CHAPTER 1

# Random Utility

## 1.1 THE ANALYST AND THE AGENT

The main character in this book is the *analyst*. She is a researcher: an econometrician, an experimental economist, and so on. The analyst has access to data about the behavior of an *agent* (or a population of agents) summarized by a stochastic choice function (SCF) $\rho$. The analyst wants to understand $\rho$ to predict the agent's behavior in a new situation, for example, forecast demand for a new product. A benevolent analyst wants to be able to measure the agent's welfare.[1]

This book focuses on nonstrategic situations, where the data are, for example, the occupational choices in a population or response frequencies in a laboratory or field experiment. Our agents don't play games with each other or with the analyst. Of course, strategic interactions are prevalent in economics, but it's worthwhile to first see how much we can understand about individual behavior. We assume that the analyst is passively studying the agent. The analyst's decisions (which will be unmodeled here) may ultimately impact the agent as new products get introduced or new contracts or mechanisms get designed, but our agents are not strategic enough to take this into account.[2]

Many analysts model the agent as a utility-maximizing creature and make various other more specific assumptions. Each model puts some restrictions on the class of behaviors that are allowed. We will try to understand these restrictions and the ways the various classes connect to each other.

Understanding the relationships between models is interesting in its own right but can also serve some practical purposes. The analyst often has to pick a particular model and it's good to know what the possible trade-offs between these models are.

[1] I will refer to the analyst as "she/her," or sometimes "us." I use "they/them" for the agent(s).

[2] In fact, the situation of the mechanism designer is similar to the situation of our analyst: She has some information about behavior in various situations and picks a situation (mechanism) to induce the agent to behave in a desired way.



https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

## 1.2 DETERMINISTIC CHOICE

We start with deterministic choice because this will be the basis for much of what is to come in this book. This will also establish notation used throughout.

Let $X$ be the set of all possible alternatives that our agent might be facing. Typical elements are denoted $x, y \in X$ and may stand for things such as brand choices, employment status, number of children, market entry decisions, or choosing which perceptual stimulus is stronger in a lab experiment.

The analyst observes the agent's choices in multiple-choice situations. The data of the analyst is a *choice function* that says what the agent does in each situation. We will treat the choice function as observable to the analyst – we will assume that she can collect this data by observing how people behave in real life or by designing a lab or field experiment.

In decision theory and consumer theory, a choice situation is typically summarized by the *menu* (a subset of $X$) the agent is choosing from (e.g., the actual menu at the restaurant, or the set of insurance plans an employer offers, or the budget set in consumer theory).

Let $\mathcal{A}$ be the collection of all nonempty and finite subsets of $X$, with typical elements $A, B, C$, which we call *menus*.[3] A single-valued *choice function* is a mapping $\chi : \mathcal{A} \to X$ such that $\chi(A) \in A$. That is, for each menu $A \in \mathcal{A}$, the analyst observes what is chosen. The condition $\chi(A) \in A$ just means that the agent cannot choose items outside of the menu.

The "revealed preference" exercise of Samuelson (1938) seeks to rationalize such observations by preference maximization and to uncover the preference relation from the observed data.

A binary relation $\succsim$ on $X$ is a *preference* if it is:

- complete ($x \succsim y$ or $y \succsim x$ for all $x, y \in X$) and
- transitive ($x \succsim y$ and $y \succsim z$ implies $x \succsim z$ for all $x, y, z \in X$).

Moreover, the relation is a *strict preference* if it also satisfies the following property:

- $x \succsim y$ and $y \succsim x$ implies $x = y$ for all $x, y \in X$.

The last requirement (called *antisymmetry*) means that the agent is never indifferent between two distinct options.

We say that a strict preference $\succsim$ *represents* $\chi$ whenever, for each $A \in \mathcal{A}$, $\chi(A)$ is the highest ranked element of $A$ according to $\succsim$. The key here is that

[3] In introductory microeconomics and consumer theory, $X$ is typically an infinite set of consumption bundles ($X = \mathbb{R}^n_+$, where $n$ is the number of goods) and the agent is choosing how much of each good to consume (the menu is an infinite set). The analysis then quickly assumes differentiability and convexity and characterizes optimality by first-order conditions. In discrete choice theory, the analysis is somewhat different: The menu is finite (discrete) and the optimality conditions are a set of inequalities instead of equalities. We allow $X$ to be infinite, but the menu will always be finite, although there is some work on stochastic choice with infinite menus (see, e.g., Bandyopadhyay, Dasgupta, and Pattanaik (1999)).

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

the agent maximizes *the same* preference on $X$ irrespective of which menu they are facing. If the preference is allowed to depend on the menu, we can explain every possible choice function and our model is not falsifiable (so there is no way of testing if it's true).

There are $\chi$s that cannot be represented by any $\succsim$; they are sometimes called "irrational," "behavioral," or "boundedly rational."

The key test for deterministic preference maximization is known under many names, such as Sen's $\alpha$ condition (Sen, 1971), Arrow's IIA (Arrow, 1959), or Chernoff's condition (Chernoff, 1954). The axiom imposes consistency conditions on choices from various menus.

**Axiom 1.1 (Sen's $\alpha$).** If $x \in A \subseteq B$, then $x = \chi(B)$ implies $x = \chi(A)$.

This axiom says that if alternative $x$ beats all things in a menu, it must also beat all things in a subset of the menu.

**Proposition 1.2.** *A choice function $\chi$ satisfies Sen's $\alpha$ if and only if there exists a strict preference relation that represents it. Moreover, $\succsim$ is unique.*

A simple proof is, for example, in Osborne and Rubinstein (2020). The assumption that $\mathcal{A}$ contains all menus can be relaxed as long as it contains all pairs and triples.

Decision theorists are attracted to results like Proposition 1.2 because they provide an exact translation between two languages:

- what is observable (the choice function $\chi$) and
- what is a mathematical *representation* (the preference $\succsim$).

This exact translation helps us understand the connections between the two ways of describing choice. It also offers a test of "rationality": If our agent violates Sen's $\alpha$, then they cannot be maximizing a complete and transitive preference.

To deal with indifferences, economists often consider a multivalued *choice correspondence* $\chi : \mathcal{A} \to \mathcal{A}$ such that $\chi(A) \subseteq A$. The idea behind multivalued choice is that from any given menu, the agent sometimes chooses one alternative and sometimes another (the set of those choices must be a subset of the menu). The analyst records both of these choices and interprets this as indifference. For choice correspondences, an additional condition, known as Sen's $\beta$, is needed to characterize preference maximization. Conditions $\alpha$ and $\beta$ combined are called weak axiom of revealed preferences (WARP). For details, see Chapter 2 of Kreps (1988) and Chapter 1 of Mas-Colell, Whinston, Green, et al. (1995). We will not deal with choice correspondences because the theory of stochastic choice provides a more precise way of modeling the situation where the agent makes different choices from the same menu.

So far, we have two languages: the observables (choice function $\chi$) and the representation (preference relation $\succsim$). To make the math easier, economists often use yet another language to represent choices – the utility functions. This

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

allows them to use familiar tools from optimization theory, such as first- and second-order conditions, Hamilton–Jacobi–Bellman equations, and so on.

A preference $\succsim$ is represented by a *utility function* $U : X \to \mathbb{R}$ whenever

$$x \succsim y \text{ if and only if } U(x) \geq U(y).$$

We will interchangeably write $U : X \to \mathbb{R}$ and $U \in \mathbb{R}^X$ for the same object, thanks to the useful notation in mathematics that says that if $X$ and $Y$ are sets, then $Y^X$ is the collection of all functions from $X$ to $Y$.

If $\succsim$ is complete and transitive and $X$ is finite or countable, then a utility representation of $\succsim$ always exists. A classic counterexample when $X$ is uncountable are lexicographic preferences. Since we often have to deal with uncountable $X$, for example, consumption bundles (as in price theory) or lotteries (Chapter 4), typically continuity is assumed to get a representation.[4]

For any preference, there is a multitude of utility representations: If $U$ represents $\succsim$, then any monotone transform $\phi(U)$ also represents $\succsim$.

**Proposition 1.3.** *Functions $U_1, U_2$ represent the same preference $\succsim$ on $X$ if and only if there exists a strictly increasing function $\phi : R_1 \to \mathbb{R}$ such that $U_2(x) = \phi(U_1(x))$ for all $x \in X$, that is, $U_2 = \phi \circ U_1$. Here $R_1$ is the range of $U_1$ defined by $\{U_1(x) : x \in X\}$.*

This is called *ordinal uniqueness*, that is, utility is unique up to the ordering of alternatives but its scale does not have any meaning. In particular, if $u(x) - u(y) > u(z) - u(w)$, then we are tempted to say that $x$ is preferred to $y$ "more intensely" than $z$ is to $w$, but this statement does not have any meaning in terms of choices because we can always take a different utility function that represents the same preferences where the inequality is reversed. Later in Section 4.1, we will see stricter "cardinal" uniqueness results.

## 1.3 STOCHASTIC CHOICE

As mentioned earlier, if the agent is alternating choices from the same menu, the classical approach is to ignore the frequency of such choices and treat them as indifferent. This means that a person who chooses $x$ from menu $\{x, y\}$ 99% of the time and another person who chooses $y$ 99% of the time are classified as the same type.

In this book we will take the choice frequencies seriously and try to extract information from them. To do this, we need to enrich the set of observables: For each menu $A$ and item $x \in A$, let $\rho(x, A)$ be the frequency with which a choice of $x$ from $A$ was observed.[5] In reality, we will have a finite sample of $n$ observations, but we will think of $\rho(x, A)$ as the limiting frequency as

[4] For the finite and countable cases, see Propositions 3.2 and 3.3 of Kreps (1988). For uncountable $X$, see Theorems 3.5 and 3.7 of Kreps (1988) or Chapter 9 of Ok (2014).

[5] The recent paper of Ok and Tserenjigmid (2022) compares the choice-correspondence approach to the choice-frequency approach.

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

$n \to \infty$.[6] A stochastic choice function collects these limiting frequencies as a function of the menu.

For any finite set $Z$, let $\Delta(Z)$ denote the set of *probability distributions* over $Z$, that is, functions $p : Z \to [0, 1]$ such that $\sum_{z \in Z} p(z) = 1$. For each menu $A$, the values of $\rho(\cdot, A)$ form a probability distribution over $A$, so we can think of the SCF as a map that takes a menu $A$ and maps it into $\Delta(A)$.

**Definition 1.4.** An *stochastic choice function* (SCF) is a mapping

$$\rho : \mathcal{A} \to \Delta(X)$$

such that $\sum_{x \in A} \rho(x, A) = 1$ for all $A \in \mathcal{A}$.

Sometimes not all menus are observed, in which case the domain of $\rho$ is smaller. For example, in experiments often there are just binary menus. To simplify notation in this case, we will write $\rho(x, y) := \rho(x, \{x, y\})$ when $x \neq y$ and define $\rho(x, x) := 0.5$.

In discrete choice econometrics the menu is often fixed but what varies are the attributes of these alternatives. The first three parts of the book focus on menu variation and the last part of the book focuses on attribute variation. However, the distinction between the two approaches is not clear cut; for example, for lotteries, each alternative is characterized by a vector of attributes (probabilities of each payoff).

If our analyst is observing a single individual who faces the problem repeatedly (as it happens in some within-subject experiments), then $\rho(x, A)$ is the fraction of times the agent chose $x$ from $A$. Stochastic choice functions can also capture population-level data. For example, McFadden (1974) studied transportation choices of the Bay area population. In this situation, $\rho(x, A)$ is the fraction of the population choosing $x$ from $A$. In such applications, choice has two sources of stochastic variation: Individual randomness (how much choice varies if a given person is sampled over and over again) and heterogeneity of preferences (how much choice varies across people).

While it's easy to imagine that preference heterogeneity leads to nontrivial choice frequencies in the aggregate data, it's less obvious why the choices of a single individual should be stochastic. Yet, stochastic choice is routinely observed. This was established first and foremost, in the context of discrimination between perceptual stimuli (Fechner, 1860; Thurstone, 1927). The following example discusses perception of weight, but similar experiments are used in the study of other senses: hearing, touch, vision, and so on.[7]

[6] Taking limiting frequencies as a primitive is routine in econometrics for the purpose of estimation and identification of parameters. We will talk about this more in Chapter 2.

[7] For example, in some experiments, in each trial the subject faces a screen where a fraction of dots is moving in a coherent direction (left or right), while others are moving randomly, and the agent is incentivized to guess the correct direction of motion (see, e.g., Newsome, Britten, and Movshon (1989), Bogacz, Brown, Moehlis, Holmes, and Cohen (2006), and Drugowitsch, Moreno-Bote, Churchland, Shadlen, and Pouget (2012)). A similar design was used by Dean and Neligh (2023).

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

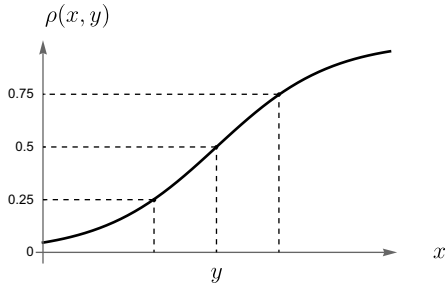


Figure 1.1 An S-shaped psychometric function.

**Example 1.5 (Perception Task).** Let $X = \mathbb{R}_+$ be a collection of weights (the weights all look the same, or the experimental subject's view is obstructed). The subject is facing a series of binary menus $A_i := \{x_i, y_i\}$, $i = 1, \ldots, n$, where $x_i, y_i$ are drawn i.i.d. from some distribution $\pi \in \Delta(X)$. The subject is tasked with picking the heavier of the two objects: There is a positive payoff for a correct guess and zero for incorrect. The analyst records the subject's choice over many i.i.d. trials. In the limit, we get $\rho(x, y)$.

It is interesting to examine $\rho(x, y)$ as a function of $x$ for a fixed value of the reference weight $y$. This is called a *psychometric function*; in fact we have a family of psychometric functions indexed by $y$. △

Numerous experiments in psychology and psychophysics can be summarized by the following stylized facts (see Woodrow (1933) and Gescheider (1997)). First, psychometric functions are typically S-shaped. This means that if $x$ is close to $y$, it is hard for the subject to discriminate between them and accuracy is low. If $x$ is far from $y$, the accuracy improves. It is typical to use the cumulative distribution function (CDF) of the normal distribution $\Phi$ to model psychometric functions.

Another stylized fact is *diminishing sensitivity*: A given weight difference between $x$ and $y$ may be big enough for the subject to notice when both $x$ and $y$ are small, but not big enough when $x$ and $y$ are both large. One way to state this stylized fact is to say that the family of psychometric functions $\rho(\cdot, y)$ gets flatter as $y$ grows.[8] Diminishing sensitivity has been incorporated into many psychological theories, such as Prospect Theory (Kahneman and Tversky, 1979) and Salience Theory (Bordalo, Gennaioli, and Shleifer, 2012).

Yet another stylized fact is *payoff-monotonicity*, which says that the error rate diminishes if the payoff for guessing correctly increases. There is some debate about this between economists (who think that incentives matter) and psychologists (who think they don't).

[8] This is often operationalized as the requirement that the interquartile range, depicted in Figure 1.1 is an increasing function of $y$. This is related to the Weber–Fechner law, which was originally formulated in terms of *just noticeable differences*, a theoretical construct that is inconsistent with our first stylized fact (S-shaped psychometric functions).

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

The final stylized fact is *frequency-dependence*, which says that $\rho(x, y)$ depends on the distribution $\pi$ of weights across trials. Intuitively, this is because the agent gets attuned to the range of weight variation, so that the same weight difference can be perceptible if all weights in the experiment are in some small range but may go unnoticed if the weights vary a lot from trial to trial. Notice that frequency-dependence implies that we should more accurately be talking about $\rho^{\pi}(x, y)$, where $\pi$ is fixed in a given batch of trials and the analyst runs several batches each with a different $\pi$.

While it may be unsurprising that perception of physical stimuli is random, there is a body of experimental evidence showing that economic choices are random as well. Mosteller and Nogee (1951) were first to show that choices between lotteries show substantial switching. This is true whether trials are separated by days (Tversky, 1969; Hey and Orme, 1994) or minutes (Camerer, 1989; Ballinger and Wilcox, 1997; Agranov and Ortoleva, 2017; Agranov, Healy, and Nielsen, 2023). This is true even in questions that offer dominated options.

We will now discuss various reasons why individual choices fluctuate. Each of them corresponds to a particular *representation* of $\rho$.

## 1.4 REPRESENTATIONS

The easiest case is population heterogeneity. For example, in the Hotelling (1929) model, consumers' or voters' bliss points are distributed along a line. More generally, we are given a probability distribution over utility functions that specifies the frequency of each utility in the population. This is called a *random utility* representation and our formal analysis of stochastic choice will begin with it. Each individual's utility function is deterministic, but choices appear random to the analyst as she only observes aggregate data. This model is at the heart of discrete choice econometrics. The heterogeneity of tastes is important for firms (e.g., to choose the product mix, which is something they can't do based on knowing just the average demand) and to policymakers (who care about distributional effects).

What about stochastic choices of a single agent? Here there are more possible mechanisms, all of which will be discussed in detail later on:

1. *Random utility*. Instead of a distribution of utilities in the population, we now have a distribution of utility realizations for a fixed agent.[9] In perception tasks, perceptions are random. For example, Thurstone (1927) assumed that the perceived stimulus equals true stimulus plus a normally distributed error, which leads to what is now known as the *probit model*. In choice tasks, the tastes of the agent fluctuate from trial to trial.

2. *Learning*. Here the agent's tastes are fixed, but their information evolves as they learn new things. The agent gets a noisy signal of the true state of the world and updates their beliefs using the Bayes rule. The agent's information

[9] This is similar to Harsanyi's purification in game theory (Harsanyi, 1973a).

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

is private and unobservable to the analyst, so observed choices are stochastic. The main two variants of the model are when information is exogenous (passive learning) or chosen by the agent (active learning), also known as "rational inattention."

3. *Random consideration*. The agent's tastes and information might be fixed, but they may not always be paying attention to the same objects in the menu. If the attention process is random, it will lead the agent to consider different subsets of the menu (called consideration sets) from trial to trial, thereby generating random choices.

Notice that 2 and 3 offer two different models of attention (endogenously choosing the information vs. being exogenously restricted to a subset of the menu). We will treat them in separate chapters.

In all of these stories above, choices are actually deterministic from the point of view of the agent. They know what their craving is today, or what they learned so far, or which options they are considering. Observed choices appear stochastic to the analyst as a result of the informational asymmetry between the two characters. In the following two stories, choices are random even in the agent's eyes, so both our characters are on the same footing.

4. *Trembling hands*: The agent cannot perfectly control their choice: There is a random implementation error or decision error. In some models, this error is exogenous; however, in others, the agent may control mistakes at a cost. Observed randomness is then the result of a balance between the importance of choosing correctly and the cost of doing so.

5. *Deliberate randomization*. The agent likes to randomize. They view each menu $A$ as the set of probability distributions $\Delta(A)$ and pick a favorite distribution according to some preference that may capture nonlinear probability weighting, a wish to hedge their bets, or aversion to regret.

This book starts with random utility. This is by far the most popular model to study population-level data: Almost all of discrete choice econometrics and demand system theory stem from this model. Moreover, much of the classical decision theory work on stochastic choice is about random utility. A good understanding of this model is also a prerequisite for the other models.

## 1.5 RANDOM UTILITY

There are three equivalent ways to formulate the model mathematically: (1) a probability distribution over preferences, (2) a probability distribution over utility functions, and (3) a random utility function. It may seem like excessive formalism to define all three here but going forward it will be convenient to seamlessly switch between them, depending on the application or context, so I want you to get comfortable with all three.

Let $\mathcal{P}$ be the set of all strict preferences over a finite set $X$. Let $\mu \in \Delta(\mathcal{P})$ be a probability distribution over strict preferences. Depending on our interpretation of $\rho$, $\mu$ is either the distribution of preferences in the population or

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

the probability that governs the evolution of the preferences of the individual. For any $A \in \mathcal{A}$ and $x \in A$, let

$$N(x, A) := \{\succsim \in \mathcal{P} : x \succsim y \text{ for all } y \in A\}$$

be the set of preferences that rationalize the choice of $x$ from $A$.

**Definition 1.6.** $\rho : \mathcal{A} \to \Delta(X)$ is represented by a *distribution over preferences* if there exists $\mu \in \Delta(\mathcal{P})$ such that $\rho(x, A) = \mu(N(x, A))$ for all $A \in \mathcal{A}$ and $x \in A$.

Notice that if we observe choices from only one menu, then any $\rho$ has such a representation. For all $x \in A$, we can just define the probability that $x$ is ranked highest in $A$ to be equal $\rho(x, A)$; the relative ranking of non-top items does not matter. It is the nontrivial menu variation that gives content to the representation.

### 1.5.1 Invariance of $\mu$

The key assumption is that the distribution $\mu$ does not depend on the menu $A$ – it is a structural invariant of the model. If $\mu$ is allowed to depend on the choice set in an arbitrary way, then any SCF $\rho$ can be trivially explained (why?).

A possible complication occurs if the invariance assumption is actually satisfied by the data-generating process but violated in the observed sample because of the way the sample is collected. For example, the distribution of preferences between two brands of orange juice can be different depending on whether the menu of choices is Whole Foods or Walmart because of self-selection: Different people choose to go to these stores.

For now, we will assume that the data-generating process and our sample are free of such effects. This assumption will let the analyst estimate $\mu$ based on choices from some incomplete set of menus $\mathcal{A}^*$ and predict choice from a new menu $A \notin \mathcal{A}^*$, for example, when a new product is introduced.

### 1.5.2 Equivalent Definitions

A slightly different object than a distribution over preferences is a distribution over *utilities*. Our set $N$ becomes

$$\begin{aligned} N(x, A) &:= \{U \in \mathbb{R}^X : U(x) \geq U(y) \text{ for all } y \in A\} \\ &= \left\{U \in \mathbb{R}^X : U(x) = \max_{y \in A} U(y)\right\}. \end{aligned}$$

Now $N$ stands for the set of utility functions that rationalize the choice of $x$ from $A$.

When $X$ is finite, it is without loss of generality to consider discrete measures over $\mathbb{R}^X$, but sometimes it is convenient to use continuous distributions that admit a density. In general, let $\Delta(\mathbb{R}^X)$ be the set of *Borel probability measures* over $\mathbb{R}^X$. (For the purpose of understanding this book, you can just think of this as containing all discrete and continuous distributions.)

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

**Definition 1.7.** $\rho : \mathcal{A} \to \Delta(X)$ is represented by a *distribution over utilities* if there exists $\mu \in \Delta(\mathbb{R}^X)$ such that $\rho(x, A) = \mu(N(x, A))$ for all $A \in \mathcal{A}$ and $x \in A$.

Yet another way to model this is to let utility be a random variable. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, $\mathcal{F}$ is a $\sigma$-algebra and $\mathbb{P}$ is a probability measure. (If you are not familiar with measure-theoretic probability, you can rely on your intuitive understanding of random variables.) Utility is a random function, that is, $\widetilde{U} : \Omega \to \mathbb{R}^X$ is $\mathcal{F}$-measurable. I will try to put a tilde on every random variable (function, element, etc.). We can think of $\Omega$ as things that are observable to the agent but unobservable to the analyst. The event $N$ is now written as

$$\begin{aligned} N(x, A) &:= \{\omega \in \Omega : \widetilde{U}_\omega(x) \geq \widetilde{U}_\omega(y) \text{ for all } y \in A\} \\ &= \left\{\omega \in \Omega : \widetilde{U}_\omega(x) = \max_{y \in A} \widetilde{U}_\omega(y)\right\}. \end{aligned}$$

**Definition 1.8.** $\rho : \mathcal{A} \to \Delta(X)$ has a *random utility* representation if there exists a random variable $\widetilde{U} : \Omega \to \mathbb{R}^X$ such that $\rho(x, A) = \mathbb{P}(N(x, A))$ for all $A \in \mathcal{A}$ and $x \in A$.

I have not made a distinction between the three different definitions of the set $N(x, A)$ and I will not do so in the future. I do make a notational distinction between $\mathbb{P}$, which is the probability measure on the probability space $\Omega$ that caries the random utility $\widetilde{U}$, and $\mu$, which is the probability distribution (a.k.a. the law) of the random variable $\widetilde{U}$.

The following is an easy adaptation of Theorem 3.1 in Block and Marschak (1960) (see also Regenwetter and Marley (2001)).

**Proposition 1.9.** *The following are equivalent for a finite X:*

*(i)* *$\rho$ is represented by a distribution over preferences,*
*(ii)* *$\rho$ is represented by a distribution over utilities, and*
*(iii)* *$\rho$ has a random utility representation.*

Given this result, we will write $\rho \sim RU$ whenever any of the conditions above holds.

**Proof.** $(i) \Rightarrow (ii)$: Suppose that $\rho$ is represented by a distribution over preferences $\mu \in \Delta(\mathcal{P})$. For each preference $\succsim$, pick a utility function $U_{\succsim}$ that represents $\succsim$. Define the distribution over utilities $\hat{\mu} \in \Delta(\mathbb{R}^X)$ by setting $\hat{\mu}(U_{\succsim}) := \mu(\succsim)$ for all $\succsim \in \mathcal{P}$ and $\hat{\mu}(U) := 0$ otherwise. We have

$$\begin{aligned} \rho(x, A) &= \mu(\{\succsim \in \mathcal{P} : x \succsim y \text{ for all } y \in A\}) \\ &= \hat{\mu}(\{U_{\succsim} \in \mathbb{R}^X : \succsim \in \mathcal{P} \text{ and } U_{\succsim}(x) = \max_{y \in A} U_{\succsim}(y)\}) \\ &= \hat{\mu}(\{U \in \mathbb{R}^X : U(x) = \max_{y \in A} U(y)\}). \end{aligned}$$

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

(*ii*) $\Rightarrow$ (*iii*): Suppose that $\rho$ is represented by a distribution over utilities $\mu \in \Delta(\mathbb{R}^X)$. Define $\Omega := \mathbb{R}^X$, $\mathcal{F} := \mathcal{B}$ (the Borel $\sigma$-algebra), $\mathbb{P} := \mu$, and $\widetilde{U}$ be the identity function, that is, $\widetilde{U}_\omega(x) := \omega(x)$ for all $\omega \in \mathbb{R}^X$. Thus,

$$\begin{aligned}\rho(x,A) &= \mu(\{U \in \mathbb{R}^n : U(x) = \max_{y\in A} U(y)\}) \\ &= \mathbb{P}(\{\omega \in \Omega : \widetilde{U}_\omega(x) = \max_{y\in A} \widetilde{U}_\omega(y)\}).\end{aligned}$$

(*iii*) $\Rightarrow$ (*i*): Suppose that $\rho$ is represented by a random utility $(\Omega, \mathcal{F}, \mathbb{P}, \widetilde{U})$. Suppose that with positive probability there is a tie between $x$ and $y$; then

$$\begin{aligned}&\rho(x,\{x,y\}) + \rho(y,\{x,y\}) \\ &\quad = \mathbb{P}(\{\widetilde{U}_\omega(x) \geq \widetilde{U}_\omega(y)\}) + \mathbb{P}(\{\widetilde{U}_\omega(y) \geq \widetilde{U}_\omega(x)\}) \\ &\quad = \mathbb{P}(\{\widetilde{U}_\omega(x) > \widetilde{U}_\omega(y)\}) + 2\mathbb{P}(\{\widetilde{U}_\omega(x) = \widetilde{U}_\omega(y)\}) \\ &\qquad + \mathbb{P}(\{\widetilde{U}_\omega(y) > \widetilde{U}_\omega(x)\}) > 1,\end{aligned}$$

which violates the definition of SCF. So it's without loss of generality to assume that there are no ties. For each strict preference $\succsim \in \mathcal{P}$, define the event $E_{\succsim} := \{\omega \in \Omega : U_\omega \text{ is represented by } \succsim\}$. Notice that $E_{\succsim} \in \mathcal{F}$ because the set of utility functions $\mathcal{U}_{\succsim}$ that represents $\succsim$ is an open set – an intersection of open sets of the form $\{U \in \mathbb{R}^X : U(x) > U(y)\}$ – and $E_{\succsim}$ is the inverse image of $\mathcal{U}_{\succsim}$ under a measurable function $\widetilde{U}$.

For any $\succsim \in \mathcal{P}$, define $\mu(\succsim) := \mathbb{P}(E_{\succsim})$. Since there are no ties, $\mu \in \Delta(\mathcal{P})$. Therefore, we have

$$\begin{aligned}\rho(x,A) &= \mathbb{P}(\{\omega \in \Omega : \widetilde{U}_\omega(x) = \max_{y\in A} \widetilde{U}_\omega(y)\}) \\ &= \mu(\{\succsim \in \mathcal{P} : x \succsim y \text{ for all } y \in A\}).\end{aligned}$$ □

Proposition 1.9 holds for countable $X$ under appropriate definitions (see Cohen (1980)). The equivalence between (ii) and (iii) holds for uncountable $X$ under appropriate technical conditions. For uncountable $X$, condition (i) is typically modified because preferences are usually assumed to be continuous, which implies that they have nontrivial indifference curves. We will talk more about the infinite case later.

## 1.6 TIE BREAKING*

Material with an asterisk may be omitted at first reading. If $\rho$ is represented by a distribution over preferences, then ties are ruled out by construction because only strict preferences are realized with positive probability. On the other hand, distribution over utilities and RU in principle allow for ties. But in fact, for choice probabilities to be well-defined, ties must occur with zero probability. To see that, let $T^{xy} := \{\omega \in \Omega : \widetilde{U}_\omega(x) = \widetilde{U}_\omega(y)\}$ be the event in which there is a tie between $x$ and $y$. As we saw in the proof of Proposition 1.9 if $\rho$ has an RU representation, then it must be that $\mathbb{P}(T^{xy}) = 0$ for all $x \neq y$;

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

otherwise, $\rho(x,y) + \rho(y,x) > 1$ because we are double-counting the event $T^{xy}$. This means that RU with ties does not lead to a legitimate SCF. I will refer to those RU without ties as as *proper* RU. Formally, $\widetilde{U}$ is *proper* if for any menu $A \in \mathcal{A}$, with probability one $\widetilde{U}$ has a unique maximizer on $A$.[10]

For various reasons it is sometimes convenient to allow for ties. Let's take a $\widetilde{U}$ that is not proper. One possible way to define $\rho$ based on $\widetilde{U}$ is to use a tiebreaker. For instance, we could assume that the agent uniformly randomizes over the maximal elements of each menu (*uniform tiebreaking*). This two-stage procedure (maximize $\widetilde{U}$, then break ties uniformly) gives us a well-defined SCF.

A more general notion of tiebreaking was introduced by Gul and Pesendorfer (2006) in the supplement to their paper. A *GP-tiebreaker* is a random utility function $\widetilde{W} : \Omega_W \to \mathbb{R}^X$ that itself is proper. In a random utility representation with a GP tiebreaker, the agent first maximizes $\widetilde{U}$ and then uses $\widetilde{W}$ to break the ties. The state space is now $\Omega \times \Omega_W$ because the tie breaker needs its own state space, as the original one may not be rich enough to allow for a proper $W$.

**Proposition 1.10.** *The following are equivalent when X is finite:*

*(i)* $\rho$ *has a proper RU representation*
*(ii)* $\rho$ *has an improper RU representation with uniform tiebreaking*
*(iii)* $\rho$ *has an improper RU representation with a GP-tiebreaker.*

**Proof.** $(i) \Rightarrow (ii)$: If $\rho$ has an RU representation, then ties occur with probability zero, so it doesn't matter how we break them.

$(ii) \Rightarrow (iii)$: Uniform tie breaking is equivalent to GP tiebreaking where $\widetilde{W}$ represents a uniform distribution over all strict orders over $X$.

$(iii) \Rightarrow (i)$: First, rescale $\tilde{U}$ so that the utility gaps between any two distinct items are larger than one. That is,

$$\widetilde{U}_\omega(x) \neq \widetilde{U}_\omega(y) \Rightarrow |\widetilde{U}_\omega(x) - \widetilde{U}_\omega(y)| \geq 1.$$

Then break any ties according to a rescaled version of $\widetilde{W}$ so that we don't exceed these gaps, that is, for any $\omega$, the maximum difference between two values of $\tilde{U}_\omega$ is strictly less than 1. Finally, note that

$$\rho(x,A) = \mathbb{P}\left(\left\{\omega \in \Omega : \tilde{U}_\omega(x) + \tilde{W}_\omega(x) \geq \tilde{U}_\omega(y) + \tilde{W}_\omega(y)\ \forall y \in A\right\}\right). \quad \square$$

This result makes it sound like it is impossible to know whether randomness in choice reflects the true preference variation or just tiebreaking. In Chapter 8 we will see that it is possible to draw a meaningful distinction between the two in a dynamic model because the two sources of randomness enter differently into the agent's option value calculation (taste variation provides flexibility whereas tiebreaking does not).

[10] This property is sometimes called *noncoincidence* (Falmagne, 1983) or *regularity* (Gul and Pesendorfer, 2006).

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

Instead of using tiebreakers, other papers allow for indifferences by changing the primitive and studying stochastic choice correspondences or capacities: Barberá and Pattanaik (1986); Gul and Pesendorfer (2013); Lu (2016); Lin (2018); Piermont and Teper (2018). To a large extent this approach is "morally equivalent" to assuming tie breakers and I view the choice between them as a matter of convenience.

The issue of ties gets even more subtle when $X$ is "large." Notice that there is another way to define ties: Let $T := \{\omega \in \Omega : \widetilde{U}_\omega(x) = \widetilde{U}_\omega(y) \text{ for some } x \neq y\}$. This is the event that there is a tie between *some* elements of $x$. Note that $T = \bigcup_{x \neq y} T^{xy}$ so if $X$ is finite then $\mathbb{P}(T) = 0$ iff $\mathbb{P}(T^{xy}) = 0$ for all $x \neq y$. But with uncountable $X$, this new definition is too strong. For example, when $X$ is multidimensional and all utilities considered are continuous, then we are forced to have $\mathbb{P}(T) > 0$ because all continuous preferences have well-behaved indifference curves, so for any fixed utility function there will be many points that are indifferent to each other. However, for any two specific points, the probability that they will be indifferent could well still be zero.

## 1.7 ADDITIVE RANDOM UTILITY

There is an equivalent way of writing random utility, called *additive random utility* (ARU). This involves writing $\tilde{U}(x) = v(x) + \tilde{\epsilon}(x)$, where $v : X \to \mathbb{R}$ is a deterministic utility function, called the "representative utility" or "systematic utility" and $\tilde{\epsilon} : \Omega \to \mathbb{R}^X$ is a "random utility shock," which is private information of the agent.

ARU is the workhorse model in discrete choice econometrics, where the focus is on estimating the function $v$ based on observations of $\rho$. In game theory, ARU is used as a model of *smoothed best responses* (Fudenberg and Levine, 1998; Hofbauer and Sandholm, 2002).

If $X$ is finite, then I will say that the distribution of $\tilde{\epsilon}$ is *smooth* if it has a density. For infinite $X$, it is smooth if for any menu $A = \{x_1, \ldots, x_n\}$ the joint distribution of $(\tilde{\epsilon}(x_1), \ldots, \tilde{\epsilon}(x_n))$ has a density. The following definition is based on McFadden (1973).

**Definition 1.11.** $\rho : \mathcal{A} \to \Delta(X)$ has an *additive random utility* (ARU) representation if it has an RU representation with $\widetilde{U}(x) = v(x) + \tilde{\epsilon}(x)$, where $v : X \to \mathbb{R}$ is deterministic and the distribution of $\tilde{\epsilon}$ is smooth.

Note well that the distribution of $\tilde{\epsilon}$ is independent of the menu: For each $A$ we just select the corresponding coordinates.

The smoothness assumption guarantees that we have a proper RU representation, as it implies that ties occur with probability zero. It is worthwhile to notice though that there *are* proper RU representations which are of the form $\widetilde{U}(x) = v(x) + \tilde{\epsilon}(x)$ where $\epsilon$ has a discrete distribution (take, for example, the one constructed in the proof of $(i) \Rightarrow (ii)$ in Proposition 1.9). McFadden's (1973) general definition does not require the existence of a density, but as the following result shows, this assumption is without loss of generality. That

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

is, even if we have a discrete distribution over utilities, we can "smoothify" it without affecting the choice probabilities.

**Proposition 1.12.** *If $X$ is finite then $\rho \sim RU$ if and only if $\rho \sim ARU$.*

The construction used in the following proof shows that it is also without loss of generality to assume that $\tilde{\epsilon}$ has finite moments.

**Proof.** Let $\rho \sim ARU$, then by definition $\rho \sim RU$. Conversely, assume now that $\rho \sim RU$. By Proposition 1.9, there exists a probability distribution $\mu$ over strict preferences $\mathcal{P}$ such that $\rho(x, A) = \mu(N(x, A))$. Let $n$ be the cardinality of $X$ and for any $\succsim \in \mathcal{P}$ and $i = 1, \ldots, n$ let $x_{\succsim}(i)$ denote the $i$th ranked element of $X$.

Define $v(x) = 0$ for all $x \in X$. We need to find a probability measure $\mathbb{P}$ over $\mathbb{R}^X$ such that $\mathbb{P}(A_{\succsim}) = \mu(\succsim)$ for each event of the type

$$A_{\succsim} = \{\epsilon \in \mathbb{R}^X : \epsilon(x_{\succsim}(1)) > \epsilon(x_{\succsim}(2)) > \cdots > \epsilon(x_{\succsim}(n))\}.$$

To do so, for each $\succsim$ take a probability measure with finite moments and density $\gamma_{\succsim}$ and support equal to the closure of $A_{\succsim}$, for example, a truncated Normal probability distribution. Define our probability measure $\mathbb{P}$ by its density

$$\gamma(\cdot) = \sum_{\succsim \in P} \mu(\succsim)\gamma_{\succsim}(\cdot).$$

This measure has finite moments and a density. □

ARU representations derive their strength from several powerful parametric special cases where the distribution of $\epsilon$ is i.i.d. The most predominant is the extreme value distribution, which leads to *logit*.

**Definition 1.13.** $\rho : \mathcal{A} \to \Delta(X)$ has a *logit representation* if it has a ARU representation where $\tilde{\epsilon}(x)$ are i.i.d. across $x$ with the Type I Extreme Value (TIEV) distribution, with CDF $G(\epsilon) = \exp(-\exp(-\epsilon))$.[11]

Another well-known model is *probit*, where the distribution of $\tilde{\epsilon}$ is Normal. We will look more at i.i.d. parameterizations in Chapter 3.

Often times it is assumed that the density in an ARU representation not only exists, but is everywhere positive. This ensures that all items are chosen with a positive probability (because arbitrarily large shocks can elevate even dominated alternatives).

**Axiom 1.14 (Positivity).** $\rho(x, A) > 0$ for any $x \in A$.

This property is important since keeping all probabilities positive leads to a nondegenerate likelihood function, which facilitates estimation of $v$.

[11] TIEV, which is also known as the Gumbel distribution, is actually a whole class of distributions with mean and variance parameters. However, in economics TIEV typically means this particular member of the family.

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

Moreover, as argued by McFadden (1973), positivity cannot be refuted based on any finite data set.[12]

There are two interpretations of $\epsilon$: (1) preference heterogeneity that is unobserved by the analyst (after conditioning on observable characteristics of the agent), or (2) mistakes/errors on the part of the agent. The difference between the interpretations is that in the first case the preference shocks are embraced by the agent (her tastes do actually change from time to time), while in the second case these shocks lead to choices that the agent disagrees with. While in predictive applications of the static model the two interpretations are largely equivalent, they differ when it comes to normative evaluations and have different predictions for dynamic behavior.

A case that is somewhat in between the two is one of imperfect perception. In the following example the agent sometimes makes mistakes, but they are doing the best they can given their imperfect information. As we will see in Chapter 5 this behavior is Bayes-optimal, so the shocks are embraced by the agent (ex ante) despite sometimes leading to errors.

**Example 1.15 (Law of Comparative Judgment).** Recall Example 1.5 with weight perception. Thurstone (1927) introduced the probit model to capture such behavior. Suppose that for each weight $x$ the agent forms a subjective, imperfect, and random perception $\gamma(x) + \tilde{\epsilon}(x)$, where $\gamma$ is a strictly increasing function (typically assumed to be logarithmic) and $\tilde{\epsilon}(x) \sim \mathcal{N}(0, \sigma_\epsilon^2)$ are i.i.d. across $x$. Faced with items $x$ and $y$, the agent chooses item $x$ if $\gamma(x) + \epsilon_x \geq \gamma(y) + \epsilon_y$ and chooses $y$ otherwise. A simple calculation reveals that

$$\rho(x, y) = \Phi\left(\frac{\gamma(x) - \gamma(y)}{\sigma_\epsilon \sqrt{2}}\right).$$

Thus, Thurstone's model leads to S-shaped psychometric functions. It is easy to see that by setting $\gamma(x) = \log x$ the model explains diminishing sensitivity. However, it does not explain frequency-dependence because the distribution of $\epsilon$ is independent of the distribution of menus $\{x, y\}$.

Finally, the model cannot explain payoff-monotonicity either. This is because $\gamma(x)$ is not the payoff from choosing $x$, but instead a subjective perception of $x$. The magnitude of the payoff for guessing correctly does not enter Thurstone's formula. One can view his model as "probit in perceptions" as opposed to "probit in payoffs."[13] △

[12] Positivity does not imply that $\epsilon$ has positive density (see Example A.1 in the Appendix). Li (2021) shows how to strengthen Positivity to ensure that there exists a ARU representation with positive density.

[13] One could imagine a "probit in payoffs," where it's the payoffs that get distorted. Let $v > 0$ be the payoff of guessing correctly. Then for $x > y$ we have $\tilde{U}(x) = v + \epsilon_x$ and $\tilde{U}(x) = 0 + \epsilon_y$. This leads to payoff-monotonicity, but induces a psychometric function that is a step function (as opposed to S-shaped), so we cannot capture the first two stylized facts. We will need a more fancy model to capture all the stylized facts simultaneously.

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press

While in the most general case RU and ARU coincide, they can lead to very different predictions if the utility is restricted to some specific family. Suppose that our analyst has a theory that the utility function belongs to some class $\mathcal{U}$. RU and ARU suggest different approaches to building a stochastic model. We can either randomize over utilities $u \in \mathcal{U}$ or fix a deterministic utility $v \in \mathcal{U}$ and add stochastic $\tilde{\epsilon}$, where which belongs to some class of distributions $\mathcal{E}$. As we saw before, with $\mathcal{U}$ and $\mathcal{E}$ unrestricted, these two approaches lead to the same class of $\rho$. But when $\mathcal{U}$ and $\mathcal{E}$ have more structure, often the two induced classes of $\rho$ are disjoint because $v+\epsilon$ does not belong to $\mathcal{U}$. We will see several instances of this: In Chapter 4, we will show that ARU with i.i.d. $\epsilon$ leads to "unreasonable" comparative statics in the risk aversion parameter. In Chapter 8, we will show it leads to "unreasonable" option value. In Chapter 10, we will see that it leads to "unreasonable" patterns of substitution.

## 1.8 SOCIAL SURPLUS

Our analyst often wants to evaluate the agent's welfare. Under RU the natural way to do this is to set

$$V(A) := \mathbb{E}\left[\max_{x \in A} \widetilde{U}(x)\right].$$

This function captures the expected utility from the best item in the menu. McFadden (1973) called it the *social surplus*.[14]

This function is key for dynamic optimization, where the agent evaluates how their current actions impact their own future welfare (Chapters 8 and 12). It also enters into nested logit, where decisions are similarly made in stages (Section 3.4).

Under ARU, we have

$$V(A) := \mathbb{E}\left[\max_{x \in A} v(x) + \tilde{\epsilon}(x)\right].$$

This formula makes sense only if we interpret $\epsilon$ as unobservable preference shocks. If we think of them as decision errors of the agent, then there is no reason for them to enter welfare. In this case, it may be more appropriate to treat them as just driving behavior, but evaluate welfare using the undistorted preferences. For example, a formula a la Strotz (1955) would look like:

$$V(A) = \sum_{x \in A} v(x)\rho(x, A).$$

A theory of stochastic choice along these lines has been developed by Ke (2018).

[14] This should be called "consumer surplus" because the social surplus also includes the firms.

https://doi.org/10.1017/9781009512749.003 Published online by Cambridge University Press