustive.

States and Acts Determine Outcomes

In our representation of your decision with a decision matrix, we will need an outcome for each entry in the matrix—which means that we will need to have a unique outcome for each choice of act and state. We saw above that this constrains which *acts* we can use. But it also constrains your choice of *states*.

For instance, it would be a mistake to model the decision in example 32 with the following decision matrix:

	Die lands low (1, 2, 3)	Die lands high (4, 5, 6)
Accept bet	???	???
Reject bet	???	???

This would be a mistake precisely because the states in the column, together with the acts in the rows, do not determine a unique outcome.

There's no uniquely correct choice of states; though some will be more complicated than they need to be. For instance, we could model example 32 with these states:

	Die lands 1	Die lands 2	Die lands 3	Die lands 4	Die lands 5	Die lands 6
Accept bet	win $10	break even	break even	break even	break even	lose $1,000,000
Reject bet	break even	break even	break even	break even	break even	break even

But since the differences between 2, 3, 4, and 5 landings don't make any difference to the outcome we get for each act, we could instead use the simpler states *Die lands 1*, *Die lands 2–5*, and *Die lands 6*.

States and Acts are Independent

For the decision theory we're going to be studying in this part of the course, it's important that you use states and acts which are *independent from each other*.

Example 33. *The king has declared war, and tomorrow you march out to battle. You can either use your life savings to buy armor or else march out to battle unprotected.*

Here, it would be a *mistake* to use the states *you survive the battle* and *you do not survive the battle*, and model your decision with the following decision matrix

	You survive the battle	You do not survive the battle
You buy armor	Lose life savings	Lose life & life savings
You don't buy armor	Lose nothing	Lose life

Looking at this decision matrix, you might reason as follows: *whether I survive the battle or not, I'll be better off if I don't buy the armor.* This would be a fallacious bit of reasoning; and the reason it would be fallacious is precisely that whether you buy armor *makes a difference* to whether you survive the battle.

For this part of the course, we will be assuming that acts and states are *independent*, in the sense that which act you choose does not make a difference to which state you're in. (Later on in the course, we'll see that this notion of 'act-state independence' is ambiguous between two different kinds of independence: *probabilistic* independence and *causal* independence. But for now, we'll only be considering states which are *both* probabilistically independent *and* causally independent of the act propositions.)

In Summation

In sum, then: when you are formalizing a decision with a decision matrix, you should:

1. Figure out everything you care about which your choice might make a difference to, and include a collection of outcomes which specify every possible way things might turn out with respect to those relevant things you care about.
2. Find a partition of acts available to you. These must be acts which are certain to be under your control.
3. Find a partition of states. These states must be independent of which act you choose.
4. Your partitions of acts and states must be detailed enough that, for each act and each state, there is a unique outcome which will result, if you choose that act in that state.

If you reach step 4 and find that your act and state partitions do *not* determine a unique outcome for each pair of act and state, then you will need to go back to steps 2 and 3 and refine your act and state partitions until they are detailed enough that they *do* determine a unique outcome.

10 | Decision Making Under Uncertainty

Goal: learn about, and critically evaluate, several methods for making decisions using the decision matrix alone, *without* using any probabilities.

In the last chapter, we saw how to formalize a decision using a *decision matrix*. In this chapter, we're going to consider whether we can decide what to do using a decision matrix alone. We will see several proposals, but I will be drawing a largely negative conclusion from this survey. While there are *some* decisions in which you can figure out what to do with the decision matrix alone, in general, you will need more information than a decision matrix alone provides. In subsequent chapters, I'll argue that what's missing is what we spent the first part of the course learning about: *probabilities*.

We can draw a distinction between decision-making under *uncertainty* and decision-making under *risk*. *Risk* is when you don't know which state you're in, but you do know exactly what probabilities to assign to each state. *Uncertainty*, in contrast, is when you don't know which state you're in and you don't have any probabilities for those states, either.

This week, we're going to be learning about various proposals for decision-making under *uncertainty*. But, as a preliminary, let's spend a bit of time talking about the *value function* that we'll be using to represent your preferences between outcomes.

Coda: Ordinal and Cardinal Value Functions

As I mentioned last time, we will represent your preferences between outcomes with a *value function*, V. You hand V an outcome, and it hands you back some real number which represents *how strongly* you want that outcome. Going forward, I'm going to start filling the decision matrix with these numbers, rather than the outcomes they correspond to.

It will be important for us to distinguish between two different ways that we could use numbers to represent the strength of your desire. In the first place, we could care only about the *ordering* of the numbers. In that case, we will say that V is an *ordinal* value function. When we call V an *ordinal* value function, we're just saying that all you need to pay attention to is the order of the values. For instance, if a, b, and c are the only outcomes,

	V_1	V_2	V_3	V_4	V_5
a	1	−1	112	8	4
b	2	−2	76	7	3
c	4	−4	16	0	2
d	3	−3	24	1	1

Table 10.1: The numbers beneath a value function are the numbers it assigns to the outcome in that row.

then V_1, V_2, V_3, and V_4 below are *ordinally equivalent* value functions:

$V_1(a) = 1$	$V_2(a) = -500$	$V_3(a) = 0$	$V_4(a) = 1/4$
$V_1(b) = 3$	$V_2(b) = 20$	$V_3(b) = 1000$	$V_4(b) = 1/2$
$V_1(c) = 2$	$V_2(c) = 19$	$V_3(c) = 1$	$V_4(c) = 1/3$

Each of these value functions places b higher than c, and c higher than a. And that's all that it takes for these value functions to be *ordinally equivalent*.

Ordinal Equivalence What it is for two value functions, V and V', to be *ordinally equivalent* is for them to agree about the order of every pair of outcomes. That is: if V and V' are ordinally equivalent, then, for any two outcomes a and b,

$$V(a) \geqslant V(b) \qquad \text{iff} \qquad V'(a) \geqslant V'(b)$$

When we call V an *ordinal* value function, we just mean "you can use this one, or you could instead use any other ordinally equivalent one—I'm not wedded to the precise numbers, so long as the order stays the same".

Exercise 23. *Suppose that there are only four outcomes*: a, b, c, *and* d. *Then, which of the value functions in table* 10.1 *are ordinally equivalent*?

Contrast this with other measurement scales, like *temperature*. In the case of temperature scales like Fahrenheit and Celsius, we don't only use numbers to represent which temperatures are *hotter* and *colder* than which others. We care about more than just the *order* of the numbers. What more do we care about? As a first pass: we additionally care about the *magnitude of the difference between temperatures*. There's something to this first-pass thought, but it's easy to see that it's not quite right. Think about Celsius and Fahrenheit. In Celsius, the temperature at which water freezes is 0° and the temperature at which water boils is 100°—the magnitude of the difference between them is 100. In Fahrenheit, then temperature at which water freezes is 32° and the temperature at which water boils is 212°—the magnitude of the difference between them is 180, *not* 100. So different temperature scales don't preserve magnitudes of differences—what *do* they preserve?

The answer is this: they preserve *ratios* of magnitudes of differences. Consider a third temperature: the one at which mercury freezes. This is about the point at which Celsius and Fahrenheit meet: it is about -40° in either scale. Now, consider the *ratio* of these temperature *differences* in Celsius,

$$\frac{C(\textit{water boils}) - C(\textit{water freezes})}{C(\textit{water boils}) - C(\textit{mercurcy freezes})} = \frac{100 - 0}{100 - (-40)} = \frac{100}{140} = \frac{5}{7}$$

and consider the same *ratio* of temperature *differences* in Fahrenheit,

$$\frac{F(\textit{water boils}) - F(\textit{water freezes})}{F(\textit{water boils}) - F(\textit{mercury freezes})} = \frac{212 - 32}{212 - (-40)} = \frac{180}{252} = \frac{5}{7}$$

They are the same; and that's not an accident. Scales like temperature care about more than just ordering, and the additional something that they care about is precisely the ratios of differences.

For most applications in decision theory, we're going to care not just about the *order* of your preferences between outcomes; we're also going to care about the kind of additional information that you find in temperature scales—the ratios of differences. A value function which encodes this kind of information is often called (by economists) a *cardinal* value function. (Mathematicians and philosophers, on the other hand, tend to call it an *interval* scale.)

If you have two value functions which *both* agree about the order of outcomes *and* agree about the ratios of differences between outcomes, then we will call those value functions *cardinally equivalent*.

Cardinal Equivalence (v1) What it is for two value functions, V and V', to be *cardinally equivalent* is for them to (1) agree about the order of every pair of outcomes, and (2) agree about the ratios of differences between every quadruple of outcomes. That is: if V and V' are cardinally equivalent, then, for any four outcomes a, b, c, and d,

1. $V(a) \geqslant V(b)$ iff $V'(a) \geqslant V'(b)$; and
2. $\frac{V(a)-V(b)}{V(c)-V(d)} = \frac{V'(a)-V'(b)}{V'(c)-V'(d)}$

When we call V a *cardinal* value function, we just mean "you can use this one, or you could instead use any other cardinally equivalent one—I'm not wedded to the precise numbers, so long as both the order and the ratios of differences stay the same."

Exercise 24. *Suppose that there are only four outcomes*: a, b, c, *and* d. *Then, which of the value functions in table* 10.2 *are cardinally equivalent*?

There's another way we can characterize what it is for two value functions to be cardinally equivalent, which will be somewhat easier to work with.

	V_1	V_2	V_3	V_4	V_5
a	1	10	−30	−1	−100
b	2	20	0	−2	− 50
c	3	30	30	−3	0
d	4	40	60	−4	50

Table 10.2: The numbers beneath a value function are the numbers it assigns to the outcome in that row.

Cardinal Equivalence (v2) What it is for two value functions, V and V', to be *cardinally equivalent* is for there to be some positive number $\alpha > 0$ and some number (positive, negative, or zero) β such that, for every outcome o,

$$V'(o) = \alpha \cdot V(o) + \beta$$

Exercise 25. *For each of the value functions in table* 10.2 *which are cardinally equivalent, V and V', find the values of α and β which make it so that, for any o, $V'(o) = \alpha \cdot V(o) + \beta$.*

For some of the decision rules we're going to consider below, the only thing about your values which will matter are the *orders* of the outcomes; so we'll only need to care about the numbers *up to ordinal equivalence* (any other ordinally equivalent value function would serve just as well). But for other of the rules, we will also need to care about the ratios of differences; so we'll need to care about the numbers *up to cardinal equivalence* (any other *cardinally* equivalent value function would serve just as well).

Dominance

Example 34 (Prisoner's Dilemma)**.** *You and an accomplice cheated on a problem set using AI. You are each called into the dean's office and placed in separate rooms. You are then informed of the following: you are strongly suspected of cheating along with your accomplice. We are now going to give you both an opportunity to confess. If neither of you confess, then you will both fail the assignment, but neither will fail the course. If you confess and your accomplice does not, then you will not be punished at all, and your accomplice will be expelled. If your accomplice confesses and you do not, then they will not be punished, and you will be expelled. If you both confess, then you will both fail the course.*

In this decision, let's suppose that there are four relevant outcomes for you: it could be that you are not punished at all, it could be that you fail the assignment, it could be that you fail the course, and it could be that you are expelled. There are two available acts: you could either confess or not confess. And there are two relevant states: either your accomplice confesses or they do not.

	Accomplice confesses	Accomplice does not confess
You confess	You fail the course	You are not punished
You do not confess	You are expelled	You fail the assignment

Let's assume that your preferences can be represented with the following (ordinal) value function,

$$V(\textit{you are not punished}) = 4$$
$$V(\textit{you fail the assignment}) = 3$$
$$V(\textit{you fail the course}) = 2$$
$$V(\textit{you are expelled}) = 1$$

Then, we could represent your decision with the following matrix:

	Accomplice confesses	Accomplice does not confess
You confess	2	4
You do not confess	1	3

Notice that, in this decision, no matter what your accomplice does, *you* will be better off confessing. If your accomplice confesses, then *you* confessing will save you from being expelled. And, if your accomplice doesn't confess, then *you* confessing will save you from any punishment at all. Either way, confessing leaves you better off. So it seems like confessing is the rational choice.

This kind of reasoning is so common in decision theory that we have special terminology for it. If one act is *guaranteed* to get you a better outcome than the second, no matter what, then we say that the first option *dominates* the second.

> **Dominance** If A leads to a better outcome than B does in *every* state, then A *dominates* B.

Consider the following decision matrix (it won't matter what the acts and states are):

	State S	State T
Act A	2	1
Act B	1	1

In this decision, A is not *guaranteed* to get a better outcome than B, but A *might* get you a better outcome than B does, and A certainly won't get you a *worse* outcome than B does. For this reason, we'll say that A *weakly dominates* B.

> **Weak Dominance** If A leads to a better outcome than B does in *some* state, and it doesn't lead to a worse outcome than B in any state, then A *weakly dominates* B.

Notice that dominance implies weak dominances—if A dominates B, then A weakly dominates B.

Then, we may consider two principles of choice:

> **Dominance Principle** If an act is dominated, then it is not a rational choice.
>
> **Weak Dominance Principle** If an act is weakly dominated, then it is not a rational choice.

(Notice that, in the weak dominance principle, the *weakness* is talking about the kind of dominance, not the strength of the principle. The weak dominance principle is *stronger* than the dominance principle, in the sense that the weak dominance principle *entails* the dominance principle.)

The dominance principle is the least controversial of the principles we're going to talk about in this section. But even it is not without controversy. In the first place, consider what the principle says about the following example.

Example 35. *God informs you that you may have as many (finite) days in heaven as you wish. Just name any (finite) number, and you'll get to reside in heaven that many days.*

In this example, let's suppose that you know for sure that God is telling the truth; so there's just one state, and you know exactly which outcome you'd get, given each act. There are infinitely many possible acts. You could ask for 1 day in heaven, 2 days in heaven, 3 days in heaven, 4 days in heaven, and so on—for every natural number n, you could ask for n days in heaven.

You ask for 1 day	You get 1 day in heaven
You ask for 2 days	You get 2 days in heaven
You ask for 3 days	You get 3 days in heaven
⋮	⋮
You ask for n days	You get n days in heaven
⋮	⋮

In this situation, if we assume that you prefer more days in heaven to fewer days in heaven, then *every* option is dominated. Asking for 1 day is dominated by asking for 2. Asking for 2 is dominated by asking for 3. And, in general, asking for n days is dominated by asking for $n + 1$ days (as well as asking for $n + 2$, $n + 3$, and so on). So the dominance principle implies that, in this decision, no choice is rational.

Perhaps, then, we should revise the dominance principle. We could do so in several ways. In the first place, we could restate the dominance principle so that it doesn't say that dominated acts are *irrational*. Instead, we could just say that a dominated act is *less rational* than an act that dominates it.

Dominance Principle (v2) If one act, A, (weakly) dominates another act, B, then A is a more rational choice than B is.

(Like the original dominance principle, this one comes in two flavors, depending upon whether it is talking about dominance or *weak* dominance.)

What is the connection between this comparative notion, "is a more rational choice than", and the absolute notion, "is a rational choice"? The standard story is that an option is *rational* (full stop) whenever it is one of the *most* rational options.

Absolute from Comparative If there's no act which is more rational than A is, then A is a rational choice.

In example 35, the 2nd version of the dominance principle says that every act has some other act that's more rational than it is. So the principle *Absolute from Comparative* doesn't tell us anything about which acts are rational choices in this decision—the 'if' part of the claim isn't satisfied by anything. But neither does it tell us that any act is *irrational* in this decision. The principle simply remains silent.

Alternatively, we could draw a different lesson from example 35. Notice that, in this example, even though each act is dominated, it is dominated by an act *that is itself dominated*. After making this observation, we might propose a different modification of the dominance principle:

Undominated Dominance Principle If an act, A, is (weakly) dominated by an act, B, which is not itself (weakly) dominated, then A is not a rational choice.

We'll see other reasons to worry about the dominance principle later in the course.

Exercise 26. *To apply the dominance principle, do we need a cardinal value function, or is an ordinal value function enough*?

As I said, the dominance principle is the least controversial of the principles we're going to talk about—but in many interesting cases, the dominance principle falls silent. Consider

Example 36.

	State S	State T
Act A	1	0
Act B	−1	2

Here, neither option (weakly) dominates the other; so the dominance principles will tell us nothing about how to choose in example 36.

Maximin and Maximax

In contrast, the principle of *maximin* will tell us precisely what to do in example 36. *Maximin* tells you to *maximize* the *minimum* value that your choice might lead to. In example 36, A has a minimum value of 0 (in state T), and B has a minimum value of -1 (in state S). Since 0 is greater than -1, maximin says that you are required to choose act A.

Let's say that the *worst case outcome* for a given act is the most dispreferred outcome that might result, if you choose that act. Then, maximin says that you have to choose an act whose worst case outcome is best.

> **Maximin Principle** If the worst case outcome for an act A is better than the worst case outcome for an act B, then A is a more rational choice than B.

Example 37 (Rawls' Veil of Ignorance)**.** *You are deciding how to allocate society's resources, without knowing who in society you are* (*you are standing behind the 'veil of ignorance'*)*. There are three possible people you could be*: *Al, Betty, or Carl. And there are four distributions between which you are choosing*: *the* equal *distribution, which gives the same to each of Al, Betty, and Carl, and the* unequal *distributions, which give more to one person, and less to everyone else. You only care about what happens to you* (*whoever you are*)*; you do not care at all about what happens to the other people in society.*

	You are Al	You are Betty	You are Carl
Equality	2	2	2
Unequal in Al's favor	4	1	1
Unequal in Betty's favor	1	4	1
Unequal in Carl's favor	1	1	4

The worst-case outcome for equality is 2, whereas the worst-case outcome for each of the unequal acts is 1. So, according to Minimax, equality is a more rational choice than inequality.

This example is modeled on a thought experiment of John Rawls'. He thought that a distribution of society's resources was *just* if it was a distribution that we would all agree to from behind the veil of ignorance. And he then appealed to the maximin principle in order to justify his contention that, from behind the veil of ignorance, people would only allow inequality if it made things better for the worst off. Therefore, Rawls concluded that justice requires us to only have inequality if this inequality makes things better for the worst off.

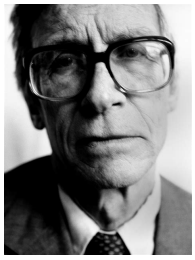


Figure 10.1: John Rawls

Decision theorists are far less sanguine about the maximin principle than Rawls was. Think about what the principle says about this decision:

Example 38. *There is an urn which contains* 10 *marbles, each of which is either red or green. You do not know how many of these marbles are red, nor how many are green* (*nor do I*). *A single marble will be drawn. I offer you a bet which will pay out* $1,000,000 *if the drawn marble is green, and which will lose you* 1¢ *if the drawn marble is red.*

The drawn marble is:	red	green
You accept the bet	−0.01	1,000,000
You reject the bet	0	0

The worst case outcome for accepting the bet is that you lose 1¢; whereas the worst case outcome for rejecting it is that you break even. So the maximin principle says that rejecting the bet is the only permissible option. To many, this seems absurdly overcautious—given that the potential upside of accepting the bet is so good, and the potential downside is negligible, it doesn't seem at all irrational to take the best.

There's a dual principle to maximin, which tells you to *maximize* the *maximum* value that your choice might lead to. This principle is known as *maximax*. For instance, in Rawls' Veil of Ignorance, the maximum value that equality might lead to is 2; whereas the maximum value that inequality might lead to is 4. Since $4 > 2$, the maximax principle says that you are rationally required to choose inequality.

Let's say that the *best case outcome* for a given act is the most preferred outcome that might result, if you choose that act. Then, maximin says that you have to choose an act whose best case outcome is best.

Maximax Principle If the best case outcome for an act A is better than the best case outcome for an act B, then A is a more rational choice than B.

In example 38, maximax says that accepting the best is the only rationally permissible option.

Maximax faces objections which are very similar to the ones faced by maximin. Consider, for instance,

Example 39. *There is an urn which contains* 10 *marbles, each of which are either red or green. You do not know how many of these marbles are red, nor how many are green* (*nor do I*). *A single marble will be drawn. I offer you a bet which will pay out* 1¢ *if the drawn marble is red, and which will lose you* $1,000,000 *if the drawn marble is green.*

The drawn marble is:	red	green
You accept the bet	0.01	−1,000,000
You reject the bet	0	0

The best case outcome for accepting the bet is that you win 1¢; whereas the best case outcome for rejecting the bet is that you break even. Since 1¢ is better than breaking even,

maximax says that you are required to accept the bet. To many, this seems absurdly *under*cautious—given that the potential downside of accepting the bet is so bad, and the potential upside is negligible, it doesn't seem at all irrational to turn down this bet.

Exercise 27. *What do Maximin and Maximax say about the decision below*?

	State S	*State T*	*State U*
Act A	0	1,000,000	1,000,000
Act B	0	0	1,000,000

Exercise 28. *What does the weak dominance principle say about the decision from exercise* 27*? From this, what can you conclude about the logical relationship between Maximin, Maximax, and the weak dominance principle*?

Exercise 29. *To apply Maximin and Maximax, do we need a cardinal value function, or is an ordinal value function enough*?

Hurwicz's Principle

Leonid Hurwicz suggested that, instead of using either minimax or maximax to make decisions under uncertainty, we instead *split the difference* between them by taking an average of each act's best and worst case outcomes. On Hurwicz's proposal, each person has some number between zero and one—call it their 'risk coefficient', and use the letter 'r' for it—which represents how *risky* they are. If $r = 0$, then the person is maximally risk-averse. If $r = 1$, then the person is maximally risk-seeking. If $r = 1/2$, then they are somewhat risk-averse and somewhat risk-seeing. Given your risk coefficient, r, you should evaluate each available act A with the weighted sum

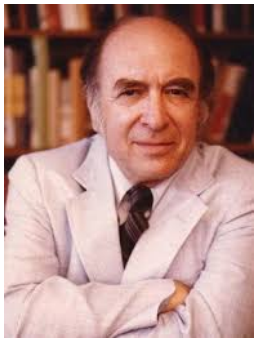


Figure 10.2: Leonid Hurwicz

$$r \cdot \textit{best case outcome for } A \;+\; (1 - r) \cdot \textit{worst case outcome for } A$$

and an act with the highest score is a rational choice.

Hurwicz's Principle If your risk-coefficient is r, and

$$r \cdot \textit{best case outcome for } A \;+\; (1 - r) \cdot \textit{worst case outcome for } A$$

is greater than

$$r \cdot \textit{best case outcome for } B \;+\; (1 - r) \cdot \textit{worst case outcome for } B$$

then A is a more rational choice than B.

If $r = 0$, then Hurwicz's Principle will say the same thing as Maximin (*why?*), and if $r = 1$, then Hurwicz's Principle will say the same thing as Maximax (*why?*).

Exercise 30. *Suppose that your risk coefficient is* 3/4*. Then, say what Hurwicz's principle tells you to do in examples* 38 *and* 39*.*

Exercise 31. *What does Hurwicz's principle tell you to do in the decision from exercise* 27 (*reproduced below*)?

	State S	*State T*	*State U*
Act A	0	1,000,000	1,000,000
Act B	0	0	1,000,000

Does this advice depend upon your risk coefficient? *From reflection on this decision, what can you conclude about Hurwicz's principle and the weak dominance principle*?

Exercise 32. *To apply Hurwicz's principle, do we need a cardinal value function, or is an ordinal value function enough*?

Minimax Regret

In exercise 31, we learnt that Hurwicz's principle is incompatible with the weak dominance principle. There, act A weakly dominates act B, but Hurwicz's principle says that they are equally good. Let's consider a different principle, which will tell us that act A is a more rational choice than act B in the decision from exercise 31. According to this principle, you should be attempting to minimize the maximum *regret* that you'll feel, once your choice is made.

To appreciate this notion of *regret*, think about what will happen if, in the decision from exercie 31, you choose act B in the state T. Then, you'll end up with a value of 0, when you could have had a value of 1,000,000 instead. In this case, let's say that your *regret* is 1,000,000. And, in general,

> The *regret* of choosing A in the state S is the difference between the outcome you get choosing A in S and the best outcome it was possible for you to get in state S.

For instance, consider the decision matrix in table 10.3. Given this decision matrix, table 10.4 shows the *regret* of each act in each state. Given the decision matrix from table 3, the maximum possible regret you'll feel for choosing act A is 90 (in state S), the maximum possible regret you'll feel for choosing act B is 10 (in state T), and the maximum regret you'll feel for choosing act C is 200 (in state S). Since 10 is less than 90, which is less than 200, the principle of minimax regret says that B is a more rational choice than A, which is in turn a more rational choice than C.

In general, say that the *worst case regret* for an act is the maximum amount of regret that that act might have. Then,

	State S	State T	State U
Act A	10	30	3
Act B	100	20	2
Act C	−100	5	1

Table 10.3: A Decision Matrix

	State S	State T	State U
Act A	90	0	0
Act B	0	10	1
Act C	200	25	2

Table 10.4: The regret of each act in each state, given the decision matrix from table 10.3.

Minimax Regret Principle If the worst case regret for an act A is less than the worst case regret for an act B, then A is a more rational choice than B is.

Exercise 33. *Which acts does the minimax regret principle say are rational to choose in this decision?*

	State S	*State T*	*State U*
Act A	9	0	0
Act B	0	10	0

Exercise 34. *Suppose that we add the act C to the decision from exercise 33,*

	State S	*State T*	*State U*
Act A	9	0	0
Act B	0	10	0
Act C	20	10	−100

Now, which acts does minimax regret says are rational? Is there anything surprising about your answer?

Exercise 35. *To apply the minimax regret principle, do we need a cardinal value function, or is an ordinal value function enough?*

1. Create a decision matrix for the following decisions.

   (a) There is an urn which contains 30 red marbles, and 60 other marbles which are either yellow or green (we don't know how many of the 60 remaining marbles are yellow and how many are green). A marble will be drawn at random from the urn. You can take one (but no more than one) of the following bets: Bet *A* will get you $100 if the drawn marble is red. Bet *B* will get you $100 if the drawn marble is yellow.

   - For bonus points: say which choice you think you'd make, if you were facing this decision, and say something about why.

   (b) There is an urn which contains 30 red marbles, and 60 other marbles which are either yellow or green (we don't know how many of the 60 remaining marbles are yellow and how many are green). A marble will be drawn at random from the urn. You can take one (but no more than one) of the following bets: Bet *C* will get you $100 if the drawn marble is red or green. Bet *D* will get you $100 if the drawn marble is yellow or green.

   - For bonus points: say which choice you think you'd make, if you were facing this decision, and say something about why.

   (c) We will roll a 100-sided die. You can take one (but not more than one) of the following bets. Bet *A* will get you $1,000,000 no matter what. Bet *B* will get you $0 if the die lands on a 1, $5,000,000 if the die lands on a number between 2 and 10, and $1,000,000 if the die lands on a number between 11 and 100.

   - For bonus points: say which choice you think you'd make, if you were facing this decision, and say something about why.

   (d) We will roll a 100-sided die. You can take one (but not more than one) of the following bets. Bet *C* will get you $1,000,000 if the die lands on a number between 1 and 10, and will get you nothing if the die lands on a number between 11 and 100. Bet *D* will get you $5,000,000 if the die lands on a number between 2 and 10, and will get you nothing if the die lands on any other number.

   - For bonus points: say which choice you think you'd make, if you were facing this decision, and say something about why.

   (e) There are three horses racing: Secretariat, Mudskipper, and War Admiral. The betting house only allows you to place a single $1 bet on any particular horse. They are offering $1:3$ odds on each horse (that is: if you bet on any horse, then you'll lose $1 if that horse loses the race, and you'll win $3 if that horse wins the race). You are at the betting house with $3 and face a decision of which bets to make (if any).

   - For bonus points: say which choice you think you'd make, if you were facing this decision, and say something about why.

2. Suppose that there are only four possible outcomes (a, b, c, and d) and consider the four value functions V_1, V_2, V_3, and V_4 shown in table 10.5.

	V_1	V_2	V_3	V_4
a	3	5	1	-5
b	5	15	2	-15
c	9	35	100	-35
d	-3	-25	-100	25

Table 10.5: The numbers beneath a value function are the numbers it assigns to the outcome in that row.

   (a) Are V_1 and V_2 ordinally equivalent? Why or why not?

   (b) Are V_1 and V_2 cardinally equivalent? If so, then show that they are cardinally equivalent by finding a value of $\alpha > 0$ and a value of β so that, for every outcome o, $V_2(o) = \alpha \cdot V_1(o) + \beta$. If not, then show that there is no such α and β.

   (c) Are V_2 and V_3 ordinally equivalent? Why or why not?

   (d) Are V_2 and V_3 cardinally equivalent? If so, then show that they are cardinally equivalent by finding a value of $\alpha > 0$ and a value of β so that, for every outcome o, $V_3(o) = \alpha \cdot V_2(o) + \beta$. If not, then show that there is no such α and β.

   (e) Are V_2 and V_4 ordinally equivalent? Why or why not?

   (f) Are V_2 and V_4 cardinally equivalent? If so, then show that they are cardinally equivalent by finding a value of $\alpha > 0$ and a value of β so that, for every outcome o, $V_4(o) = \alpha \cdot V_2(o) + \beta$. If not, then show that there is no such α and β.

3. For the decision in part 1(a), assume that your value function is linear with dollars—that is, $V(\$n) = n$. (So, in particular, $V(\$0) = 0$ and $V(\$100) = 100$). Then, say which choice is rational (or, which choices are rational) according to

   (a) the maximin principle

   (b) the maximax principle

   (c) the Hurwicz principle (with a risk coefficient of $1/2$)

   (d) the minimax regret principle

4. For the decision in part 1(c), assume that your value function is linear with dollars ($V(\$n) = n$). Then, say which choice is rational (or which choices are rational) according to

   (a) the maximin principle

   (b) the maximax principle

(c) the Hurwicz principle (with a risk coefficient of 1/2)

(d) the minimax regret principle

5. For the decision in part 1(d), assume that your value function is linear with dollars ($V(\$n) = n$). Then, say which choice is rational (or which choices are rational) according to

   (a) the maximin principle

   (b) the maximax principle

   (c) the Hurwicz principle (with a risk coefficient of 1/2)

   (d) the minimax regret principle

6. Suppose that you are deciding how to allocate society's resources, without knowing who in society you are (you are standing behind the 'veil of ignorance'). There are three possible people you could be: Al, Betty, or Carl, and society's resources consists of hamburgers to be distributed between these three people. There are four distributions between which you are choosing, in the rows of the decision matrix below.

	You are Al	You are Betty	You are Carl
Equality	4 hamburgers	4 hamburgers	4 hamburgers
Unequal in Al's favor	9 hamburgers	1 hamburger	1 hamburger
Unequal in Betty's favor	1 hamburger	9 hamburgers	1 hamburger
Unequal in Carl's favor	1 hamburger	1 hamburger	9 hamburgers

   Suppose that you have a value function according to which $V(n \textit{ hamburgers}) = \sqrt{n}$. That is,

$$V(1 \textit{ hamburger}) = 1$$
$$V(4 \textit{ hamburgers}) = 2$$
$$V(9 \textit{ hamburgers}) = 3$$

   Then, say which choice is rational (or, which choices are rational) according to

   (a) the maximin principle

   (b) the maximax principle

   (c) the Hurwicz principle (with a risk coefficient of 1/2)

   (d) the minimax regret principle

11 | Expected Monetary Values

Goal: understand *expected monetary value* and how it can be used to find the 'fair price' for a game of chance.

The Problem of the Points

Puzzle (The Problem of the Points)**.** *Suppose that Harold and Thelma are playing a game. They flip a coin repeatedly. Each time the coin lands heads, Harold gets a point. Each time the coin lands tails, Thelma gets a point. The first person to reach 3 points wins. Both Harold and Thelma have put in $50, and the winner of the game walks away with the pot—all $100. The coin has been flipped three times so far, and it has landed heads twice and tails once. So Harold has two points and Thelma has one. But then, the game is interrupted by the morality police. Harold and Thelma now need to distribute the $100 between them. What is the most fair way to do so?*

You might think: perhaps Harold should just receive all of the $100—after all, he's ahead, right? Even if that's somewhat persuasive in this version of the case, think about other versions: suppose that Harold and Thelma were playing so that the first to 1,000,000 points won, and Harold was ahead with 2 points to Thelma's 1. True, Harold is *ahead*, but Thelma's surely not *so* far behind that she's not entitled to *any* of the money.

This puzzle has a distinguished lineage. It was debated by mathematicians for almost a century before receiving the currently accepted answer by Blaise Pascal and Pierre de Fermat. Finding the answer to this puzzle opened the door to the theory of rational choice that we will be learning in the second part of this course. So let's start with the puzzle; understanding its solution will help us understand the theory of rational choice.

Figure 11.1: Luca de Pacioli

The puzzle appeared in a 1494 textbook (*Summa de arithmetica, geometrica, proportioni et proportionalità*) by the Italian mathematician Luca Bartolomeo de Pacioli. Pacioli also proposed a solution:

Pacioli's Proposal If the game is interrupted, Harold has h points and Thelma has t points, then Harold should receive $h/h+t$ of the prize money and Thelma should receive $t/h+t$ of the prize money.

In our version of the puzzle, Pacioli said that Harold should get $2/3$ of the \$100 (around \$66) and Thelma should get $1/3$ of the \$100 (around \$33).

But this solution faces two objections. In the first place, consider a case in which Harold and Thelma are playing to 1,000,000 points, and the coin is only flipped once before the game is interrupted. In that case, Pacioli says that *all* of the money should go to Harold. But again, this seems like the wrong result. Thelma still has every chance of winning that \$100; she should be entitled to *some* of it.

In the second place, consider a game in which Harold and Thelma are playing to 1000 points, Harold has 999 and Thelma has 900. Then, Pacioli says that the money should be split *about evenly*—with Harold receiving $999/1899 \approx 53\%$. But surely, in this case, Harold is *way* far ahead of Thelma. This game is not about even; Thelma has basically *no chance* of winning.

Other proposals were made by the mathematicians Nicolo Tartaglia and Gerolamo Cardano. They each proposed that we not look *backwards* at the number of points each player had *already* accumulated, but rather look *forwards* at the number of points each player *needed* to win.[1]

Figure 11.2: Nicolo Tartaglia

Tartaglia's Proposal If Harold and Thelma are playing to n points, Harold has h points and Thelma has t points, and the game is interrupted, then Harold should receive \$50 (his initial stake) *plus* a fraction, $(h-t)/n$, of the other \$50. Likewise, Thelma should receive \$50 + \$50 $\cdot(t-h)/n$.

In our example, $n = 3$, $h = 2$, and $t = 1$, so Tartaglia's proposal says that Harold should get \$50 + \$50 $\cdot 1/3 \approx$ \$66, and Thelma should get \$50 + \$50 $\cdot -1/3 \approx$ \$33, same as Pacioli's proposal.

Tartaglia's method is in *some* sense forward-looking, but it also looks backwards to the past to some extent. Suppose, for instance, that Harold and Thelma are both playing to 100 points, Harold has 99, and Thelma has 98. Then, the forward-looking situation is the same for both Harold and Thelma as it is in our original puzzle: Harold needs 1 point to win, and Thelma needs 2 points to win. But *now*, Tartaglia will give far less money to Harold. If the game is interrupted, then Tartaglia says that Harold deserves \$50 + \$50 $\cdot 99\text{-}98/100$ = \$50.5, and Thelma deserves \$50 + $98\text{-}99/100$ = \$49.5. Tartaglia conjectured that the puzzle was unsolvable, insofar as any proposed answer was contestable. He wrote that "in whatever way the division is made there will be cause for litigation".

[1]Cardano's proposal was that, if Harold and Thelma are playing to n points, Harold has h and Thelma

(a) Blaise Pascal

(b) Pierre de Fermat

Figure 11.3

In 1654, a young Blaise Pascal write to the elder Pierre de Fermat with his own solution to the problem of the points. (He had been introduced to the problem by the Chevalier de Méré, whose eponymous paradox we discussed back in lecture 4.) Pascal's proposal is now widely regarded as the correct resolution of the puzzle.

Pascal's Proposal Let p be the probability that Harold would win the game, were it not interrupted. (So $1-p$ is the probability that Thelma would win the game, were it not interrupted.) If the game is interrupted, Harold should receive \$100p, and Thelma should receive the remainder, \$100$(1-p)$.

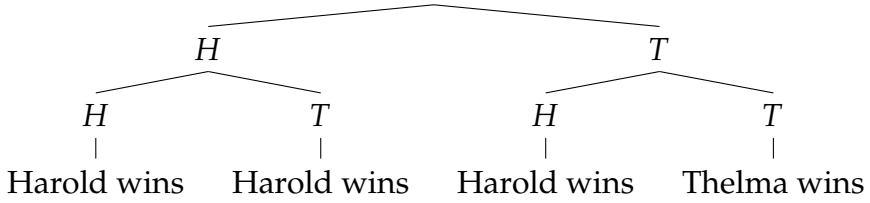
In our example, the probability that Harold goes on to win is $3/4$, and the probability that Thelma goes on to win is $1/4$ (*why?*). So Pascal's proposal says that Harold should get \$75 and Thelma should get \$25.[2]

That's the proposal; why should we take it seriously? Pascal offered the following argument: since they're playing to 3 points, Harold has 2 point, and Thelma has 1, the game will certainly be over after 2 more flips. Let's think about how those flips could go:

---

has t (so that Harold *needs* $n-h$ and Themla *needs* $n-t$), then the \$100 should be split between the players in the ratio $(n-h)\cdot(n-h+1) : (n-t)\cdot(n-t+1)$.

[2]Pascal and Fermat corresponded about how to actually calculate these probabilities in general. If you take a more mathematically rigorous course in probability and statistics, you'll learn about the techniques that Pascal ended up developing to treat that problem (it involves Pascal's triangle); our interest here is on the conceptual innovation that it is the *probability of winning* that should guide the fair division of the money.



In three fourths of these outcomes, Harold wins; whereas, in one fourth of them, Thelma wins. So, Pascal reasoned, when the game is called off, Harold should be entitled to three fourths of the pot.

In his response to Pascal, Fermat offered another justification for the proposal: think about what could happen on the next flip of the coin: it could be that it lands *tails*. If it lands tails, then Harold and Thelma will be tied. By the symmetry of their situation, surely, if the game were to be interrupted at this point, they should each receive $50. However, if the coin lands *heads*, then Harold will instantly win. So Harold is in a position where he has a 50% shot of ending up in a position worth $50, and a 50% shot of ending up in a position worth $100. So he should be entitled to $100/2 + $50/2, or $75.

In one sense, neither of these justifications are particularly strong. For they both simply *presuppose* the very thing Pascal had proposed: namely, that the money should be divided according to the probability that Harold wins. If you didn't *already* find this idea compelling, the arguments Pascal and Fermat gave wouldn't persuade you otherwise. However, the idea *is* incredibly compelling, and most people in fact *were* inclined to agree with Pascal and Fermat's reasoning. For this reason, Pascal's proposal has been largely accepted within probability and statistics, and it forms the basis of the theory of rational choice which we will be learning about in this course (and which is foundational throughout the social and behavioral sciences).

Let's think about whether we can give a better justification of Pascal's proposal. Firstly, let's think about the connection between *probability* and *frequency*. If the probability of the coin landing heads is 1/2, if the outcomes of successive flips are (mutually) independent, and if the coin is flipped a large number of times, then likely the coin will land heads about half of the time. So, suppose that Harold and Thelma were to play out the conclusion of their game 100 times. Since the probability of Harold winning is 3/4, You'd expect that Harold would win about 3/4 of the games. So, on average, he'd walk away with about $75. And, on average, Thelma would walk away with about $25. So another way of thinking about Pascal's proposal is like this: if the game is called off, then Harold is entitled to the amount he *would* win, *on average*, if the conclusion of the game were to be played out a large number of times. If we think that this is fair, then we have reason to accept Pascal's proposal.

Expected Monetary Value

Here's another way of putting Pascal's proposal: if the game is called off, then each player should be compensated with the *expected monetary value* of their position in the game. The *expected monetary value* of Harold's position is the amount he would expect to win, *on average*, were the game to be played out a large number of times. Or, putting the same point a different way: his expected monetary value is the probability-weighted average of his possible winnings.

Expected Monetary Value The *expected monetary value* of a game of chance is the probability-weighted average of its possible winnings.

$$\sum_x \Pr(\textit{you win } \$x) \cdot x$$

(Recall, we're using '$\sum_i$' as a shorthand for 'take the sum over all values of i'. So the above expression says: take each dollar amount you might win, multiply it by the probability that you win that amount, and then add up all of those products. The result is the expected monetary value of the game of chance.)

Example 40. *We will roll a three sided die twice, and you will win $\$1$ every time the die lands on 3. What is the expected monetary value of this game*?

There are three possible outcomes of this game: you could win nothing (if the die never lands on 3). You could win \$1 (if the coin lands on 3 once). And you could win \$2 (if the coin lands on 3 twice). The probability that the die never lands on 3 is $4/9$. The probability that the die lands on 3 once is $4/9$. And the probability that the die lands on 3 twice is $1/9$. So the expected monetary value of the game is

$$\begin{aligned}
&\Pr(\textit{you win } \$0) \cdot 0 \;+\; \Pr(\textit{you win } \$1) \cdot 1 \;+\; \Pr(\textit{you win } \$2) \cdot 2 \\
=&4/9 \cdot 0 + 4/9 \cdot 1 + 1/9 \cdot 2 \\
=&0 + 4/9 + 2/9 \\
=&6/9 = 2/3
\end{aligned}$$

Exercise 36. *We will flip a fair coin twice, and you will win $\$1$ every time the coin lands on a heads. What is the expected monetary value of this game*?

A natural generalization of Pascal's proposal says that the expected monetary value of a game should be your guide to whether or not that game is fair. Suppose the price

to pay the game is *greater than* the game's expected monetary value. Then, if the game were played a large number of times, you would lose money on average. In that case, the game is *unfair*—it is biased against you. Suppose the price to play the game is *less than* the game's expected monetary value. Then, if the game were played a large number of times, you would gain money on average. In that case, the game is unfair in a different way—it is biased in your favor. Finally, suppose the price to play the game is *equal to* its expected monetary value. Then, if the game were played a large number of times, you would break even on average. In that case, the game is *fair*.

Corresponding to Pascal's proposal for the problem of the points, then, there's a corresponding proposal for evaluating whether or not to play games of chance:

> **Maximize Expected Monetary Value** If the expected monetary value of a game of chance is greater than or equal to the price to play, then it is rational to play. If, however, the expected monetary value of a game of chance is less than the price to play, then it is irrational to play.

Example 41. *The slot machine costs $1 to play. It has a 1/25 chance of paying out $10, and a 24/25 chance of paying out nothing. According to* Maximize Expected Monetary Value*, is it rational to play this game*?

We can start by calculating the expected monetary value of the game, and then checking to see whether this is greater than or equal to $1. The expected monetary value is

$$\begin{aligned} &\Pr(\textit{you win \$10}) \cdot 10 + \Pr(\textit{you win nothing}) \cdot 0 \\ =&{}^{1}/_{25} \cdot 10 + {}^{24}/_{25} \cdot 0 \\ =&{}^{10}/_{25} = {}^{2}/_{5} \end{aligned}$$

Since $^2/_5$ is less than 1, **Maximize Expected Monetary Value** says that it is irrational to play this game.

Example 42. *The probability of winning the jackpot in California's lottery is* $1/300{,}000{,}000$.[3] *Each ticket costs $5. According to* Maximize Expected Monetary Value*, how high does the jackpot have to be in order for it to be rational to play*?

Let's call the jackpot 'j'. Then, the expected monetary value of a lottery ticket is greater than $5 exactly when

$$\begin{aligned} \Pr(\textit{you win the jackpot}) \cdot j + \Pr(\textit{you lose}) \cdot 0 &\geqslant 5 \\ {}^{1}/_{300{,}000{,}000} \cdot j + {}^{299{,}999{,}999}/_{300{,}000{,}000} \cdot 0 &\geqslant 5 \end{aligned}$$

[3] The actual odds of winning the jackpot are slightly better: 1 in 290,472,336.

$$\begin{aligned} j/_{300,000,000} &\geqslant 5 \\ j &\geqslant 1,500,000,000 \end{aligned}$$

So, in order for it to be rational to play, the jackpot must be at least 1.5 billion dollars.

Exercise 37. *We will flip a fair coin three times. If the coin lands heads on the first flip, then the game is over and you will win $2. If the coin lands tails on the first flip and heads on the second, then the game is over and you win $4. If the coin lands tails on the first two flips but heads on the third, then the game is over and you win $8. If the coin lands tails three times in a row, then the game is over and you win nothing. According to* Maximize Exepected Monetary Value*, what's a fair price to play this game*?

Expected Monetary Value and Decision Matrices

We can integrate the theory *Maximize Expected Monetary Value* with the decision matrices we encountered earlier.

For instance, suppose that Harold and Thelma are playing the coin-tossing game—each time the coin lands heads, Harold gets a point, each time the coin lands tails, Thelma gets a point, and the first to three points wins. Harold is ahead with 2 points to Thelma's 1. And Thelma—having read Pacioli's textbook—offers Harold the following deal: if we stop playing right now, I'll give you $67 of the prize money, and I'll take $33.

Harold now faces a decision: should he accept Thelma's offer or not? Let's make the decision matrix: there are three relevant states: it could be that the next flip will land heads; it could be that the next flip will land tails but the one after that will land heads; or it could be that the next two flips will both land tails. There are two available acts: Harold will either play the game out or else he'll take the deal and walk away. And there are three possible outcomes: Harold gets $100, Harold gets $0, and Harold gets $67. Here's the decision matrix, which tells us which outcome Harold will get, depending upon which act he chooses and which state he's in.

	Heads	Tails, then heads	Tails, then tails
Harold plays on	$100	$100	$0
Harold takes the deal	$67	$67	$67

The probability of the first state is $1/2$, the probability of the second state is $1/4$, and the probability of the final state is $1/4$. So *Maximize Expected Monetary Value* says that we should calculate the expected monetary value of each of Harold's options and then choose whichever is greater. In this case, we get that the expected monetary value of playing on is

$$\begin{aligned} &\Pr(\textit{Heads}) \cdot 100 + \Pr(\textit{Tails then heads}) \cdot 100 + \Pr(\textit{Tails then tails}) \cdot 0 \\ =&1/2 \cdot 100 + 1/4 \cdot 100 + 1/4 \cdot 0 \\ =&50 + 25 + 0 \end{aligned}$$

$=75$

whereas the expected monetary value of taking the deal is only

$$\begin{aligned}
&\Pr(\textit{Heads}) \cdot 67 + \Pr(\textit{Tails then heads}) \cdot 67 + \Pr(\textit{Tails then tails}) \cdot 67 \\
=&1/2 \cdot 67 + 1/4 \cdot 67 + 1/4 \cdot 67 \\
=&67 \cdot (1/2 + 1/4 + 1/4) \\
=&67 \cdot 1 = 67
\end{aligned}$$

Since the expected monetary value of playing on is greater than the expected monetary value of taking the deal, *Maximize Expected Monetary Value* says that it would be irrational for Harold to take the deal.

This example illustrates another way we can calculate expected monetary values, given a decision matrix together with a probability distribution over states. First, some notation: let '$o_{A,S}$' be the outcome you get if you choose act A in the state S. And let '$V_\$(o)$' be the monetary value of the outcome o. Then,

Expected Monetary Value The *expected monetary value* of an act A is

$$\sum_S \Pr(S) \cdot V_\$(o_{A,S})$$

12 | Expected Utility

Goal: understand *utility* and its relationship to monetary value, and understand how the concept of utility can be used to dissolve the 'Saint Petersburg Paradox'

Nicolaus Bernoulli was born thirty years after the correspondence between Pascal and Fermat, in 1687. In 1713, he was living in Saint Petersburg when he thought about the following game of chance.

Puzzle (The Saint Petersburg Paradox)**.** *We will flip a fair coin* until it lands heads. *If the first heads comes on the first flip, then you win* $\$2$. *If the first heads comes on the second flip, you win* $\$4$. *If the first heads comes on the third flip, then you win* $\$8$. *In general, if the first heads comes in the* n*th flip, then you win* $\$2^n$.

According to *Maximixe Expected Monetary Value*, what's a fair price to play this game?

$$\begin{aligned}\sum_x \Pr(\textit{you win } \$x) \cdot x &= 1/2 \cdot 2 \;+\; 1/4 \cdot 4 \;+\; 1/8 \cdot 8 \;+\; 1/16 \cdot 16 \;+\; \cdots + 1/2^n \cdot 2^n \;+\; \cdots \\ &= 1 + 1 + 1 + 1 + \cdots + 1 + \cdots \\ &= \infty\end{aligned}$$

So according to *Maximized Expected Monetary Value*, you should be willing to pay *any amount* of money to play this game—indeed, you should be willing to hand over an *infinite* number of dollars to play.

There are at least two odd things about this. Firstly, almost nobody is willing to pay more than $\$30$ to play this game, even after carrying out the calculation. Are we all making a systematic error? Secondly, this game is guaranteed to pay out a *finite* number of dollars. The amount it will pay out is *unbounded* (meaning that, for any number N, there's some probability that the game will pay out more than $\$N$). But there's no possibility in which it pays out $\$\infty$. So if you were to hand over $\$\infty$, your net winnings would be $\$2^n - \infty$, for some n. But that's an *infinite loss*.

But then, it looks like the principle *Maximize Expected Monetary Value* is in conflict with the principle of *Dominance* we discussed earlier. For suppose you face two options: pay

(a) Nicolaus Bernoulli

(b) Daniel Bernoulli

Figure 12.1

$\$\infty$ to play this game or do not play. *Maximize Expected Monetary Value* says that you are permitted to pay, since $\$\infty$ is a fair price. But notice that this choice is *dominated* by not playing.

	H_1	$T_1 \& H_2$	$T_1 \& T_2 \& H_3$	$\cdots$	$T_1 \& \cdots \& T_{n-1} \& H_n$	$\cdots$
Pay $\$\infty$ to play	$\$2 - \infty$	$\$4 - \infty$	$\$8 - \infty$	$\cdots$	$\$2^n - \infty$	$\cdots$
Do not play	\$0	\$0	\$0	$\cdots$	\$0	$\cdots$

So it seems that the principle *Maximize Expected Monetary Value* is giving us some very bad advice about this game of chance.

Utility

Nicolaus Bernoulli wrote to his brother, Daniel Bernoulli, about the problem. And Daniel came up with an ingenious solution. His thought was this: what you really *care about* isn't *money*. Rather, what you care about is your own *happiness*. And your happiness does not increase linearly with dollars. When you're very poor, an additional \$100 can bring you a large increase in happiness. But when you're very rich, \$100 brings you almost no increase in happiness. It's hardly worth Jeff Bezos' time to lean over and pick up a \$100 bill. But much poorer people will put in a full days' work for \$100.

Daniel Bernoulli introduced the notion of *utility* to capture this idea. As Bernoulli was thinking about it, *utility* was a measure of *happiness*. And he thought that each *new* dollar gave you less happiness than the last one. We can represent this mathematically with a function, U, which you hand a dollar amount, and which hands you back some number representing how much happiness that dollar amount would bring you.

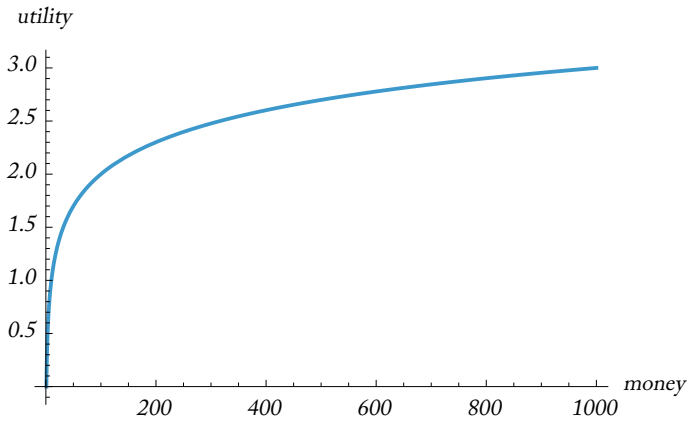


Figure 12.2: Along the x-axis, money (measured in dollars). Along the y-axis, utility. The blue curve shows the utility function $U(\$x) = \log(x)$.

Bernoulli conjectured that utility was *logarithmic* in money. That is, $U(\$x) = \log(x)$. (Let's take this to be the logarithm base 10; so $\log(x)$ is the exponent y such that $10^y = x$. For instance, $\log(1) = 0$, $\log(10) = 1$, $\log(100) = 2$, and $\log(1000) = 3$.) This gives us the function shown in figure 12.2. Notice that going from \$1 to \$100 gives a *large* boost in utility (from zero to two), whereas going from \$100 to \$1000 gives you a smaller boost in utility (from two to three). So Bernoulli's conjecture was that increasing amounts of money brought less and less happiness.

Expected Utility

Bernoulli suggested that, instead of evaluating games of chance in terms of their *expected monetary value*, as Pascal and Fermat had suggested, we should instead evaluate them in terms of their *expected utility*.

Expected Utility The *expected utility* of a game of chance is the probability-weighted average of *the utility of* its possible winnings.

$$\sum_x \Pr(\textit{you win } \$x) \cdot U(\$x)$$

(Recall, we're using '$\sum_x$' as a shorthand for 'take the sum over all values of x'. So the above expression says: take each dollar amount you might win, figure out the *utility* of winning that dollar amount, multiply that by the probability of you winning that dollar amount, and then add up all of those products. The result is the expected utility of that game of chance.)

In the Saint Petersburg Paradox, if we suppose that $U(\$x) = \log(x)$, then the expected monetary value of the game will be equal to about 0.6, which is the utility you'd get from

about \$4.[1] And this is about what people are willing to pay to play the game!

Example 43. *Suppose we're going to flip a fair coin, and I offer you a gamble that will pay out \$100 if the coin lands heads, and nothing if the coin lands tails. Suppose that your utilities are given by the function $U(\$x) = \sqrt{x}$. What is the expected utility of this gamble?*

The expected utility of the gamble is given by

$$\begin{aligned}\Pr(\textit{you win \$100}) \cdot U(\$100) \;+\; \Pr(\textit{you win \$0}) \cdot U(\$0) &= \Pr(\textit{heads}) \cdot \sqrt{100} \;+\; \Pr(\textit{tails}) \cdot \sqrt{0} \\ &= 1/2 \cdot 10 + 1/2 \cdot 0 \\ &= 5\end{aligned}$$

This means that the expected utility of the gamble is equal to the utility of \$25, since $\sqrt{25} = 5$.

Exercise 38. *Suppose we're going to roll a fair 6 sided die, and I offer you a gamble that will pay out \$100 if the die lands on 1 or 2, \$64 if the die lands on 3 or 4, and \$0 if the die lands on 5 or 6. And suppose that your utilities are given by $U(\$x) = \sqrt{x}$. What is the expected utility of this gamble?*

Bernoulli's suggestion gives us a different way of deciding whether or not to play a game of chance:

> **Maximize Expected Utility** If the expected utility of a game of chance is greater than or equal to the utility of the price to play, then it is rational to play. If, however, the expected utility of the game is less than the utility of the price to play, then it is irrational to play.

Example 44. *The slot machine costs \$1 to play and has a 1/25 chance of paying out \$10 (and a 24/25 chance of paying out nothing). Suppose that your utility for winning \$x is x^2. According to* Maximize Expected Utility, *is it rational to play this game?*

[1] It doesn't matter where this comes from, but if you're curious, the expected utility of the Saint Petersburg game will be

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \cdot \log(2^n) = \sum_{n=1}^{\infty} \frac{n}{2^n} \cdot \log(2) = \log(2) \cdot \sum_{n=1}^{\infty} \frac{n}{2^n} = \log(2) \cdot 2$$

Here, we used the fact that $\sum_{n=1}^{\infty} n/2^n$ is equal to 2. If you take a course in calculus, you'll learn how to find the value of infinite sums like this, but I'll just take it for granted. And since $\log(2) \approx 0.3$, we get that the expected utility of the game is about 0.6.

We can calculate the expected utility of the game, and then check to see whether this is greater than, less than, or equal to the utility of the price to play.

The expected utility of the game is

$$\begin{aligned}\Pr(\textit{you win \$10}) \cdot U(\$10) \;+\; \Pr(\textit{you win \$0}) \cdot U(\$0) &= {}^{1}/_{25} \cdot 10^2 \;+\; {}^{24}/_{25} \cdot 0^2 \\ &= {}^{1}/_{25} \cdot 100 \;+\; {}^{24}/_{25} \cdot 0 \\ &= 4\end{aligned}$$

whereas the utility of the price to play is $U(\$1) = 1^2 = 1$. Since the expected utility of the game is *greater* than the utility of the price to play, *Maximize Expected Utility* says that it is rational for you to play.

Exercise 39. *We will flip a fair coin twice. I offer you the following gamble: if the coin lands heads twice in a row, then you will win \$2. Otherwise, you won't win anything. Your utilities are given by $U(\$x) = 2x$. According to* Maximize Expected Utility*, what is a fair price for you to pay for this gamble?*

Expected Utility and Decision Matrices

We can integrate the theory *Maximize Expected Utility* with the decision matrices we encountered earlier. For instance, we can represent the decision you face in example 2 with the following decision matrix:

	Machine pays out	Machine doesn't pay out
Play	*win \$9*	*lose \$1*
Do not play	*win \$0*	*win \$0*

And we can calculate the *expected utility* of playing by multiplying the probability of each column state by the utility of the corresponding outcome the 'play' row:

$$\Pr(\textit{Machine pays out}) \cdot U(\textit{win \$9}) \;+\; \Pr(\textit{Machine doesn't pay out}) \cdot U(\textit{lose \$1})$$

Since we're assuming that your utility for winning \$x is x^2, and your utility for losing $\$x$ is $-x^2$, this is

$$\begin{aligned}&= {}^{1}/_{25} \cdot 9^2 - {}^{24}/_{25} \cdot 1^2 \\ &= {}^{81}/_{25} - {}^{24}/_{25} \\ &= {}^{57}/_{25}\end{aligned}$$

whereas the expected utility of not playing is

$$\begin{aligned}&\Pr(\textit{Machine pays out}) \cdot U(\textit{win \$0}) \;+\; \Pr(\textit{Machine doesn't pay out}) \cdot U(\textit{win \$0}) \\ &= {}^{1}/_{25} \cdot 0^2 + {}^{24}/_{25} \cdot 0^2\end{aligned}$$

$$= 0 + 0$$
$$= 0$$

This example illustrates another way we can calculate expected utility, given a decision matrix together with a probability distribution over states. Some notation: let '$o_{A,S}$' be the outcome you get if you choose the act A in the state S. And let '$U(o)$' be the utility of the outcome o. Then,

Expected Utility The *expected utility* of an act A is

$$\sum_S \Pr(S) \cdot U(o_{A,S})$$

Exercise 40. *We will flip a fair coin twice. I offer you the following gamble: if the coin lands heads twice in a row, then you will win* **$2**. *Otherwise, you won't win anything. Your utilities are given by* $U(\$x) = 2x$. *Create a decision matrix and say which choice or choices are rational, according to* Maximize Expected Utility.

1. You've won a cash prize of $74, but it's on the other side of town, and you can only collect it if you arrive before 5pm. There's a 65% probability of rain. If it rains, then the roads will be closed and you won't be able to get to the prize before 5pm. If it doesn't rain, then the roads will be open and you will be able to collect the prize. The trip will cost you $25 in gas, whether or not it rains. You face a decision of whether to get in the car and drive to the other side of town in an effort to collect your prize or whether to instead stay at home. [140pts]

   (a) Make a decision matrix for for this decision (assume that the only thing you care about is money) [20pts]

   (b) What is the expected monetary value of driving? [20pts]

   (c) What is the expected monetary value of staying at home? [20pts]

   (d) According to *Maximize Expected Monetary Value*, which choice is rational? [20pts]

   (e) Suppose that your utility for money is given by the function $U(\$x) = \sqrt{x}$ (if you're gaining money, the utility is positive; if you're losing money, the utility is negative). Then:

      i. What is the expected utility of driving? [20pts]

      ii. What is the expected utility of staying at home? [20pts]

      iii. According to *Maximize Expected Utility*, which choice is rational?[20pts]

2. Suppose that you are deciding how to allocate society's resources, without knowing who in society you are (you are standing behind the 'veil of ignorance'). There are three possible people who you could be: Al, Betty, and Carl. Society's resources consist of hamburgers to be distributed between these three people. There are four distributions between which you are choosing, in the rows of the decision matrix below.

	You are Al	You are Betty	You are Carl
Equality	4 hamburgers	4 hamburgers	4 hamburgers
Unequal in Al's favor	9 hamburgers	1 hamburger	1 hamburger
Unequal in Betty's favor	1 hamburger	9 hamburgers	1 hamburger
Unequal in Carl's favor	1 hamburger	1 hamburger	9 hamburgers

   Suppose that your utilities for hamburgers are given by the function $U(n \textit{ hamburgers}) = \sqrt{n}$. And suppose that your probabilities are divided evenly between these states—that is, your probability that you are Al is $1/3$, your probability that you are Betty is $1/3$, and your probability that you are Carl is $1/3$. Then, which choice or choices maximize expected utility? [60pts]

3. Suppose we're going to flip a fair coin twice, and I offer you a gamble which will pay out \$74 if the coin lands heads on both flips and nothing otherwise. In exchange for this gamble, I ask for \$25. Your choice is whether or not to pay to play.

   (a) Make a decision matrix for this decision. [20pts]

   (b) Suppose that your utilities are given by the function $U_1(\$x) = \sqrt{x}$ (positive if money is gained, negative if money is lost). Which choice or choices maximize expected utility? [20pts]

   (c) Suppose that your utilities are given by the function $U_2(\$x) = x^2$ (positive if money is gained, negative if money is lost). Which choice or choices maximize expected utility? [20pts]

   (d) Are U_1 and U_2 ordinally equivalent? Why or why not? [20pt]

   (e) Considering your answers to parts (b), (c), and (d), say whether we should use the decision rule *Maximize Expected Utility*, if we are only able to measure your utilities on an *ordinal* scale. [20pts]

13 | Measuring Utility

Goal: Understand how utility can be measured, and understand the assumptions we rely on when making these measurements.

Last time, we saw Daniel Bernoulli's solution to the Saint Petersburg paradox. He suggested that you shouldn't evaluate gambles in terms of their expected *monetary value*. Instead, he suggested that you evaluate them in terms of their expected *utility*.

We introduced this notion of *utility*, and used it to solve some problems. But what kind of thing is utility? We know how to measure *money*, and we know what the *units* of money are—dollar, euros, *etc*. But how do we measure utility? And what are the *units* of utility?

Philosophers and economists sometimes use the term 'utile' for a unit of utility, but introducing this name raises more questions than it answers. What is a utile? How is it measured?

Measuring Utility

Frank Ramsey had an idea of how utility was to be measured. (Actually, his idea was much grander than this—he showed how we could *simultaneously* measure *both* your probabilities *and* your utilities. But we're going to try to keep things simple here and just focus on the measurement of utility.)

Figure 13.1: Frank Ramsey

To start, we can get *ordinal* information about your utilities, about which things you prefer to which other things, by just giving you a choice between them and seeing which ones you choose. For instance, suppose we give you a choice between coffee and tea, and you choose coffee. Then, we can infer that your utility for coffee is *greater than* your utility for tea: $U(\textit{tea}) < U(\textit{coffee})$. And suppose we give you a choice between soda and tea, and you choose tea. Then, we can infer that $U(\textit{soda}) < U(\textit{tea})$. Putting these two choices together, we can infer that $U(\textit{soda}) < U(\textit{tea}) < U(\textit{coffee})$.

This is all the information we need to construct an *ordinal* utility function. (Recall, when we call a utility function *ordinal*, we are saying: 'I don't care about the particular numbers

I'm using here; any other numbers would work just as well, so long as they keep outcomes in the same order'.) However, in order to meaningfully evaluate gambles in terms of their *expected utility*, we will need more than an ordinal utility function. We will need a *cardinal* utility function. Let's spend a bit of time on this point, before coming back to Ramsey's idea for measuring utility.

Expected Utility Only Makes Sense with Cardinal Utility Functions. Why? Well, consider the two *ordinally* equivalent utility functions, U_1 and U_2, shown in table 13.1.

	U_1	U_2
coffee	1000	2
tea	0	0
soda	−2	−1000

Table 13.1: The numbers beneath a utility function are the numbers it assigns to the outcome in that row.

And consider a gamble which will give you coffee with probability 1/2 and soda with probability 1/2. Is the expected utility of this gamble greater than or less than the utility of tea? If we use U_1, then the expected utility of the gamble will be

$$\begin{aligned}\Pr(\textit{soda}) \cdot U_1(\textit{soda}) + \Pr(\textit{coffee}) \cdot U_1(\textit{coffee}) &= 1/2 \cdot 1000 + 1/2 \cdot (-2) \\ &= 500 - 1 \\ &= 499\end{aligned}$$

which greater than the utility of the tea, which is zero. However, if you use U_2, then the expected utility of the gamble will be

$$\begin{aligned}\Pr(\textit{soda}) \cdot U_2(\textit{soda}) + \Pr(\textit{coffee}) \cdot U_2(\textit{coffee}) &= 1/2 \cdot 2 + 1/2 \cdot (-1000) \\ &= 1 - 500 \\ &= -499\end{aligned}$$

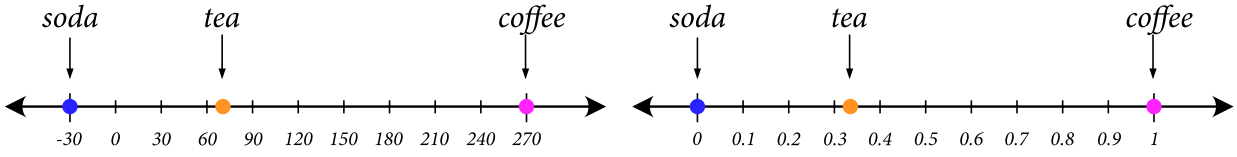
which is less than the utility of the tea, which is zero.

The lesson: if we only have *ordinal* information about your utilities, then we won't be able to evaluate gambles in terms of their expected utility. With only ordinal information about utilities, we won't be able to say whether the gamble between coffee and soda is better or worse than the tea.

It turns out to be true that a *cardinal* utility function *will be* enough to evaluate gambles in terms of the expected utility. You can show this in general, but let's just see how it works with a particular example. Consider the *cardinally* equivalent utility functions U_3 and U_4, shown in table 13.2 and figure 13.2.

	U_3	U_4
coffee	270	1
tea	70	$1/3$
soda	-30	0

Table 13.2: The numbers beneath a utility function are the numbers it assigns to the outcome in that row.



(a) Utility as measured by U_3

(b) Utility as measured by U_4

Figure 13.2

Exercise 41. *Show that U_3 and U_4 are cardinally equivalent by finding a positive number $\alpha > 0$ and a number β which makes it so that, for every outcome o, $U_3(o) = \alpha \cdot U_4(o) + \beta$.*

Now, consider a gamble which gives you the coffee with a probability of $1/3$ and the soda with a probability of $2/3$. What is the expected U_3 utility of this gamble?

$$\begin{aligned} \Pr(\textit{soda}) \cdot U_3(\textit{soda}) + \Pr(\textit{coffee}) \cdot U_3(\textit{coffee}) &= 2/3 \cdot (-30) + 1/3 \cdot (270) \\ &= -20 + 90 \\ &= 70 \end{aligned}$$

So the expected U_3 utility of the gamble is exactly the U_3 utility of the tea! What about if we instead used the (cardinally equivalent) utility function U_4? Then, the expected U_4 utility of the gamble is

$$\begin{aligned} \Pr(\textit{soda}) \cdot U_4(\textit{soda}) + \Pr(\textit{coffee}) \cdot U_4(\textit{coffee}) &= 2/3 \cdot (0) + 1/3 \cdot (1) \\ &= 1/3 \end{aligned}$$

So the expected U_4 utility of the gamble is *also* exactly the U_4 utility of the tea! And that's not just how it worked out in this particular case. That's how it will work out in general. Changing between cardinally equivalent utility functions won't make any difference with respect to which gambles have greater expected utilities than which others.

Utility from Preferences Between Gambles. Ramsey's basic thought was this: if you are indifferent between tea and a gamble with a $1/3$ probability of getting you coffee, and a $2/3$ probability of getting you soda, then your *utility* for tea must be one third of the way

between your utility for coffee and your utility for tea. So if you're willing to exchange tea for the gamble, and you're willing to exchange the gamble for the tea, then this tells us enough to quantify your utilities for soda, tea, and coffee on a cardinal scale.

Suppose that there's some outcome which is your *most* preferred outcome. Call that outcome 'b' for 'best' And suppose that there's some outcome which is your *least* preferred outcome. Call that outcome 'w', for 'worst'. Then, we can create a cardinal utility function, U, for which $U(b) = 1$ and $U(w) = 0$. This is sometimes called the 'zero-one' utility function, for obvious reasons.

Let's suppose we want to measure your utilities on the zero-one scale. And we're interested in finding out what your utility for a Caribbean vacation is. Then, Ramsey's procedure goes like this. First, given you a choice between the vacation and a gamble that gives you the best outcome b with probability $1/4$ and the worst outcome w with probability $3/4$. Suppose that, presented with this choice, you tell us that you prefer the gamble. Then, we know that your utility for the vacation (on the zero-one scale) has to be less than $1/4$. Why? Because you prefer the gamble, the expected utility of the gamble must be greater than the utility of the vacation. So we must have

$$\begin{aligned} U(\textit{vacation}) &< \textit{expected utility of gamble 1} \\ &= 1/4 \cdot 1 + 3/4 \cdot 0 \\ &= 1/4 \end{aligned}$$

So $U(\textit{vacation}) < 1/4$. Suppose we next offer you a choice between the vacation and a gamble which gets you b with a probability of $1/8$ and w with a probability of $7/8$. And suppose, presented with this choice, you'd rather have the vacation. Then, we know that your utility for the vacation (on the zero-one scale) has to be greater than $1/8$. Why? Because you prefer the vacation, so the utility of the vacation must be greater than the expected utility of the gamble. So we must have

$$\begin{aligned} U(\textit{vacation}) &> \textit{expected utility of gamble 2} \\ &= 1/8 \cdot 1 + 7/8 \cdot 0 \\ &= 1/8 \end{aligned}$$

So $U(\textit{vacation}) > 1/8$. So we now know that your utility for the vacation is strictly between $1/8$ and $1/4$. Perhaps we next offer you a choice between the vacation and a gamble which gets you b with probability $3/16$ and w with probability $13/16$. And at this point, you tell us that you are *indifferent* between the vacation and the gamble—you could go either way. Then, we would know that your utility for the vacation must be equal to $3/16$. Why? Because you are indifferent between the vacation and this new gamble, so the expected utility of the gamble must be *equal* to the utility of the vacation. So we must have

$$U(\textit{vacation}) = \textit{expected utility of gamble 3}$$

$$= 3/16 \cdot 1 + 13/16 \cdot 0$$
$$= 3/16$$

Example 45. *Sabeen prefers lavender ice cream to honeycomb ice cream. And, when she's offered a choice between macha ice cream and a gamble which gets her lavender ice cream with 2/3 probability and honeycomb ice cream with 1/3 probability, she says that she could go either way (she's indifferent between the macha ice cream and the gamble). Give a (cardinal) utility function for Sabeen's preferences between these three ice cream flavors.*

Let's use 'l' for the outcome in which Sabeen gets lavender ice cream, 'h' for the outcome in which she gets honeycomb, and 'm' for the outcome in which she gets macha ice cream. We could assign any numbers we want to $U(l)$ and $U(h)$, but it will make things easier if we use $U(l) = 1$ and $U(h) = 0$. With these choices made, we know that Sabeen is indifferent between m and a gamble which gets her a $2/3$ chance of l and a $1/3$ chance of h. So we must have

$$\begin{aligned} U(m) &= 2/3 \cdot U(l) + 1/3 \cdot U(h) \\ &= 2/3 \cdot 1 + 1/3 \cdot 0 \\ &= 2/3 \end{aligned}$$

So this is one utility function for Sabeen's preferences between ice cream flavors: $U(h) = 0$, $U(m) = 2/3$, and $U(l) = 1$.

But we didn't have to start with the values 0 and 1 for the utility of lavender and honeycomb. We could instead have started with the values $2/3$ and 3. Then, we'd be using a *different* (cardinal) utility function—call it 'U^*', for which $U^*(l) = 3$ and $U^*(h) = 1$. We would then have that

$$\begin{aligned} U^*(m) &= 2/3 \cdot U^*(l) + 1/3 \cdot U^*(h) \\ &= 2/3 \cdot 3 + 1/3 \cdot 1 \\ &= 2 + 1/3 \\ &= 5/3 \end{aligned}$$

Exercise 42. *Suppose you have a cardinal utility function for Sabeen, U, and you already know that, in this scale, her utility for lavender ice cream is 1, her utility for macha ice cream is 2/3, and her utility for honeycomb ice cream is 0. You then learn that, when Sabeen is offered a choice between cherry ice cream and a gamble which gets her macha ice cream with a probability of 1/2 and honeycomb ice cream with a probability of 1/2, she is indifferent, and could go either way. What is Sabeen's utility for cherry ice cream (as measured by the function U)?*

Part III

The Philosophy of Decision Theory

14 | Risk and Ambiguity Aversion

Goal: Understand two objections to *Maximize Expected Utility*: 1) it says that a certain kind of *risk-aversion* is irrational and 2) it says that *ambiguity-aversion* is irrational.

Last week, we saw a way of integrating the notion of *expected utility* into decision theory. If we have a decision matrix, and we have a probability distribution over the *states* in that decision matrix, then we can define the *expected utility* of an act, A (which lies along the rows of the matrix), like so:

Expected Utility The *expected utility* of an act A is

$$EU(A) \stackrel{\text{def}}{=} \sum_S \Pr(S) \cdot U(o_{A,S})$$

Here, the sum is being taken over all of the states in the columns of the decision matrix, and '$o_{A,S}$' is the outcome that the act A leads to in the state S.

Given this definition, we have the following rule for making decisions:

Maximize Expected Utility An act, A, is a rational choice if and only if there's no other act available which has a greater expected utility than A does.

That is, A is a rational choice iff, for every available act B, $EU(A) \geqslant EU(B)$.

Today, we're going to look at a consequence of this theory, and discuss two reasons to reject this consequence. (Of course, a reason to reject this consequence will *also* be a reason to reject the theory.) The first objection will be that expected utility maximization rules out a pervasive kind of *risk-aversion*. The second objection will be that expected utility maximization rules out a certain kind of *ambiguity-aversion* (we'll come back to what that means later on).

But before getting to the objections, we will start by discussing one way in which expected utility maximizers *can* be risk-averse.

Risk-Aversion for Expected Utility Maximizers

Consider a decision between a) a gamble which gets you \$10 with 50% probability and \$0 with 50% probability, and b) a guaranteed \$5. (Notice that \$5 is the expected *monetary* value of the gamble.) Expected Utility maximizers will make this choice by figuring out whether the utility of \$5 is greater than, less than, or equal to the expected utility of the gamble.

Whether they will prefer the guaranteed \$5 or the gamble will vary, depending upon the shape of their utilities (as a function of money). Consider, for instance, the graphs in figure 14.1. The *expected* utility of the gamble will be the midpoint between $U(\$0)$ and $U(\$10)$ on the y-axis (shown in yellow), whereas the utility of the \$5 is shown in purple. If your utilities are a *concave* function of money—like in figure 14.1a—then you will get more utility from the guaranteed \$5 than you will get from the gamble. Whereas, if your utilities are a *convex* function of money—like in figure 14.1b—then you will get more utility from the gamble than you will get from the guaranteed \$5.

More generally, if your utilities are concave function of money, then you will prefer a gamble's expected monetary value to the gamble. So you'll need extra enticement in order to purchase a gamble. This is a form of *risk aversion*. You're averse to risk in the sense that you're willing to sacrifice expected money just in order to have less risk. And, in general, if your utilities are convex function of money, then you will prefer a gamble to the gamble's expected monetary value. So you'll need *less* enticement in order to purchase a gamble. This is a form of *risk seeking*. You seek out risk in the sense that you're willing to sacrifice expected money just in order to have more risk.

Exercise 43. *Consider a gamble which gives you \$16 with probability 5/12 and \$4 with probability 7/12.*

(*a*) *What is the expected monetary value of this gamble*?

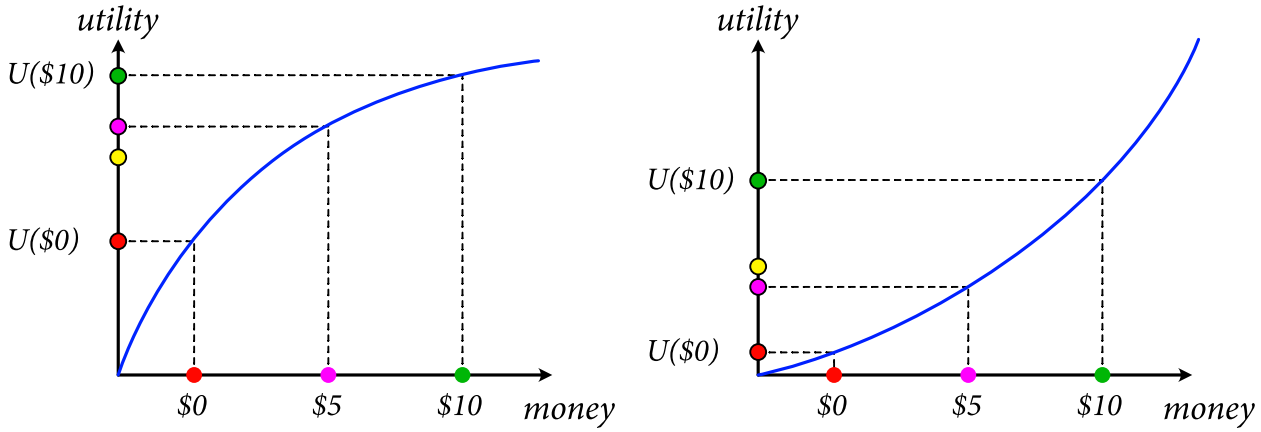
(*b*) *Suppose that your utilities* (*as a function of dollars*) *are given by* $U(\$x) = \sqrt{x}$. *Then, what is the expected utility of the gamble*?

(*c*) *What is the expected utility of the gamble's expected monetary value*?

(*d*) *Given these utilities, if you were given a choice between the gamble and its expected monetary value, which would you prefer*?

(*e*) *Given these utilities, are you risk-averse, risk-seeking, or neither*?

So there's a form of risk-aversion which is perfectly rational, according to *Maximize Expected Utility*. But it turns out that there is another form of risk-aversion which *Maximize Expected Utility* deems to be *irrational*. Before getting to that, let's explore an important consequence of *Maximize Expected Utility*, which we'll call 'The Sure Thing Principle'.



(a) Risk-averse Utilities

(b) Risk-seeking utilities

Figure 14.1: If your utility is a concave function of money (as in figure 14.1a), then you will prefer a guaranteed $5 to a gamble which gets you $10 with probability ½ and $0 with probability ½. However, if your utility is a convex function of money (as in figure 14.1b), then you will prefer the gamble to a guaranteed $5. (Along the y-axis of each figure, the expected utility of the gamble is shown in yellow, whereas the utility of a guaranteed $5 is shown in purple.)

The Sure Thing Principle

Consider the following decision: there are two available acts, A and B, and there are two possible states, S and T. In state T, both A and B lead to the same outcome, with a utility of x. And in state S, A leads to an outcome with utility 100, whereas B leads to an outcome with utility 50.

	State S	State T
Act A	100	x
Act B	50	x

Question: according to *Maximize Expected Utility*, do we need to know the value of x in order to figure out which choice is rational?

To figure out the answer, let's see what happens if we try to discover whether $EU(A)$ is greater than, less than, or equal to $EU(B)$ without knowing the value of x. We start by asking ourselves whether $EU(A) > EU(B)$:

$$EU(A) \stackrel{?}{>} EU(B)$$

$$\Pr(S) \cdot 100 + \Pr(T) \cdot x \stackrel{?}{>} \Pr(S) \cdot 0 + \Pr(T) \cdot x$$

Now, notice that we have exactly the same term, '$\Pr(T) \cdot x$' showing up on both sides of the inequality. So we can subtract this from both sides, getting

$$\Pr(S) \cdot 100 \stackrel{?}{>} \Pr(S) \cdot 50$$

$$100 \overset{\checkmark}{>} 50$$

So it turns out that, according to *Maximize Expected Utility*, A is a more rational choice than B is, *no matter what value x takes on*. We don't need to know what x is in order to know that A is more rational than B.

In this example, we *also* didn't need to know the probability of state S—for, in this example, *A weakly dominates B*. Similar reasoning will show that, in general, whenever one act weakly dominates another, the first act will be at least as rational as the second. But the reasoning we went through will apply more generally. Take this decision, for instance,

	State R	State S	State T
Act A	100	0	x
Act B	50	100	x

According to *Maximize Expected Utility*, we don't need to know the value of x in order to know whether A is more, less, or equally as rational as B.

For, in order to figure out whether A is more rational than B, we have to figure out whether $EU(A) > EU(B)$,

$$EU(A) \overset{?}{>} EU(B)$$

$$\Pr(R) \cdot 100 + \Pr(S) \cdot 0 + \Pr(T) \cdot x \overset{?}{>} \Pr(R) \cdot 50 + \Pr(S) \cdot 100 + \Pr(T) \cdot x$$

But notice that we have exactly the same term, $\Pr(T) \cdot x$, showing up on both sides of the inequality. So we can subtract it from both sides, getting

$$\Pr(R) \cdot 100 + \Pr(S) \cdot 0 \overset{?}{>} \Pr(R) \cdot 50 + \Pr(S) \cdot 100$$

Now, until I know what the probability of R and S are, I won't know whether the expected utility of A is greater than the expected utility of B or not. But the important thing is that I *don't* need to know what x is. If I know that both A and B get me precisely the same outcome in the state T, then I don't need to know what that outcome *is* in order to know whether A is more, less, or just as rational as B.

Here's another way of putting the same point: suppose I faced a *different* choice, between two acts C and D, with the same three states, the same probabilities, and the same payouts in states R and S, like this:

	State R	State S	State T
Act C	100	0	y
Act D	50	100	y

In all relevant respects, C and D are just like A and B—*but for* the different outcome in state T. But since we've already seen that this outcome *makes no difference* to whether the expected utility of A is greater than, less than, or equal to the expected utility of B, changing x to y can't make any difference with respect to whether C is more, less, or just as rational as D. From expected utility theory, we can conclude that:

C is more rational than *D* if and only if *A* is more rational than *B*

And likewise, *C* is *less* rational than *D* if and only if *A* is less rational than *B*. And *C* is *just as* rational a choice as *D* if and only if *A* is just as rational a choice as *B*.

In general, the lesson is this:

> **The Sure Thing Principle** If *A* and *B* lead to exactly the same outcome in a state, then it doesn't matter what that outcome is. Changing that outcome won't change whether *A* is more, less, or just as rational as *B*.

Think of the Sure Thing Principle like this: if you're gonna get the same outcome in a state, no matter which act you select, then that outcome in that state is a *sure thing*, and it shouldn't make any difference to your preference between the acts.

The Allais 'Paradox'

Maurice Allais discussed a case in which people tend to have preferences that violate the Sure Thing Principle. Since the Sure Thing Principle is a consequence of *Maximize Expected Utility*, this is a case in which people tend to have preferences incompatible with maximizing expected utility.

Figure 14.2: Maurice Allais

To understand Allais's case, imagine that we are going to roll a 100-sided die. I then offer you a choice between two gambles. Gamble *A* is guaranteed to get you $1,000,000, no matter how the die lands. (So it's not really much of a *gamble* at all; but we can think of it as a trivial or degenerate gamble, where the only probabilities are zeros and ones.) Gamble *B*, on the other hand, will get you nothing if the die lands on 1, it will get you $5,000,000 if the die lands on a number between 2 and 10, and it will get you $1,000,000 if the die lands on a number between 11 and 100.

	Die lands 1	Die lands 2–10	Die lands 11–100
A	$1,000,000	$1,000,000	$1,000,000
B	$0	$5,000,000	$1,000,000

Here, given the choice, many people say that they would take gamble *A* over gamble *B*. They think that they'd rather have the guaranteed $1,000,000 than take a chance of getting nothing.

Contrast this choice with another: again, we will roll a 100-sided die, and I offer you a choice between two *different* gambles. Gamble *C* will get you $1,000,000 if the die lands 1–10, and will get you nothing otherwise. Gamble *D*, on the other hand, will get you $5,000,000 if the die lands 2–10, and will get you nothing otherwise.

	Die lands 1	Die lands 2–10	Die lands 11–100
C	\$1,000,000	\$1,000,000	\$0
D	\$0	\$5,000,000	\$0

Given this choice, many people say that they would take gamble D over gamble C. They think: I'm already taking a risk, and given that I'm taking the risk, I'm willing to sacrifice a 1% chance of winning in order to raise the amount I stand to win an additional \$4,000,000.

So people choose A over B, and choose D over C. But notice: the only difference between these two pairs of gambles is what happens if the die lands 11–100. And, in each pair, both gambles in the pair get you exactly the same outcome if the die lands 11–100. Both decisions look like this, with a different choice of x:

	Die lands 1	Die lands 2–10	Die lands 11–100
A/C	\$1,000,000	\$1,000,000	\$$x$
B/D	\$0	\$5,000,000	\$$x$

For the choice between A and B, x is 1,000,000. And for the choice between C and D, x is 0. But the Sure Thing Principle says that *it shouldn't matter what x is*. Changing the value of x won't change whether the first gamble is more, less, or just as rational as the second. So if A is a more rational choice than B, then C should be a more rational choice than D. And if D is a more rational choice than C, then B should be a more rational choice than A.

What this teaches us is that actual people can display a kind of risk-aversion which is not captured by *Maximize Expected Utility*. And it shows us that many real-world people are not in fact expected utility maximizers.

There are two potential problems for *Maximize Expected Utility* here, corresponding to use ways we could understand the theory—as a *descriptive* theory about how people actually choose, or as a *normative* theory about how people *ought* to choose. The problem for the descriptive theory is that people don't actually seem to make the kinds of choices that *Maximize Expected Utility* predicts. So we can't understand people as expected utility maximizers. (Recall that our method for measuring people's utility was predicated on the assumption that people are maximizing expected utility. So insofar as people have patterns of risk-aversion which are incompatible with the Sure Thing Principle, this is a reason to think that we cannot measure people's utilities.)

The second problem is for the *normative* interpretation of *Maximize Expected Utility*. Insofar as we think that the preferences between Allais gambles are *rational*, this is a reason to think that it can be rational to fail to maximize expected utility.

In defense of the normative interpretation of *Maximize Expected Utility*: it's worth noting that there are many cases in which people seem to have uncontroversially irrational preferences.

Start by considering this decision: there are 600 people in the hospital with a terminal case of the flu. You run the hospital and have two treatment options: the *sure* option and the *risky* one. If you take the sure option, then 200 people will certainly be saved from

death. However, if you take the risky option, then you'll have a $1/3$ probability of saving 600 people, and a $2/3$ probability of saving nobody. Given this choice, many people opt for the sure option: better to save 200 for sure than take a chance that nobody gets saved.

Next, consider this decision: there are 600 people in the hospital with a terminal case of the flu. You run the hospital and have two treatment options: the *sure* option and the *risky* option. If you take the sure option, 400 people will certainly die. However, if you take the risky option, then there will be a $1/3$ probability that nobody dies, and a $2/3$ probability that all 600 die. Given *this* choice, many people opt for the risky option: better to take a shot at nobody dying than to guarantee 400 deaths.

But notice that this isn't two different decisions. It's *exactly the same* decision, first framed in terms of how many people will live, and next framed in terms of how many people will die. Surely these kinds of superficial differences in presentation shouldn't change which choice is rational. But you might suspect that exactly the kind of reasoning which leads people to fall prey to these kinds of framing effects also leads them to adopt the Allais preferences. So you might hold onto the *normative* version of *Maximize Expected Utility* in spite of Allais's objection to the Sure Thing Principle.

The Ellsberg 'Paradox'

Daniel Ellsberg noted another odd consequence of the Sure Thing Principle. Suppose that we have an urn which contains nine marbles. 3 of the marbles are red, and the remaining 6 marbles are either yellow or green—but you don't know how many are yellow and how many are green. We are going to draw one marble from the urn randomly. And you're offered a choice between two gambles, A and B. Gamble A will get you \$100 if the drawn marble is red, and will get you nothing otherwise. Gamble B, on the other hand, will get you \$100 if the drawn marble is yellow, and will get you nothing otherwise.

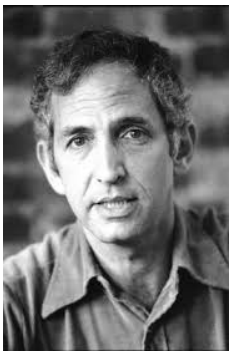


Figure 14.3: Daniel Ellsberg

	Marble is red	Marble is yellow	Marble is green
A	\$100	\$0	\$0
B	\$0	\$100	\$0

Given this choice, many people say that they would take gamble A over gamble B. After all, you know for sure that gamble A gets you a $1/3$ chance of \$100, but who knows what chance of \$100 gamble B gives you. It could be a probability of zero and it could be a probability of $2/3$, depending upon how many of the 6 non-red marbles are yellow. Better to stick with the known chances than take a risk on the unknown chances.

Contrast this choice with another: again, we have the urn which contains 9 marbles overall, 3 red and the remaining either yellow or green (you don't know how many of the non-red ones are green and how many are yellow). And I offer you a choice between two

different gambles: gamble C will get you \$100 if the drawn marble is either red or green (and nothing if it's yellow). And gamble D will get you \$100 if the drawn marble is either yellow or green (and nothing if it's red).

	Marble is red	Marble is yellow	Marble is green
C	\$100	\$0	\$100
D	\$0	\$100	\$100

Given this choice, many people say that they would take gamble D over gamble C. They think: I know for sure that gamble D has a $2/3$ chance of getting me \$100. But who knows what chance of \$100 gamble A is giving me. It could be a chance of $1/2$ and it could be a chance of 1, depending upon how many marbles in the urn are green. Better to stick with the known chances than take a risk on the unknown chances.

Exercise 44. *Are these preferences consistent with the Sure Thing Principle? Why or why not?*

People with these preferences don't seem to be averse to *risk*—both of the gambles are risky. Instead, they seem to be averse to a certain kind of *ambiguity*—they prefer gambles in which the chances are known to gambles in which the chances are unknown. People with preferences like these are said to be *ambiguity-averse*.

1. Consider the three utility functions, U_1, U_2, and U_3, shown in figure 14.4.

	U_1	U_2	U_3
a	55	80	−100
b	30	−20	−50
c	−65	−400	140
d	105	200	−200

Figure 14.4: The numbers under each utility function show the value it assigns to the outcome in that row.

   (a) Are U_1 and U_2 ordinally equivalent?

   (b) Are U_1 and U_2 cardinally equivalent? (If so, say which choice of α and β show that they are cardinally equivalent.)

   (c) Are U_1 and U_3 ordinally equivalent?

   (d) Are U_1 and U_3 cardinally equivalent? (If so, say which choice of α and β show that they are cardinally equivalent.)

2. Suppose that Tabitha has several shirts in front of her: a burgundy one, a chartreuse one, a grey one, and a navy one. Let's use the following letters for these outcomes:

   b = Tabitha wears the burgundy shirt

   c = Tabitha wears the chartreuse shirt

   g = Tabitha wears the grey shirt

   n = Tabitha wears the navy shirt

   Suppose that Tabitha tells you the following information:

   Given a choice between b and c, she'd prefer c

   Given a choice between b and n, she'd prefer b

   Given a choice between n and g, she'd prefer n

   Given a choice between b and a gamble which gives her c with probability $1/3$ and g with probability $2/3$, she's indifferent, and could go either way.

   Given a choice between n and a gamble which gives her b with probability $1/2$ and g with probability $1/2$, she's indifferent, and could go either way.

Then, using Ramsey's procedure, construct a utility function for Tabitha's preferences between shirts. To make it easier on yourself, use the 'zero-one' utility scale, which gives a utility of one to Tabitha's most preferred option (of the ones listed) and which gives a utility of zero to Tabitha's least preferred option (of the ones listed).

3. Consider a gamble which gives you \$100 with probability $1/10$ and \$1 with probability $9/10$.

   (a) What is the expected monetary value of this gamble?

   (b) Suppose that your utilities (as a function of dollars) are given by $U(\$x) = \log(x)$ (where this is the logarithm base 10—that is, the exponent, y, that 10 has to be raised to in order for 10^y to be equal to x). Then, what is the expected utility of the gamble?

   (c) Given this utility function, what is the utility of the gamble's expected monetary value?

   (d) Given this utility function, if you were given a choice between the gamble and its expected monetary value, which should you prefer? (I mean: according to the theory *Maximize Expected Utility*.)

   (e) Given these utilities, are you risk-averse, risk-seeking, or risk-neutral? (I mean: as the expected utility maximizer understands risk-aversion.)

4. Suppose that you will roll a fair six-sided die, and you give Johann and Filipa a choice between the two gambles A and B, which will yield the following dollar amounts, depending upon how the die lands:

Die lands on:	1	2	3	4	5	6
A	\$2	\$5	\$6	\$2	−\$3	−\$12
B	\$4	\$1	\$6	\$4	−\$2	−\$5

   Johann says that he prefers A to B, and Filipa says that she prefers B to A.

   Next, you give Johann and Filipa a choice between the two gambles C and D, which yield the following dollar amounts, depending upon how the die lands:

Die lands on:	1	2	3	4	5	6
A	\$2	\$5	\$1	\$2	−\$3	−\$12
B	\$4	\$1	\$1	\$4	−\$2	−\$5

   Johann says that he prefers D to C, and Filipa says that she prefers C to D.

   (a) Do Johann's preferences violate the Sure Thing Principle? If so, why? If not, why not?

   (b) Do Filipa's preferences violate the Sure Thing Principle? If so, why? If not, why not?

15 | Infinity

Goal: Understand several issues with expected utility maximization having to do with infinity.

Pascal's Wager

Recall that Blaise Pascal, along with Pierre de Fermat, forged the foundations of expected utility theory at the age of 31. He died at the age of 39. After his death, a collection of his writings were published under the name *Pensées* ('Thoughts' in French). These writings contained within them a *decision-theoretic* argument for believing in God:

> *Either God exists, or He does not. To which view shall we incline? Reason cannot decide for us one way or the other: we are separated by an infinite gulf. At the extremity of this infinite distance a game is in progress, where either heads or tails may turn up. What will you wager?...*
>
> *Let us weigh the gain and the loss involved in wagering that God exists. Let us estimate these two probabilities; if you win, you win all; if you lose, you lose nothing. Wager then, without hesitation, that He does exist.*

Let's think a bit more carefully about this style of argument; and let's reformulate it using the decision-theoretic tools we've been developing in the course.

Pascal is imagining that you have two options available to you: you may either believe in God, or not. And there are two relevant states: either God exists, or God does not exist. So the decision matrix for your decision looks like this:

	God exists	God does not exist
Believe in God		
Do not believe in God		

Which outcomes lie in this table? Well, if you believe in God and he exists, then you will be rewarded with eternal happiness in heaven. What is the *utility* of eternal happiness in heaven? Well, presumably, it is *infinite*. So the utility of belief in the state where God exists is ∞. What about if you don't believe in God and God exists? We might think that failure

to believe in God will result in *negative* infinite utility, since it results in eternal suffering in hell. So the utility of disbelief in the state where God exists is $-\infty$. What about if God does not exist? Well, however good or bad belief or disbelief may be, they will be *finite*. Let's represent the utility of some good but finite outcome with the number 10 and the utility of some bad but finite outcome with the number -10.

	God exists	God does not exist
Believe in God	∞	-10
Do not believe in God	$-\infty$	10

Suppose that the probability of God existing is positive. Call that probability 'p'. Then, the expected utility of believing in God will be

$$\begin{aligned} p \cdot \infty + (1 - p) \cdot (-10) &= \infty - 10 + 10p \\ &= \infty \end{aligned}$$

(since ∞ times anything positive is still ∞ and ∞ minus anything finite is still ∞.) And the expected utility of not believing in God will be

$$\begin{aligned} p \cdot (-\infty) + (1 - p) \cdot 10 &= -\infty + 10 - 10p \\ &= -\infty \end{aligned}$$

(since $-\infty$ times anything positive is still $-\infty$ and $-\infty$ plus anything finite is still $-\infty$.)

So the expected utility of believing in God is greater than the expected utility of not believing in God. So Pascal concludes that it is rational for you to believe in God.

Is This The Right Kind of Reason for Belief?. There's something odd about the kind of argument Pascal's giving here. He's *not* arguing that you should believe in God because there's strong *evidence* that God exists. Instead, he's saying that you should believe in God because of what the *practical benefits* of that belief will be.

Philosophers sometimes distinguish between *practical* rationality and *theoretical* rationality. Practical rationality has to do with our reasons for action—when do we have most reason to make this choice or that one? Theoretical rationality has to do with our reasons for belief—when do we have most reason to believe this thing or that?

Decision theory is a theory of practical rationality. It says that you have more reason to make a choice the higher that choice's expected utility.

It's natural to think that theoretical rationality is a matter of what *evidence* you have. That is, it's natural to think that the only reasons for holding a belief are reasons to think that that belief is *true*. But Pascal is not appealing to evidence like this. He is not giving reasons to think that belief in God would be true. Instead, he is appealing to *practical benefits* of belief in God. It's natural to think that these are just the *wrong kinds of reasons* for belief. Not believing in climate change might make you happy, but that's no reason

to not believe in climate change. It is still (theoretically) *irrational* to not believe in climate change, even if that belief might have various practical benefits for you.

On the other hand, some philosophers think that the distinction between practical and theoretical rationality is artificial. They contend that practical reasons to adopt a belief are just as good as evidence that the belief is true. This kind of position is known as *pragmatism*. (Note: there are many different philosophical views which have this name; we're just talking about one particular brand of pragmatism here.) So perhaps we should understand Pascal as implicitly committed to a brand of pragmatism.

Are The Acts Under Your Control?. Recall, when we learned how to construct a decision matrix, we learned about four rules that we had to follow. Rule #2 said:

2. Find a partition of acts available to you. These must be acts which are certain to be under your control.

But you might think Pascal's decision matrix doesn't satisfy this requirement. 'Believe in God' and 'Do not believe in God' are a partition—that's fine. But it doesn't look like it's under your control what you believe.

That's a problem for the argument. But perhaps we can fix it. Perhaps there are certain steps you can take which will make it more likely that you will end up believing in God. For instance, people tend to adopt the beliefs of their social group. So you could start going to Church and start associating with people who believe in God. Doing so would make it more likely that you end up adopting the belief.

So perhaps we should instead have the acts 'Go to church and try to believe in God' and 'Do not try to believe in God'. With these new acts, we run into the problem that our acts and our states won't any longer determine outcomes. Recall rule #4:

4. Your partitions of acts and states must be detailed enough that, for each act and each state, there is a unique outcome which will result, if you choose that act in that state.

If you don't know whether you'd end up believing, if you went to church regularly, then you won't know whether you'd be saved, if you started going to church.

If regular church attendance would lead you to believe in God, then let's say that you're 'impressionable'. Then, we will need three states: *God exists and you are impressionable*, *God exists and you are not impressionable*, and *God doesn't exist*. (Note that these states form a partition.)

	God exists and you're impressionable	God exists and you're not impressionable	God does not exist
Go to church			
Do not go to church			

What outcome would each act lead to in each of these states? Well, let's continue assuming that, if God exists and you believe, you get $+\infty$ utiles, and if God exists and you do not belive, you get $-\infty$ utiles. If God doesn't exist, then your utility will be finite—let it be ± 10, it won't matter.

	God exists and you're impressionable	God exists and you're not impressionable	God does not exist
Go to church	∞	$-\infty$	-10
Do not go to church	$-\infty$	$-\infty$	10

But now, the expected utility calculations are well-defined. For instance, if p is your probability that God exists, and q is your probability that you're impressionable (and if we suppose that these two propositions are probabilistically independent), then the expected utility of going to church will be

$$pq \cdot \infty + p(1-q) \cdot (-\infty) + (1-p)(-10) = \infty - \infty - 10 \cdot (1-p)$$

But $\infty - \infty$ is not well-defined. So we can't calculate the expected utility of going to church. And the theory *Maximize Expected Utility* falls silent.

But wait, anything minus itself is going to be zero, right? Not so! Consider the following two infinite sums:

$$S_1 = 1 + 2 + 3 + 4 + 5 + 6 + 7 + \cdots$$
$$S_2 = 1 + 1 + 1 + 1 + 1 + 1 + 1 + \cdots$$

Both of these infinite sums have the same value: ∞. Now, in fact, subtracting S_2 from S_1 is not a well-defined operation. But let's think for a bit about what things would be like if it *were* a well-defined operation. Then, subtracting S_2 from S_1 would give us this:

$$\begin{aligned} S_1 - S_2 &= (1-1) + (2-1) + (3-1) + (4-1) + (5-1) + (6-1) + (7-1) + \cdots \\ &= 0 + 1 + 2 + 3 + 4 + 5 + 6 + \cdots \\ &= S_1 \end{aligned}$$

but we know that the infinite sum S_1 is ∞. So, even if we suppose that this kind of subtraction is well-defined, we shouldn't think that $\infty - \infty$ will equal 0. And, more importantly, this kind of subtraction is *not* well-defined. Part of the reason this sum is not well-defined is that it will matter how we group our terms. I'll come back to this a bit later, below.

Do Acts and States Determine a Unique Outcome?. Let's just grant Pascal that it's entirely under your control what you believe. We can imagine versions of the case where you are presented with a belief pill—if you take the pill, then you'll believe in God instantly, and if you don't, then you won't. Even with this assumption granted, there are concerns with the argument. Recall rule #4 for constructing a decision matrix:

4. Your partitions of acts and states must be detailed enough that, for each act and each state, there is a unique outcome which will result, if you choose that act in that state.

There's a concern that the states 'God exists' and 'God doesn't exist' are not detailed enough to determine a unique outcome.

First problem: it could be that God exists and sends *everyone* to heaven (as some Christian universalists believe). Since this makes a difference to something you care about, the state 'God exists' is not detailed enough. The state must be subdivided further. Let's call God 'forgiving' if everyone gets into heaven. And let's call God 'unforgiving' if only believers get into heaven. (These are just stipulative terms, we're not taking a stand on any divine attributes by using these terms.) Then, we will have these states:

	God exists and is forgiving	God exists and is not forgiving	God does not exist
Believe in God	∞	∞	-10
Do not believe in God	∞	$-\infty$	10

Now, when we calculate the expected utility of not believing in God, we'll get $\infty - \infty$, which is not well-defined.

Second problem: there are *multiple* gods that could exist. The Catholic God could exist (this is the one that Pascal was most concerned with). But it could instead be that Allah, the god of the Qur'an, exists. And it could be that Vishnu or Gnesha, the Hindu gods, exist. Or that Zeus, the Greek god, exists. And each of these gods could be promising eternal salvation if you believe in *them* and no other gods. In that case, we might be leaving out both on importantly different states and importantly different acts. Really, our decision matrix should look like this:

	God exists	Allah exists	Vishnu exists	$\cdots$	No god exists
Believe in God	∞	$-\infty$	$-\infty$	$\cdots$	-10
Believe in Allah	$-\infty$	∞	$-\infty$	$\cdots$	-10
Believe in Vishnu	$-\infty$	$-\infty$	∞	$\cdots$	-10
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Believe in no god	$-\infty$	$-\infty$	$-\infty$	$\cdots$	10

And again, with all these positive and negative infinities, the expected values simply become undefined.

The foregoing is often presented as an objection to Pascal's argument, but I think it's better understood as an objection to *Maximize Expected Utility*. The issue is that the theory of maximizing expected utility simply breaks down when it's considering decisions like these, where the utilities become infinite.

The Saint Petersburg Paradox, Redux

Let's go back to the original puzzle that got Daniel Bernoulli thinking about *utility* in the first place, the Saint Petersburg paradox. Recall,

Puzzle (The Saint Petersburg Paradox)**.** *We will flip a fair coin* until it lands heads*. If the first heads comes on the first flip, then you win* $\$2$*. If the first heads comes on the second flip, you win* $\$4$*. If the first heads comes on the third flip, then you win* $\$8$*. In general, if the first heads comes in the* n*th flip, then you win* $\$2^n$*.*

When we calculated the *expected monetary value* of this gamble we found that it diverged to ∞:

$$\begin{aligned}\sum_x \Pr(\textit{you win } x) \cdot x &= 1/2 \cdot 2 + 1/4 \cdot 4 + 1/8 \cdot 8 + 1/16 + 16 \cdot \cdots + 1/2^n \cdot 2^n + \cdots \\ &= 1 + 1 + 1 + 1 + \cdots + 1 + \cdots \\ &= \infty\end{aligned}$$

Daniel Bernoulli suggested that the solution was to replace *dollars* with *utility*. In particular, he suggested that your utilities were *logarithmic* in dollars. For example, you could have the utility function $U(\$x) = \log(x)$, where '$\log$' is the logarithm base 10—that is, the exponent, y, that 10 has to be raised to in order to have $10^y = x$. Then, the sum above will not diverge to ∞, but will take on a finite value around 0.6 (which is about the utility of $\$4$).

But does this really solve the problem? Consider the following variant of the case:

Puzzle (The Saint Petersburg Paradox, redux)**.** *We will flip a fair coin* until it lands heads*. If the first heads comes on the first flip, then you win* $\$10^2$*. If the first heads comes on the second flip, you win* $\$10^4$*. If the first heads comes on the third flip, then you win* $\$10^8$*. In general, if the first heads comes in the* n*th flip, then you win* $\$10^{2^n}$*.*

What's the expected *utility* of this gamble? Well, it's

$$\begin{aligned}\sum_x \Pr(\textit{you win } \$x) \cdot x &= 1/2 \cdot \log(10^2) + 1/4 \cdot \log(10^4) + 1/8 \cdot \log(10^8) + \cdots + 1/2^n \cdot \log(10^{2^n}) + \cdots \\ &= 1/2 \cdot 2 + 1/4 \cdot 4 + 1/8 \cdot 8 + 1/16 + 16 \cdot \cdots + 1/2^n \cdot 2^n + \cdots \\ &= 1 + 1 + 1 + 1 + \cdots + 1 + \cdots \\ &= \infty\end{aligned}$$

Perhaps you could worry that there's no bank in the world capable of paying out this amount of money. But consider this decision:

Puzzle (Divine Saint Petersburg Paradox)**.** *God comes to you and tells you*: *I will flip a coin until it lands heads. If the first heads comes on the first flip, I will give you two days in heaven. If*

the first heads comes on the second flip, I will give you four days in heaven. If the first heads comes on the third flip, I will give you 8 days in heaven. And, in general, if the first heads comes on the nth flip, then I will give you 2^n days in heaven.

God then asks: how many days of suffering on earth are you willing to exchange for this gamble?

It certainly seems plausible that each new day in heaven should have as much utility as the one that came before it. (And, in any case, we can always pull the same trick we pulled above, where 2^n days is replaced with $\log(10^{2n})$ days.) But when we evaluate the expected utility of this gamble, we'll again say that it is infinite, and so we'll say that it is worth sacrificing any finite number of days on earth for the gamble.

But if you agree to ∞ days of suffering on earth, in exchange for a gamble which is certain to only give you a *finite* number of days in heaven, it seems like you've made a mistake. Even if you're only agreeing to, for instance, $10^{10^{10}}$ days of suffering on earth, it can seem like a mistake to agree to this in exchange for a gamble which is *overwhelming* likely to get you no more than 100 days in heaven.

The Pasadena Paradox

While he was working at Cal Tech in Pasadena, the philosopher Alan Hájek and co-author Harris Nover came up with another interesting puzzle for expected utility theory. This puzzle has come to be known as the *Pasadena Paradox* (or the Pasasdena game).

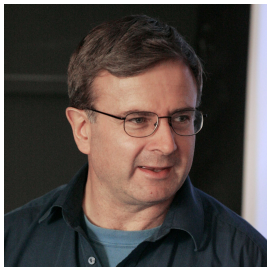


Figure 15.1: Alan Hájek

Background: Conditionally Convergent Infinite Sums. To understand their puzzle, let's start off by talking a bit about how infinite sums are rigorously defined. For instance, take the infinite sum

$$1/2 + 1/4 + 1/8 + 1/16 + \cdots + 1/2^n + \cdots$$

How do we make sense of an infinite sum like this? We start off by considering the *partial sums*. These are the sums we get if we just stop adding summands after a certain point. For instance, the first partial sum is just $1/2$ (where we've only added the first summand). The next partial sum is $1/2 + 1/4$ (where we've added the first two summands). And the partial sums continue in this way.

1st partial sum:	$1/2 = 1/2$
2nd partial sum:	$1/2 + 1/4 = 3/4$
3rd partial sum:	$1/2 + 1/4 + 1/8 = 7/8$
4th partial sum:	$1/2 + 1/4 + 1/8 + 1/16 = 15/16$
$\vdots$	$\vdots$
nth partial sum:	$1/2 + 1/4 + 1/8 + 1/16 + \cdots + 1/2^n = (2^n - 1)/2^n$

Hopefully it's clear that, as n gets larger and larger, these partial sums will get closer and closer to the value 1. For this reason, we say that the *limit* of the sequence of partial sums is 1. And we define the infinite sum to be this limit of partial sums, if it exists.

In other cases, the sequence of partial sums will not have a limit. There are a few ways that this could happen. In the first place, the partial sums could just get larger and larger without bound. For instance, consider the infinite sum

$$1 + 2 + 3 + 4 + 5 + 6 + \dots$$

This infinite sum has the following partial sums:

$$\begin{aligned}
\text{1st partial sum:} &\qquad 1 = 1 \\
\text{2nd partial sum:} &\qquad 1 + 2 = 3 \\
\text{3rd partial sum:} &\qquad 1 + 2 + 3 = 6 \\
\text{4th partial sum:} &\qquad 1 + 2 + 3 + 4 = 10 \\
\vdots &\qquad \vdots \\
n\text{th partial sum:} &\qquad 1 + 2 + 3 + 4 + \cdots + n = \frac{n(n+1)}{2}
\end{aligned}$$

Hopefully its clear that, as n gets larger and larger, then nth partial sum will also get larger and larger, without bound. So it will eventually blow past any possible limit. In that case, we say that the infinite sum *diverges to* ∞. (We can also get infinite series that diverge to $-\infty$. Just consider the same sum, but make all the numbers negative.)

In the second kind of case where the sequence of partial sums fails to have a limit, the partial sums will remain small, but they won't converge to any particular value. For instance, consider the infinite sum below.

$$1 - 1 + 1 - 1 + 1 - 1 + 1 - 1 + \dots$$

This infinite sum has the following partial sums:

$$\begin{aligned}
\text{1st partial sum:} &\qquad 1 = 1 \\
\text{2nd partial sum:} &\qquad 1 - 1 = 0 \\
\text{3rd partial sum:} &\qquad 1 - 1 + 1 = 1 \\
\text{4th partial sum:} &\qquad 1 - 1 + 1 - 1 = 0 \\
\vdots &\qquad \vdots
\end{aligned}$$

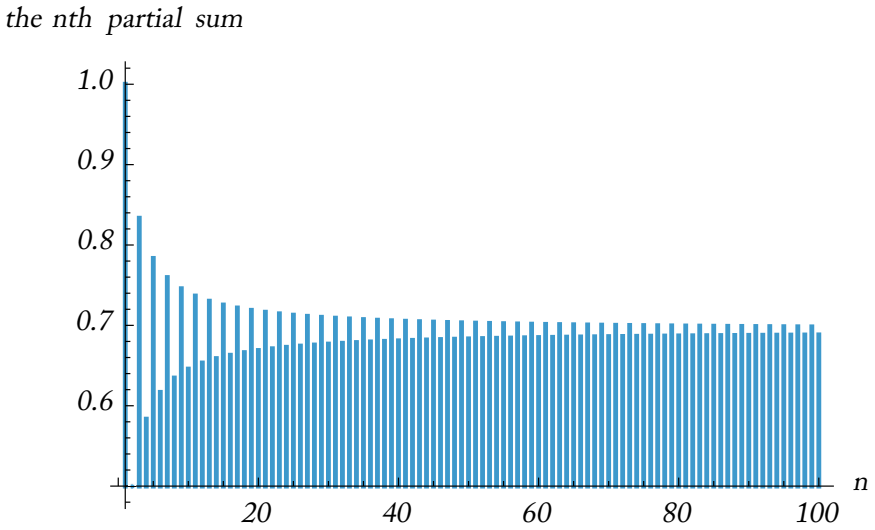


Figure 15.2: The first 100 partial sums of the infinite sum $1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + \cdots$.

The nth partial sum will be 1 if n is odd and 0 if n is even. So the sequence of partial sums won't approach any limit at all. It'll just bounce back and forth between 0 and 1, forever. In this case, we say that the infinite sum is *undefined*.

There's another kind of case worth thinking about. Sometimes, *whether* a sequence of partial sums converge—and *what* they converge to—can depend upon the order in which we take the sum. For instance, consider the following infinite sum:

$$1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + 1/7 - 1/8 + \cdots$$

In general, in this infinite sum, if n is odd, then the nth term is $+1/n$, and if n is even, then the nth term is $-1/n$. It's not immediately clear what happens with the partial sums of this infinite sequence,

$$\begin{aligned}
\text{1st partial sum:} &\quad 1 = 1 \\
\text{2nd partial sum:} &\quad 1 - 1/2 = 1/2 \\
\text{3rd partial sum:} &\quad 1 - 1/2 + 1/3 = 5/6 \\
\text{4th partial sum:} &\quad 1 - 1/2 + 1/3 - 1/4 = 7/12 \\
\vdots &\quad \vdots
\end{aligned}$$

but it turns out that it *does* converge. You can see the result of the first 100 partial sums in figure 15.2. In fact, these partial sums converge to the value $\ln(2)$ (the natural logarithm of 2—that is, the number y such that $e^y = 2$, which is around 0.69).

However, it matters what *order* we put the summands here. Notice that we can split the summands up, depending upon whether they are the reciprocal of an odd or an even integer.

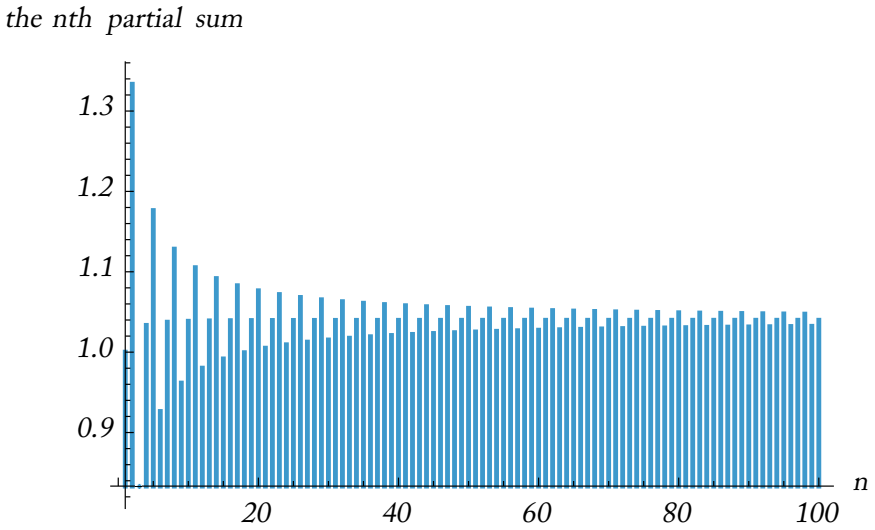


Figure 15.3: The first 100 partial sums of the infinite sum $1 + 1/3 - 1/2 + 1/5 + 1/7 - 1/4 + 1/9 + 1/11 - 1/6 + 1/13 + 1/15 - 1/8 + \cdots$.

reciprocal of odd integer:	1	$1/3$	$1/5$	$1/7$	$1/9$	$\cdots$
reciprocal of even integer:	$-1/2$	$-1/4$	$-1/6$	$-1/8$	$-1/10$	$\cdots$

Consider the following way of ordering things: we first add together the reciprocal of the first two odd numbers. Then, we subtract the reciprocal of the first even number. Then, we add together the next two odd numbers. Then, we subtract the reciprocal of the next even number. And so on,

$$1 + 1/3 - 1/2 + 1/5 + 1/7 - 1/4 + 1/9 + 1/11 - 1/6 + 1/13 + 1/15 - 1/8 + \cdots$$

With this rearrangement, the partial sums start to look different.

1st partial sum:	$1 = 1$
2nd partial sum:	$1 + 1/3 = 4/3$
3rd partial sum:	$1 + 1/3 - 1/2 = 5/6$
4th partial sum:	$1 + 1/3 - 1/2 + 1/7 = 41/42$
$\vdots$	$\vdots$

The first 100 partial sums are shown in figure 15.3. As you can see there, the partials sums are still converging, but they are converging to a *different limit*. Now, they are getting closer and closer to a limit of about 1.04.

Suppose we rearrange the terms differently: we start by taking the first *five* numbers from the even list, and then we take *one* number from the odd list, and then go back and take the next *five* from the even list, then the second number from the odd list, and so on

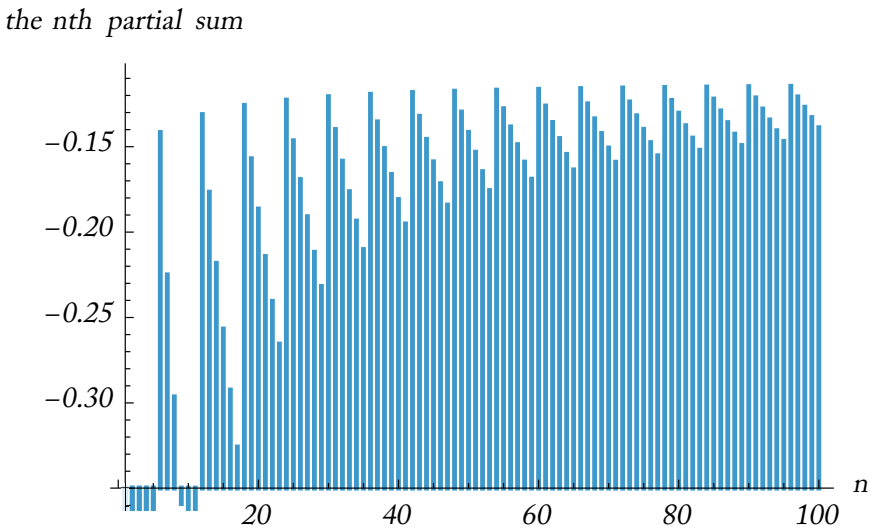


Figure 15.4: The first 100 partial sums of the infinite sum $-1/2 - 1/4 - 1/6 - 1/8 - 1/10 + 1 - 1/12 - 1/14 - 1/16 - 1/18 - 1/20 + 1/3 \cdots$.

and so forth, taking five numbers from the even list for every one number from the odd list. Then, we'll have the infinite sum

$$-1/2 - 1/4 - 1/6 - 1/8 - 1/10 + 1 - 1/12 - 1/14 - 1/16 - 1/18 - 1/20 + 1/3 \cdots$$

If we order the terms *this* way, then the partial sums will converge to a *negative* value—somewhere around -0.11. (See figure 15.4.)

In fact, there's a more general theorem, known as the Riemann rearrangement theorem, which says that by simply rearranging the terms in this infinite sum, we can make it converge to *any* value *whatsoever* in between $-\infty$ and $+\infty$ (including $\pm\infty$ themselves).

Infinite sums like these are known as *conditionally* convergent infinite sums. (Because whether they converge, and what they converge to, is conditional on how the terms in the sum are ordered.)

The Pasadena Paradox. Suppose that we're going to flip a fair coin until it lands heads for the first time. We have written on consecutive cards your winnings for each possible outcome:

(Top card)	If the first heads is on flip #1, you win \$2
(2nd card)	If the first heads is on flip #2, you lose \$2
(3rd card)	If the first heads is on flip #3, you win \$$^8/_3$
(4th card)	If the first heads is on flip #4, you lose \$4.
⋮	

In general, the nth card says that, if the coin first lands heads on the nth flip and n is odd, then you will win $\$2^n/n$, and if n is even, then you will lose $\$2^n/n$. Let's suppose that your utilities are linear in dollars. Then, what's the expected value of this gamble? Well, let's go card by card:

$$1/2 \cdot 2 - 1/4 \cdot 2 + 1/8 \cdot 8/3 - 1/16 \cdot 4 + 1/32 \cdot 32/5 - 1/64 \cdot 64/6 + \cdots = 1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + \cdots$$

which is exactly our conditionally convergent sequence from above!

So, as we saw above, the expected utility of this gamble will be $\ln(2) \approx 0.69$.

But then Hájek and Nover continue:

> By accident, we drop the cards, and after picking them up and stacking them on the table, we find that they have been rearranged. No matter, you say—obviously the game has not changed, since the pay-off schedule remains the same. The game, after all, is correctly and completely specified by the conditionals written on the cards, and we have merely changed the order in which the conditionals are presented. As it happens, the consecutive cards read:
>
> (Top card) If the first heads is on flip #1, you win \$2
> (2nd card) If the first heads is on flip #3, you win $\$8/3$
> (3rd card) If the first heads is on flip #2, you lose \$2
> (4th card) If the first heads is on flip #5, you win $\$32/5$.
> (5th card) If the first heads is on flip #7, you win $\$128/7$.
> (6th card) If the first heads is on flip #4, you lose \$4.
> (7th card) If the first heads is on flip #9, you win $\$512/9$
> ⋮

What is the expected value of the gamble now? Well, if we go card by card, we get

$$1/2 \cdot 2 + 1/8 \cdot 8/3 - 1/4 \cdot 2 + 1/32 \cdot 32/5 + 1/128 \cdot 128/7 - 1/16 \cdot 4 + \cdots = 1 + 1/3 - 1/2 + 1/5 + 1/7 - 1/4 + \cdots$$

which is just the rearrangement of the infinite sum we considered above. We saw there that this infinite sum (in this order) converges to a value greater than one.

But then suppose a gust of wind comes along and scatters the cards again—once we pick them up and rearrange them on the table, they are in this order, with five consecutive loss cards interspersed between each win card:

(Top card) If the first heads is on flip #2, you lose \$2
(2nd card) If the first heads is on flip #4, you lose \$4
(3rd card) If the first heads is on flip #6, you lose $\$2^6/6$
(4th card) If the first heads is on flip #8, you lose $\$2^8/8$.
(5th card) If the first heads is on flip #10, you lose $\$2^{10}/10$.
(6th card) If the first heads is on flip #1, you win \$2.
(7th card) If the first heads is on flip #12, you lose $\$2^{12}/12$
⋮

What is the expected value of the gamble now? Well, if we go card by card, we get

$$\begin{aligned} &-1/4 \cdot 2 - 1/16 \cdot 4 - 1/2^6 \cdot 2^6/6 - 1/2^8 \cdot 2^8/8 - 1/2^{10} \cdot 2^{10}/10 + 1 - 1/2^{12} \cdot 2^{12}/12 \cdots \\ =& -1/2 - 1/4 - 1/6 - 1/8 - 1/10 + 1 - 1/12 - 1/14 - 1/16 - 1/18 - 1/20 + 1/3 \cdots \end{aligned}$$

We saw above that the partial sums in this infinite sum converge to somewhere around -0.4. So by dropping the cards, the gamble now has *negative* expected value!

But surely all this is absurd. Rearranging the cards can't make any difference to whether it's rational or irrational to play the Pasadena game. So this looks like another kind of case where *Maximize Expected Utility* is failing to give us the right kind of guidance about which choices are rational and which are irrational.

16 | The Two Envelope Paradox

In previous lectures, we've been learning about the history of expected utility theory. Many of the innovations were driven by puzzles and paradoxes. The problem of the points led to Pascal and Fermat's methods of evaluating gambles in terms of their expected monetary value. The St. Petersburg paradox led to Daniel Bernoulli's introduction of the concept of *utility*. The Allais and Ellsberg paradoxes have each led to their own modifications and refinements of expected utility theory (though we haven't discussed these modifications and innovations here).

The paradoxes we've been discussing in the previous lecture and which we will continue discussing today are different—at the present moment, we do not have any clear consensus about how these paradoxes are to be solved. For the problems we discussed last time—Pascal's wager, the divine St. Petersburg paradox, and the Pasadena game—the problem is partly just that we don't have the right mathematical tools for the job. It's natural to have the reaction that there's nothing particularly paradoxical going on: we just need better tools, and then the theory can be extended.

Today, I want to discuss a paradox which, in my opinion, *remains* deeply paradoxical even after we've got the right mathematical tools for the job.[1]

The Two Envelope Paradox

God presents you with two envelopes:

He informs you that inside each envelope is a ticket which entitles the holder to a certain number of days in heaven. He also tells you that one of the envelopes entitles you to twice as many days in heaven as the other. But you don't know which envelope gets you more days in heaven and which gets you fewer.

[1] My presentation of the two envelope paradox is indebted to Yuval Nov's wonderful YouTube video on the paradox, which may be found here: https://www.youtube.com/watch?v=_NGPncypY68. A similar treatment of the paradox can be found in John Norton's "Where the sum of our expectations fail us: the exchange paradox" (1998) *Pacific Philosophical Quarterly* 79 (1): 34–58.

The envelopes are shuffled, ones goes to you, and the other goes to me. I then offer you the chance to switch. Suppose that your utilities are linear in days in heaven—that is, the utility of n days in heaven is n. Then, should you take me up on my offer?

Intuitively, you have no more reason to switch envelopes than you have reason to not switch. After all, the situation with respect to the two envelopes is perfectly symmetric. Surely there's nothing that your envelope has that mine lacks.

But here's an argument that you're rationally required to switch—that not switching would be irrational:

The First Fallacious Argument Suppose that your envelope will get you y days in heaven. Then, *my* envelope will either get you $2y$ days in heaven (if mine has the larger number) or $\$y/2$ (if mine has the smaller number). I'm just as likely to have the envelope with the larger number as I am to have the envelope with the smaller number. So the probability that my envelope contains $2y$ is one half, and the probability that my envelope contains $y/2$ is one half. So the expected utility of switching is

$$1/2 \cdot 2y + 1/2 \cdot y/2 \quad = \quad y + y/4$$

which is greater than the utility of what's in your envelope, y. So the expected utility of switching is greater than the expected utility of not switching. So it would be irrational to not switch.

This argument leads to a surprising conclusion. But are we so sure that it's fallacious? Perhaps our intuitive thought was too quick, and there's some asymmetry between the two envelopes which makes your more choice-worthy?

I don't think so. *If* the argument worked, it would work too well. For it would *also* tell us that switching is irrational.

The Second Fallacious Argument Let 'm' be the number of days in heaven you'd get with my envelope. Then, the number in *your* envelope is either $2m$ (if you have the larger number) or $m/2$ (if I have the smaller number). You're just as likely to have the larger number as you are to have the smaller. So the probability that your envelope gets you $2m$ days in heaven is one half, and the probability that your envelope gets you $m/2$ days in heaven is one half. So the expected utility of keeping your envelope is

$$1/2 \cdot 2m + 1/2 \cdot m/2 \quad = \quad m + m/4$$

which is greater than the utility of what's in my envelope, m. So the expected utility of switching less than the expected utility of not switching. So it would be irrational to switch.

So the same style of argument can lead us to two inconsistent conclusions. Something must be wrong with the argument—but *what* is wrong with it?

Notice that we can make another argument for the conclusion that it doesn't matter whether we switch or not:

The Valid Argument? We know that one of the envelopes gets you twice as many days in heaven as the other. So there's some x such that one envelope gets you $2x$ days in heaven and the other gets you $x/2$ days. There's a 50% probability that my envelope has the larger amount and a 50% probability that my envelope has the smaller amount. So the expected utility of switching for my envelope is

$$EU(\textit{switching}) = \tfrac{1}{2} \cdot 2x + \tfrac{1}{2} \cdot \tfrac{x}{2} \;\; = \;\; x + \tfrac{x}{2}$$

And there's a 50% probability that your envelope has the larger number and a 50% probability that your envelope has the smaller number. So the expected utility of not switching envelopes is

$$EU(\textit{not switching}) = \tfrac{1}{2} \cdot 2x + \tfrac{1}{2} \cdot \tfrac{x}{2} \;\; = \;\; x + \tfrac{x}{2}$$

So the expected utility of switching is exactly the same as the expected utility of not switching. So either choice would be rational.

This reasoning *seems* correct—but it also doesn't seem to be that different from the reasoning we were doing earlier.

Even if we're convinced by the symmetry of the two envelopes that the final line of reasoning must be correct, and that the first two arguments must be incorrect, we still have a paradox: just *where* did that reasoning go wrong? Why *can't* we reason about the decision in the way the first two arguments reasoned about it?

Do we have the right probabilities?

Perhaps we're making a probabilistic mistake when we assume that your envelope is just as likely to contain the larger number as it is to contain the smaller number. This definitely *seems* correct—what reason do we have to think that yours has more or less than mine? But let's think it through carefully by trying to get specific about what kind of probabilities could have been used to distribute the different numbers between the two envelopes.

Just to get concrete, let's suppose that number of days in heaven is allocated between the two envelopes in the following way: a fair coin is flipped until it lands heads. If it first lands heads on flip n, then there are 2^n days in one envelope S and there are 2^{n-1} days in the other. After the days in heaven are allocated to the two envelopes, we shuffle them and randomly assign one to me and one to you. So, in general, you're just as likely to have the envelope with the smaller number as you are to have the envelope with the larger number. So we can represent the possibilities for the number in *your* envelope, along with the corresponding probabilities, with the table in figure 16.1. If you're in the first row of this table, then you have the envelope with the lesser number, and the number beneath is the one in my envelope. And if you're in the second row of this table, then you have the envelope with the greater number, and the number above is the one is in my envelope.

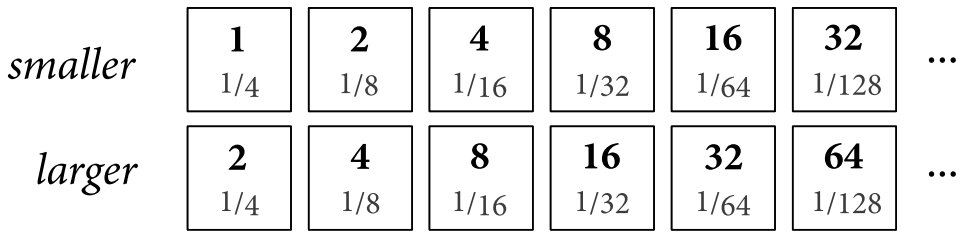


Figure 16.1

With these probabilistic assumptions in place, we can start to identify a place where the first fallacious argument went wrong. There, we picked some precise number of days in heaven that you could have in your envelope, y, and we then assumed that it was just as likely that you had the smaller number as it was that you had the larger number, given that choice of y. But looking at figure 16.1, we can see that this is not correct. For instance, suppose that you have a pass for 2 days in heaven in your envelope. Then, you're either in row 2, column 1, or you're in row 1, column 2. But the former is *twice* as likely! (Since $1/4$ is twice $1/8$.) So, conditional on your envelope having 2 days in heaven, there's a $2/3$ probability that you have the larger number, and only a $1/3$ probability that you have the smaller. So, conditional on us assuming that you have 2 days in your envelope, the expected utility of switching will be

$$\tfrac{1}{3} \cdot 4 + \tfrac{2}{3} \cdot 1 = \tfrac{4}{3} + \tfrac{2}{3} = \tfrac{6}{3} = 2$$

which is exactly the same as the utility of not switching!

So perhaps the paradox doesn't cut very deep at all—we were just being a bit sloppy with our assumptions about probability. And once we're clearer about the underlying probabilities, the paradox dissolves.

The Paradox Strikes Back

Piet Hein says that "problems worthy of attack prove their worth by fighting back". And the two-envelope paradox is out to prove its worth; it has more trouble in store for us. Our resolution of the first version of the paradox worked. But pay attention to the *reason* that it worked: it was because the ratio between the number of days in the two envelopes (2 to 1) was the same as the odds that 2 was the greater number of days in heaven (2 to 1). But there's no reason that these two ratios have to line up. We can easily change the case so that they come apart. And when we do so, the problem emerges afresh.

Let's imagine that days in heaven are allocated between the envelopes as before: a fair coin is flipped until it lands heads. But this time, if the first heads is on flip n, then there are 3^n days in one envelope and 3^{n-1} days in the other. So now, instead of saying that one of the envelopes has *twice* the number of days in heaven as the other, we are saying that one of the envelopes has *three times* the number in the other.

Just as before, the envelopes are shuffled before being randomly assigned to me and you. So you're just as likely to end up with the larger number as you are to end up with the smaller. So we can represent the possibilities for the contents of your envelope, along with their probabilities, with the table in figure 16.2. In you're in the first row of this table,

smaller	**1** 1/4	**3** 1/8	**9** 1/16	**27** 1/32	**81** 1/64	**243** 1/128	...
larger	**3** 1/4	**9** 1/8	**27** 1/16	**81** 1/32	**243** 1/64	**729** 1/128	...

Figure 16.2

then you have the envelope with the lesser number, and the number below is how much is in my envelope. And if you're in the second row of this table, then you have the envelope with the greater number, and the number above is how much is in my envelope.

Let's think about what would happen, if you were to look inside your envelope. Well, suppose that you saw 1. Then, you'd know for sure that you have the envelope with the smaller number of days, and that there's 3 days in mine. In that case, you should plainly switch. Suppose, on the other hand, that you saw 3. Then, you wouldn't know whether your 3 was the lesser or the greater number, but you would know that it's twice as likely to be the greater amount (since $1/4$ is twice $1/8$). So, if you saw 3 days in your envelope, then there would be a $2/3$ probability that my envelope has 1 day in it and a $1/3$ probability that my envelope has 9 days in it. So, if you saw 3 days in your envelope, then the *expected* utility of switching would be

$$\tfrac{2}{3} \cdot 1 + \tfrac{1}{3} \cdot 9 = \tfrac{2}{3} + 3$$

which is greater than 3 (the known number of days in your envelope). So it's irrational to not switch.

The same thing happens if you instead look in your envelope and see that it has 9 days in heaven. Since the prior probability that 9 would be the lesser number is $1/16$ and the prior probability that 9 would be the larger number is $1/8$, it's twice as likely that 9 is the larger number. So there's a $1/3$ probability that switching would leave you with 27 days in heaven, and a $2/3$ probability that switching would leave you with 9 days in heaven. In expectation, the utility of switching will be

$$\tfrac{1}{3} \cdot 27 + \tfrac{2}{3} \cdot 3 = 9 + 2 = 11$$

which is greater than the utility of keeping your envelope. So again it's irrational to not switch.

And the same thing is true in general. Suppose you look in your envelope and see that it gets you 3^n days in heaven. Then, there will be a $1/3$ probability that switching will get you 3^{n+1} days, and a $2/3$ probability that switching will get you 3^{n-1} days. So, in expectation, the utility of switching will be

$$\tfrac{1}{3} \cdot 3^{n+1} + \tfrac{2}{3}3^{n-1} = 3^n + 2 \cdot 3^{n-2}$$

which is greater than the utility of what's in your envelope: 3^n. So it'll be irrational to keep your envelope *no matter what you see in your envelope*.

So you know that, if you look in your envelope, it'll be irrational for you to keep your envelope, *no matter what you see*. Then, what's the point of even looking? It should be irrational to keep your envelope *now*, before you've looked.

But again—this is *nonsense*. There's a perfect symmetry between the two envelopes. How could it be irrational to keep yours?

And again, if the argument worked, it would work *too well*—we could use exactly the same argument to establish that it is irrational to *switch*. Suppose that you were to look in *my* envelope, instead of yours. Suppose you saw that I had 1 day of heaven in my envelope. Then, you'd know for sure that your envelope has more days, and it would be irrational to switch. Suppose, on the other hand, that you saw 3 days in my envelope. If there's a 3 in one of the envelopes, then it's twice as likely to be the larger number as it is to be the smaller. So, if you see 3 days in my envelope, then there's a $1/3$ probability that 3 is the smaller number, and you've got 9 days in your envelope, and there's a $2/3$ probability that 3 is the larger number, and you've got 1 day in your envelope. In expectation, the utility of your envelope is

$$\tfrac{1}{3} \cdot 9 + \tfrac{2}{3} \cdot 1 = 3 + 2/3$$

which is greater than the utility what's in my envelope. So it's irrational to switch.

And the same thing holds in general. Suppose you look in my envelope and see that it gets you 3^n days in heaven. Then, there will be a $1/3$ probability that you've got 3^{n+1} days in your envelope and a $2/3$ probability that you've got 3^{n-1} days in your envelope. In expectation, the utility of your envelope is

$$\tfrac{1}{3} \cdot 3^{n+1} + \tfrac{2}{3} \cdot 3^{n-1} \quad = \quad 3^n + 2 \cdot 3^{n-2}$$

which is greater than the utility of my envelope. So it's irrational to switch.

So you know that, if you look in my envelope, it'll be rational for you to switch, *no matter what you see*. Then, what's the point of even looking? It should be irrational to switch *now*, even *before* you've looked.

So we've got two seemingly valid arguments leading us to diametrically opposed conclusions. There's also the third argument, which seems to lead us to the correct conclusion—but which looks suspicously similar to the two foregoing arguments: suppose you were

to learn *how the coin landed*. If you learned that it landed heads on the first flip, then you'd know that there's 1 day in one of the envelopes, and 3 days in the other. Since it's just as likely that you have the lesser amount as it is that you have the greater amount, the expected utility of your envelope would be:

$$\tfrac{1}{2} \cdot 1 + \tfrac{1}{2} \cdot 3 \;\; = \;\; 2$$

And since it's just as likely that *I* have the lesser amount as it is that *I* have the greater amount, the expected utility of my envelope would be

$$\tfrac{1}{2} \cdot 3 + \tfrac{1}{2} \cdot 1 \;\; = \;\; 2$$

So the expected utility of your envelope and the expected utility of my envelope are the same. So it's not irrational to switch and it's not irrational to not switch.

This seems like the right conclusion, but we still have our puzzle: what's the difference between these arguments—if the last argument is good, why are the first two arguments bad?

When Our Expectations Fail Us

Perhaps we're going about things the wrong way. Instead of thinking about what the expected utility of switching *would* be, were we to learn this or that, perhaps we should think about the expected utility of switching *before* we've learned anything at all about what's in either envelope.

The expected utility of keeping your envelope will be the probability-weighted average of the number of days in heaven it might contain. Looking at figure 2, we can see that there's a $1/4$ probability that it has 1 day, a $3/8$ probability that it has 3 days in it (since $1/4$ plus $1/8$ is $3/8$), a $3/16$ probability that it has 9 days in it, a $3/32$ probability that it has 27 days in it, and so on and so forth. The general pattern is that, for every $n \geqslant 2$, there's a probability of $3/2^n$ that is has 3^{n-2} days in it. So the expected utility of your envelope is

$$\begin{aligned} EU(\textit{your envelope}) &= \tfrac{1}{4} \cdot 1 + \tfrac{3}{8} \cdot 3 + \tfrac{3}{16} \cdot 9 + \tfrac{3}{32} \cdot 27 + \tfrac{3}{64} \cdot 81 + \cdots + \tfrac{3}{2^n} \cdot 3^{n-2} + \cdots \\ &= \tfrac{1}{4} + \tfrac{9}{8} + \tfrac{27}{16} + \tfrac{81}{32} + \tfrac{243}{64} + \cdots + \tfrac{3^{n-1}}{2^n} + \cdots \\ &= \infty \end{aligned}$$

After $1/4$, each of the terms in this sum is greater than one. So the sum is at least $1+1+1+\ldots$. So the sum is ∞. So the expected utility of your envelope is ∞.

And exactly the same thing goes for *my* envelope. It *also* has an expected utility of ∞, and for exactly the same reason: it has a probability of $1/4$ of containing 1 day in heaven, a probability of $3/8$ of containing 3 days in heaven, a probability of $3/16$ of containing 9 days in heaven, and so on. So its expected utility is also

$$EU(\textit{my envelope}) = \tfrac{1}{4} \cdot 1 + \tfrac{3}{8} \cdot 3 + \tfrac{3}{16} \cdot 9 + \tfrac{3}{32} \cdot 27 + \tfrac{3}{64} \cdot 81 + \cdots + \tfrac{3}{2^n} \cdot 3^{n-2} + \cdots$$

$$= \tfrac{1}{4} + \tfrac{9}{8} + \tfrac{27}{16} + \tfrac{81}{32} + \tfrac{243}{64} + \cdots + \tfrac{3^{n-1}}{2^n} + \cdots$$
$$= \infty$$

Does this solve the puzzle? After all, when I offer you the offer to switch, you should be directly comparing the expected utility of my envelope against the expected utility of yours—if they are equal, then surely either envelope is a rational choice, right?

Not so fast. Just because both of the expected utilities are infinite, this doesn't mean that both options are rationally permissible. Suppose I give you a choice between the following two gambles, whose winnings depend upon how long it takes for a flipped coin to land heads:

	1 flip	2 flips	3 flips	4 flips	⋯
St. Petersburg	\$2	\$4	\$8	\$16	⋯
St. Petersburg+	\$3	\$5	\$9	\$17	⋯

(The general pattern is: St. Petersburg gets you $\$2^n$ if it takes n flips for the coin to land heads, and St. Peterburg+ gets you $\$2^n+1$ if it takes n flips for the coin to land heads.) Both of these gambles have exactly the same expected utility: ∞. But one of them is *guaranteed* to get you \$1 more than the other. In the terminology we learned earlier: St. Petersburg+ *dominates* St. Petersburg! Surely we shouldn't just compare the expected utility of these two options to conclude that each choice is just as rational as the other.

In the case of St. Petersburg and St. Petersburg+, we can't use the expectations of the two options' utilities to compare them. But notice that we *can* compare the two options by taking an expectation of the *difference* in their utilities. No matter how many flips it takes the coin to land heads, St. Petersburg+ will get us 1 more utile than St. Petersburg does. So if we calculate the expected *difference* in the utilities of

$$\begin{aligned} E[U(\textit{St. Petersburg}) - U(\textit{St. Petersburg+})] &= \tfrac{1}{2} \cdot (3-2) + \tfrac{1}{4} \cdot (5-4) + \tfrac{1}{8} \cdot (9-8) + \cdots \\ &= \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 1 + \tfrac{1}{8} \cdot 1 + \tfrac{1}{16} \cdot 1 + \cdots \\ &= \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \tfrac{1}{16} + \tfrac{1}{32} + \cdots \\ &= 1 \end{aligned}$$

Since the expected *difference* between the utilities of the two gambles is positive, perhaps this can tell us that St. Petersburg+ is a more rational choice than St. Petersburg.

Let me tell you about a general feature of expected values, which isn't unique to expected *monetary* values or expected *utilities*. In general, you can find the *expected* value of a quantity, Q (which could be anything—money, utility, height, whatever), by calculating $\sum_q \Pr(Q \text{ is } q) \cdot q$. And, in general, no matter what kind of quantities you're talking about—whenever the expected value of Q_1 is the finite value q_1 and the expected value of Q_2 is q_2, the expected value of $Q_1 - Q_2$ will be $q_1 - q_2$. So, when we're dealing with *finite* expected utilities, we'll have $EU(A) \geqslant EU(B)$ if and only if $E[U(A) - U(B)] \geqslant 0$. So perhaps we should reformulate our normative principle *Maximize Expected Utility* along the following lines:

Maximize Expected Utility (v2) An act, A, is a rational choice if and only if there's no other act, B, such that the expected difference between the utility of B and the utility of A is positive.

That is, A is a rational choice iff, for every available act, B, $E[U(A) - U(B)] \geqslant 0$.

This theory will agree with our old theory in all cases where expected utilities are finite; but it will give us some additional generality when it comes to acts with *infinite* expected utilities.

So perhaps, in order to evaluate the choice between sticking with your envelope and switching for mine, we should be looking at the expected *difference* between the utility of keeping your envelope and the utility of taking mine.

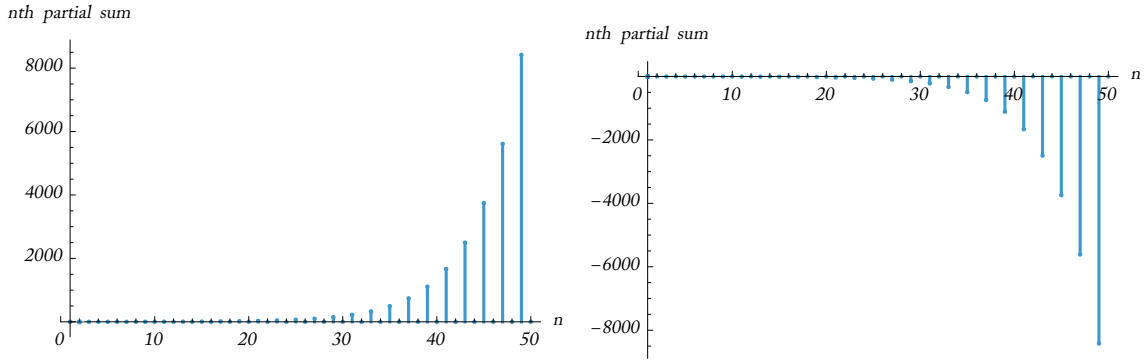
There's a $1/4$ probability that you have 1 and I have 3, and a $1/4$ probability that you have 3 and I have 1. There's a $1/8$ probability that you have 3 and I have 9, and a $1/8$ probability that you have 9 and I have 3, and so on. So the expected *difference* between the utility of your envelope and the utility of mine will be

$$\begin{aligned} E[U(\textit{mine}) - U(\textit{yours})] &= \tfrac{1}{4} \cdot (3-1) + \tfrac{1}{4} \cdot (1-3) + \tfrac{1}{8} \cdot (9-3) + \tfrac{1}{8} \cdot (3-9) + \cdots \\ &= \tfrac{1}{2} - \tfrac{1}{2} + \tfrac{3}{4} - \tfrac{3}{4} + \tfrac{9}{8} - \tfrac{9}{8} + \cdots + \tfrac{3^{n-1}}{2^n} - \tfrac{3^{n-1}}{2^n} + \cdots \end{aligned}$$

At first, you might think that this infinite sum is equal to zero, since each time we add on a value, it is instantly subtracted off again. But remember how we evaluate an infinite sum like this: we look at what happens to the partial sums. If they converge, then the value they converge towards is the value of the infinite sum. But here, the partial sums do *not* converge. The first 50 partial sums are shown in figure 16.3a. Every even partial sum is zero, but the odd partial sums are growing every higher and higher. If we had instead started with the *negative* terms, then every even partial sum would have been zero, but the odd partial sums would have gotten *smaller* and *smaller*. The first 50 partial sums of this alternate infinite sum are shown in figure 16.3b.

Recall from our discussion of the Pasadena paradox that, for some infinite sums, the *order* in which the terms in that sum are taken can make a difference to the infinite sum's value. This is the case in our infinite sum $E[U(\textit{mine}) - U(\textit{yours})]$. We can order the terms in this sum so that the infinite sum converges to *any value we like*.

If you were taking this class in one hundred years time, you'd likely learn about an extension of this method for taking infinite sums and you'd learn about how it allows us to evaluate the options of swapping and not swapping envelopes. Right now, there are several proposals for extending the tools of expected utility theory, and philosophers are debating their merits. (These methods also tell us something about how to evaluate the Pasadena game; many approaches give it an expected utility of $\ln(2)$.) It's an active research program, and it's not clear where it's heading. But suffice it to say that there are *some* proposals on which the infinite sum $E[U(\textit{mine}) - U(\textit{yours})]$ has a value of zero. So if



(a) (b)

Figure 16.3: In figure 16.3a, the first 50 partial sums of the infinite sum $\frac{1}{2} - \frac{1}{2} + \frac{3}{4} - \frac{3}{4} + \frac{9}{8} - \frac{9}{8} + \cdots$. *In figure 16.3b, the first 50 partial sums of the infinite sum* $-\frac{1}{2} + \frac{1}{2} - \frac{3}{4} + \frac{3}{4} - \frac{9}{8} + \frac{9}{8} - \cdots$.

we accept the reformulated principle **Maximize Expected Utility (v2)**, and we accept this new proposals for evaluating infinite sums, we can say that you have no reason to swap envelopes.

So where do the fallacious arguments go wrong?

So is the paradox resolved? Well, if we have a way of evaluating the infinite sums, we can use the reformulated **Maximize Expected Utility (v2)** to say that it's not irrational to switch or not switch. But what about the arguments we considered earlier? Where did they go wrong?

Recall: we saw that, if you looked in *your* envelope, then *my* envelope would be expected to get a better outcome than yours—*no matter what you saw*. Then, we reasoned that there's no point in looking—you can just infer straight away that it's irrational to not swap. There were two relevant premises:

P1. If you were to learn that contents of your envelope, then it would be irrational to not swap (no matter what you learned)

P2. If it would be irrational to not swap, were you to learn about the contents of your envelope (no matter what you learned), then you can reason to a foregone conclusion: it is irrational to not swap, even if you haven't learned what's in your envelope.

∴ C. It is irrational to not swap.

It does not seem that we can deny P1. Our calculations of the expected utility of swapping and not swapping, conditional on the contents of your envelope, were not wrong. There's no mathematical error there. And none of those calculations involved any infinite sums. So the ordinary theory of *Maximize Expected Utility* says that it would be irrational

to not swap, if you were to look inside your envelope. It does not matter what you see inside.

The argument is valid, and we have to reject the conclusion. We cannot reject the first premise without rejecting *Maximize Expected Utility*. So if we are expected utility maximizers, then we must reject the second premise. We must think that there's something wrong with reasoning to a foregone conclusion. If you were to look in your envelope, you would be rationally required to swap envelopes. But before looking, there's nothing irrational about keeping your envelope.

Similarly, we saw that, if you looked in *my* envelope, then *your* envelope would be expected to have a better outcome than mine—*no matter what you saw*. So we have the parallel argument:

P1. If you were to learn that contents of my envelope, then it would be irrational to swap (no matter what you learned)

P2. If it would be irrational to swap, were you to learn about the contents of my envelope (no matter what you learned), then you can reason to a foregone conclusion: it is irrational to swap, even if you haven't learned what's in my envelope.

∴ C. It is irrational to swap.

Again, this argument is valid, but we must reject its conclusion. So we have to reject one of its premises. But we cannot reject the first premise. So it looks like we have to reject the second premise. For some reason, we can't reason to a foregone conclusion.

Notice another consequences of rejecting the second premise: Suppose that, in exchange for swapping envelopes, I've asked you to give me some insubstantial sum of money—something which is negligible compared to a day in heaven. Then, you have the option of whether to make the choice of whether to pay and swap right here and now, or instead look inside your envelope and *then* make the choice of whether to pay and swap. The number of days inside your envelope is surely relevant information. And it's cost-free to acquire this relevant information before making up your mind about whether or not to swap envelopes. But you know for sure that, if you look inside your envelope, then you'll pay to swap with me. So when you're deciding whether or not to look inside your envelope, you're deciding between paying to swap and not swapping. And *right now* (before looking) you think that it's irrational to pay to swap with me. So you'll also think that it's irrational to look inside your envelope. So you'll have *pragmatic* reason to ignore and pass up cost-free evidence—not because you think that you'd make any *mistake*, if you were to gather this relevant cost-free evidence. You think that, if you were to look in your envelope, then you'd *rationally* pay to swap envelopes. It's just that, before you've looked, you think it would be irrational to allow yourself to end up in that position. That's a very surprising and paradoxical kind of situation to find yourself in.

17 | Act State Dependence

In the first half of this decade, a group of former Rationalists turned 'vegan anarchotranshumanists' known as the Zizians (after the adopted name of their leader, Ziz LaSota) went on a killing spree. They stopped paying their landlord, then called the landlord to fix a leak, and when he arrived, they stabbed him somewhere around 50 times, blinding him in his right eye, and left him impaled on a samuri sword. Later, he was found with his throat slit. Other Zizians killed parents and relatives, and two shot and killed a border patrol agent.

One of the things that drove the Zizians to engage in this behavior was their take on the philosophical thought experiment that we're going to talk about in this chapter.

Two Kinds of Independence

Recall: when we were learning about how to formalize a decision with a decision matrix, we said: it is important that you use states and acts which are *independent from each other*. I asked you to consider the following example:

Example 46. *The king has declared war, and tomorrow you march out to battle. You can either use your life savings to buy armor or else march out to battle unprotected.*

Here, it would be a *mistake* to use the states *you survive the battle* and *you do not survive the battle*, and model your decision with the following decision matrix

	You survive the battle	You do not survive the battle
You buy armor	Lose life savings	Lose life & life savings
You don't buy armor	Lose nothing	Lose life

Looking at this decision matrix, you might reason as follows: buying armor gets me a worse outcome than not buying armor in *every* state, so buying armor is *dominated* by not buying armor. So it's irrational to buy armor.

This reasoning would be fallacious—surely you have *excellent* reason to purchase the armor, given that it makes it much more likely that you'll survive the battle.

Consider other cases like this:

Example 47. *'Crazy Joe' Gallo offers you 'insurance' for your store, telling you "it sure would be a shame if anything happened to it". The insurance costs $20 a month, but covers nothing.*

	Store burgled	Store not burgled
You don't buy 'insurance'	−$10,000	$0
You buy 'insurance'	−$10,020	−$20

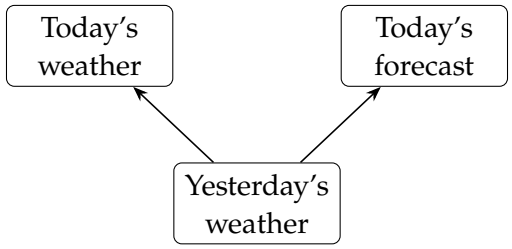
Again, you might be tempted to reason as follows: buying insurance always makes me $20 poorer, *no matter whether the store is burgled or not*. So not buying the insurance *dominates* buying it, and it would be irrational to buy the insurance.

And again, this reasoning would be fallacious—this is a *threat*. Crazy Joe is going to burgle your store unless you buy the 'insurance'. Given that, you have every reason to pay the $20.

So it's quite important that we have *acts* and *states* which are independent of each other. Today, I want to dwell for a bit on the fact that there's two importantly different senses in which acts and states can be independent. And so there's two different ways of understanding the demand that acts and states be independent:

1. Acts and states must be *probabilistically* independent. Which act you choose cannot make a *probabilistic* difference to which state you are in. That is: for every act A and state S,
$$\Pr(S \mid A) = \Pr(S)$$
2. Acts and states must be *causally* independent. That is, which act you choose cannot *causally influence* which state you are in.

Probabilistic and causal independence are different notions, because correlation is not causation. To see how they come apart, think about the relationship between the weather report and the weather. Whether it rains is probabilistically dependent upon whether there's a forecast of rain—a forecast of rain makes it more likely that it will rain. But that's not because the forecast *causes* rain, and it's not because the rain causes the forecast (the forecast comes before the rain; and there's no funny backwards causation). Instead, the forecast and the rain have a *common cause*: they are both caused by the prior weather conditions (things like atmospheric pressure). This common cause explains the correlation between the forecast and the rain.



Evidential Decision Theory

The philosopher Richard Jeffrey had a solution to problems like examples 46 and 47, which builds on the first (probabilistic) understanding of independence. He thought that we needed to incorporate the evidence an act gives about which state you're in. In general, you should incorporate new evidence by *conditioning* on it. So Jeffrey recommended that, instead of using the *unconditional* probability of each state, we should instead use the *conditional* probability of each state, *given* that you've selected each act. These conditional probabilities will tell us how your choice of act changes the probabilities of each of the states.

Figure 17.1: Richard Jeffrey

Because Jeffrey is focused on what kind of *evidence* your choice of act gives you, this theory is known as *evidential* decision theory. Evidential decision theory tells you to maximize *conditional* expected utility (where the thing you're conditioning on is which act you've selected).

Evidential Expected Utility The *evidential expected utility* of an act, A, is

$$EEU(A) \stackrel{\text{def}}{=} \sum_S \Pr(S \mid A) \cdot U(o_{A,S})$$

Recall, we're using '$o_{A,S}$' for the outcome that act A leads to in the state S. Notice that, if states are probabilistically independent of acts, evidential expected utility is the same as expected utility.

Evidential Decision Theory An act, A, is a rational choice if and only if there is no other act available which has a greater evidential expected utility than A does.

Example 48. *Let's suppose that, in example* 46*, your* conditional *probabilities are as shown below. (Here, the value in the matrix is the* conditional *probability of the column state, given that you've made the choice in the row.*)

	You survive the battle	*You do not survive the battle*
You buy armor	9/10	1/10
You don't buy armor	1/10	9/10

And let's suppose that your utilities for the various outcomes are as shown below:

	You survive the battle	*You do not survive the battle*
You buy armor	−10	−110
You don't buy armor	0	−100

Then, explain why evidential decision theory says that it is irrational to not buy the armor.

Conditional on you buying armor, odds are 9 to 1 that you survive, so the evidential expected utility of buying armor is:

$$\begin{aligned} EEU(\textit{buy armor}) &= \Pr(\textit{you survive} \mid \textit{buy armor}) \cdot U(\textit{lose life savings}) \\ &\quad + \Pr(\textit{you don't survive} \mid \textit{buy armor}) \cdot U(\textit{lose life and life savings}) \\ &= \tfrac{9}{10} \cdot (-10) + \tfrac{1}{10} \cdot (-110) \\ &= -9 - 11 \\ &= -20 \end{aligned}$$

And conditional on you not buying armor, odds are 9 to 1 that you do not survive. So the evidential expected utility of not buying armor is:

$$\begin{aligned} EEU(\textit{don't buy armor}) &= \Pr(\textit{you survive} \mid \textit{don't buy armor}) \cdot U(\textit{lose life savings}) \\ &\quad + \Pr(\textit{you don't survive} \mid \textit{don't buy armor}) \cdot U(\textit{lose life and life savings}) \\ &= \tfrac{1}{10} \cdot (0) + \tfrac{9}{10} \cdot (-100) \\ &= 0 - 90 \\ &= -90 \end{aligned}$$

Since the evidential expected utility of buying armor is higher than the evidential expected utility of not buying armor, evidential decision theory says that buying armor is the only rational choice.

Exercise 45. *Let's suppose that, in example 47, your utilities are linear in dollars ($U(\$x) = x$), and that your conditional probabilities are as shown below (again, the entries in this matrix give the conditional probability of the column state, given that you've selected the row act.)*

	Store burgled	*Store not burgled*
You buy 'insurance'	0	1
You don't buy 'insurance'	1	0

Explain why evidential decision theory says that it's irrational to not buy the 'insurance'.

Newcomb's Problem

$EEU(A)$ is a measure of the 'news value' of choosing A. It tells you how glad you'd be to learn that you selected A. Jeffrey's advice is to select whichever act has the greatest news value. As he puts it:

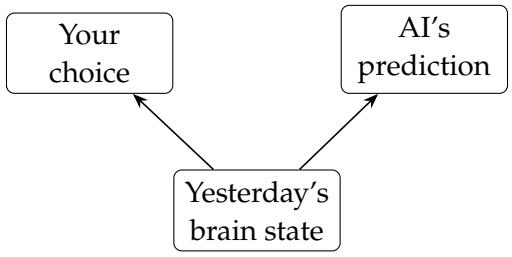
> there is no effective difference between asking whether you prefer A to B as a news item or as an act, for you make the news.[1]

[1] Richard Jeffrey, *Logic of Decision* (1965, New York, NY, USA: University of Chicago Press), p. 84.

However, some thought that there *is* an important difference between these two questions, and it is related to the distinction between our two kinds of independence: *probabilistic* and *causal* independence. We saw earlier that, when two effects have a common cause, we can have probabilistic dependence without causal dependence. So let's consider a case in which *how you choose* has a common cause with some state of the world:

Example 49 (Newcomb's Problem)**.** *There are two boxes before you*: *one transparent and one opaque* (*the 'mystery box'*). *In the transparent box is* $1000. *In the mystery box is either* $1,000,000 *or nothing. You have two options. You can either take both boxes and receive their contents* (*'two box'*), *or you can take only the mystery box* (*'one box'*). *Yesterday, we fed a scan of your brain to a superintelligent AI and it made a prediction about how you would choose in this decision. If it predicted that you would one box, then I put* $1,000,000 *in the mystery box. If it predicted that you would two box, then I left the mystery box empty.* 80% *of the time people one box, the AI correctly predicted that they would one box*; *and* 80% *of the time they two box, it correctly predicted that they would two box.*

The causal structure of Newcomb's problem is shown below. How you choose makes no difference to the AI's prediction, but because your choice and the AI's prediction have a common cause (your brain state yesterday), your choice is not probabilistically independent of the AI's prediction.



Let's suppose that your conditional probabilities and utilities for this decision are given below.

	AI predicted one-box	AI predicted two-box
One-box	1,000,000	0
Two-box	1,001,000	1000

(a) Utilities

	AI predicted one-box	AI predicted two-box
One-box	$8/10$	$2/10$
Two-box	$2/10$	$8/10$

(b) Probabilities

Exercise 46. *Explain why evidential decision theory says that it is irrational to take both boxes* (*'two box'*) *in Newcomb's Problem.*

When people first encounter Newcomb's Problem, there are two reactions they have:

1. First reaction: Taking just the one box makes it *much more likely* that I will walk away a millionaire. So I should leave the transparent box behind, and one box.

2. Second reaction: The million dollars is either there in the mystery box or it isn't; and nothing I do now will make any difference to whether it's there or not. Taking the transparent box will make be $1,000 richer no matter what. So I should take the transparent box, too (that is: two box).

In defense of the first reaction: think about what would happen, if multiple people were to play this game, over and over again, for a long period of time. You'd expect that, of the people who one-boxed, about 80% would walk out with $1,000,000 and about 20% would walk out with nothing. On average, that's a net gain of $800,000. And you'd expect that, of the people who two-boxed, about 80% would walk out with $1,000 and about 20% would walk out with $1,001,000. On average, that's a net gain of $201,000. So, on average, one boxers are making more money than two-boxers are. Shouldn't your goal be to make as much money as possible? So shouldn't you one box?

In defense of the second reaction: imagine that your mother was allowed to look inside the mystery box and give you advice about how to choose. What advice would she give? Well, if she saw that the one box contained the million, then she'd advise you to take both—taking both would get you $1,001,000, rather than $1,000,000. On the other hand, if she saw that the one box contained nothing, then she'd advise you to take both—taking both would get you $1000, rather than nothing. So your mother would advise you to two box, no matter what she saw. Why not listen to her?

Causal Decision Theory

Philosophers disagree about what the right solution to Newcomb's Problem is; but most decision theorists think that leaving a free $1,000 behind, when doing so has no effect whatsoever on the amount in the mystery box, is irrational. So many of them grew dissatisfied with evidential decision theory, and sought a replacement. (Richard Jeffrey agreed that you should two box in Newcomb's Problem; though he thought he could save evidential decision theory through other means.)

These philosophers developed an alternative theory which is known as *causal decision theory*. Papering over some differences between causal decision theorists, the key idea is that you should only maximize expected utility over *the right kind of states*—what we can call *states of nature*. A state of nature is one which is causally independent of how you choose; how you choose has no causal influence over which of these states obtain. Following standard convention, let's use the letter '*K*', rather than '*S*', to signal that we're talking about states of nature.

Causal Expected Utility The *causal expected utility* of an act, A, is

$$CEU(A) \stackrel{\text{def}}{=} \sum_K \Pr(K) \cdot U(o_{A,K})$$

where the Ks in the sum are *states of nature*

Causal Decision Theory An act, A, is a rational choice if and only if there is no other act available which has a greater causal expected utility than A does.

Example 50. *In example* 46*, the states* You survive the battle *and* You do not survive the battle *are not states of nature. Whether you buy armor or not causally influences which of these states obtains. So we need to find new states. For instance, we could use these*:

K_1 : *You would survive the battle, with or without armor*

K_2 : *You would survive the battle with armor, but not without armor*

K_3 : *You would not survive the battle with armor, but you would survive with armor*

K_4 : *You would not survive the battle, with or without armor.*

Then, if we use the same utilities from before, where the utility of losing your life savings is -10 *and the utility of losing your life is* -100*, then we will have the following decision matrix*:

	K_1: *Survive either way*	K_2: *Survive w/ armor only*	K_3: *Die w/ armor only*	K_4: *Die either way*
You buy armor	-10	-10	-110	-110
You don't buy armor	0	-100	0	-100

And let's suppose that your probability distribution over these states is given below:

$$\Pr(K_1) = 1/10 \qquad \Pr(K_2) = 8/10$$
$$\Pr(K_3) = 0 \qquad \Pr(K_3) = 1/10$$

Then, explain why causal decision theory says that it is irrational to not buy the armor.

Causal decision theory will begin by calculating the causal expected utility theory of each available act:

$$\begin{aligned} CEU(buy) &= \Pr(K_1) \cdot (-10) + \Pr(K_2) \cdot (-10) + \Pr(K_3) \cdot (-110) + \Pr(K_4) \cdot (-110) \\ &= 1/10 \cdot (-10) + 8/10 \cdot (-10) + 0 \cdot (-110) + 1/10 \cdot (-110) \\ &= -1 - 8 - 11 \end{aligned}$$

$$= -20$$

whereas

$$\begin{aligned} CEU(\textit{don't buy}) &= \Pr(K_1) \cdot (0) + \Pr(K_2) \cdot (-100) + \Pr(K_3) \cdot (0) + \Pr(K_4) \cdot (-100) \\ &= 1/10 \cdot (0) + 8/10 \cdot (-100) + 0 \cdot (0) + 1/10 \cdot (-100) \\ &= -80 - 10 \\ &= -90 \end{aligned}$$

Since the causal expected utility of buying is greater than the causal expected utility of not buying, causal decision theory says that it is irrational to not buy.

Exercise 47. *Suppose that, in example* 47*, your utilities are linear in dollars, suppose you're certain that 'Crazy Joe' Gallo will leave the store alone if you buy the insurance, and suppose you're* 90% *confident that 'Crazy Joe' Gallo will burgle the store if you don't buy it. Then, find a collection of states of nature for the decision, and explain why causal decision theory says that you should purchase the 'insurance'.*

What about Newcomb's problem? Well, notice that the two states *AI predicted one-box* and *AI predicted two-box* are *states of nature*—nothing you do now will make any difference to what the AI predicted yesterday. So causal decision theory will evaluate the options of one boxing and two boxing with:

$$\begin{aligned} CEU(\textit{one-box}) &= \Pr(\textit{AI predicted one-box}) \cdot 1,000,000 + \Pr(\textit{AI predicted two-box}) \cdot 0 \\ CEU(\textit{two-box}) &= \Pr(\textit{AI predicted one-box}) \cdot 1,001,000 + \Pr(\textit{AI predicted two-box}) \cdot 1000 \end{aligned}$$

It turns out that the causal expected utility of two boxing will be greater than the causal expected utility of one boxing, no matter how probable it is that the AI predicted you'd one box or two box. Just for illustration, suppose that you're 50-50 about the AI's predictions. Then,

$$\begin{aligned} CEU(\textit{one-box}) &= 1/2 \cdot 1,000,000 + 1/2 \cdot 0 \\ &= 500,000 \\ CEU(\textit{two-box}) &= 1/2 \cdot 1,001,000 + 1/2 \cdot 1000 \\ &= 500,500 + 500 \\ &= 501,000 \end{aligned}$$

So two boxing has a greater causal expected utility that one boxing does, and causal decision theory says that it is irrational to one box.

Summary: Evidential and Causal Decision Theory

According to *evidential* decision theory, you should maximize *evidential* expected utility. Whereas, according to *causal* decision theory, you should maximize *causal* expected utility.

Evidential Expected Utility The *evidential expected utility* of an act, A, is

$$EEU(A) \stackrel{\text{def}}{=} \sum_{S} \Pr(S \mid A) \cdot U(o_{A,S})$$

Causal Expected Utility The *causal expected utility* of an act, A, is

$$CEU(A) \stackrel{\text{def}}{=} \sum_{K} \Pr(K) \cdot U(o_{A,K})$$

where the Ks in the sum are *states of nature* (states which do not causally depend upon how you choose)

The main differences between the views:

1. For *evidential* decision theory, you use *conditional* probabilities, $\Pr(- \mid A)$. Whereas, for *causal* decision theory, you use *unconditional* probabilities, $\Pr(-)$.

2. For *causal* decision theory, you must use *states of nature*—states which are causally independent of how you choose. Whereas, for *evidential* decision theory, you can use *any* states, bar none.

Evidential and Causal Dominance

Recall that, earlier in the course, we learned about the notion of *dominance*. We said that one act, A, *dominated* another act, B, iff A leads to a better outcome than B does *in every state*.

Two-boxers often appeal to a principle of dominance to justify their choice in Newcomb's problem. After all, they allege: taking two boxes will leave you with \$1,000 more than one boxing will, *no matter which state you're in*.

	AI predicted one-box	AI predicted two-box
One-box	1,000,000	0
Two-box	1,001,000	1000

However, notice that it matters here that we used *states of nature* in our matrix representation of the decision. Think about the following, alternate decision matrix:

	AI predicted correctly	AI predicted incorrectly
One-box	1,000,000	0
Two-box	1,000	1,001,000

In *this* decision matrix, two-boxing does not dominate one-boxing. The difference between these two sets of states is that the first ones (*predicted one box* and *predicted two box*) are *causally* but not *probabilistically* independent of your choice; whereas the second ones (*predicted correctly* and *predicted incorrectly*) are *probabilistically* but not *causally* independent of your choice.

Corresponding to these two different kinds of states are two different principles of dominance:

Causal Dominance *A causally dominates B* if and only if there's a partition of states which are *causally* independent of your choice (*states of nature*) such that, for every state in that partition, *A* leads to a better outcome in that state than *B* does.

Evidential Dominance *A evidentially dominates B* if and only if there's a partition of states which are *probabilistically* independent of your choice such that, for every state in that partition, *A* leads to a better outcome in that state than *B* does.

Unfortunately, the principle of dominance doesn't settle the debate between causal and evidential decision theorists on its own. For they each interpret the principle of dominance differently.

Does it matter for the real world which view is right?

We've seen that the difference between causal and evidential decision theory makes a difference in Newcomb's Problem. But this decision seems rather recherché. Do the differences between causal and evidential decision theory matter for any other decisions?

I believe that the answer is 'yes'. Let's talk about a few of those, starting with the more supernatural and descending to the more mundane. Firstly, let's talk about Calvinist theology. The Calvinists believe that, at the beginning of time, God selected some individuals—the elect—for eternal salvation, and condemned the others—the non-elect—to eternal damnation. According to this doctrine of predestination, nothing that you do in this life makes any difference to whether or not you are elect. Your acts in this life are not causally relevant to whether or not you are elect. However, if you find yourself selecting good acts, then this is good evidence that you are among the elect. And if you find yourself selecting bad acts, then this is good evidence that you are not among the elect.

So for Calvinists, causal and evidential decision theory give radically different advice. Suppose that you could either be pious or live a life of sin, and that you'd prefer the life of sin to the life of piety. Of course, the most important thing to you is attaining eternal salvation. You'd much rather live a life of piety and thereafter achieve eternal salvation

than live a life of sin and thereafter achieve eternal damnation. Then causal decision theory says that you should live the life of sin. *You are elect* and *You are not elect* are states of nature. If you are elect, then nothing you do now will change that. And if you are not elect, then nothing you do now will change that. So the causal decision theorist can use this decision matrix:

	You are elect	You are not elect
Life of Piety	bad life, great afterlife	bad life, terrible afterlife
Life of Sin	good life, great afterlife	good life, terrible afterlife

A life of sin gets you a better outcome than a life of piety *in every state of nature*. So the causal decision theorist says that it is irrational to live a life of piety.

On the other hand, living a life of piety gives you the evidence that you are elect, and that you have a great afterlife in store, and living a life of sin gives you evidence that you are not elect, and that you have a terrible afterlife in store. Assuming the evidence is strong enough, evidential decision theorists will advise you to live a life of piety.

Less theologically: there are all kinds of decisions in day to day life in which how *you* choose can give you evidence about how others will choose, even though your choices do not causally influence the choices of others.

For instance, think about national elections. It's astronomically unlikely that a national election is going to come down to a single vote. So it's astronomically unlikely that your vote is going to make a difference to the outcome of the election. Now, from fact alone we can't conclude that voting is irrational—for several reasons. Firstly, because you might care about more than just who wins the election. You might also care about *doing your duty as a citizen*. If so, then voting can be a way of getting the thing you care about, even if you know for sure that you'll make no difference to the election's outcome. Secondly, even if the probabilities are tiny, the *utility* of making a difference to the outcome of a national election can be massive. A tiny probability times a massive utility can be worth your while.

However, let's suppose—just for illustration—that you don't care about doing your civic duty. You only care about who wins the election. And let's suppose that the utilities are small enough that they don't overcome the tiny probabilities all by themselves. Even so, one of the main things that makes a difference in elections is *turn out*. The electorate ends up being evenly split most elections; so if Trump supporters are more excited about Trump than Harris supporters are excited about Harris, then more of the Trump supporters are going to get off the couch to vote than Harris supporters, and Trump will end up carrying the day. Before election day, it's hard to know how many people are going to go out and vote. But your willingness to go out and vote gives you some indication of this. If you're a Harris voter and you're inspired enough to go vote for her, that's evidence that other people like you are also going to be inspired enough to go vote for her. If you think she's the best candidate but you let yourself get distracted by other things and don't end up voting, then that's evidence that other people like you are also going to stay home.

Of course, most of the voters don't know you and won't know whether or not you show up to the polls to vote. So we have a case of non-causal correlation. Showing up at the polls to vote is very unlikely to make any difference to which candidate wins, but it will nonetheless give you some *evidence* about who will win. Finding yourself at the polls is *good news*, even though your vote is almost certainly not going to affect the outcome of the election.

If you're an evidential decision theorist, this could be reason enough to get out to the polls on election day. If, however, you're a causal decision theorist, then it will not be enough of a reason to get out to the polls on election day.

For another case, suppose that a company's CEO supports a political cause you find objectionable, and you're considering whether to boycott the company. To make things concrete, suppose that you've learned that the CEO of Chic-Fil-A is donating money to California's Proposition 8 (which was an attempt to ban same sex marriage in California). You like Chic-Fil-A's sandwiches, so you'd like to remain a customer. And you don't have any moral objections to remaining a customer *per se*—it's not that you think it would be immoral for you to continue eating Chic-Fil-A sandwiches. But you'd like the CEO to stop donating to Proposition 8, and you think that a successful boycott would achieve this.

Of course, you are just a single solitary customer and Chic-Fil-A is a large corporation. Your individual boycott won't make any difference to them. But if *many* people join in the boycott, then it will make a difference. For the most part, whether you join in the boycott isn't going to influence whether other people do. But, if you find yourself joining in the boycott, that's excellent *evidence* that other people like you are going to join in the boycott—just like in the national election. So we have a non-causal correlation between your choices and other people's choices. And that's enough for evidential and causal decision theory to give different recommendations (assuming we fill in the details about your utilities in the right way). Evidential decision theorists will say that it is rational to boycott Chic-Fil-A, and causal decision theorists will say that it is not—not, at least, unless you care about supporting the boycott for its own sake, and not for the causal consequences of your contribution to the boycott.

Playing Games with a Twin

In general, the differences between causal and evidential decision theory come out when we consider strategic interactions between you and other people who are similar to you. In those cases, it can happen that how *you* choose gives you evidence about how *they* will choose.

For instance, think about the 'prisoner's dilemma', which we discussed earlier in the class.

Example 51 (Prisoner's Dilemma with a Twin)**.** *You and your twin cheated on a problem set using AI. You are each called into the dean's office and placed in separate rooms. You are then*

informed of the following: *you are strongly suspected of cheating along with your accomplice. We are now going to give you both an opportunity to confess. If neither of you confess, then you will both fail the assignment, but neither will fail the course. If you confess and your twin does not, then you will not be punished at all, and your accomplice will be expelled. If your twin confesses and you do not, then they will not be punished, and you will be expelled. If you both confess, then you will both fail the course.*

You and your twin are very *similar to each other. If you find yourself confessing, this is very strong evidence that your twin will confess, too. And if you find yourself not confessing, this is very strong evidence that your twin won't confess, either.*

	Twin confesses (T)	Twin does not confess ($\sim T$)
You confess (Y)	You fail the course (2)	You are not punished (4)
You do not confess ($\sim Y$)	You are expelled (1)	You fail the assignment (3)

Normally, in a prisoner's dilemma, we suppose that how you choose doesn't make any difference (probabilistically or causally) to how *your accomplice* choose. But here, we've made a different supposition: we've supposed that your twin is very similar to you. So if you find yourself deciding to confess, that makes it more likely that your twin is going to confess, too. And if you find yourself deciding to not confess, that makes it less likely that your twin is going to confess.

For the causal decision theorist, this makes no difference. The states *Twin confesses* and *Twin doesn't confess* are states of nature—how you choose doesn't make any causal difference to whether your twin confesses. So confessing *causally* dominates not confessing. So causal decision theory says that you should confess.

But for the evidential decision theorist, it makes a difference. If we use 'T' for the proposition that your twin confesses and 'Y' for the proposition that you confess, then just for illustration, let's suppose that you have the following conditional probabilities:

$$\Pr(T \mid Y) = \tfrac{4}{5} \qquad \Pr(T \mid \sim Y) = \tfrac{1}{5}$$
$$\Pr(\sim T \mid Y) = \tfrac{1}{5} \qquad \Pr(\sim T \mid \sim Y) = \tfrac{4}{5}$$

Then, the *evidential* expected utility of confessing will be

$$\begin{aligned} EEU(Y) &= \Pr(T \mid Y) \cdot U(o_{Y,T}) + \Pr(\sim T \mid Y) \cdot U(o_{Y,\sim T}) \\ &= \tfrac{4}{5} \cdot 2 + \tfrac{1}{5} \cdot 4 \\ &= \tfrac{8}{5} + \tfrac{4}{5} \\ &= \tfrac{12}{5} \end{aligned}$$

while the evidential expected utility of not confessing will be

$$\begin{aligned} EEU(\sim Y) &= \Pr(T \mid \sim Y) \cdot U(o_{\sim Y,T}) + \Pr(\sim T \mid \sim Y) \cdot U(o_{\sim Y,\sim T}) \\ &= \tfrac{1}{5} \cdot 1 + \tfrac{4}{5} \cdot 3 \end{aligned}$$

$$= \tfrac{1}{5} + \tfrac{12}{5}$$
$$= \tfrac{13}{5}$$

Since $13/5 > 12/5$, the evidential decision theorist says that it is irrational to confess.

We know that the causal decision theorist will accept the dominance reasoning and say that you should confess on this basis. But let's think through how to calculate causal expected utilities in this case. Thus far, we've only stipulated the *conditional* probabilities for your twin confessing, given that you confess. The law of total probability allows us to determine, from these, the *unconditional* probability that your twin confesses. We just need to know how likely it is that *you* will confess:

$$\Pr(T) = \Pr(T \mid Y) \cdot \Pr(Y) + \Pr(T \mid \sim Y) \cdot \Pr(\sim Y)$$

Suppose that you're 50% likely to confess, $\Pr(Y) = 1/2$. Then,

$$\Pr(T) = \tfrac{4}{5} \cdot \tfrac{1}{2} + \tfrac{1}{5} \cdot \tfrac{1}{2} = \tfrac{4}{10} + \tfrac{1}{10} = \tfrac{5}{10} = \tfrac{1}{2}$$

So we can now calculate the causal expected utility of you confessing,

$$\begin{aligned} CEU(Y) &= \Pr(T) \cdot U(o_{Y,T}) + \Pr(\sim T) \cdot U(o_{Y,\sim T}) \\ &= \tfrac{1}{2} \cdot 2 + \tfrac{1}{2} \cdot 4 \\ &= 1 + 2 = 3 \end{aligned}$$

and you not confessing,

$$\begin{aligned} CEU(\sim Y) &= \Pr(T) \cdot U(o_{\sim Y,T}) + \Pr(\sim T) \cdot U(o_{\sim Y,\sim T}) \\ &= \tfrac{1}{2} \cdot 1 + \tfrac{1}{2} \cdot 3 \\ &= \tfrac{1}{2} + \tfrac{3}{2} \\ &= 2 \end{aligned}$$

Since $3 > 2$, the causal expected utility of confessing is greater than the causal expected utility of not confessing. And causal decision theory advises you to confess.

Being the Kind of Chooser for Whom Things Go Best

There are some people—known as *functional decision theorists*—who think that both causal and evidential decision theory have gotten things wrong. Evidentialists are guided by the slogan 'give yourself the best news possible!'. Causalists are guided by the slogan 'do whatever will bring about the best outcome!' The functional decision theorists are guided by the slogan 'be the kind of chooser for whom things go best!'.

The functional decision theorists say that you should take one box in Newcomb's problem—but *not* because doing so gives you better evidence than not. Instead, they think that you

should take one box in Newcomb's problem because people who take one box tend to make more money than people who take two.

I won't go through the mathematics of functional decision theory. But to illustrate how it differs from evidential decision theory, it's worth thinking about a different version of Newcomb's Problem:

Example 52 (Newcomb's Problem without Mystery)**.** *Everything is as in Newcomb's Problem, except that there's no longer any mystery*: *both boxes are transparent*. (*In this version of the case, the AI's prediction was about what you would do when you were able to see into both boxes*; *it was not a prediction about what you would do in the original Newcomb Problem*).

This is really *two* decisions, depending upon what you see in the boxes. You could walk into the room and find a million dollars in one box and a thousand dollars in the other. In this case, you face a decision between a guaranteed million and a guaranteed million and a thousand. Alternatively, you could walk into the room and find one box empty and another box with a thousand dollars in it. In that case, you face a decision between a guaranteed thousand and a guaranteed nothing.

In *Newcomb's Problem without Mystery*, both evidential decision theorists and causal decision theorists agree that you should take both boxes—you should take all of the money in the room. But functional decision theorists think that this is a mistake. Imagine someone who was disposed to *always* take just the one box—even when they can see what's inside of it. Such a person would have been predicted to take just the mystery box. So such a person would predictably walk into a room that contains a million and a thousand dollars, and predictably walk out with a million dollars. In contrast, both causal and evidential decision theorists will predictably walk into a room that contains a thousand dollars, and predictably walk out with a thousand dollars.

So the kind of chooser who takes just one box, *even when* they can see what's inside both boxes, is the kind of chooser who ends up with the most money. For this reason, functional decision theorists say that you should take one box *even in example 52.*

In a sense, the functional decision theorists have turned that 'why ain'cha rich?' argument against the evidential decision theorists. They have pointed out that there are situations in which a different kind of chooser will predictably end up richer than evidential decision theorists will. And since they endorse the 'why ain'cha rich?' argument, functional decision theorists advise you to be *that* kind of chooser. In a slogan, functional decision theorists say that you should be the kind of chooser for whom things go best.

The Zizians were a group of people influenced by functional decision theorists, though I think it's fair to say that they radically misunderstood and misapplied that theory. Rather than asking themselves: who are the kinds of choosers for whom things go best?, they asked themselves: what's the way of choosing that would make things go best for everyone? And they said that it was rational for each individual to make *that* choice. It was partly because they thought things would work out best if everyone pre-committed to stabbing a landlord who came to collect rent that they ended up doing just that.

Instability in Causal Decision Theory

Game Theory studies decision-making in strategic situations, where two or more actors are making decisions, and the outcomes each actor gets depends upon which choices are made by the other actors. *Prisoner's Dilemma* is a classic kind of two-person game. Another classic two-person game is *Chicken*.

Example 53 (Chicken)**.** *You and your adversary are charging towards each other in your cars. You can either swerve or drive straight, and your adversary can either swerve or drive straight. If you swerve while they drive straight, then you will be ashamed and lose the respect of your beloved. If you drive straight while they swerve, then you will win glory and the love of your beloved. If you both drive straight, you will both die. If you both swerve, you will both save face.*

	Adversary drives straight (A)	*Adversary swerves* ($\sim A$)
You drive straight (Y)	*Death* (-100)	*Glory* (50)
You swerve ($\sim Y$)	*Shame* (-50)	*Status Quo* (0)

If your adversary drives straight, then it's best for you to swerve. But if your adversary swerves, then it's best for you to drive straight.

Game theory has lots of tools for dealing with decisions like these, but if we just reach for decision theory on its own, it tells us: what you should do depends upon what you think your adversary is most likely to do. If we assume that these states (*Adversary drives straight* and *Adversary swerves*) are both causally and probabilistically independent of how you choose, then we just need to compare the expected utility of Y with the expected utility of $\sim Y$.

$$
\begin{aligned}
EU(Y) &\overset{?}{\leqslant} EU(\sim Y) \\
\Pr(A) \cdot U(\textit{Death}) + \Pr(\sim A) \cdot U(\textit{Glory}) &\overset{?}{\leqslant} \Pr(A) \cdot U(\textit{Shame}) + \Pr(\sim A) \cdot U(\textit{Status Quo}) \\
\Pr(A) \cdot (-100) + [1 - \Pr(A)] \cdot (50) &\overset{?}{\leqslant} \Pr(A) \cdot (-50) + [1 - \Pr(A)] \cdot (0) \\
-100 \Pr(A) + 50 - 50 \Pr(A) &\overset{?}{\leqslant} -50 \Pr(A) \\
-100 \Pr(A) + 50 &\overset{?}{\leqslant} 0 \\
50 &\overset{?}{\leqslant} 100 \Pr(A) \\
\tfrac{50}{100} &\overset{?}{\leqslant} \Pr(A) \\
\tfrac{1}{2} &\overset{?}{\leqslant} \Pr(A)
\end{aligned}
$$

So if the probability of A is greater than $1/2$, then you should swerve. And if the probability of A is less than $1/2$, then you should drive straight.

Let's think about what happens if you play chicken, not with your adversary, whose choices are probabilistically independent of yours, but rather with *a twin* whose choices are strongly correlated with yours.

Example 54 (Chicken with a Twin)**.** *Everything is as in Chicken, except that you are facing off against your twin.*

	Twin drives straight (T)	*Twin swerves* ($\sim T$)
You drive straight (Y)	*Death* (-100)	*Glory* (50)
You swerve ($\sim Y$)	*Shame* (-50)	*Status Quo* (0)

Conditional on you swerving, the probability that your twin swerves is 90%*. And conditional on you driving straight, the probability that your twin drives straight is* 90%*.*

$$\Pr(T \mid Y) = \tfrac{9}{10} \qquad \Pr(T \mid \sim Y) = \tfrac{1}{10}$$
$$\Pr(\sim T \mid Y) = \tfrac{1}{10} \qquad \Pr(\sim T \mid \sim Y) = \tfrac{9}{10}$$

Exercise 48. (*a*) *Explain why evidential decision theory tells you to swerve.*

(*b*) *Explain why the states T and $\sim T$ are* states of nature*.*

(*c*) *Suppose that you are* 75% *likely to swerve. Then, use the law of total probability to explain why the probability of T is* 7/10*, or* 70%*.*

(*d*) *Use your answer from part* (*c*) *to explain why causal decision theory says that it is irrational for you to swerve.*

(*e*) *Suppose that you listen to causal decision theory's advice from* (*d*)*, and you are now only* 25% *likely to swerve. Assume that this doesn't change the conditional probabilities of each state, given each act. Then, use the law of total probability to explain why the probability of T is now* 3/10*, or* 30%*.*

(*f*) *Suppose you listen to causal decision theory's advice from* (*d*)*, and you are now only* 25% *likely to swerve. Use your answer from part* (*e*) *to explain why causal decision theory now says that you should drive straight.*

This exercise illustrates an interesting feature of causal decision theory in decisions like these: if you *listen* to its advice, then it can *change* the advice it gives you. In decisions like *Chicken with a Twin*, this leads causal decision theory to waffle back and forth endlessly. If you start out confident that you'll swerve, it tells you to drive straight. If you listen and become confident you'll drive straight, then it tells you to swerve. If you listen and become confident you'll swerve, it tells you to drive straight. And so on and so forth. In the 8th problem set, you'll be asked to think about what causal decision theory says about some other decisions with this feature.

1. Consider the following decision,

   Smoking Gene Susan is debating whether or not to smoke. She knows that smoking is strongly correlated with lung cancer, but only because there is a common cause – a genetic condition that always causes cancer, and frequently causes smoking. Once we fix the presence or absence of this gene, there is no additional correlation between smoking and cancer. Susan prefers smoking without cancer to not smoking without cancer, and prefers smoking with cancer to not smoking with cancer.

   Suppose that the only relevant outcomes are:

   Smoking and getting cancer (with a utility of -90)

   Smoking and not getting cancer (with a utility of 10)

   Not smoking and getting cancer (with a utility of -100)

   Not smoking and not getting cancer (with a utility of 0)

   Consider the following states:

   C: Susan has cancer

   $\sim C$ Susan doesn't have cancer

   and the following acts:

   S: Susan smokes

   $\sim S$: Susan doesn't smoke

   Suppose that Susan has the following conditional probabilities:

   $$\Pr(C \mid S) = \tfrac{8}{10} \qquad \Pr(C \mid \sim S) = \tfrac{1}{10}$$
   $$\Pr(\sim C \mid S) = \tfrac{2}{10} \qquad \Pr(\sim C \mid \sim S) = \tfrac{9}{10}$$

   (a) Create a decision matrix for Susan's decision, and fill the cells of the matrix with the utilities of the possible outcomes

   (b) Explain why evidential decision theory says that it is irrational to smoke.

   (c) Explain why the states C and $\sim C$ are *states of nature*.

   (d) Suppose that the probability that Susan smokes is $1/3$. Then, use the law of total probability to explain why the probability of C is $1/3$ and the probability of $\sim C$ is $2/3$.

(e) Use your answer from part (d) to explain why causal decision theory says that it is irrational for Susan to not smoke.

2. Consider the following decision, due to the philosopher Andy Egan:

Psychopath Button Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying.

Suppose that the only relevant outcomes are:

Live in a world with psychopaths (with a utility of 0)

Live in a world without psychopaths (with a utility of 40)

Die along with all other psychopaths (with a utility of −100)

Consider the following states of nature:

K_1 : Paul is a psychopath, and pushing the button would kill him (along with all other psychopaths)

K_2 : Paul is not a psychopath, and pushing the button would kill all psychopaths, but not him

and there are two available acts:

P : Paul pushes the button

$\sim P$: Paul does not push the button

Suppose that Paul has the following conditional probabilities:

$$\Pr(K_1 \mid P) = \tfrac{9}{10} \qquad \Pr(K_1 \mid \sim P) = \tfrac{1}{10}$$
$$\Pr(K_2 \mid P) = \tfrac{1}{10} \qquad \Pr(K_2 \mid \sim P) = \tfrac{9}{10}$$

(a) Create a decision matrix for Paul's decision, and fill the cells of the matrix with the utilities of the possible outcomes.

(b) Explain why evidential decision theory says that it is irrational to push the button.

(c) Suppose that the the probability that Paul presses the button is $1/10$. Then, use the law of total probability to explain why the probabilities of K_1 is $9/50$ (or 18%) and the probability of K_2 is $41/50$ (or 82%).

(d) Use your answer from part (c) to explain why causal decision theory says that it is irrational to not push the button.

(e) Suppose that Paul listens to causal decision theory's advice from (d), and he is is now 90% likely to push the button. Assume that this doesn't change the conditional probabilities of each state, given each act. Then, use the law of total probability to explain why the probability of K_1 is now $41/50$ (or 82%) and the probability of K_2 is now $9/50$ (or 18%).

(f) Suppose that Paul listens to causal decision theory's advice from (d), and he is now 90% likely to push the button. Use your answer from part (e) to explain why causal decision theory now says that it is irrational to push the button.

3. Consider the following decision,

The Hunter-Richter Problem Before you are three boxes, labeled 'A', 'B', and 'C'. You may take one of the boxes, if you wish. Or you may instead take none of the boxes (call this option 'N'). Yesterday, we used an advanced AI to make a prediction about what you would do in this decision. If it predicted that you would take box A, then we left \$100 in box A and we left an invoice for \$100 in boxes B and C (meaning that, if you take one of these boxes, then you will have to pay us \$100). If it predicted that you would take box B, then we left \$100 in box B and we left an invoice for \$100 in boxes A and C. If it predicted that you would take box C, then we left \$100 in box C and we left an invoice for \$100 in boxes A and B. If it predicted you'd take none of the boxes, then we left an invoice for \$100 in all of the boxes. The AI has never made a mistake; you are certain that it has correctly predicted your choice.

Suppose that the only relevant outcomes are how much money you end up with, and suppose that your utilities, as a function of money, are given by $U(\$x) = \sqrt{x}$ (if x is positive, and $U(\$x) = -\sqrt{x}$ if x is negative).

Consider the following states of nature:

K_A: The AI predicted that you'd take box A

K_B: The AI predicted that you'd take box B

K_C: The AI predicted that you'd take box C

K_N: The AI predicted that you'd take none of the boxes ('N')

And the four available acts:

A : Take box A

B : Take box B

C : Take box C

N : Take none of the boxes

Suppose that you have the following conditional probabilities:

$$\begin{array}{cccc}
\Pr(K_A \mid A) = 1 & \Pr(K_A \mid B) = 0 & \Pr(K_A \mid C) = 0 & \Pr(K_A \mid N) = 0 \\
\Pr(K_B \mid A) = 0 & \Pr(K_B \mid B) = 1 & \Pr(K_B \mid C) = 0 & \Pr(K_B \mid N) = 0 \\
\Pr(K_C \mid A) = 0 & \Pr(K_C \mid B) = 0 & \Pr(K_C \mid C) = 1 & \Pr(K_C \mid N) = 0 \\
\Pr(K_N \mid A) = 0 & \Pr(K_N \mid B) = 0 & \Pr(K_N \mid C) = 0 & \Pr(K_N \mid N) = 1
\end{array}$$

(a) Create a decision matrix for your decision, and fill out the cells of the matrix with the utilities of the possible options.

(b) Explain why evidential decision theory says that you should take one of the boxes.

(c) Suppose that you are equally likely to make any of the choices:

$$\Pr(A) = \Pr(B) = \Pr(C) = \Pr(N) = \tfrac{1}{4}$$

Then, use the law of total probability to explain why each of the states of nature are equally likely,

$$\Pr(K_A) = \Pr(K_B) = \Pr(K_C) = \Pr(K_N) = \tfrac{1}{4}$$

(d) Use your answer from part (c) to explain why causal decision theory says that it is irrational to take one of the boxes.

(e) Suppose you ignore causal decision theory's advice, and decide to go for one of the boxes. After some further deliberation, you end up 80% likely to take box A, 10% likely to take box B, and 10% likely to take box C. Assume that this doesn't change the conditional probabilities of each state, given each act. Then, use the law of total probability to explain why the probabilities of the states of nature are now

$$\Pr(K_A) = \tfrac{8}{10} \quad \Pr(K_B) = \tfrac{1}{10} \quad \Pr(K_C) = \tfrac{1}{10} \quad \Pr(K_N) = 0$$

(f) Suppose you ignore causal decision theory's advice, and decide to go for one of the boxes. After some further deliberation, you end up 80% likely to take box A, 10% likely to take box B, and 10% likely to take box C. Use your answer from part (e) to explain why causal decision theory now says that you must take box A.

Part IV

The Philosophy of Probability

18 | The Philosophy of Probability

Probabilities have been playing a large role in our theory of rational decision-making, and we've been using probabilities to think about inductive inference and questions about how well some piece of evidence supports a hypothesis.

In the first part of the course, we learned how to *reason about* probabilities, using the probability rules. These rules allowed us to work out what some probabilities were, given some *other* probabilities. For instance, the conjunction rule told us how to work out $\Pr(A\&B)$, if we already knew $\Pr(A \mid B)$ and $\Pr(B)$. But none of those rules told us that probabilities *were* in the first place.

In this part of the course, we're going to start asking philosophical questions about what kind of thing probabilities are. In particular, we'll focus on two questions:

1. What do we *mean* when we make a claim about probability?

2. What determines which probabilities are the correct ones?

There are roughly three families of answers to these questions. The answers in the first group all say that claims about probabilities are claims about some *objective* quantity and that these quantities may be determined *a priori*, or *from the armchair*—you don't need to do any experimentation in order to work out what the probability of some proposition is. You just need to think about it hard enough. The answers in the second group agree that probabilities are objective, but they maintain that probabilities can only be determined *empirically*, through experimentation. (In the terminology of philosophers: probabilities are *a posteriori*.) Finally, the answers in the final group say that probabilities are *subjective*, and they can vary from person to person.

In the final section of the course, we're going to spend most of our time thinking about this final understanding of probabilities, according to which they are subjective and represent an individual's *degree of belief* or *degree of confidence* in a given proposition. On this view, when I say that it's unlikely to rain, I'm just saying that I am more confident that it will not rain than I am that it will rain. To understand why people were led to this way of understanding claims about probabilities, it's worth taking a brief historical tour through the other possible answers to our two questions.

The Classical View: Probabilities are Objective and A Priori

The Marquis de Laplace thought that probabilities were objective, in the sense that they didn't vary from person to person, and that they could be worked out *a priori*, without having to rely upon experiments. According to Laplace, when we say something like "the probability that the coin lands heads is $1/2$", all we mean is that there are two *equally possible* ways for the coin to land, heads and tails. Since half of these 'equally possible' ways for the coin to land are ones in which the coin lands heads, the probability of a heads landing is one half. Laplace wrote:

Figure 18.1: Laplace

> *The theory of chances consists in reducing all events of the same kind to a certain number of equally possible cases, that is to say, to cases whose existence we are equally uncertain of, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all possible cases is the measure of this probability.*

Think about rolling a die. There are six equally possible ways for the die to land: it could land on 1, 2, 3, 4, 5, or 6. Since each of these outcomes are equally possible, and since half of them are outcomes in which the die lands on an even number, Laplace says that the probability of the die landing on an even number must be $1/2$.

Or think about drawing a marble from an urn. If there are 6 red marbles, 6 yellow marbles, and 6 green marbles, then there are 18, equally possible, outcomes. Since 6 of these 18 outcomes are ones in which a red marble is drawn, the probability of drawing a red marble from such an urn is $6/18$, or $1/3$.

In general, Laplace endorsed something known as

The Principle of Indifference If A and B are equally possible, then the probability of A equals the probability of B

This view about probability has come to be known as *the classical interpretation*.

What should we make of the classical interpretation of probability? Is it correct? Is this what we're talking about when we talk about probability? And is Laplace correct that we should use the principle of indifference to determine what probabilities are?

There are reasons for doubt. Firstly, it's not clear what the principle of indifference is saying, since it's not clear what Laplace means by 'equally possible'. Does he just mean 'equally *probable*?' If so, then the principle is unobjectionable, but completely uninformative—if that's how we interpret it, then it's just the tautologous claim that, if A and B have the same probability, then they have the same probability.

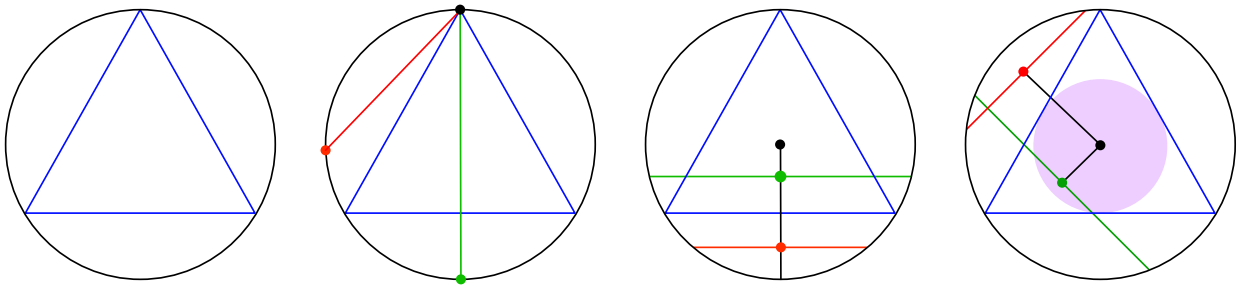


Figure 18.3: Bertrand's Paradox. An equilateral triangle is inscribed inside of a circle, as in figure 18.3a. A chord of the circle is selected at random. What is the probability that this chord has a length greater than the side length of the triangle?

Think about the case of flipping a coin. It's of course *possible* that the coin lands on its edge, though we usually ignore this possibility. Is the coin landing on its edge *equally* possible as it landing on heads? Laplace had better say 'no'. If he says 'yes', then we'll have to say that the probability of a fair coin landing heads is actually $1/3$, and not $1/2$! (After all, there would then be *three* equally possible outcomes: heads, tails, and edge.) But surely we know that the probability of a flipped coin landing heads is not $1/3$—it's got to be somewhere very close to $1/2$. So an edge landing cannot be *equally possible* as a heads landing. But why not? How are we going to understand this talk of *degree of* possibility so that 1) an edge landing is *less possible* than a heads landing, and 2) we don't just mean that the edge landing is *less probable* than a heads landing?

A second reason for doubt comes from a paradox due to Joseph Bertrand. Bertrand pointed out that the principle of indifference seems to lead to inconsistent probability judgments. He considered the question: suppose we have a circle with an equilateral triangle inscribed inside of it, like in figure 18.3a. And suppose that a chord of the circle (a line segment with ends on the circle) is selected at random. Then, what's the probability that the chord has a length greater than the side length of the triangle?

Figure 18.2: Joseph Bertrand

By applying the principle of indifference to this problem, we can arrive at three different, incompatible answers. Firstly, let's fix one of the endpoints of the chord. Just to visualize the comparative lengths better, we can rotate the circle so that this fixed endpoint lies at the top of the triangle, as in figure 18.3b. Then, the question is: when will the *second* endpoint determine a chord with a length greater than the side length of the triangle? We can easily see by looking at figure 18.3b that the triangle splits the circle's circumference into three segments, and that it is only when the second endpoint lies in the

furthest segment from the first endpoint that the chord will be longer than the triangle's side length. It is equally possible that the second endpoint lie in any one of these regions of the circle. So, applying the principle of indifference, the probability that the chord is longer than the side length of the triangle is equal to $1/3$.

On the other hand, notice that every chord of the circle has a midpoint that lies on a radius of the circle. Pick one such radius. Then, just to help with visualizing things, rotate the triangle so that the radius is perpendicular to one of the triangle's sides, as in figure 18.3c. Notice that the triangle's side will then cut the radius into two equally sized segments. It's just as possible that the chord lies inside the triangle as it is that the chord lies outside the triangle. And the chord will be longer than the triangle's side iff the midpoint of the chord along this radius is inside of the triangle. So, applying the principle of indifference, the probability that the chord is longer than the side length of the triangle is $1/2$.

Finally, notice that any given *point* in the circle will determine a unique chord. Just pick the radius that the point lies along, and take the chord which passes through that point and is perpendicular to that radius, as in figure 18.3d. The chord will be longer than the triangle's side length iff the randomly selected *point* is inside of the circle inscribed within the triangle. So, to work out the probability that the randomly selected chord is at least as long as the side of the triangle, you need to work out what proportion of the circle's area is taken up by that inscribed circle. When you work through the calculations, you find that the inscribed circle has an area which is one fourth of the area of the larger circle. It is equally possible that the midpoint of the chord lies in any given area of the circle. So, applying the principle of indifference, the probability that the chord's length is at least the length of the triangle's side is $1/4$.

The philosopher Bas C. van Fraassen has another example to illustrate the same point. He imagines that there is a factory which produces cubes, and that you know nothing about these cubes, other than that they have a side length between 0cm and 2cm. What, then, is the probability that the cubes have a side length greater than 1cm? Well, you might think: it's just as possible that they have side lengths between 0 and 1 as it is that they have side lengths between 1 and 2. So, applying the principle of indifference, the probability must be $1/2$.

Imagine now a second factory, which produces cubes with *face areas* between 0cm^2 and 4cm^2. What is the probability that the cubes have face areas greater than 1cm^2? Well, you might think: it's just as possible that they have face areas between 0 and 1 cm^2 as it is that they have face areas between 1 and 2 cm^2, which is just as possible as them having face areas between 2 and 3 cm^2, which is just as likely as they having face areas between 3 and 4 cm^2. So, by the principle of indifference, the probability that the face areas are greater than 1 cm^2 should be $3/4$.

But this isn't a *second* factory at all—it's just exactly the same factory, redescribed a different way! After all, the face area will be between 1 and 4 cm^2 if and only if the side

length is between 1 and 2 cm.

The lesson from both examples is the same: the principle of indifference ends up giving us different, and contradictory, probability assignments, depending upon how we approach and describe the problem.

The Frequency View: Probabilities are Objective and A Posteriori

Dissatisfaction with the classical interpretation's principle of indifference led people to consider alternative ways of understanding probabilities. According to the *frequentist*, claims about probabilities are just claims about *relative frequencies*. When we say that the probability of the coin landing heads is 1/2, we just mean that one half of the time the coin is flipped, it lands heads.

$$\Pr(\textit{a flipped coin lands heads}) = \frac{\textit{\# of flips that land heads}}{\textit{\# of flips}}$$

Or consider rolling a die. To determine the probability that a die lands on 1, you just consider *how often* that die lands on 1 when its rolled. If the frequency of 1 landings is 1/6, then 1/6 is the probability of the die landing 1 on any given roll.

$$\Pr(\textit{a rolled die lands 1}) = \frac{\textit{\# of rolls that land 1}}{\textit{\# of rolls}}$$

Or think about drawing a marble from an urn. Imagine that there are 6 red marbles, 6 yellow marbles, and 6 green marbles. Then, the probability that you get a red marble on any given draw is equal to the total number of times a red marble is drawn, divided by the total number of times a marble is drawn.

$$\Pr(\textit{a red marble is drawn}) = \frac{\textit{\# of times red is drawn}}{\textit{\# of draws}}$$

And, in general, for any given properties F and G, the frequency view will say that the probability of an F thing being G is just the proportion of Fs which are G.

$$\Pr(\textit{an } F \textit{ is } G) = \frac{\textit{\# of } F\textit{s which are } G}{\textit{\$ of } F\textit{s}}$$

Notice that, according to the Classical view, we didn't have to do any *experimentation* in order to work out what the probability of heads was. Whereas, according to the frequentist, we need to actually *check* what's happening with the coin, the die, or the urn in order to know something about what the probabilities are. It could be that the coin lands heads way more often than it lands tails. If so, then the probability of the coin landing heads need not be 1/2. So while the classical view said that probabilities were *a priori* (knowable

independently of experience and experimentation), the frequency view says that they are *a posteriori* (only knowable on the basis of experience and experimentation).

The frequency view has advantages over the classical view. It seems that our views about probabilities should be affected by experience—if a die keeps landing on 1, that should make us think that the die is biased. And this is easy to explain on the frequency view—when we see the die land on 1 over and over again, this gives us evidence that the frequency of die rolls that land on 1 is greater than one sixth.

However, there are also a number of reasons to worry about the frequency view. Firstly, notice that lots of our claims about probability concern *one off* events—things that have only happened once, and will only happen once. In 2016, there was tons of talk about the probability of Trump beating Clinton. Nate Silver suggested that the probability was somewhere around 50%, whereas the New York Times was suggesting that the probability was much lower, somewhere around 1%. But there is only one election between Trump and Clinton. So when we apply the frequency view, we're told that the probability of Trump beating Clinton is

$$\Pr(\textit{Trump beats Clinton in an election}) = \frac{\#\textit{ of times Trump beats Clinton}}{\#\textit{ of times Trump is in an election with Clinton}}$$

With the benefit of hindsight, we can see that this ratio is 100%. But even without the benefit of hindsight, we would be able to know that it is *either* 100% *or* 0%—since we know that the denominator is 1, and the numerator is either 0 or 1.

So, if probabilities are just frequencies, then both Nate Silver and the New York Times were wrong in a pretty embarrassing way—both of them were making probability claims that had *no shot* at being true!

This problem recurs for any kind of one off event. When we ask about the probability of human extinction due to nuclear war, or the probability that the dinosaurs were killed by an asteroid impact, or the probability of Trump going to war with Iran, all of these are events of a kind that only happen *once*, if at all. So it looks like the frequency view won't allow us to make sense of claims like 'it's about 50% likely that humanity kills itself off with nuclear weapons in the 21st century', or 'it's over 90% probable that dinosaurs went extinct because of an asteriod impact, though there's a 10% probability that it was due to volcanic eruptions in the Deccan Traps.'

There are other worries about the frequency view which stem from the fact that, in order to determine a *frequency*, we need to have some *reference class*, relative to which the frequency is taken. For instance, suppose that you're an insurance company who's interested in the probability that I get cancer in the next 20 years. According to the frequency view, this probability is just a relative frequency of people in my condition—people like me—who get cancer between 40 and 60. But which people are the people *like me*? Which actuarial tables should the insurance company consult? They might ask which proportion of *men* get cancer in any 20 year period (using the reference class *men*). Or they might instead ask which proportion of *men age 40* end up getting cancer before they are 60 (using

the reference class *men age 40*). Or they might instead ask which proportion of *men age 40, living in Los Angeles, who do not smoke* get cancer before they are 60. All of these difference references classes will determine different frequencies.

So on the frequency view, there's not *just one* probability that I get cancer in the next 20 years. There's the probability that I get cancer in the next 20 years *as a member of the reference class of all men*, and another, different probability that I get cancer in the next 20 years *as a member of the reference class of all men 40 or over*.

It can seem that the frequency view is giving us *too many* probabilities, especially given the use to which we want to put probabilities in decision theory. The insurance company needs to know what the expected value of insuring me is going to be. If there's 100 different probabilities of me getting cancer, then there's going to be 100 different expected values for insuring me—which one should the insurance company use to make its decision?

The Bayesian View: Probabilities are Subjective

The final view, which we're going to spend the most time investigating, says that probabilities are entirely *subjective*. When I say that the probability of the coin landing heads is $1/2$, what I am saying is that I am as confident that the coin lands heads as I am that it doesn't. When I say that the probability that a rolled die lands on 1 is $1/6$, I am reporting that I'm one fifth as confident that it will land on one as I am that it won't land on one. And when I say that the marble drawn from an urn has a probability of $1/3$ of being red, I am reporting that I am one half as confident that the marble will be red as I am that it won't be red.

This view is named after the Reverend Thomas Bayes. This is the same Bayes who is responsible for Bayes' Theorem, which we learned about earlier in the course. Bayes wrote a paper on probability in which he interpreted the probabilities subjectively, and for this reason the view has been named after him.

Figure 18.4: Thomas Bayes

There's a potential problem for the Bayesian view. It's not at all clear that people's *degrees of confidence* or *degrees of belief* are going to satisfy the rules of probability. For instance, Daniel Kahneman and Amos Tversky gave people the following question:

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*
>
> *Which is more probable?*
>
> 1. *Linda is a bank teller.*

2. *Linda is a bank teller and is active in the feminist movement.*

And they found that most people answered that 2 is more probable than 1. If we take this at face value, then it looks as though these people have a higher degree of belief in the *conjunction* 'Linda is a bank teller and is active in the feminist movement' than they have in its *conjunct* 'Linda is a bank teller'.

But there is no probability function which gives a higher probability to 'Linda is a bank teller' than it gives to 'Linda is a bank teller and is active in the feminist movement'. Let's use 'B' for 'Linda is a bank teller' and 'F' for 'Linda is active in the feminist movement'. Then, notice that $B\&F$ *entails* B—there's no possibility in which $B\&F$ is true but B is false. Think about the Euler diagram from figure 18.5. There can't be more mud sitting on top of

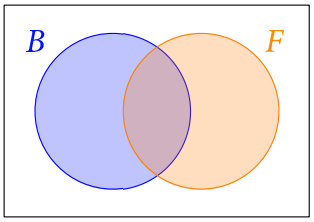


Figure 18.5

the *overlap* between the B and F circles (where $B\&F$ is true) than there is sitting on top of the B circle! In general, whenever one proposition *entails* another, the probability of the second proposition must be at least the probability of the first. Probability cannot decrease along entailment. But people's degrees of belief seem to sometimes decrease along entailment. So probabilities can't just be degrees of belief.

This much seems right: not all degrees of belief are probabilities. But the Bayesian doesn't have to say that they are. Instead, they might say that all *rational* degrees of belief are probabilities. On this view, when we talk about what the probability of some proposition is, we're talking about how confident *to be* in that proposition. We're talking about what degree of confidence is *rational*, given our evidence.

If the Bayesian takes this stance, then they face a challenge: *why* should our degrees of belief be probabilities? Notice that this isn't a challenge faced by either the classical or the frequentist views about probability. On both the classical and the frequentist view, probabilities are just a certain kind of *proportion*. And proportions are automatically going to satisfy the probability rules. But things are different for the Bayesian. On the Bayesian view, there's not going to be any automatic or straightforward connection between people's subjective attitudes and the probability rules; and, moreover, we have lots of good evidence that people are instinctively *very bad* at probability. So if the Bayesian is going to posit some connection between your subjective attitudes and the probability rules—if they're going to say that your subjective degrees of confidence *should obey* the probability rules, then they had better say something about *why*.

We'll see that the Bayesian has two main arguments for this conclusion. The first argument is *pragmatic*: if your degrees of belief *aren't* probabilities, then you could be sold a collection of bets, the combination of which is *guaranteed* to lose you money, no matter what. (A collection of bets like this is known as a 'Dutch book'.) The idea is that, if your degrees of belief can lead you to purchase a combination of bets like this—a *sure loss*—then those degrees of belief cannot have been particularly rational.

The second argument is *alethic* (having to do with truth): if your degrees of belief *aren't* probabilities, then you will be *accuracy dominated*, in the sense that there will be some other degrees of belief which are guaranteed to be closer to the truth than yours are, no matter what. The idea behind this argument is that you want your degrees of belief to be as close to the truth as possible. If it rains, then it's better to be 90% confident in rain than it is to be 70% confident in rain. And if it doesn't rain, then it's better to be 70% confident in rain than it is to be 90% confident in rain. And if your degrees of belief are *accuracy dominated*, then you will know that there are some other degrees of belief which are going to be better than yours are *no matter what*. If your goal is to be as close to the truth as possible, then non-probabilistic degrees of confidence will be irrational.

We'll spend the rest of the course going through those arguments more carefully.

19 | Credences and Bets

The *Bayesian* says that probabilities are rational degrees of belief. Degrees of belief are often called *credences*, as in: 'I don't put much credence in anything John says'. The Bayesian thinks that much of our talk about probability is actually talk about rational credence. When we say "The Republicans are unlikely to keep control of the House in the midterms", the Bayesian understands that as the claim that you shouldn't be confident that the Republicans keep control of the house in the midterms.

The Bayesian endorses two norms about rational credence. Firstly, that rational credences are probabilities. And secondly, that rational credences evolve over time by conditioning on newly acquired evidence. These two commitments are called

Probabilism If you are rational, then your credences will obey the probability rules.

Conditionalization If you are rational, then your new, posterior, credences will come from your old, prior, credences by conditioning them on all of your newly acquired evidence.

The Bayesian endorses these norms, but not everyone agrees with them. In particular, Arthur Dempster and Glenn Shafer have a theory of degrees of belief which allows you to have a credence of zero in both A and $\sim A$. As they are thinking about things, if you have no reason to think that Trump is going to win, then your *degree of belief* that Trump wins should be zero. And if you have no reason to think that Harris is going to win, then your *degree of belief* that Harris is going to win should likewise be zero. Of course, you have conclusive reason to think that one of them is going to win, so your credence that *either* Trump *or* Harris is going to win should be 100%.

The Bayesian thinks that Dempster and Shafer are wrong. They say that, if your degree of belief in *either* Trump *or* Harris winning is 100%, then your degree of belief that Trump wins and your degree of belief that Harris wins must add up to 100%. This follows from *The Disjunction Rule*. Since there will be at most one winner, the propositions *Trump wins* and *Harris wins* are mutually exclusive. So

$$\Pr(\textit{Trump wins} \vee \textit{Harris wins}) = \Pr(\textit{Trump wins}) + \Pr(\textit{Harris wins})$$

What reasons can the Bayesian offer to think that these two norms are correct—that credences *should* be probabilities, and that you *should* respond to learning that E by conditioning your prior credences on E?

One kind of reason is *pragmatic*. The basic idea is that non-probabilistic credences will lead you to make bad *decisions*. To get clear on this argument, let's start by thinking more about the relationship between *credence* and *action*, and how we can use your choice dispositions to reveal something about your credences.

Credences and Bets

The Italian mathematician and philosopher Bruno de Finetti had a particular way of thinking about credences which he used to argue for probabilism. On de Finetti's understanding, your credences were importantly related to which *bets* you were willing to make.

Figure 19.1: Bruno de Finetti

Suppose that you are offered a ticket which entitles the holder to \$1 if USC beats UCLA, and nothing if USC doesn't beat UCLA.

\$1	if USC wins
\$0	else

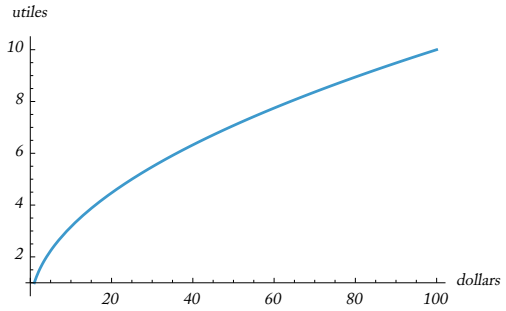
If 'USC wins' turns out to be true, then this ticket is worth \$1. Whereas, if 'USC wins' turns out to be false, then this ticket is worth nothing. A ticket like this is what we'll call a *bet* on A.

What are you willing to pay for this bet? de Finetti's idea was that, if your credence that USC wins is x, then you must be willing to pay up to \$$x$ for the ticket. For instance, suppose that your credence that USC wins is 50%. Then, you are willing to pay up to 50¢ (or \$0.5) for it.

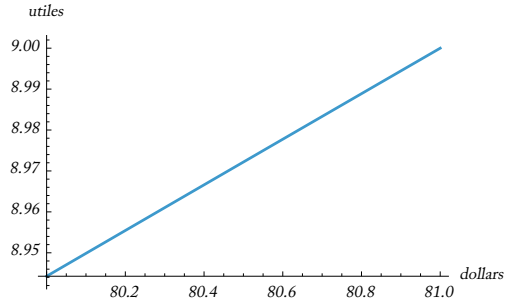
> **de Finetti's First Assumption** If your credence in a proposition, A, is x, then you are willing to buy a \$1 bet on A for all and only prices at or below \$$x$.

What should we make of this assumption? One natural objection is that it's too tied to *money*. Bernoulli taught us that it's not *expected money* that rational people try to maximize, but rather *expected utility*. So shouldn't the assumption be talking about *utility*, rather than *money*?

de Finetti is aware of this concern, but he thinks it can be side-stepped. Think about the utility function $U(\$x) = \sqrt{x}$ (shown in figure 19.2a). Even though this curve isn't linear—it goes up less and less the further out you get along the x-axis—if you just stick to *small* changes in money, it is basically linear. For instance, figure 19.2b shows the function



(a) $U(\$x) = \sqrt{x}$, from \$0 to \$100



(b) $U(\$x) = \sqrt{x}$, from \$80 to \$81

Figure 19.2

$U(\$x) = \sqrt{x}$ from \$80 to \$81. It's basically a straight line. So de Finetti thought that, so long as we're talking about something like a *dollar* bet on a proposition, we can effectively ignore risk-aversion.

Then, de Finetti's assumptions tell us that, if the price of the \$1 bet on A is *less* than your credence in A, then that price is *favorable*, and you will certainly purchase the bet at that price. If the price is *equal* to your credence in A, then the price is *fair*, and you may or may not purchase the bet at that price. Finally, if the price is *greater* than your credence in A, then the price is *unfavorable*, and you will not purchase the bet at that price.

Notice that we could justify this assumption if we assumed that you were an expected utility maximizer. For instance, give our assumption that your credences are linear in dollars, your expected utility for the ticket is your credence that USC wins times 1 plus your credence that USC *doesn't* win times 0,

$$Cr(USC\ wins) \cdot 1 + Cr(\sim USC\ wins) \cdot 0 = Cr(USC\ wins)$$

So: the expected utility of the ticket is going to be equal to your credence that USC wins. Assuming that you value a bet with its expected utility, you should value the bet more than any dollar amount below $\$Cr(USC\ wins)$.

Notice that I'm writing '$Cr(USC\ wins)$' instead of '$\Pr(USC\ wins)$'. That's because I'm *not* taking for granted that your credences are probabilities. They may be; but they may not be. Part of de Finetti's goal is to argue that they *should be*. But in order to argue for that claim, he's going to consider what happens when your credences are *not* probabilities.

Let's spend a bit more time on this point. Consider someone—call them 'Bob'—whose credence that USC wins is 60%, and whose credence that USC doesn't win is 80%. Bob's credences are *not* probabilities, since the negation rule requires 'USC wins' and 'USC doesn't win' to add up to 1.

Before, we saw two different ways of using expected utilities to think about whether you should purchase a bet. Firstly, we could calculate the expected utility *of the bet itself*, and ask whether it is greater than, less than, or equal to its price. Secondly, we could cal-

culate the expected utility *of the act of purchasing the bet*, and then ask whether this is greater than, less than, or equal to the *act* of not purchasing. When we were taking for granted the rules of probability, these two methods were always going to agree. The expected utility of the bet will be greater than the price if and only if the expected utility of purchasing the bet is greater than the expected utility of not purchasing the bet. But if we're considering people whose credences are not probabilities, then the two methods come apart.

Let's walk through an example. Bob's credence that USC wins is 60%, his credence that USC doesn't win is 80%, and he's offered a $1 bet on whether USC wins for 50¢. That is, we tell him that, if he gives us 50¢, then we will give him a ticket worth $1 in the event that USC wins, and worth nothing otherwise.

$1	if USC wins
$0	else

price: 50 ¢

Since Bob's credence that USC wins is 60%, and his utilities are linear in dollars, Bob's expected utility for this ticket is

$$\begin{aligned} & Cr(\textit{USC wins}) \cdot 1 + Cr(\sim \textit{USC wins}) \cdot 0 \\ =& 0.6 \cdot 1 + 0 \\ =& 0.6 \end{aligned}$$

Since $0.6 > 0.5$, the ticket is worth more than its price. So Bob should be willing to pay the price.

On the other hand, suppose that we calculate Bob's expected utility for the *act* of purchasing the ticket. Bob has two acts available to him: he could pay $0.5 and get the ticket, or he could decline. And there are two relevant states (that are causally and probabilistically independent of how he chooses): USC could win or they could not win.

	USC wins	USC doesn't win
Pay for the ticket	gain 50¢	lose 50¢
Do not pay for the ticket	break even	break even

Since Bob's utilities are linear in dollars, his expected utility for paying is

$$\begin{aligned} & Cr(\textit{USC wins}) \cdot 0.5 + Cr(\sim \textit{USC wins}) \cdot (-0.5) \\ =& 0.6 \cdot 0.5 + 0.8 \cdot (-0.5) \\ =& 0.3 - 0.4 \\ =& -0.1 \end{aligned}$$

whereas the utility of not paying is a guaranteed 0. Since $0 > -0.1$, the expected utility of *the act* of not paying for the ticket is greater than the expected utility of *the act* of paying for it.

Because Bob's credences are not probabilities, these two ways of applying expected utility theory come apart. If we compare the expected utility of the bet to the price of the bet, de Finetti's assumption follows from the idea that you maximize expected utility. But if we instead compare the expected utility of *the act* of taking the bet to the expected utility of *the act* of not taking the bet, then de Finetti's assumption will contradict the idea that you maximize expected utility.

Exercise 49. *Suppose that Tina's credence that it rains tomorrow is* 40%*, and her credence that it doesn't rain tomorrow is* 30%*. Her utilities are linear in dollars, and she is offered a ticket which pays out* \$1 *if it rains, and which costs* 50¢*.*

\$1	*if it rains*
\$0	*else*

price to buy: 50¢

1. *Given de Finetti's assumption, is Tina willing to buy this ticket*?

2. *Create a decision matrix for Tina's decision of whether or not to buy. Which of the available acts maximizes her expected utility*?

It should seem very strange that these two different ways of figuring out whether Tina should buy the ticket lead us to different answers. Indeed, we're going to see later on that one way of understanding de Finetti's argument for probabilism is precisely that non-probabilism leads your appraisal of bets to change, depending upon how they are presented to you. But we'll come to that later on.

Rather than thinking of de Finetti's first assumption as being *derived from* expected utility maximization, I think that it's better to think of it as a *stipulative definition* of what de Finetti *means* in the first place by 'credence'.[1] On his view, part of *what it is* for you to have a credence of x in A is for you to be willing—in the right circumstances—to pay up to \$$x$ for a \$1 bet on A.

Relatedly, de Finetti thought that you should be willing to *sell* \$1 a bet on A for any value *greater than* your credence in A.

de Finetti's Second Assumption If your credence in a proposition, A, is x, then you are willing to sell a \$1 bet on A for any amount greater than \$$x$.

If the price of the \$1 bet on A is *greater* than your credence in A, then selling the bet at that price is *favorable* to you, and you are willing to sell the bet at that price. If the price is *equal* to your credence in A, then the price is *fair*, and you may or may not sell the bet at that

[1] In fact, de Finetti used the word 'prevision' for what we're calling credence, and he had a more general theory of prevision than we're going to have the space to get into here.

price. Finally, if the price is *greater* than your credence in A, then the price is *unfair*, and you will not sell the bet at that price.

Again, if we are assuming that your credences are probabilities, then de Finetti's second assumption *follows from* expected utility maximization. But if we are not assuming that your credences are probabilities, then matters are more complicated.

Consider again Bob, whose credence that USC wins is 60% and whose credence that USC *doesn't* win is 80%. And consider again the $1 bet on USC winning. Should Bob be willing to sell this ticket for 50¢? If we're just approaching this question using the rule of expected utility maximization, it will depend upon whether we compare the expected utility of the bet to its price, or whether we instead compare the expected utility of *purchasing the bet* to the expected utility of *not* purchasing the bet.

If we compare the expected utility of the bet to its price, we'll see that Bob will *not* be willing to sell it for 50¢. After all, Bob's expected utility for the bet will be 60¢, since

$$0.6 \cdot 1 + 0.8 \cdot 0 = 0.6$$

Since 60¢ is greater than the 50¢ that he'd get by selling it, Bob won't be willing to sell.

On the other hand, if we compare the expected utility of the *acts* of selling and not selling, we'll see that Bob *is* willing to see it for 50¢. Here is Bob's decision table,

	USC wins	USC doesn't win
Sell the ticket	gain 50¢	gain 50¢
Keep the ticket	gain $1	gain nothing

Since Bob's credence in USC winning is 60% and his credence in USC not winning is 80%, the expected utility of keeping the ticket is

$$Cr(\textit{USC wins}) \cdot 1 + Cr(\sim \textit{USC wins}) \cdot 0 = Cr(\textit{USC wins}) = 0.6$$

What about the expected utility of selling the ticket? If we calculate things using the usual formula, we'll get

$$\begin{aligned} & Cr(\textit{USC wins}) \cdot 0.5 + Cr(\sim \textit{USC wins}) \cdot 0.5 \\ =& 0.6 \cdot 0.5 + 0.8 \cdot 0.5 \\ =& 0.3 + 0.4 \\ =& 0.7 \end{aligned}$$

Since $0.7 > 0.6$, it seems that Bob should sell the ticket.

Exercise 50. *Suppose that Tina's credence that it rains tomorrow is* 40%*, and her credence that it doesn't rain tomorrow is* 30%*. Her utilities are linear in dollars, and she is offered* 50¢ *in exchange for her ticket which pays out* $1 *if it rains.*

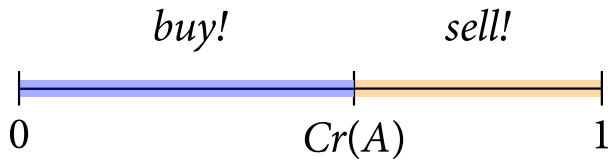


Figure 19.3: Your credence in A divides the possible prices on a $1 bet on A into three groups. If the price is less than your credence in A, then you should buy it and not sell it. If the price is greater than your credence in A, then you should sell it, and not buy it. If the price is equal to your credence in A, then you are permitted to either buy or sell.

$1	*if it rains*
$0	*else*

1. *Given de Finetti's assumption, will Tina sell the ticket for* 50¢?

2. *Create a decision matrix for Tina's decision of whether or not to sell. Which of the available acts maximizes expected utility*?

Again, it should seem very strange that these two different ways of figuring out what Bob should do are leading us to different conclusions. Notice another oddity of this particular example: we *know for sure* that selling the ticket will get Bob 50¢. But when we calculate expected values using Bob's non-probabilistic credences, we get a value *greater* than 50¢. Again, the oddity of all this is going to be part of de Finetti's argument for having probabilistic credences. But we can also use it as a reason to favor de Finetti's assumptions about the connection between credences and betting. If we were instead calculating the expected utilities of the *acts*, then the values we get would depend upon which *states* we used. In contrast, de Finetti's assumptions don't depend upon that at all.

Putting together de Finetti's first and second assumptions, we get that your credence in A determines exactly what you will do with respect to any given bet on A. If the price of that bet is less than your credence in A, then you should buy it, and not sell it, at that price. If the price is greater than your credence in A, then you should sell it, and not buy it, at that price. If the price is exactly equal to your credence in A, then you are permitted to either buy or sell. See figure 19.3.

Part of de Finetti's idea here is that *what we mean* when we say that Bob's credence that USC wins is 60% *just is* that he is willing to purchase $1 bets on USC winning for anything up to 60¢, and he is willing to sell $1 bets on USC winning for anything more than 60¢. Let's call this de Finetti's *betting interpretation of credence*.

de Finetti's Betting Interpretation of Credence What it is for you to have a credence of x in the proposition A is for you to be willing to buy a $1 bet on A for anything up to $\$x$, and for you to be willing to sell a $1 bet on A for anything more than $\$x$.

What de Finetti is suggesting here is something very different from the standard expected utility theory we've been learning about up to this point. The theory of expected utility, as we have been understanding it, says that *if* your probability in A is x, and *if* your utilities are linear in dollars, then *it will be rational* for you to buy a \$1 bet on A for anything up to $\$x$ and sell it for anything more than $\$x$. de Finetti, by contrast, is saying that being willing to buy and/or sell bets like these at these prices *makes it the case* that your credence in A is x.

That is, before, we were thinking about *probabilities* and *utilities* as two different and pre-existing things which combine together to determine which choices are rational. In the first place, de Finetti is not talking about what choices are *rational*—he's just talking about which choices you are actually inclined to make. In the second place, de Finetti is reversing the order of things. He's *starting* with facts about which bets you're willing to take, and *from there* saying something about what your credences, or subjective probabilities, are.

Next time, we'll see how de Finetti can leverage this interpretation of credences to argue that credences should be probabilities. He'll argue that, if they are not probabilities, then you will be willing to make a combination of trades which are *guaranteed* to leave you worse off, no matter what.

20 | The Dutch Book Argument

To review, last time we saw de Finetti's betting interpretation of credence, according to which

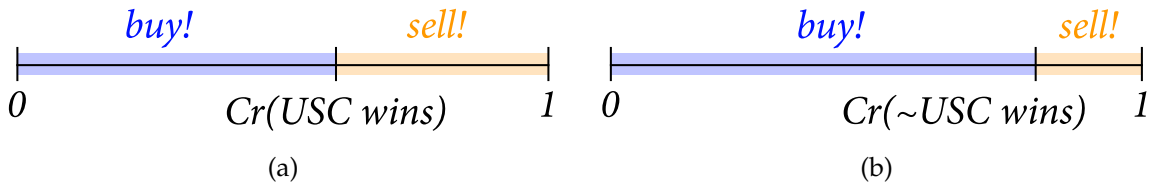
> **de Finetti's Betting Interpretation of Credence** What it is for you to have a credence of x in the proposition A is for you to be willing to buy a \$1 bet on A for anything up to \$$x$, and for you to be willing to sell a \$1 bet on A for anything more than \$$x$.

The betting interpretation of credence plays an integral role in de Finetti's argument for probabilism—the thesis that your credences ought to be probabilities.

> **Probabilism** If you are rational, then your credences will obey the probability rules.

Dutch Books

To appreciate how this argument goes, let's return to Bob, whose credences were not probabilities. Bob's credence that USC wins the game against UCLA is 60% and his credence that USC doesn't win the game is 80%. Let's think about what this means, in terms of Bob's willingness to buy and sell certain bets. We can visualize Bob's betting dispositions with figure 20.1. Figure 20.1a. If you offer Bob a \$1 bet on USC's winning, he'll buy it for anything up to 60¢. And Bob will sell such a bet for anything more than 60¢.



(a) (b)

Figure 20.1: Bob's betting dispositions. In figure 20.1a, Bob's dispositions with respect to a \$1 bet on USC wins. *In figure 20.1b, Bob's dispositions with respect to a \$1 bet on* ~USC wins.

Let's reflect the diagram from figure 20.1b around 1/2 and stack it beneath the diagram form figure 20.1a. Then, we can visualize Bob's betting dispositions for both of the propositions *USC wins* and *~USC wins* with the diagram in figure 20.2 Notice that the two blue

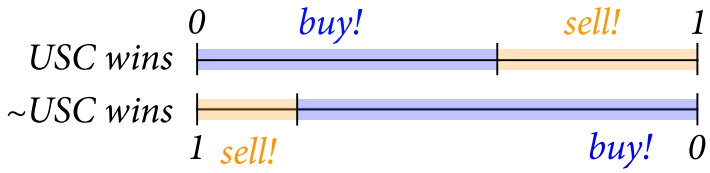


Figure 20.2: Bob's betting dispositions for a $1 bet on USC wins *(on top), and Bob's betting dispositions for a $1 bet on* USC doesn't win *(on bottom, reflected about 1/2).*

'buy!' segments overlap each other in the middle. Probabilistic credences would never overlap in this way, by the negation rule. The negation rule tells us that the probability of A is always 1 minus the probability of $\sim A$. So if Bob's credences were probabilities, he'd be willing to buy a bet on *USC wins* at x¢ iff he was willing to sell a bet on *USC doesn't win* at $1-x$¢. So, if Bob's credences were probabilities, and we reflected his betting dispositions for *USC doesn't win* around 1/2, the blue 'buy!' segment for *USC doesn't win* would align perfectly with the orange 'sell!' segment for *USC wins*.

de Finetti then notes that, because the two blue 'buy!' segments overlap in this way, Bob will be willing to buy a $1 bet on *USC wins* for 50¢ and also buy a $1 bet on *~USC wins* for 75¢ . That is, we can provide Bob with the following two bets, and he'll happily purchase both of them:

Bet 1	
$1	if USC wins
$0	else

price: 50¢

Bet 2	
$1	if USC doesn't win
$0	else

price: 75¢

But now think about what Bob has done. He's handed over $1.25 (50¢+ 75¢= 125¢), and what he's gotten back in return is two tickets which are *certainly* worth only $1. If USC wins, then the first ticket is worth $1 and the second ticket is worthless. And if USC doesn't win, then the first ticket is worthless and the first ticket is worth $1. So Bob has just made himself 25¢ poorer. Surely that's not a rational choice!

Let's go through that again. Consider the possible outcomes: either USC will win or it won't. Here's Bob's net profit from purchasing each of the two bets above, in each of those two cases:

	USC wins	USC doesn't win
Net profit from buying bet 1	50¢	−50¢
Net profit from buying bet 2	−75¢	25¢
Overall net profit	−25¢	−25¢

A collection of bets like this is known as a *Dutch Book*. It's not entirely clear why it's so-called. A 'book' is just a collection of bets. There are a number of explanations for why a book like this is called 'Dutch'. One explanation I've heard is that there was a particular bookie who excelled at constructing *sure loss* books like this, and whose nickname was 'Dutch'.

Recall Tina. Tina's credence that it rains tomorrow is 40% and her credence that it doesn't rain tomorrow is 30%. Doing the same thing that we did for Bob, reflecting Tina's betting dispositions for *no rain* around 1/2, we can visualize these two dispositions with figure 20.3. Here, the blue 'buy!' segments do not overlap. But the orange 'sell!' segments

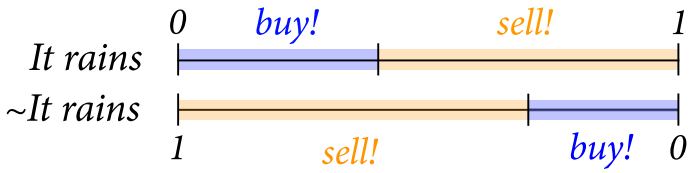


Figure 20.3: Tina's betting dispositions for a $1 bet on it rains *(on top), and Tina's betting dispositions for a $1 bet on* it doesn't rain *(on bottom, reflected about 1/2)*

do overlap. This is also something that would never happen with probabilistic credences. By the negation rule, the probability of *it doesn't rain* has to be one minus the probability of *it rains*. So, if Tina's credences were probabilistic, then she'd be willing to buy a bet on *it rains* at x¢ iff she's willing to sell a bet on *it doesn't rain* at $1 - x$¢.

de Finetti notes that, because the two 'sell!' segments overlap in this way, Tina will be willing to sell a $1 bet on *it rains* for 50¢and sell a $1 bet on *it doesn't rain* for 40¢. That is, we can ask Tina to sell us the followiong two bets, and she'll happily hand them over in exchange for the listed prices.

Bet 3	
$1	if it rains
$0	else

price: 50¢

Bet 4	
$1	if it doesn't rain
$0	else

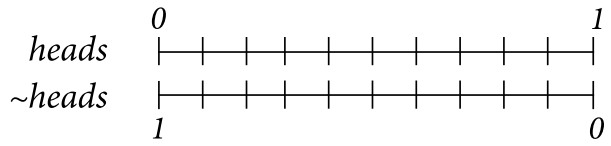
price: 40¢

But now think about what Tina has done. She's handed over two tickets which are certainly worth $1, and what she's gotten back in return is a mere 90¢. Tina has just made herself 10¢ poorer. Surely that's not a rational choice!

Let's go through that again. Consider the possible outcomes: either it will rain or it won't. Here's Tina's net profit from selling each of the two bets above, in each of these two cases:

	It rains	It doesn't rain
Net profit from selling bet 3	−50¢	50¢
Net profit from selling bet 4	40¢	−60¢
Overall net profit	−10¢	−10¢

Exercise 51. *We're going to flip a coin. Consider Hubert, whose credence in* heads *is* 75%*, and whose credence in* not heads *is* 50%*. Given de Finetti's betting interpretation of credence, write down Hubert's betting dispositions below.*



Notice where these betting dispositions overlap in ways that probabilistic credences would not. Then, price the following bets so that, with these prices, Hubert will happily buy the bets and will incur a sure loss. (*To help yourself, fill out the table provided.*)

Bet 5

$1	*if heads*
$0	*else*

price:

Bet 6

$1	*if* ~ *heads*
$0	*else*

price:

	heads	~ *heads*
Net profit from buying bet 5		
Net profit from buying bet 6		
Overall net profit		

Exercise 52. *Consider Bilal, whose credence in* Either it will rain or it will not rain *is* 90%*. Are Bilal's credences probabilistic*? *Which probability rule do his credences violate*? *Can you build a 'Dutch book' against Bilal*?

The Dutch Book Argument

In fact, something similar will happen *whenever* somebody has non-probabilistic credences. This was proven by de Finetti, and it's known as 'the Dutch Book Theorem':

The Dutch Book Theorem If your credences are not probabilistic, then a Dutch Book can be constructed against you.

You might try to use this theorem in an *argument* that rational credences will always be probabilistic.

P1) If you are rational, then you will not be susceptible to a Dutch Book

P2) If your credences are not probabilities, then you will be susceptible to a Dutch Book

$\therefore$ C) If you are rational, then your credences will be probabilities

However, you might worry about the first premise. We've shown that non-probabilistic credences are Dutch-bookable, but we haven't shown that probabilistic credences *aren't* Dutch-bookable. Perhaps there's just no way to avoid being Dutch-bookable.

In fact, we can allay this kind of concern. For there's another theorem—known as the *converse* Dutch book theorem—which shows that having probabilistic credences is always enough to avoid Dutch-bookability.

The Converse Dutch Book Theorem If your credences are probabilistic, then a Dutch Book cannot be constructed against you.

What should we make of the argument? There's a kind of concern which we encountered in our discussion of Pascal's wager. You might think that the argument is simply offering us the *wrong kinds of reasons*. Recall we distinguished between *practical* rationality and *theoretical* rationality. Practical rationality has to do with our reasons for action, whereas theoretical rationality has to do with our reasons for belief. And you might think that credences, like beliefs, are on the side of *theoretical* rationality. And it's natural to think that, when it comes to theoretical rationality, the only relevant kinds of reasons have to do with things like *evidence* and *truth*. But with this argument, de Finetti isn't appealing to anything about *truth* or about your *evidence* in favor of various propositions. Instead, he is appealing to the *practical benefits* of having probabilistic credences. And you might think that these are just the *wrong kinds of reasons* to hold degrees of belief.

As we mentioned with our discussion of Pascal, we could decide to be *pragmatists* who deny the distinction between practical and theoretical rationality. But for those of us committed to that distinction, there's reason to worry about de Finetti's argument in its current form.

In response to this kind of worry, many have tried to 'depragmatize' the Dutch book argument. Roughly, their thought is that the Dutch book argument reveals an *evaluative inconsistency* in non-probabilistic credences. They take there to be a link between your credences, on the one hand, and your *evaluation* of bets, on the other.

Evaluation-Credence Link If your credence in a proposition A is x, then you should evaluate the bet

\$1	if A
\$0	else

price: y¢

as valuable if $x > y$, disvaluable if $y < x$, and neutral if $x = y$.

Then, the 'depragmatized' version of the Dutch book argument goes like this:

P1) If your credences are not probabilities, then you'll evaluate each of a collection of bets as individually valuable, and you'll evaluate their collection as disvaluable.

P2) If your credences are rational, they will not lead you to evaluate each of a collection of bets as individually valuable and yet evaluate their collection as disvaluable.

$\therefore$ C) If your credences are rational, then they will be probabilities.

The thought behind (P1) is this: when Bob looks at bets 1 and 2 individually, he thinks that they're both good bets—they're favorable to him, and so he wants to take them. But when he looks at their combination, he is able to recognize that purchasing both is just losing a quarter no matter what. Bob prefers more money to less, so he doesn't value handing over a quarter no matter what. So he doesn't value the combination of bets 1 and 2, even though he evaluates both of them positively *on their own*.

This 'depragmatized' Dutch book argument is related to some of the pathologies of Bob's credences that we observed last time. We notice that there were two methods for deciding whether or not to buy a bet. On the first method, Bob should compare the bet's expected utility to the utility of the bet's price. If the price is less than the bet's expected utility, then it will be rational for Bob to buy the bet. On the second method, Bob should compare the expected utlity of the act of buying to the expected utility of the act of not buying. If Bob's credences were probabilistic, then both methods would lead to exactly the same conclusion. But because Bob's credences were not probabilistic, the two methods led to radically different evaluations. The 'depragmatized' Dutch book argument is suggesting that inconsistent evaluations like these are the real problem with non-probabilistic credences.

What should we make of the argument's second premise? We could argue for it by appealing to two further assumptions:

Equivalence Principle If two bets are guaranteed to get you the same amount of money in every possible state of the world, then you should value them in exactly the same way.

Package Principle If you regard two bets as individually valuable, then you should also regard the *package* of both of them as valuable.

Applied to the case of Bob above: the package principle says that, since Bob regards bets 1 and 2 and individually valuable, he should regard their combination as valuable, too. And the combination of these bets is just equivalent to the guaranteed loss of a quarter. Since Bob doesn't value the loss of a quarter, the equivalence principle says that Bob should not value the combination of bets 1 and 2. So, putting the two principles together, we get the conclusion that Bob should evaluate the combination of bets 1 and 2 as both valuable and as not valuable. And that's an inconsistency in his evaluations.

Non-probabilists have objected to the package principle. They say that it is suspiciously similar to the disjunction rule which it is being used to argue for. Notice that the disjunction rule says that, since A and $\sim A$ are mutually exclusive, your credence for A and your credence for $\sim A$ have to 'add up' to your credence in $A \vee \sim A$ (which is 100%, by the tautology rule). But the package principle similarly says that your evaluation of two individual bets have to 'add up' to your evaluation of their package. Perhaps non-probabilists should simply deny the package principle?

A. DECISION-MAKING UNDER UNCERTAINTY. [10%] Consider the following decision matrix. The entries in the matrix give the utility of the outcome you get choosing the row act in the column state.

	State S	State T	State U
Act A	30	100	100
Act B	30	30	100
Act C	0	0	150

1. Which act or acts are rational according to the principle of maximin? [2%]

2. Which act or acts are rational according to the principle of maximax? [2%]

3. Which act or acts are rational according to Hurwicz's principle (assuming a risk coefficient of $1/2$)? [2%]

4. Which act or acts are rational according to the principle of minimax regret? [2%] For this question, you must fill out the following table with the *regret* of the row act in the column state:

	State S	State T	State U
Act A			
Act B			
Act C			

B. ORDINAL VS CARDINAL UTILITY. [10%] Consider the following two utility functions.

	U_1	U_2
a	1	0
b	9	360
c	10	405

Table 20.1: The numbers beneath a utility function are the numbers it assigns to the outcome in that row.

1. Explain why these two utility functions are ordinally equivalent. Be sure to give the definition of ordinal equivalence in your answer. [2%]

2. Prove that these two utility functions are cardinally equivalent. Be sure to give the definition of cardinal equivalence in your answer. [8%]

C. EXPECTED MONETARY VALUE/EXPECTED UTILITY. [15%] Consider the following gamble.

We will draw two cards from a well-shuffled deck, one after the other (we will not put the first card back before drawing the second). The gamble pays $17 if we draw two diamonds; and it pays $0 otherwise.

(a) Use the probability rules to show that the probability of drawing two diamonds is $1/17$,and the probability of not drawing two diamonds is $16/17$. (To get credit, you must write down a correct formula involving the probability function, without any numbers, and explain which probability rules you are using.)

(b) Write down the formula for the expected monetary value of a gamble, and use it to show that the expected monetary value of this gamble is $1. [5%] (You must write down the formula, without any numbers, to get credit.)

(c) Suppose that your utilities, as a function of dollars, are given by $U(\$x) = x^2$. Then, write down the formula for the expected *utility* of a gamble, and use it to show that the expected utility of this gamble is 17 utiles. [5%] (You must write down the formula, without any numbers, to get credit.)

D. MEASURING UTILITY. [10%] Ernest is deciding between vacations. Consider the following outcomes:

a = Ernest goes to Africa

c = Ernest goes to camp

j = Ernest goes to jail

Suppose that Ernest tells you that he prefers going to Africa to going to jail. And, given a choice between going to camp and a gamble that gets him a trip to jail with probability $1/10$ and a trip to Africa with probability $9/10$, he's indifferent and could go either way.

Explain how to build a (cardinal) utility function for Ernest's preferences between these vacations. (You might find it easier to if you use the 'zero-one' utility function, but that's up to you.) Be sure to clearly explain where the numbers in this utility function are coming from. If you just write down the numbers, you will not receive full credit.

E. THE SURE THING PRINCIPLE. [10%] Suppose that we're going to roll a fair, 3-sided die. And consider the following four gambles.

	Die lands 1	Die lands 2	Die lands 3
Gamble *A*	\$100	\$45	\$23
Gamble *B*	\$68	\$100	\$23
Gamble *C*	\$100	\$45	\$−100
Gamble *D*	\$68	\$100	\$−100

Given the choice between gambles *A* and *B*, Sun says that she prefers *A*. And given the choice between gambles *C* and *D*, Sun says that she prefers *D*. Explain why Sun's preferences are deemed irrational by the Sure Thing Principle. (Be sure to say what the Sure Thing Principle *is*, and explain how it applies to Sun's preferences.)

F. INFINITY. Give the decision matrix for Pascal's wager, and show the expected utility calculation that Pascal uses to argue that it is rational to believe in God.

G. EVIDENTIAL AND CAUSAL DECISION THEORY. [20%]

Consider the following decision:

> *Before you are two boxes, one on the left ('Lefty') and one on the right ('Righty'). You can either take Lefty or you can take Righty. These are your only options. Your fairy godmother has made a prediction about which box you would take. If she predicted that you would take Lefty, then she left you a prize of $1,000 in Lefty and nothing in Righty. If she predicted that you would take Righty, then she left you a prize of $800 in Righty and nothing in Lefty. The predictions of your fairy godmother are very accurate. You are 90% sure that she has correctly predicted your choice.*

Consider the following states:

K_L: Your fairy godmother predicted you take Lefty

K_R: Your fairy godmother predicted you take Righty

And the following acts:

L: You take Lefty

R: You take Righty

Suppose that your utilities are linear in dollars ($U(\$x) = x$). And suppose that you have the following conditional probabilities:

$$\Pr(K_L \mid L) = \tfrac{9}{10} \qquad \Pr(K_L \mid R) = \tfrac{1}{10}$$
$$\Pr(K_R \mid L) = \tfrac{1}{10} \qquad \Pr(K_R \mid R) = \tfrac{9}{10}$$

(a) Explain why K_L and K_R are states of nature. [2%]

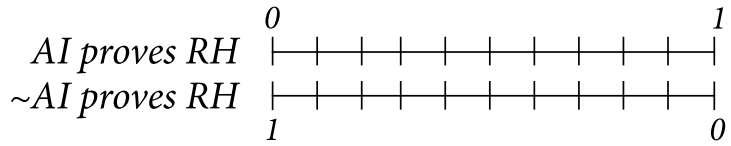
(b) Create a decision matrix for this decision, and fill the cells of the matrix with the utilities of the possible outcomes. [3%]

(c) Explain why evidential decision theory says that it is irrational to take Righty. [5%]

(d) Suppose that the probability that you take Righty is $3/4$. Then, use the law of total probability to explain why the probability of K_L is $3/10$. [5%]

(e) Use your answer from part (d) to explain why causal decision theory says that it's irrational to choose Lefty. [5%]

H. DUTCH BOOKS. [10%] Scott's credence that AI will prove the Riemann hypothesis is 80%, and his credence that AI will not prove the Riemann hypothesis is 30%. Given de Finetti's betting interpretation of credence, fill out the diagram below to indicate the prices at which Scott is willing to buy or sell a $1 bet on AI proving the Riemann hypothesis (in the top line), and the prices at which Scott is willing to buy or sell a $1 bet on AI not proving the Riemann hypothesis (in the bottom line). Notice that the bottom line is reflected, so that zero is on the right and 1 is on the left.



Now, price the following bets so that, with these prices, Scott will happily buy the bets from you and will incur a sure loss. Fill out the provided table to verify that buying these bets will ensure Scott suffers a sure loss.

Bet 1

$1	if AI proves the RH
$0	else

price:

Bet 2

$1	if ~ AI prove the RH
$0	else

price:

	AI proves the RH	~ AI proves the RH
Net profit from buying bet 1		
Net profit from buying bet 2		
Overall net profit		