Formal Models: Section 1*

Zachary Lorico Hertz

21 January 2026

What are we doing here?

This course is a required part of the Berkeley methods sequence. Why? It can help to think of the research process as containing two distinct processes/problems: a *statistical task* (inferring estimands from data) and a *modeling task* (picking estimands that identify parameters). To this point, most of your methods training has lived in the world of *statistical inference* and *identification*: how do we use our sample to learn about the population? In incredibly simple and abstract terms, you could distill the entirety of your previous coursework as dedicated to estimate 'the effect of X on Y.'[1]

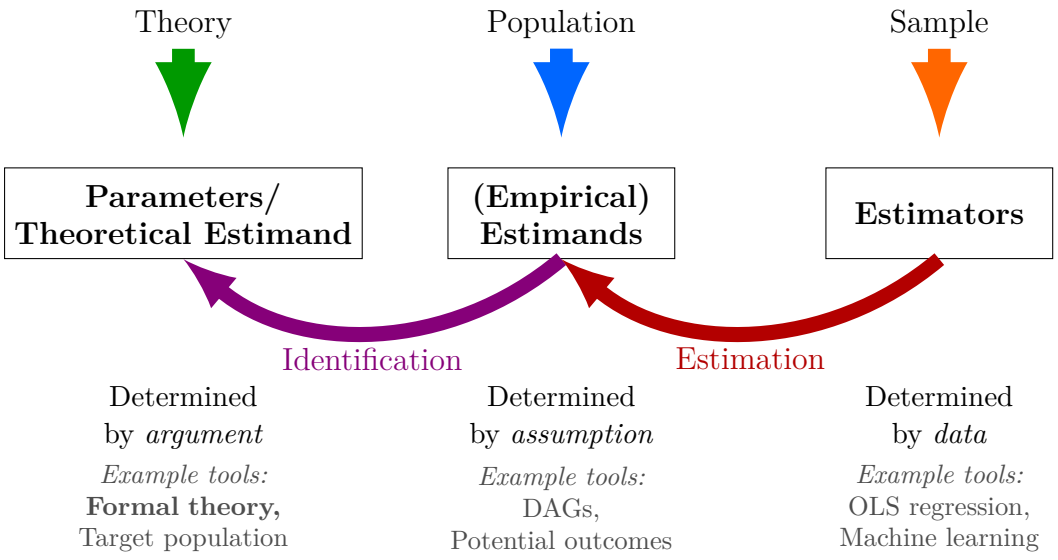

Figure 1: Based on a slide from Peter Hull, incorporating ideas from Lundberg et al. 2021.

But while we have spent much time discussing *whether* and *how* you might estimate something, less attention has been paid to the process of defining *what* the hell you are trying to estimate in

*The notation and definition section draws on materials from Tak-Huen Chau, who himself drew on materials from David Foster and Jieun Kim. All writing (including its poor sense of humor and any errant mistakes) are mine.

[1]This is, of course, a brute oversimplification that I often chafe at. Furthermore, sometimes even this apparently simple question is a rich research question!

the first place. What this means, in terms of the above diagram, is focusing on the modeling task: the link between *theory* and **parameters**. We might be good at this point at estimating effects of X on Y, but is this effect in a partial or general equilibrium? Under what sets of constraints are these effects observed?

This is where formal theory comes in. Without a clear theoretical model, you're just estimating correlations, or average effects, checking for statistical significance and calling them insights - a sort of computationally expensive stargazing.[2] We as researchers need to be precise about the question, its context, and the true parameters of interest. Formal theory helps us define these models in such a way that we can formalize social structures and examine the effects of said structure on individual decisions. Hopefully at this point the link between this course and your previous training in empirical methodology is now more clear; though it is not a main focus of the class, we will return to this at future points in the course.

But why formal models?

I just told you that a moment ago.[3] Nevertheless, in political science we deal often with *strategic decisions* at the individual, party, organization, and even state level: candidates compete for office; parties seek governing coalitions to pass their platforms; organizations seek to enact pet policies; states seek to trade their goods and avoid international conflict. From here on out I'll use the term **"actor"** loosely to refer to the **level of individual we are looking at in a particular situation.**[4]

What does it mean for these actors to make *strategic* decisions? To understand how these *actors*, well, act, we perform a brilliant[5] abstraction relying on a few key assumptions. In plain language, **strategic decisions** means that actors assess outcomes as a function of both their own actions **and** the actions of the other relevant actors. No man is an island — all actors must think about what the other actors will do.

Second, in political science (and social science writ large) we commonly acknowledge that individuals make choices that are influenced not just by other individuals, but by their entire social settings. We call this **structure,** referring broadly to the factors that produce decisions' consequences. In real life, gaining analytical traction over structure is difficult due to its breadth: structure includes states' military capacities, whether individuals believe they have the ability to run for office, the ability of certain parties to form coalitions, and more. This is the second act of abstraction key to our work; in game theory, we specify the structure of a game through actors' choice sets, how their choices map to outcomes, and how actors evaluate these outcomes.

Ultimately, the way we specify these structures will reflect different social theories. Note that this is itself an *analytical choice*! We should **not** expect formal models to tell us whether these theories are ultimately correct. This is not a tool to adjudicate between theories of the world; that is best done by producing a strand of research over time. What we *should* expect to gain from formal models is a way to elucidate the assumptions particular theories make, communicate what consequences we should expect in the real world as a result of these theories, and predictions for actors' behavior as expected by these theories.

[2] Obviously this does not apply if you create your regression tables using the `modelsummary` package instead.
[3] See this somewhat-dated pop cultural reference.
[4] I'll be keeping a running list of terms and their definitions for you to refer to throughout the semester.
[5] In my view, anyways.

Theory is a language

As anything else, theory is written in a… language. *Formal* theory is just theory written in a *formal language.* Any language can be decomposed into the following components:

- an **Alphabet**: the finite set of symbols, characters, or tokens from which the strings of the language are formed. This alphabet could include digits, letters, or any other symbols. In formal theory, our alphabet includes:
  - *Mathematical symbols* such as $\in$, $\notin$, $\forall$
  - *Variables* such as x, y, α
  - *Numbers and sets* such as $\mathbb{R}$
  - *Relation symbols* such as $\succ$
  - *Punctuation* such as parentheses, commas, and brackets
- **Strings**: A string in a formal language is a finite sequence of symbols from the alphabet. For instance, if the alphabet is $\{0, 1\}$, then 0101 and 100 are strings in this language. Some strings are meaningful, others are not.
  - For example, $x \in \mathbb{R}$ is meaningful, while $\in x\mathbb{R}$ is not.
  - The component x refers to the *variable* x; the component $\in$ refers to the mathematical symbol which corresponds to the statement "which is in"; the component $\mathbb{R}$ refers to the set of all real numbers.
  - You can intuit this by speaking this out loud: putting together each component, the first string represents "x in the set of all real numbers" while the second string translates to "in x; the set of all real numbers".
- **Formation Rules**: These are the rules that define which strings of symbols are valid or well-formed in the language. These rules can be very simple or extremely complex, depending on the language.
  - For example, our meaningful statement $x \in \mathbb{R}$ contains an element (x), then uses the relation symbol $\in$ to explan the element's relation to the set $\mathbb{R}$.
- **Syntax and Grammar**: A formal grammar is a complete system of rules for generating strings in a language. These rules describe how to form strings from the language's alphabet that are valid according to the language's syntax.
  - For example, quantifier statements like $\forall$ must specify a domain.

Just as in natural language, where grammar rules let us distinguish "The cat sat on the mat" from "The cat the mat on sat," our mathematical grammar guides us in composing valid formal statements. We will spend the rest of this section introducing this formal language and taking our first steps towards reading, writing, and constructing valid theoretical arguments.[6]

[6] At this point, Tak-Huen flags a recommended companion reading: Velleman, Daniel J. *How to prove it: A structured approach.* Cambridge University Press, 2019. I think this is useful for learning more about proofs but overall goes into too much depth for the scope of this course; I may pull some tips from it in the future and guide motivated students towards reading it.

Sets: Definitions and proper notation

Definition (Sets): A *set* is a well-defined collection of distinct objects. The objects that make up a set can be anything: numbers, people, letters, etc. These objects are called the *elements* or *members* of the set. Formally, a set S is defined as follows:

- An object x is either an element of S or not; this must be determinable *without ambiguity.*
- If an object x is an element of S, we say that "x is in S," which we write as $x \in S$.
- If an object x is not an element of S, we say that "x is not in S" and write as $x \notin S$.

Generally, we write the set by separating elements using commas. If the order of the elements in the set matters, we enclose the set in parentheses. If the order of the elements in the set *does not matter*, we enclose the elements inside a pair of curly brackets (also known as "braces"). So, if I were to define a set S as the even, positive integers that are greater than zero and less than ten, we would write $S = \{2, 4, 6, 8\}$. If the total number of elements in a set is a whole, countable number, we say the set is a *finite set*; if a set is not a finite set we say that it is an *infinite set.* In this class, one special type of set that we will analyze is the set of *real numbers.*

Definition (Real Numbers): The set of *real numbers*, or $\mathbb{R}$, is the set of all rational and all irrational numbers. This includes both positive and negative integers, all fractions and decimals.[7]

- So, 1 is a real number; using set theory language to talk about this, we might say $1 \in \mathbb{R}$.
- $-\pi$ is *also* a real number, and we could also say $-\pi \in \mathbb{R}$.
- $-\frac{1}{67}$ is *also* a real number; how do you think we might write that statement formally?

We may also partition $\mathbb{R}$ into what is called an *interval*, which is a special kind of set.

Definition (Interval): An *interval* is the set of all real numbers that are contained between one lower endpoint, called a lower bound a, and one upper endpoint, called an upper bound b.

- Intervals can contain *neither* the upper nor the lower bound, in which case they are referred to as *"open intervals"*; either the upper or the lower bound, in which case they are referred to as *"half-open intervals"*; or both the upper and the lower bound, in which case they are referred to as *"closed intervals"*.
- While we use curly brackets to enclose most sets, we use square brackets or parentheses to denote that a set is an interval. To indicate that an endpoint is included, we use a square bracket; to indicate that an endpoint is excluded, we use a parentheses.
- So, for example, consider the interval of all real numbers greater than 0 and less than 1. This interval does not include 0 or 1, so we would consider this interval as the set of all x where $0 < x < 1$, and write this open interval as $(0, 1)$.
- If instead we were interested in the interval of all real numbers greater than *or equal to* 0 and less than *or equal to* 1, because this interval includes both 0 and 1 we can define it as the set of all x where $0 \leq x \leq 1$ and write the closed interval as $[0, 1]$.
- If we were interested in an interval of all real numbers greater than 0 and less than *or equal to* 1, we define it as the sest of all x where $0 < x \leq 1$ and write the half-open interval as $(0, 1]$.

At this point, I want to emphasize that sets are defined entirely by the elements that compose the set. Only one set contains no elements whatsoever: this is called the "empty set."

[7] Pause for understanding: is $\mathbb{R}$ a finite or infinite set?

Definition (Empty Set): The *empty set*, or $\emptyset$, is the set containing no elements; its size (the number of elements in a set) is zero.

We can then use the property that sets are defined entirely by their composite elements to *compare* two sets. This is the basis for *set theory*! Consider two sets, which we shall refer to as A and B. We can compare these two sets and conclude that they are *equal* if **and only if** we were to go through every element of A; confirmed that each and every element of A exists in B; then went through every element of B and confirmed that each element of B exists in A.

Note that this is essentially a sentence:

Definition (Set equality): "A equals B if and only if for all elements a in A, a exists in B **and** for all elements b in B, b exists in A."

We can write this definition using our new formal language and only a few unfamiliar symbols:

$$A = B \iff \forall a \in A, a \in B \land \forall b \in B, b \in A$$

Of course, two sets do not have to be equal (and, in fact, usually are not). One special condition occurs when we compare two sets and notice that *every* element in the first set appears somewhere in the second set. If this is true of two sets A and B, we might say that A is a **subset** of B.

Definition (Subsets): We say that A is a **subset** of B if every element of A is also an element of B. We denote this by stating $A \subseteq B$.

Here, we note a few things that follow from this definition of subsets:

- Recall our previous definition of equal sets. If $A = B$, then it is also true that $A \subseteq B$ (can you prove to yourself why this is true?).
- In fact, $A = B$ **if and only if** A is a subset of B and vice versa. I suggest pausing again to justify to yourself why this is true.
- Sometimes, we are interested in subsets *but not sets that are equal*. If A is a subset of B and **they are not equal**, we say A is a **strict or proper subset** and write $A \subset B$.
- Because the empty set contains no elements, it is technically true[8] that all of the elements of the empty set are contained in any other set you compare it to; thus, the empty set is a subset of every other non-empty set. You don't need to understand entirely why this is true, but I encourage you to work through it on your own.

These notes have already begun to use a number of symbols and operators with which we can construct "sentences" in our formal language to talk about sets. I'll introduce a few rapid-fire here. Note that in this class, we will focus only on decisions made on *finite sets* or decisions made on $\mathbb{R}$.[9] Thus, some of the definitions below (e.g. closed, open intervals) are applicable only to $\mathbb{R}$.

[8] Specifically, this is an example of what is formally referred to as a "vacuous truth": a statement that is true because the antecedent cannot be satisfied (that is, the premise(s) are false or cannot be checked). If you're looking for a fun game, my friends have an ongoing bit where we try to slip one into conversation without it being caught.

[9] Two other sets of note, which we will not spend much time with, are the set of *integers* $\mathbb{Z}$ (all whole numbers, both positive and negative) and the set of *natural numbers* $\mathbb{N}$ (or "counting numbers", the set of positive integers not including 0). Rational numbers $\mathbb{Q}$ are formally defined as any number that can be expressed in the form: $\frac{a}{b}$, where $a, b \in \mathbb{Z}$.

Notation reference

- $\mathbb{R}$: the set of real numbers (e.g. 1, -π, $e^{\pi} + 3.2$)
- $\in$: a symbol for inclusion, signifying some element "is in" a set (e.g. $1 \in \mathbb{N}$)
- $\notin$: a symbol for non-inclusion, signifying some element "is not in" a set (e.g. $-1 \notin \mathbb{N}$)
- $\forall$: "for all elements of," indicating that some statement applies to *every single element* in a given set without exception (e.g. $\forall x \in \mathbb{R}$)
- $\exists$: within a set, there exists some particular element (e.g. $\exists x \in \mathbb{N}$ s.t. $x = 1$)
- $\subseteq$: something is a subset of something else (e.g. $\mathbb{N} \subseteq \mathbb{R}$)
- $\subsetneq$: is not a subset of (e.g. $\mathbb{Q} \subsetneq \mathbb{N}$)
- $\subset$: is a strict subset of (e.g. $\mathbb{N} \subset \mathbb{Z}$)
- $\not\subset$: is not a strict subset of (e.g. $\mathbb{Q} \not\subset \mathbb{Q}$)
- Open interval: a set that includes $x \in (a, b) : a < x < b$
- Closed interval: $x \in [a, b] : a \leq x \leq b$
- $\wedge$: "and", used to link logical statements

Pause for understanding

Complete Worksheet Exercises 1 and 2.

Preference relations

Now at this point you might say "Zach, we've spent a lot of time talking about set theory. But I thought we were taking a class on game theory?" So let's take a moment and connect the empirical work to our theory. Recall that in this course, we are focused on formal theory as the *mathematical analysis of strategic decision-making.*

But it's not enough for us to assume actors are *strategic* and make their decisions based on the decisions of other actors; we must also assume that actors are *rational*. What do we mean by rational? Colloquially, the term encompasses behavior that is seen as reasonable, reflective, just, sane, thoughtful. Defining the term from common parlance, one might conclude that rationality is in the eye of the beholder!

Instead, following in the tradition of Smith, Downs, Olson and their kin we take a much narrower definition of rationality. We start by assuming actors have a set of actions they can choose, and have specified outcomes or consequences that they hope to manifest. Our actors are *rational* if, among their possible actions, they choose the action that best acheives their desired outcome.

Definition (Rational choice): Our theory of rational choice will depend on two main tenets:

1. Individuals have well-defined and coherent preferences over potential outcomes.
2. Individuals will act upon these preferences to reach their preferred outcome.

It should become clear at this point that in order to reach a computational solution, we will need to be able to determine how our actors evaluate their preferences across the set of all possible outcomes. To do this, we'll start applying the set theory we've learned to this point! For any problem, we can group possible outcomes into a set. Recall that means any individual outcome is thus an *element* of that set.

To make this more clear, I'll use a specific example. Imagine a road trip with three people, Alex, Bob, and Casey. For this road trip, these three people have to decide who is on aux setting the music. So there are three possible outcomes for who controls the music during this drive: Alex is on aux, Bob is on aux, or Casey is on aux.[10] We could then define a set of outcomes for the music on this road trip as a set M where $M = \{\text{Alex is on aux}, \text{Bob is on aux}, \text{Casey is on aux}\}$. To determine individuals' preferences between any set of outcomes, or more generally to compare individual elements of a set to each other, we'll have to start by taking what is called the *Cartesian product* of the set of outcomes with itself.

Definition (Cartesian Product): The Cartesian product $X \times X$ of a set X with itself is the set of all ordered pairs (x, y), where x and y are elements of X. This set includes every possible pair of elements from X. Formally,

$$X \times X = \{(x, y) \mid x \in X \text{ and } y \in X\}.$$

In our example, because:

$$M = \{\text{Alex is on aux}, \text{Bob is on aux}, \text{Casey is on aux}\}$$

then, the Cartesian product $M \times M$ can be written as:

$$\begin{aligned} M \times M = \{&(\text{Alex is on aux}, \text{Alex is on aux}), (\text{Alex is on aux}, \text{Bob is on aux}), (\text{Alex is on aux}, \text{Casey is on aux}), \\ &(\text{Bob is on aux}, \text{Alex is on aux}), (\text{Bob is on aux}, \text{Bob is on aux}), (\text{Bob is on aux}, \text{Casey is on aux}), \\ &(\text{Casey is on aux}, \text{Alex is on aux}), (\text{Casey is on aux}, \text{Bob is on aux}), (\text{Casey is on aux}, \text{Casey is on aux})\} \end{aligned}$$

Recall that our goal is to relate each individual outcome to each other individual outcome, so we can represent someone's preferences over the whole set of outcomes. So some of these pairings are redundant to each other! Ultimately, what we are looking to define is a subset of the complete Cartesian product known as a "*binary relation*". You can think of a binary relation as a type of function that relates one element to another, though of course we should define this more formally:

Definition (Binary relation): Let X be a set. A binary relation R on X is a subset of $X \times X$. For elements $a, b \in X$, if $(a, b) \in R$, we say that a is related to b under R. This can be denoted by aRb.

Here, we also pause to make another comment on notation. Note that we use $\{\}$ brackets to indicate that elements in a set are order-invariant. For example, $\{a, b, c\} = \{b, c, a\}$. However, we use parentheses to indicate that binary relations are order-variant. For example, $(a, b) \neq (b, a)$.

You may note that the definition of binary relations is not enough to provide us with the preference relation we are trying to prove. Specifically, **preference relations** are special types of binary relations that satisfy two key properties: they must be *complete* and they must be *transitive*. To indicate that something is a preference relation, we use the special symbol $\succsim$.

Definition (Preference relation): A preference relation $\succsim$ over domain X is a binary relation that is both **complete** and **transitive**.

[10] At this point, you might say "Zach, aren't there other solutions? Couldn't they just create a playlist that all three contribute to equally, or start a Spotify jam, or [some other arbitrary other solution]"? And indeed this is our first applied experience with *structure*. The analyst has set the terms of the game; whether it is realistic or complete remains a different question.

For a binary relation to be *complete*, the binary relation must relate every pair of two elements within the set. More formally:

Definition (Completeness): A binary relation P on domain X is *complete* if and only if for all $(\forall)$ $a, b \in X$, either aPb or bPa (or both).

To intuit this, consider the binary relation "is greater than or equal to", used to analyze a set of numbers $S = \{1, 2, 3\}$. In this example, there are a total of 9 ordered pairs $R = \{(1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3)\}$. Working through it in order, we can see that $1 \geq 1$, $1 \not\geq 2$, $1 \not\geq 3$, $2 \geq 1$, $2 \geq 2$, $2 \not\geq 3$, $3 \geq 1$, $3 \geq 2$ and $3 \geq 3$.

So even though there are three ordered pairs for which the binary relation $\geq$ does not apply ($1 \not\geq 2$, $1 \not\geq 3$ and $2 \not\geq 3$), because we still relate those pairs successfully ($2 \geq 1$, $3 \geq 1$ and $3 \geq 2$) the binary relation $\geq$ is **complete**.

For a binary relation to be *transitive*, if the binary relation relates a to b and relates b to c, it must also relate a to c. More formally:

Definition (Transitivity): A binary relation P on domain X is *transitive* if and only if for all $(\forall)$ $a, b, c \in X$, $aPb, bPc \Rightarrow aPc$.

To intuit this, consider a non-mathematical binary relation, "is an ancestor of" on a set of three people $G = \{\text{Amy}, \text{Becca}, \text{Charlotte}\}$. If it is true that Amy is an ancestor of Becca, and it is also true that Becca is an ancestor of Charlotte, it must be true that Amy is an ancestor of Charlotte. Thus, by our definition of transitivity, we can conclude that the binary relation "is the ancestor of" is transitive.

As a counter-example, consider the similar yet distinct binary relation "is the birth mother of" on the same set of three people. If it is true that Amy is the birth mother of Becca, and it is also true that Becca is the birth mother of Charlotte, it is not necessarily true that Amy is the birth mother of Charlotte (and in fact, it is necessarily false!). Thus, the non-mathematical binary relation "is the birth mother of" is *not transitive*.

Putting this together, and returning to our example of friends on a road trip, we can hopefully intuit at this point why preference relations must be complete and transitive. Defining the preference relation as a binary relation "is preferred, rather than", we can reconsider as a question of what a fourth friend Dave prefers.

For the binary relation "is preferred, rather than" to be a preference relation, we must have some information about what David thinks about every possible pair of friends to be on aux, even if the answer is simply that he is indifferent. In other words, it must be complete.

For the binary relation "is preferred, rather than" to be a preference relation, we must also be able to arrange David's thoughts about the aux situation. Consider a situation where David's opinions were non-transitive: Alex is on aux is preferred, rather than Bob is on aux; Bob is on aux is preferred, rather than Casey is on aux; Casey is on aux is preferred, rather than Alex is on aux. In this case, we'd never find a solution! We can avoid these situations by choosing outcomes that the actors can distinguish.

At this point, I note that our use of $\succsim$ indicates what is called the **weak preference relation**. For weak preference relations, we say that the actor "weakly prefers" some first outcome to some second outcome. Maybe Dave prefers that Alex is on aux to Bob being on aux, but we also leave open the possibility that Davis is *indifferent*.

Specifically, indifference is the preference relation when the actor considers two options and concludes that each is "as good as" the other — in short, they are indifferent between the two. Mathematically we use the symbol $\sim$ to refer to indifference. Formally:

Definition (Indifference): The indifference relation $\sim$ over domain X from $\succsim$ is defined as: for $x, y \in X$, $x \sim y$ if and only if $x \succsim y$ and $y \succsim x$.

At times, we might want to consider preferences without allowing for indifference. For this we use the **strict preference relation.** This occurs when the actor "strongly prefers" some first outcome to some second outcome. If this is true, the actor considers one option to be "better than" another option. Consider that instead Alex might prefer that Alex is on aux to Bob being on aux, and is not indifferent between the two consequences. Formally:

Definition (Strict preference): The strict preference relation $\succ$ over domain X from $\succsim$ is defined as: for $x, y \in X$, $x \succ y$ if and only if $x \succsim y$ and $y \not\succsim x$.

Pause for understanding

Complete Worksheet Exercise 3.

Expected utility and utility functions

To this point, we've discussed the key components of the modeling that we will be doing in this class, but I want to pause and make it explicit. Every decision problem, theory, and model that we will approach in this class can be described formally with the following four components:

1. A set of **choices** or acts that actors pick from, that we call A. One of these actions a will be chosen as the actor's decision.
2. A set of **states of the world** that influence the decisions but cannot be controlled by the actors, that we call S. These states s are "mutually exclusive," meaning only one state of the world s_i can occur, and "exhaustive," meaning one state $s_i \in S$ must occur.
3. A set of **outcomes**, that we call C, where an individual outcome is the consequence to a given pair of actions a and state of the world s.
4. A **preference relation** that actors hold over the outcomes, that we call P. These preferences are by assumption complete, transitive, and fixed (ie. they do not change).

In this class, we will generally discuss preferences over outcomes: things like policy positions, election results, or the distribution of goods. But, of course, often outcomes are uncertain. This is in large part because of the **state of the world**.

Definition (State of the world): All factors in a given game/decision problem that have an influence on the outcome (either its likelihood of occurring, or its potential payoffs) yet are not controlled by any of the actors in the model.

The state of the world is important: it determines the costs and benefits to particular consequences, the likelihood that certain actions lead to certain consequences, and thus adds considerable uncertainty to our models. Ideally (for an actor) they would be able to observe the state of the world; if they knew the state of the world, actors could know with certainty the consequence to every action and precisely pick actions that lead to their most-preferred outcome.

But, in reality, the state of the world is usually unknown. Consequences remain opaque and uncertain to actors, who then must assess their possible actions, the actions' possible consequences, and decide which action is likeliest to lead to the best final outcome. Given the complexity in these situations, a simple binary preference relation that gives us ordinal preferences over outcomes is no longer sufficient. For our actors to remain rational (ie, to act consistent with their preferences), given the level of uncertainty, we must first find a mathematical function that can tell us the probability for which we should expect each outcome to actually occur. This is where **lotteries** come in.

Definition (Lottery): A *lottery* is a probability distribution over outcomes. If X is a set of outcomes, a lottery p assigns a probability $p_i \geq 0$ to each outcome $x_i \in X$, where $\sum_i p_i = 1$.

For example, consider a lottery that gives you \$100 with probability $P(\text{Win } \$100) = 0.3$ and \$0 with probability $P(\text{Win nothing}) = 0.7$. We might write this as $l = (0.3 \cdot \$100, 0.7 \cdot \$0)$.

Now, if we know someone's preferences over *certain* outcomes (represented by utility function u), how should they rank lotteries? The **expected utility** of a lottery is simply the probability-weighted average of the utilities of its outcomes:

Definition (Expected Utility): Let u represent preferences over outcomes X. The *expected utility* of a lottery p that yields outcome x_i with probability p_i is:

$$U(p) = \sum_{i=1}^{n} p_i u(x_i)$$

This is just a weighted average! If you have a 30% chance of getting \$100 and a 70% chance of getting \$0, and your utility function is $u(x) = \sqrt{x}$, then your expected utility is $(0.3 \cdot \sqrt{100}) + (0.7 \cdot \sqrt{0}) = 0.3 \cdot 10 + 0 = 3$.

Remark: In Problem Set 1, you'll work with expected utility to analyze decision-making under uncertainty. For now, what you should take away from this is that expected utility lets us rank lotteries, or outcomes under uncertainty, by taking the probability-weighted average of outcome utilities.

Risk attitudes

One important factor is how much actors actually *know* about the state of the world. There are three conditions for actors when they make decisions. The first, *certainty*, refers to situations in which actors know the state of the world before they make any decisions. Under certainty, states of the world are fixed. As a result, decisions under certainty are, from the perspective of decision theory, trivial.

The second condition is called risk. Under risk, there is a known probability that each state will occur; the probability of each state occurring is derived from many iterations or known frequencies. We model decisions under risk using lotteries and the expected utility framework described above.

The third, *uncertainty*, refers to situations in which the probability of states of the world occurring is either unknown, or lacks a stable, re-occurring probability. Under uncertainty, states of the world are probabilistic and unclear. As a result, each actor may have differing beliefs about the underlying probability distribution determining states of the world.

When we face situations of risk — in short, when probabilities are known — expected utility theory allows us to do more than just identify optimal choices. It also lets us characterize how different actors respond to risk. Perfectly-rational actors might still have completely different preferences in identical lotteries depending on their attitudes about risk. Consider three types of decision-makers facing the same lottery:

Definition (Risk Neutrality): An actor is *risk neutral* if they care only about the actual expected value of the payoff. Risk neutral actors have a *linear* utility function. Formally, this means $u(x) = a + bx$ for some constants a, b (with $b > 0$). For a risk-neutral person, a 50-50 lottery between \$0 and \$100 is exactly as good as receiving \$50 for certain.

Definition (Risk Aversion): An actor is *risk averse* if they prefer a certain payoff to a lottery with the same expected value. Formally, this corresponds to a *concave* utility function (one where $u''(x) < 0$, like $u(x) = \sqrt{x}$). A risk-averse person would prefer \$50 for certain over a 50-50 lottery between \$0 and \$100, even though both have expected value \$50.

Definition (Risk Loving): A decision-maker is *risk loving* (or *risk acceptant*) if they prefer a lottery to a certain payoff with the same expected value. This corresponds to a *convex* utility function ($u''(x) > 0$, like $u(x) = x^2$). These folks would rather take the 50-50 gamble than receive \$50 with certainty.

You'll work with these concepts in the problem set, where you'll calculate certainty equivalents and explore how different utility functions generate different attitudes toward risk.

A deeper technical note (for later)

You might be wondering: why *should* we use expected utility to rank lotteries? What assumptions about preferences make this the "right" way to evaluate uncertain outcomes? The answers involve two key axioms:

Definition (Continuity): $\succsim$ on lotteries $p \in P$ satisfies continuity iff: For any 3 lotteries, $p, p', p'' \in P$, where $p \succ p' \succ p''$, $\exists \alpha \in [0, 1]$ such that $p' \sim \alpha p + (1 - \alpha)p''$.

Definition (Independence of irrelevant alternatives): $\succsim$ on lotteries $p \in P$ satisfies the independence of irrelevant alternatives property iff: For any 3 lotteries, $p, p', p'' \in P$, and any $\alpha \in (0, 1)$, $p \succsim p' \Leftrightarrow \alpha p + (1 - \alpha)p'' \succsim \alpha p' + (1 - \alpha)p''$.

Combining these axioms lead us to a key theorem:

Definition (von Neumann-Morgenstern utility theorem for finite outcomes): Let u represent a preference over outcomes X. Let $\succsim$ be a rational preference (which implies completeness and transitivity) over lotteries, P. Then $\succsim$ satisfies continuity (C) and independence of irrelevant alternatives (IIA) if and only if $\succsim$ is representable by a function:

$$U(P) = p_1 u(x_1) + \cdots + p_n u(x_n) = \sum_{i=1}^{n} p_i u(x_i)$$

In plain English: if an actor's preferences over lotteries satisfy some reasonable-sounding assumptions (completeness, transitivity, continuity, and independence), then the actor *must* be acting to

maximize expected utility. This is key to most of game theory, and we'll return to it later in the course when we have more time to appreciate its implications and limitations.[11]

For now, focus on *using* expected utility to solve problems; we'll worry about *justifying* it later.

Optional extra reading

For brevity and because this section is less relevant to the immediate lecture or problem set material, I've cut this section from last year's notes. This also means I'm spending minimal time adding explanation to the subsequent math. If you are interested in taking the material one step further, however, I've left it included in the lecture note document along with some minimal explication to help those of you who may be so inclined.

Transitivity and acyclicity

Remark (Negation): For a lot of proofs (by counterexample or by contradiction), we need to negate the statement in question. We will practice a few in future sections, but Chapter 1 of Velleman will provide good reference for that. More generally, Velleman p.376-379 provides an excellent reference on how to construct a proof.

Definition (Acyclicity): A binary relation P on a domain X is said to be *acyclic* if there are no finite sequences of distinct elements $a_1, a_2, \dots, a_n \in X$ (with $n > 1$) such that $a_1 P a_2, a_2 P a_3, \dots, a_{n-1} P a_n, a_n P a_1$. In other words, the relation P does not allow for cycles among distinct elements. Formally,

$$\nexists a_1, a_2, \dots, a_n \in X \text{ with } n > 1 \text{ such that } a_1 P a_2 \wedge a_2 P a_3 \wedge \dots \wedge a_{n-1} P a_n \wedge a_n P a_1.$$

Summary

To understand how political actors make decisions, we start by considering a set of consequences. We assume that actors are rational and strategic: they will act in response to all other relevant actors, and in such a way that they try to reach an outcome that they prefer over the set of all possible outcomes. As we conduct our analyses, actors' action choices will be sets; we'll have to compare those! Similarly, consequences (outcomes) will be sets; we'll have to see how people rank / compare them.

As analysts, we choose a set of outcomes that is appropriate for the question we are modeling. This means we avoid incomplete and nontransitive preferences, at times by imposing structure on our model.

In our analysis, we take individual actors' preferences to be fixed. This allows us to make the key conclusion that changes in behavior are caused by changes either in the structure of the model, or by changes in the information that actors have access to.

There are a number of key misconceptions about the definition of rationality here that I want to respond to before we finish. First, we do not assume that actors are literally calculating utilities for

[11] If you're curious about the proof or want to dive deeper now, reach out and we can discuss during office hours. But again, fair warning: you don't need this depth for Problem Set 1!

hypothetical outcomes. We are not building a cognitive theory of political processing[12] but instead simply aim to understand political actions and consequences in a way that respects the complexity of cognition and the variation in individual choice. Second, it is important that this analysis starts with outcomes, takes preferences over outcomes to be fixed, and aims to explain actions; it is not a way to recover preferences over outcomes.

Concepts covered:

- Common notations
- Binary relations and preference relations
- A primer on expected utility and risk attitudes

[12] At least, not in this class!