Lecture 9: Probability

Juan Dodyk

WashU

Plan

Probability

1. Probability spaces
2. Epistemic interpretation
3. Basic combinatorics
4. Conditional probability
5. Bayes rule
6. Independence

Random variables

1. Binomial
2. Continuous

Why Probability?

We need probability to quantify uncertainty. We have *a lot* of uncertainty in empirical research, so this is really necessary.

Also, if we see a pattern in the data, we want to say that it's unlikely to have occurred by chance. In order to be able to say this, we need to understand what's likely to occur by chance.

In addition, we may intentionally randomize a variable in an experiment. And we may treat events as "natural experiments," or "as if" random. So we need to understand randomness.

Also, it's useful to model people's beliefs in terms of probability, even if there is nothing random involved.

Frequentist Interpretation of Probability

A process generates a random *state* ω from a finite set Ω.

Each state ω occurs with probability $p(\omega) \in [0, 1]$. Meaning: if we run the process "infinitely many times," it will generate the state ω a fraction $p(\omega)$ of the time.

The probabilities should sum to 1: $\displaystyle\sum_{\omega \in \Omega} p(\omega) = 1$.

Example. A fair coin generates a state $\omega \in \{\text{Heads}, \text{Tails}\}$ with uniform probability:

$$p(\text{Heads}) = p(\text{Tails}) = \frac{1}{2}.$$

Example. A die generates a state $\omega \in \{1, 2, 3, 4, 5, 6\}$ with uniform probability:

$$p(1) = \cdots = p(6) = \frac{1}{6}.$$

Events

An **event** E is a set of states. For example, in the last example $E = \{1, 3, 5\}$ is an event. We assign probabilities to events:

$$\Pr(\boldsymbol{E}) \equiv \sum_{\omega \in E} p(\omega).$$

We denote the set of events by $\mathbf{2^{\Omega}}$. This is the set of subsets of Ω.

We obtain a function $\Pr : 2^{\Omega} \to [0, 1]$ such that:

- $\Pr(\Omega) = 1$,
- if $E \cap F = \varnothing$ then $\Pr(E \cup F) = \Pr(E) + \Pr(F)$.

We define the *complement* of an event $\boldsymbol{E^c} = \Omega \setminus E$. We have $\Pr(E^c) = 1 - \Pr(E)$.

What about Infinite Spaces?

Suppose that a process generates a number X in $[0, 1]$ such that each number is "equally likely."

What is the probability that $X = 0.5$? There are infinite numbers in $[0, 1]$, each one should have the same probability, and the sum should be 1. We can't make this work.

But what is the probability that X is in $[0, 0.5]$? "Half" of the numbers are in $[0, 0.5]$, and "half" are in $(0.5, 1]$, so the events $X \in [0, 0.5]$ and $X \in (0.5, 1]$ should have the same probability, $\frac{1}{2}$.

In general, it makes sense to say that $\Pr([a, b]) = b - a$ for $0 \leqslant a \leqslant b \leqslant 1$. The probability that $X \in [a, b]$ should be proportional to the length of the interval.

So, even though we can't define $\Pr(E) = \sum_{x \in E} p(x)$, we have a reasonable definition of $\Pr(E)$ for at least some events E.

Probability

A (finitely additive) *probability space* is a nonempty set Ω (called the *state* or *sample space*) equipped with a set of events $\mathcal{E} \subset 2^{\Omega}$ that satisfies

- $\varnothing \in \mathcal{E}$,
- for all $E \in \mathcal{E}$ we have $E^c \in \mathcal{E}$,
- for all $E, F \in \mathcal{E}$ we have $E \cup F \in \mathcal{E}$,

and a function $\Pr : \mathcal{E} \to [0, 1]$ such that

- $\Pr(\Omega) = 1$,
- for all $E, F \in \mathcal{E}$ such that $E \cap F = \varnothing$ we have $\Pr(E \cup F) = \Pr(E) + \Pr(F)$.

An actual probability space requires that we can take countably infinite unions, and additivity holds for countably infinite events. We won't need this for now.

Additivity

Additivity is the property that for all $E, F \in \mathcal{E}$ such that $E \cap F = \varnothing$ we have

$$\Pr(E \cup F) = \Pr(E) + \Pr(F).$$

Let's see some implications:

- Suppose that E_1, E_2, E_3 are pairwise disjoint, meaning that $E_1 \cap E_2 = \varnothing$, $E_1 \cap E_3 = \varnothing$ and $E_2 \cap E_3 = \varnothing$.

  Then $E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3) = \varnothing$, so

  $$\Pr(E_1 \cup E_2 \cup E_3) = \Pr(E_1) + \Pr(E_2 \cup E_3) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3).$$

  In general, if $E_1, \ldots, E_n$ are pairwise disjoint then

  $$\Pr(E_1 \cup \cdots \cup E_n) = \Pr(E_1) + \cdots + \Pr(E_n).$$

- We have $\varnothing \cap \varnothing = \varnothing$, so $\Pr(\varnothing) = \Pr(\varnothing \cup \varnothing) = \Pr(\varnothing) + \Pr(\varnothing)$. Therefore, $\Pr(\varnothing) = 0$.
- If $E \subset F$, which means "every state in E is in F," then we can partition F into E and the rest: $F = E \cup (F \smallsetminus E)$.

  Now, $E \cap (F \smallsetminus E) = \varnothing$, hence

  $$\Pr(F) = \Pr(E \cup (F \smallsetminus E)) = \Pr(E) + \Pr(F \smallsetminus E).$$

  This has two implications. First, $\Pr(F) \geqslant \Pr(E)$, or

  $$\Pr(E) \leqslant \Pr(F),$$

  because $\Pr(F \smallsetminus E) \geqslant 0$.

  Second,

  $$\Pr(F \smallsetminus E) = \Pr(F) - \Pr(E).$$

  Notice that this works for $E \subset F$.

Epistemic Interpretation of Probability

There is a set of *possible worlds* Ω. One possible world is the *actual world*, but we don't know which one it is.

A *proposition* is a statement about the world. E.g., E can be "tomorrow is going to rain." In some possible worlds it rains tomorrow, in others it doesn't.

We can see a proposition E as the set of possible worlds in which the proposition is true. So, $E \subset \Omega$.

If E, F are propositions, the proposition "E or F" corresponds to $E \cup F$, and "E and F" corresponds to $E \cap F$. The negation of E corresponds to $E^c = \Omega \setminus E$.

Notice that $E \subset F$ means that the proposition E implies the proposition F in the sense that in every possible world if E then F. (This is different from the proposition "E implies F," which is $E^c \cup F$.)

Epistemic Interpretation of Probability

We may assign a *credence*, or *degree of belief*, to each proposition E, i.e., a number in $[0,1]$ denoted by $\Pr(E)$.

Bigger $\Pr(E)$ means we are more certain that E is true. $\Pr(E) = 1$ means certainty, and $\Pr(E) = 0$ means we are certain that E is not the case.

We should have $\Pr(\Omega) = 1$ and $\Pr(\varnothing) = 0$.

Also, it makes sense to require that if $E \cap F = \varnothing$, i.e., we know that E and F cannot both be true, then $\Pr(E \cup F) = \Pr(E) + \Pr(F)$. In other words, we should assign credence $\Pr(E) + \Pr(F)$ to the proposition "E or F."

Epistemic Interpretation of Probability

In sum: we have subjective beliefs over a set of propositions $\mathcal{E} \subset 2^{\Omega}$. If we have a credence for two propositions E, F, we also have a credence for the propositions "not E" and "E or F."

(And, therefore, also "E and F," because it's "not (not E or not F)".)

Therefore, $(\Omega, \mathcal{E}, \Pr)$ is a probability space.

The exact same mathematical structure has this different interpretation.

Notice that it's truly different: there is no randomness here. Just uncertainty. The world may be deterministic. We use "probabilities" to quantify our uncertainty.

Random Variables

A **random variable** is a number that depends on the state. Formally: a function $X : \Omega \to \mathbb{R}$.

For example, suppose that X is a number such that on some states (or possible worlds) is 0 and on others is 1. We can define the event "$X = 0$" as the set of states (or possible worlds) E in which $X = 0$, i.e., $E = \{\omega \in \Omega : X(\omega) = 0\}$.

So, we can define $\Pr(X = 0) = \Pr(\{\omega \in \Omega : X(\omega) = 0\})$.

Example

Suppose that X_1, X_2, X_3 are random variables taking values in $\{0,1\}$ and such that each possible value of (X_1, X_2, X_3) is equally likely.

There are 8 possible values: $(0,0,0), (0,0,1), \ldots, (1,1,1)$. Each one has probability $1/8$.

Example. We want $\Pr(X_1 + X_2 = 1)$. There are four values of (X_1, X_2, X_3) that work: $(1,0,0), (1,0,1), (0,1,0), (0,1,1)$. Each one has probability $1/8$. Therefore

$$\Pr(X_1 + X_2 = 1) = \frac{4}{8} = \frac{1}{2}.$$

Question 1

What is $\Pr(X_1 + X_2 + X_3 = 1)$?

Basic Combinatorics

Consider these questions:

- If we toss 100 fair coins, what is the probability that we get 40 heads?
- Suppose that there are K Republicans in a population of N individuals. If we take a random subset of n individuals, what is the probability that k of them are Republicans?

We need a bit of combinatorics to answer them.

Basic Principle

If A is a finite set let $|A|$ be the number of elements.

Suppose we choose a_1 from a set A_1, then a_2 from a set A_2, and so on, until we choose a_n from a set A_n. How many possible choices do we have?

Answer: $|A_1| \times |A_2| \times \cdots \times |A_n|$.

Example. If I have 10 shirts and 5 pants, my choices of outfit are $10 \times 5 = 50$.

Number of Subsets

If A is a set of n elements, how many subsets does it have?

For each element we have two choices: we either include it in the subset, or we don't. That's 2 choices per element, and there are n elements. So the number of subsets is

$$\underbrace{2 \times \cdots \times 2}_{n \text{ times}} = 2^n.$$

Number of k-element Subsets

How many subsets of size k does A have?

First, how many lists of k distinct elements of A are there? We have n options for the first item, $n-1$ options for the second one, etc, and $n-k+1$ options for the last one. So, the number is

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

Now, suppose that $\{a_1, \ldots, a_k\}$ is a k-element subset of A. There are $k!$ ways of ordering its elements in order to form a list of k distinct elements: we have k choices for the first item, $k-1$ choices for the second, etc, until we get to the last one and we only have one choice. So, if C is the number of k-element subsets, $C \times k!$ is the number of lists of k distinct elements of A, which is $\frac{n!}{(n-k)!}$. Therefore,

$$\binom{n}{k} \equiv C = \frac{n!}{k!(n-k)!}.$$

Exercise

Question 2
There are 50 senators and 4 of them form a committee. How many possible choices are there?

Question 3
How many different permutations of the string "AAAABBB" are there?

Question 4
How many different permutations of the letters of the word MISSISSIPPI are there?

Applications

> If we toss 100 fair coins, what is the probability that we get 40 heads?

Each possible result is equally likely. There are 2 results for each coin, so 2^{100} in total. So, the probability of each result is $1/2^{100}$.

How many results have 40 heads? We have to choose the 40 coins that are heads. The rest will be tails. There are $\binom{40}{100}$ choices for these coins.

Answer: $\displaystyle\binom{40}{100}\frac{1}{2^{100}}$.

> Suppose that there are K Republicans in a population of N people. If we take a random subset of n people, what is the probability that k of them are Republicans?

There are $\displaystyle\binom{n}{N}$ subsets of n people. Since each subset is equally likely (implicit), each one has $\dfrac{1}{\binom{n}{N}}$ probability of being chosen.

How many n-person subsets have k Republicans? First, we have to choose k Republicans from the set of K Republicans. Then, we have to choose the remaining $n-k$ people, who have to be non-Republicans, and there are $N-K$ of those. Therefore, the number of subsets is $\displaystyle\binom{k}{K}\binom{n-k}{N-K}$. Each one has probability $\dfrac{1}{\binom{n}{N}}$, so the total probability is

$$\frac{\displaystyle\binom{k}{K}\binom{n-k}{N-K}}{\displaystyle\binom{n}{N}}.$$

Conditional Probability

Suppose that we learn that the proposition E is true, i.e., the actual world is in E.

How should we update our beliefs? Let $\Pr'$ be the updated beliefs.

We should have $\Pr'(E) = 1$, because we now know that E is true.

Given F, we should have $\Pr'(F) = \Pr'(F \cap E)$, because worlds not in E should have zero probability.

A natural thing to do is to simply rescale the previous beliefs so that $\Pr'(E) = 1$. That is: $\Pr'(F) = \frac{\Pr(F \cap E)}{\Pr(E)}$.

This is exactly the definition of **conditional probability**:

$$\Pr(\boldsymbol{F}|\boldsymbol{E}) = \frac{\Pr(F \cap E)}{\Pr(E)}.$$

It can be read as the probability of F given E.

Example

Pan, Jennifer, and Kaiping Chen. 2018. "Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances." APSR.

We have a set of complaints.

- A: set of complaint that alleges wrongdoing by prefecture officials,
- B: set of complaint forwarded by prefecture to upper-level leaders.

Counts in the data:

	B	B^c
A	17	61
A^c	573	761

If we take a random complaint, what are $\Pr(B|A)$ and $\Pr(B|A^c)$?

Counts in the data:

	B	B^c
A	17	61
A^c	573	761

$\Pr(B|A)$ is the probability that the complaint is forwarded given that the complaint alleged wrongdoing by officials. It's

$$\frac{\Pr(B \cap A)}{\Pr(A)} = \frac{17}{17+61} \approx 0.22.$$

$\Pr(B|A^c)$ is the probability that the complaint is forwarded given that the complaint did not allege wrongdoing by officials. It's

$$\frac{\Pr(B \cap A^c)}{\Pr(A^c)} = \frac{573}{573+761} \approx 0.43.$$

A lot larger! Weird.

Exercise

Question 5

What is $\Pr(A|B)$? Calculate it. What's the difference with $\Pr(B|A)$?

Law of Total Probability

We have $\Pr(A \cap B) = \Pr(A|B)\Pr(B)$.

Suppose that the alternatives $B_1, \ldots, B_n$ are mutually exclusive and exhaustive, meaning that $B_1 \cup \cdots \cup B_n = \Omega$ and $B_i \cap B_j = \varnothing$ for every $i \neq j$. Then

$$\Pr(A) = \Pr(A|B_1)\Pr(B_1) + \cdots + \Pr(A|B_n)\Pr(B_n).$$

Question 6

If a country is a democracy, it has 0.8% chances of having a coup in a given year. If it's an autocracy, it has 5.7% chances. 44% of countries are democracies. What is the probability that a randomly chosen country has a coup in a given year?

Bayes Rule

If a country is a democracy, it has 0.8% chances of having a coup next year. If it's an autocracy, it has 5.7% chances. 44% of countries are democracies. If we know that a random country had a coup in a given year, what are the chances that it is a democracy?

$$\begin{aligned}
\Pr(\text{Democracy} \mid \text{Has Coup}) &= \frac{\Pr(\text{Democracy} \cap \text{Has Coup})}{\Pr(\text{Has Coup})} \\
&= \frac{\Pr(\text{Has Coup} \mid \text{Democracy})\Pr(\text{Democracy})}{\Pr(\text{Has Coup})} \\
&= \frac{0.8\% \times 44\%}{3.5\%} = 10\%.
\end{aligned}$$

In general, if we know $\Pr(A|B_i)$ and $\Pr(B_i)$ then

$$\Pr(B_i|A) = \frac{\Pr(A \cap B_i)}{\Pr(A)} = \frac{\Pr(A|B_i)\Pr(B_i)}{\Pr(A|B_1)\Pr(B_1) + \cdots + \Pr(A|B_n)\Pr(B_n)}.$$

Independence

We say that events E and F are **independent** if $\Pr(E \cap F) = \Pr(E)\Pr(F)$.

In other words, if

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E),$$

or, equivalently, $\Pr(F|E) = \Pr(F)$.

This means: knowing that F doesn't tell you anything about whether E is true or not, and vice versa.

We say that the events $E_1, \dots, E_n$ are **mutually independent** if

$$\Pr(E_{i_1} \cap \cdots \cap E_{i_k}) = \Pr(E_{i_1}) \dots \Pr(E_{i_k})$$

for any $i_1 < \cdots < i_k$. I.e., the condition holds for any subset of $\{E_1, \dots, E_n\}$.

Independent Random Variables

We say that two random variables X, Y are independent if the events $X \in A$ and $Y \in B$ are independent for any sets A, B.

In other words, if knowing anything about X doesn't tell you anything about Y, and vice versa.

We say that the random variables $X_1, \ldots, X_n$ are independent if the events $X_1 \in A_1, \ldots X_n \in A_n$ are mutually independent for any sets $A_1, \ldots, A_n$. In words: if knowing anything about any of the variables doesn't tell you anything about any others.

Binomial Random Variables

A random variable X is **Bernoulli(p)** if it is binary, i.e., X takes values in $\{0, 1\}$, and $\Pr(X = 1) = p$. Think of it as a biased coin.

Suppose that $X_1, \ldots, X_n$ are independent Bernoulli(p) random variables. Let

$$S = X_1 + \cdots + X_n.$$

We say that S is **Binomial(n, p)**. Notice that S takes values in $\{0, \ldots, n\}$.

What is $\Pr(S = k)$?

Example. Take $n = 3$, $k = 2$. There are three possibilities for $X_1 + X_2 + X_3 = 2$:

- $X_1 = 1, X_2 = 1, X_3 = 0$, which occurs with probability

$$\begin{aligned}\Pr(X_1 = 1, X_2 = 1, X_3 = 0) &= \Pr(X_1 = 1)\Pr(X_2 = 1)\Pr(X_3 = 0)\\ &= p \cdot p \cdot (1 - p) = p^2(1 - p),\end{aligned}$$

- $X_1 = 1, X_2 = 0, X_3 = 1$, which occurs with probability

$$\begin{aligned}\Pr(X_1 = 1, X_2 = 0, X_3 = 1) &= \Pr(X_1 = 1)\Pr(X_2 = 0)\Pr(X_3 = 1)\\ &= p \cdot (1 - p) \cdot p = p^2(1 - p),\end{aligned}$$

- $X_1 = 0, X_2 = 1, X_3 = 1$, which occurs with probability

$$\begin{aligned}\Pr(X_1 = 0, X_2 = 1, X_3 = 1) &= \Pr(X_1 = 0)\Pr(X_2 = 1)\Pr(X_3 = 1)\\ &= (1 - p) \cdot p \cdot p = p^2(1 - p).\end{aligned}$$

Therefore, $\Pr(X_1 + X_2 + X_3 = 2) = 3p^2(1 - p)$.

In general, there are $\binom{n}{k}$ ways of choosing which of the k variables are 1.

Each one adds probability $p^k(1-p)^{n-k}$.

Therefore, $\Pr(S = k) = \Pr(X_1 + \cdots + X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Question 7

In some population 35% of the people support a carbon tax. We randomly sample 100 individuals with replacement (meaning that, though unlikely, we could sample the same individual multiple times). Find a formula for the probability that the proportion of individuals that support a carbon tax in the sample is between 30% and 40%.

Continuous Random Variables

We want a random variable to take any real number in some interval.

Suppose that $f : \mathbb{R} \to [0, +\infty)$ is such that $\int_{-\infty}^{+\infty} f(x)\, dx = 1$.

Then we can define a random variable X taking values in $\mathbb{R}$ such that

$$\Pr(X \leqslant a) = \int_{-\infty}^{a} f(x)\, dx$$

and $\Pr(X = a) = 0$ for any $a \in \mathbb{R}$.

We say that f is the **density** (PDF) of X, and $F(a) = \Pr(X \leqslant a)$ is its **cumulative distribution function** (CDF).

- $\Pr(X < a) = \Pr(X \leqslant a) - \Pr(X = a) = F(a)$,
- $\Pr(a \leqslant X \leqslant b) = \Pr(X \leqslant b) - \Pr(X < a) = F(b) - F(a)$.

Uniform Distribution

The **uniform distribution** on $[a,b]$, denoted by $\boldsymbol{\mathcal{U}[a,b]}$, has density

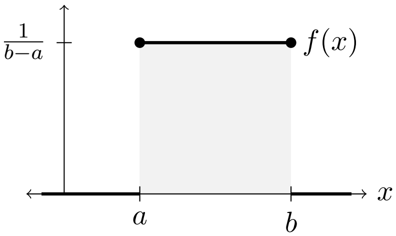
$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b], \\ 0 & \text{otherwise.} \end{cases}$$

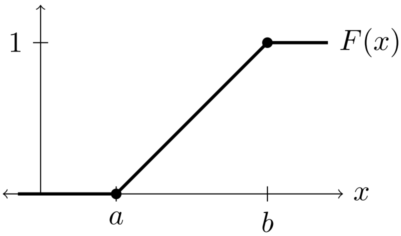
We see that

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \left[\frac{x}{b-a}\right]_a^b = \frac{b}{b-a} - \frac{a}{b-a} = 1,$$

so it is in fact a density.

PDF and CDF of $\mathcal{U}[a, b]$





Exercise

Question 8

Consider a continuous random variable X with PDF

$$f(x) = \begin{cases} e^{-x} & \text{if } x \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

Verify that f is a valid density, and calculate $\Pr(X > 1)$.