# Basic Probability Theory

This material comes from Hansen's *Probability and Statistics for Economists* (PSE) and Len Goff's (former instructor of ECON 8070 at UGA) lecture notes along with some of my own comments.

## Introduction

PSE 1.1

Probability is the mathematical language for dealing with **uncertainty** or **random** events (will define these more precisely soon, but it is fine to think about them informally for now). Some things are obviously uncertain and connected with probability; e.g., flipping a coin, rolling dice, card games or other games of chance. Other, more complicated applications that seem obviously connected to probability are things like forecasting; for example, it seems natural to think of tomorrow's weather or the stock price of some company one year in the future as being uncertain/random. We will use these sorts of examples some during the first few weeks of the course, but these sorts of applications are not what most researchers in economics and other business disciplines (at least those working with data) are most interested in studying.

Instead, in econometrics (and statistics more generally), the most common type of problem that we'll be interested in is that there is some feature of a large population that we'd like to know about, and we want to use data (that is, random draws from the population) to learn something about the feature of the population that we are interested in. A good example of this would be if you were a labor economist interested in studying the unemployment rate in the United States. In principle, there is a "true" unemployment rate in the U.S. But it would be hard to figure out exactly what it is; it would be costly to track down every person in the U.S. and figure out their employment status. And, even if you could do it, it would be challenging to do it quickly enough so that some people did not change their employment status since the first time you asked them. Instead, in order to track the unemployment rate, the Census Bureau takes a survey; that is, they (attempt to) randomly select a subset of individuals in the U.S. (I think about 60,000 per month but I'm not 100% sure) their employment status. From this sample, they report an estimate of the unemployment rate along with a measure of **statistical uncertainty** (roughly: how different would it be reasonable to think that their estimate could be from "true" unemployment rate).

Going back to uncertainty, for a particular person, it may not feel like their employment status is random (in particular, a person knows whether or not they are employed). But if you take the view of the researcher, employment status is random. Supposing that we have some way to randomly "draw" a person from the population of people in the U.S., whether this person is employed or not is random to us (sort of like the examples of weather and stock prices above).

We might also be interested in how unemployment is correlated with people's characteristics. Does it vary by age or race or region of the country? These sorts of questions are also related to some ideas in probability.

We will come back to issues related to data in a big way over the course of the next few weeks.

For now, we'll concern ourselves learning some useful tools concerning probability.

## Outcomes and Events

PSE 1.2

To start with, let's define some terms:

- An **outcome** is the result of a random process. For example, the outcome of flipping a coin is whether we actually flipped a heads or tails.
- The **sample space** is the set of all possible outcomes; we will denote the sample space by $S$. For flipping a coin, the sample space is $S = \{H, T\}$ (for heads and tails). For rolling a die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$.

  As examples of economic variables, the sample space for a person's yearly income: $S = [0, \infty)$; the sample space for a firm's industry: $S = \{\text{manufacturing}, \text{service}, \text{other}\}$.
- An **event** $A$ is a subset of outcomes in $S$. For example, the event of rolling an even number on a single dice roll is $A = \{2, 4, 6\}$. The event of rolling a 1 is $A = \{1\}$. Another possible event is $A = S$ (i.e., the whole sample space). Soon we'll define probabilities of events.

Before we get to defining probability itself, following the textbook, let's define a few more things:

### Review of Operations on Sets (Definition 1.1)

1. $A$ is a **subset** of $B$, written $A \subset B$, if every element of $A$ is an element of $B$. In math: $x \in A \implies x \in B$.
   - Two sets are said to be **equal**, that is $A = B$, if $A \subset B$ and $B \subset A$.
2. The event with no outcomes $\emptyset = \{\}$ is called the **empty set**.
3. The **union** $A \cup B$ is the collection of all outcomes that are in either $A$ **or** B (or both). In math: $A \cup B = \{x : x \in A \text{ or } x \in B\}$. Note: you should read the ":" as "such that".
4. The **intersection** $A \cap B$ is the collection of elements that are in both $A$ **and** $B$. In math: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.
5. The **complement** $A^c$ of $A$ are all outcomes in $S$ which are not in $A$. In math: $A^c = \{x : x \notin A\}$.
6. The events $A$ and $B$ are **disjoint** if they have no outcomes in common; that is, if $A \cap B = \emptyset$
7. The events $A_1, A_2, ...$ are a **partition** of $S$ if they are mutually disjoint and their union is $S$; that is, if $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{\infty} A_i = S$.

**Example: Rolling a die**

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Consider the events $A = \{1\}$, $B = \{2\}$, and $C = \{2, 4, 6\}$. Then

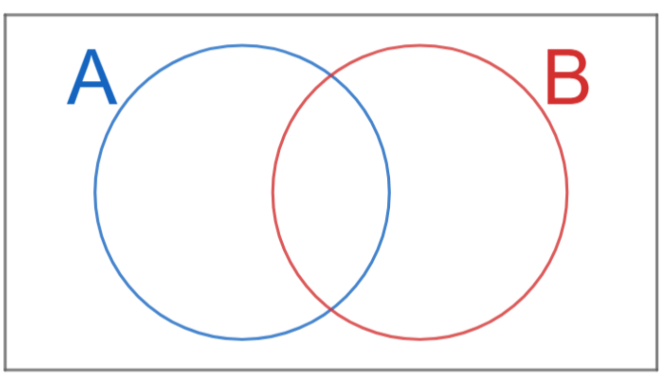
$$A \cup C = \{1, 2, 4, 6\}, \qquad B \cap C = \{2\},$$
$$A \cap C = \emptyset, \qquad C^c = \{1, 3, 5\}.$$

**Properties of Set Operations** For any three events $A$, $B$, $C$ defined on sample space $S$, the following properties hold:

- **Commutative:** $A \cup B = B \cup A$ and $A \cap B = B \cap A$. The order of the two sets does not matter.
- **Associative:** $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$. With one repeated operation the grouping does not matter, so we can write $A \cup B \cup C$ and $A \cap B \cap C$ without parentheses.
- **Distributive:** $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$. An intersection pushes through a union term by term (and vice versa), like $a(b + c) = ab + ac$ in algebra — except that for sets it works in both directions.
- **DeMorgan's Law:** $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$. Taking a complement swaps "or" and "and": "not ($A$ or $B$)" means "not $A$ and not $B$". (Think of eligibility rules: you qualify if you meet requirement $A$ and requirement $B$, so you fail to qualify if you miss $A$ or miss $B$.)

Two habits from ordinary algebra do *not* carry over to sets: there is no operation that undoes a union (no "subtracting" a set), and you cannot cancel — $A \cup B = A \cup C$ does not imply $B = C$ (take $A = S$).

It's relatively straightforward to write proofs of all of these. We'll write a proof of DeMorgan's Law momentarily. Before we do that, let me mention **Venn Diagrams**. Typically, they look something like this



In my view, writing down a Venn diagram is not a proof, but it can be useful to clarify your thinking when working with sets. For example, it is fairly easy to map DeMorgan's law into the picture above (you can try this as practice).

Now for a proof of DeMorgan's law. To start with, let's define two new sets, $C = (A \cup B)^c$ and $D = A^c \cap B^c$; we want to show that $C = D$, and we will show this by showing (i) $C \subset D$ and (ii) $D \subset C$. To show the first part, notice that $x \in C \implies x \notin (A \cup B) \implies x \notin A$ and $x \notin B \implies x \in A^c$ and $x \in B^c \implies x \in (A^c \cap B^c) \implies x \in D$. Thus, $C \subset D$.

For the second part, notice that $x \in D \implies x \in A^c$ and $x \in B^c \implies x \notin A$ and $x \notin B \implies x \notin (A \cup B) \implies x \in (A \cup B)^c \implies x \in C$. Thus, $D \subset C$, and we have completed the proof.

## Probability Function

PSE 1.3-1.4

A **probability function** P is a function that takes in events and returns probabilities (numbers from 0 to 1). Strictly speaking, P is not defined on *every* subset of $S$ but on a collection of subsets called a **$\sigma$-algebra** (or **$\sigma$-field**), and it is those subsets that we call *events*; the optional appendix at the end of these notes explains what this means and why it is needed (see also PSE 1.14). For everything we do this semester you can safely read "event" as "subset of $S$" — the distinction only rules out some pathological cases that arise when there is an uncountably infinite number of possible outcomes.

Here is a more formal definition. A **probability function** $\mathrm{P}(\cdot)$ is a function from events (formally, from a $\sigma$-algebra $\mathscr{F}$ of subsets of $S$) to $\mathbb{R}$, satisfying the following properties (which are referred to as the **Axioms of Probability**):

1. $\mathrm{P}(A) \geq 0$. In words: the probability of any event (and therefore probabilities in general) is non-negative

2. $\mathrm{P}(S) = 1$. In words: the probability of some outcome in the sample space (recall: this is the set of all possible outcomes) is equal to 1.

3. If $A_1, A_2, ...$ is a countable collection of disjoint sets, then

$$\mathrm{P}\left(\bigcup_j A_j\right) = \sum_j \mathrm{P}(A_j)$$

In words: Probabilities are additive on disjoint events.

**Side-Comment**

The textbook also has an interesting discussion of the meaning of probabilities. To briefly summarize, there are two main ways to ascribe meaning to probabilities. Perhaps the most common one, called *frequentism*, is based on the long run frequencies (e.g., the probability of flipping a heads is 50% because if you did it a tremendously large number of times, the "long run probability" of flipping a heads is 50%). The other main view is called *subjectivism* which basically refers to an individual's "degree of belief" about the uncertainty of an event.

The first two parts of the definition of the probability function seem straightforward. For the third part of the definition, consider rolling a die. Suppose we are interested in the event $A = \{2, 4, 6\}$ (i.e., that we roll an even number). Notice that we can re-write $A = \{2\} \cup \{4\} \cup \{6\}$, and that these are all disjoint events (this is because if you roll a 2, it means that you could not have rolled a 4). The third part of the definition implies that we can write $\mathrm{P}(A) = \mathrm{P}(\{2\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{6\}) = 1/6 + 1/6 + 1/6 = 1/2$.

These axioms imply several intuitive properties of probability. As an example, let's start by showing: $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$. In words, this says that the probability of all outcomes that are not in $A$ is equal to one minus the probability of $A$.

To show this, notice that $A$ and $A^c$ are disjoint sets. Moreover, $A \cup A^c = S$. Thus, $1 = \mathrm{P}(S) = \mathrm{P}(A) + \mathrm{P}(A^c)$ where the first equality holds by the second property of a probability function and the second equality holds by the third property of a probability function. This implies the result.

**Practice**

Using the definition of a probability function, show the following

1. Show that if $A \subseteq B$: $\mathrm{P}(A) \leq \mathrm{P}(B)$.
2. Use the above to show that $\mathrm{P}(A \cap B) \leq \min\{\mathrm{P}(A), \mathrm{P}(B)\}$.
3. Derive the expression: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.
4. Derive the expression: $\mathrm{P}(A \cap B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cup B)$. *Hint:* use $(A \cap B)^c = A^c \cup B^c$.

See also Theorem 1.2 in the textbook (as well as Section 1.15) for more examples.

## Counting

PSE 1.5

A number of classical probability problems are versions of counting problems. These types of problems showed up often in the context of gambling, which often served as motivation for early work on probability. Although these are a classic probability topic, I will not cover them in too much detail as they are not especially relevant for most econometrics topics.

When the sample space $S$ is finite and every outcome is equally likely, the probability of an event $A$ is

$$\mathrm{P}(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{|A|}{|S|}$$

so computing a probability just involves *counting*, i.e., determining how many outcomes are in $A$, and how many are in $S$. When the lists are too long to write out by hand, the techniques below do the counting for us, but they are all the same idea.

**The multiplication rule.** If each outcome is built up as a sequence of $k$ parts—$n_1$ possibilities for the first part, $n_2$ for the second, and so on—then the number of outcomes is $n_1 \times n_2 \times \cdots \times n_k$.

*Example.* Rolling two dice: there are 6 options for the first die and 6 for the second, so $|S| = 6 \times 6 = 36$. The event "the two dice sum to 4" is $\{(1,3),(2,2),(3,1)\}$, so its probability is $3/36 = 1/12$.

**Permutations** (arrangements in which order matters) are just the multiplication rule with the number of options running down by one at each step. In particular, the number of ways to line up $k$ of $n$ distinct objects is

$$n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$$

Taking $k = n$ gives $n!$, the number of orderings of all $n$ objects.

*Example.* With 8 runners in a race, the number of possible gold–silver–bronze finishes is $8 \times 7 \times 6 = 336$.

**Combinations** (choices in which order does *not* matter) are permutations divided by the amount of overcounting. Choosing $k$ of $n$ objects and then lining them up can be done in $n!/(n-k)!$ ways; but each unordered choice of $k$ objects is counted $k!$ times there, once for each way of ordering it. So the number of unordered choices is

$$\binom{n}{k} = \frac{1}{k!} \cdot \frac{n!}{(n-k)!} = \frac{n!}{k!\,(n-k)!}$$

*Example.* The number of 3-person committees that can be formed from 8 people is $\binom{8}{3} = 336/6 = 56$ — the same 336 as the medal example, now divided by the $3! = 6$ orderings of each committee.

(Drawing $k$ items *with* replacement, with order mattering, is the multiplication rule once more: $n$ options every time, so $n^k$.)

**Example.** A committee of 3 is drawn at random from a group of 10 people, of whom 6 are economists and 4 are statisticians. There are $|S| = \binom{10}{3} = 120$ equally likely committees. The number made up of 3 economists is $\binom{6}{3} = 20$, so

$$\mathrm{P}(\text{all 3 economists}) = \frac{\binom{6}{3}}{\binom{10}{3}} = \frac{20}{120} = \frac{1}{6}$$

The number with exactly 2 economists is $\binom{6}{2}\binom{4}{1} = 15 \cdot 4 = 60$ (choose which 2 of the 6 economists, then which 1 of the 4 statisticians — the multiplication rule again), so $\mathrm{P}(\text{exactly 2 economists}) = 60/120 = 1/2$.

## Joint events, conditional probability, and independence

PSE 1.6-1.8

At the beginning of these notes, we briefly mentioned that one might be interested in how events are related to each other. One of the examples in the textbook is where $A$ is the event of "making an A on an econometrics exam" and $B$ is the event of "studying 12 hours a day", and where a natural question is how event $B$ affects the probability of event $A$. Another example is the one from earlier in the notes where one event might be "being unemployed" and another event might be "being younger than 35". To think along these lines, we'll define conditional probabilities; that is, the probability that event $A$ occurs conditional on event $B$ occurring.

**Definition.** Given an event $B$ such that $\mathrm{P}(B) > 0$, the **conditional probability** of event $A$ given $B$ is defined as

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

A natural question at this point is why we define a conditional probability this way. Recall that the intersection of two events $A \cap B$ can be interpreted as the event that *both* of events $A$ and $B$ occur — this corresponds to the term in the numerator. And you might be tempted to think that this is the conditional probability. The textbook refers to dividing by $\mathrm{P}(B)$ as a normalization to the "new" sample space $B$; as an additional comment, it seems helpful to think about a relatively extreme example. Consider the example of studying for an econometrics exam above. Suppose that all students that study for 12 hours a day make an A (this immediately suggests that the conditional probability of making an A conditional on 12 hours per day studying should be equal to 1). However, suppose that only 5% of students study for 12 hours a day. In this case, $\mathrm{P}(A \cap B) = \mathrm{P}(B) = 0.05$ (which is not equal to 1). However, $\mathrm{P}(A \cap B)/\mathrm{P}(B) = \mathrm{P}(B)/\mathrm{P}(B) = 1$ which is as it should be.

Another very important concept in probability is when events are **independent**. Intuitively, events are independent if knowing the outcome of one event does not provide any information about the probability that another event occurred. In math, we say that two events $A$ and $B$ are independent if

$$\mathrm{P}(A|B) = \mathrm{P}(A)$$

or, in other words, the conditional probability of $A$ conditional on $B$ is equal to the unconditional probability of $A$.

An implication of two events being independent (often, this is actually given as the definition of events being independent though I think the previous one is a bit more straightforward to understand) is that

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$$

To see this, notice that

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(A|B)\mathrm{P}(B) \\ &= \mathrm{P}(A)\mathrm{P}(B) \end{aligned}$$

where the first line holds by rearranging the definition of conditional probability (as a side-comment, this is a useful/common step so it's one to pay attention to) and the second equality holds because $A$ and $B$ are independent.

It's natural to think about things like outcomes of two flips of a coin or rolls of dice as being independent though most of the examples that you'd find interesting in economics/business/etc. are unlikely to be independent. Section 1.8 in the textbook walks through some examples that you might find helpful.

The discussion above defines independence for *two* events. With more than two events, we can define different notions of independence. For a collection of events $A_1, A_2, \dots, A_n$, we say they are **(mutually) independent** if for *every* sub-collection $A_{i_1}, A_{i_2}, \dots, A_{i_k}$ (with $k \geq 2$),

$$\mathrm{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}\right) = \mathrm{P}(A_{i_1})\,\mathrm{P}(A_{i_2}) \cdots \mathrm{P}(A_{i_k})$$

A weaker condition is **pairwise independence**: $\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\mathrm{P}(A_j)$ for every $i \neq j$. Although there are some results that only require pairwise independence, most often we will need the stronger notion of mutual independence. Moreover, mutual independence is probably the more natural way to think about independence, and you can imagine confusing yourself by using the weaker notion of pairwise independence (see the example below).

**Example: Mutual vs. Pairwise Independence**

Flip a fair coin twice and define

$$\begin{aligned} A &= \{\text{first flip is heads}\} \\ B &= \{\text{second flip is heads}\} \\ C &= \{\text{the two flips are the same}\} \end{aligned}$$

Each of these events has probability 1/2. Every *pair* is independent; for example,

$$\mathrm{P}(A \cap C) = \mathrm{P}(\{HH\}) = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(C)$$

and the same check works for $A, B$ and for $B, C$. But the three events are *not* mutually independent. What is happening here is that $C$ is independent of $A$ and $C$ is independent of $B$, but $C$ is not simultaneously independent of both $A$ and $B$ because $A$ and $B$ together force $C$ to occur:

$$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(\{HH\}) = \frac{1}{4} \neq \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C) = \frac{1}{8}$$

## Law of total probability and Bayes' rule

PSE 1.9-1.10

Another useful result for working with probabilities and multiple events is called the **law of total probability**: If $B_1, B_2, \dots$ is a partition of the sample space and $\mathrm{P}(B_i) > 0$ for all $i$, then for any event $A$: $\mathrm{P}(A) = \sum_i \mathrm{P}(A|B_i)\mathrm{P}(B_i)$.

*Proof:*

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(A \cap S) \\
&= \mathrm{P}\left(A \cap (\cup_i B_i)\right) \\
&= \mathrm{P}\left(\cup_i (A \cap B_i)\right) \\
&= \sum_i \mathrm{P}(A \cap B_i) \\
&= \sum_i \mathrm{P}(A|B_i)\mathrm{P}(B_i)
\end{aligned}$$

where the first line holds since any event $A \subseteq S$, so that $A = A \cap S$ and thus $\mathrm{P}(A) = \mathrm{P}(A \cap S)$. The second line holds because $B_1, B_2, \ldots$ forms a partition of $S$. The third equality holds because $A \cap (\bigcup_i B_i) = \bigcup_i (A \cap B_i)$. The fourth line holds because the events $(A \cap B_i)$ are disjoint for different values of $i$ (since each is a subset of $B_i$) and by applying the third part of the definition of probability. And the last equality holds by the definition of conditional probability.

Finally, we will cover a famous result in probability called **Bayes' rule**. Bayes' rule says that, if $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$, then

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B|A)\mathrm{P}(A) + \mathrm{P}(B|A^c)\mathrm{P}(A^c)}$$

*Proof:*

The proof essentially follows from repeatedly applying the definition of conditional probability and from the law of total probability.

$$\begin{aligned}
\mathrm{P}(A|B) &= \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \\
&= \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)} \\
&= \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B|A)\mathrm{P}(A) + \mathrm{P}(B|A^c)\mathrm{P}(A^c)}
\end{aligned}$$

where the first equality uses the definition of conditional probability, the second equality uses the (rearranged version of) the definition of conditional probability, and the third equality uses the law of total probability (notice that $A$ and $A^c$ form a partition of the sample space).

## Appendix: $\sigma$-algebras and Measurability (optional)

This appendix provides some additional discussion on some technical topics that we did not fully cover above related to $\sigma$-algebras and measurability. The discussion provided here is still high-level with the goal being to introduce you to some terminology that you may see when you read econometrics papers.

When we defined the probability function P, we described it as a function "from events to $[0,1]$" and left "event" loosely specified as "a subset of $S$." If $S$ is finite or countable, we really can assign a probability to every subset and there is nothing more to say. But if $S$ is uncountable (e.g., $S = [0,1]$ or $S = \mathbb{R}$), it is impossible to assign a sensible probability to every subset while keeping countable additivity (axiom 3). The fix is to require P to be defined only on a well-behaved collection of subsets.

A collection $\mathcal{F}$ of subsets of $S$ is a **$\sigma$-algebra** (or **$\sigma$-field**) if

1. $S \in \mathcal{F}$;
2. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (closed under complements);
3. if $A_1, A_2, ... \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions).

These three conditions also imply $\emptyset \in \mathcal{F}$ and closure under countable intersections. The sets in $\mathcal{F}$ are exactly what we call **events** — the subsets of $S$ to which we are entitled to assign a probability. The axioms of probability from earlier are really statements about a function $\mathrm{P} : \mathcal{F} \to [0,1]$, and the disjoint sets in axiom 3 are understood to be members of $\mathcal{F}$.

Putting the pieces together, a **probability space** is a triple $(S, \mathcal{F}, \mathrm{P})$: a sample space, a $\sigma$-algebra of events, and a probability function defined on that $\sigma$-algebra. This triple is the formal starting point for all of probability theory.

On $S = \mathbb{R}$, the standard choice of $\mathcal{F}$ is the **Borel $\sigma$-algebra**: the smallest $\sigma$-algebra that contains every interval. It contains essentially every set you would ever write down. Subsets of $\mathbb{R}$ that are *not* Borel (equivalently, not **measurable**) do exist, but constructing one requires the axiom of choice and they never arise in applied work — so for our purposes "event" and "measurable set" can both be read as "any subset you can actually describe."

We say an event $A$ occurs **almost surely** (a.s.), or **with probability one**, if $\mathrm{P}(A) = 1$ — equivalently, if the event $A^c$ on which it fails has probability zero. When $S$ is uncountable, "probability zero" is not the same as "impossible": if $X$ is drawn uniformly on $[0,1]$, then $\mathrm{P}(X = 0.5) = 0$ even though $X = 0.5$ is a perfectly possible outcome. This terminology is used heavily in the statistics half of the course — for example, the strong law of large numbers says a sample average converges to the population mean almost surely.