Mathematical Definition of Discrete Random Variables

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU Courant Institute of Mathematical Sciences

NYU Data Science

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Goal

Model uncertain quantities that can take discrete values

- Number of students attending a class
- Number of goals scored in a soccer game
- Number of earthquakes in San Francisco over a year

We represent them using random variables

Notation

Deterministic variables: a, b, x, y

Random variables: $\tilde{a}$, $\tilde{b}$, $\tilde{x}$, $\tilde{y}$

Deterministic variables represent fixed values

Random variables represent uncertain values

They are described probabilistically, we don't say

the random variable $\tilde{a}$ equals 3

but rather

the probability that $\tilde{a}$ equals 3 is 0.5

What is a random variable?

Data scientist:

An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

Rolling a die twice

Probability space representing two rolls of a six-sided die

Outcomes?

$$\omega := \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} \qquad \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}$$

Quantity of interest: Result of first roll

Key insight: It can be represented as a function of the outcome

Functions of Outcomes

A random variable is a function that maps outcomes to real numbers

$$\tilde{a}(\omega) := \omega_1$$

The range of a random variable is the set of values that it can take

Range of $\tilde{a}$? $\{1, 2, 3, 4, 5, 6\}$

Functions of Outcomes

We can define many random variables in the same probability space

Value of second roll $\tilde{b}(\omega) := \omega_2$

Sum of rolls $\tilde{c}(\omega) := \omega_1 + \omega_2$

The outcome fixes the values of all random variables simultaneously

Very useful to represent dependencies between uncertain quantities

Probability mass function

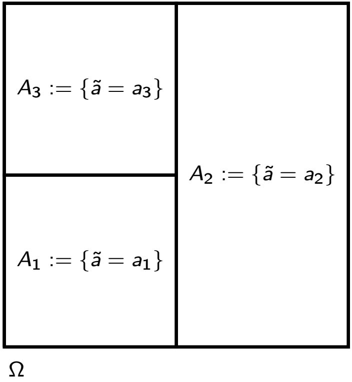
The probability mass function (pmf) $p_{\tilde{a}} : \mathbb{R} \to [0, 1]$ of $\tilde{a}$ is the probability that $\tilde{a}$ equals each of its possible values a_1, a_2, ...:

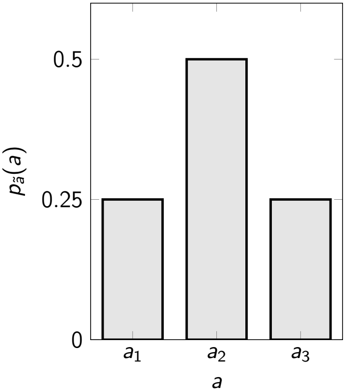
$$p_{\tilde{a}}(a_i) := \mathrm{P}(\{\omega \mid \tilde{a}(\omega) = a_i\})$$

We say that $\tilde{a}$ is distributed according to $p_{\tilde{a}}$

Wait, *are we sure we can assign probabilities to these events?*

Probability mass function





Formal definition

Probability space $(\Omega, \mathcal{C}, \mathrm{P})$

Function $\tilde{a} : \Omega \to \mathbb{R}$ maps Ω to discrete set $\{a_1, a_2, \ldots\}$

The function $\tilde{a}$ is a discrete random variable if the sets

$$A_i := \{\omega \mid \tilde{a}(\omega) = a_i\} \qquad i = 1, 2, \ldots$$

are in the collection $\mathcal{C}$ so that the probability

$$\mathrm{P}(\tilde{a} = a_i) := \mathrm{P}(A_i) \qquad i = 1, 2, \ldots$$

is well defined

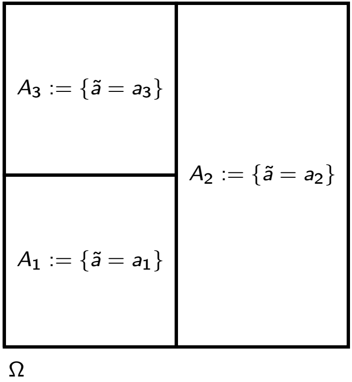
Such functions are called measurable

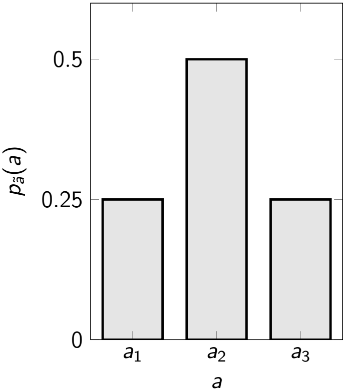
In practice

We never define random variables as functions of outcomes

Instead, we define them through their pmf

Important: Preimages form a partition of Ω





Preimages form a partition

Probability space $(\Omega, \mathcal{C}, \mathrm{P})$

Function $\tilde{a} : \Omega \to \mathbb{R}$ maps Ω to discrete set $\{a_1, a_2, \ldots\}$

The events

$$A_i := \{\omega \mid \tilde{a}(\omega) = a_i\}, \qquad i = 1, 2, \ldots$$

form a partition of Ω

Computing probabilities

Probability that $\tilde{a}$ is in any set $S \subseteq \{a_1, a_2, \ldots\}$

$$\begin{aligned} \mathrm{P}(\tilde{a} \in S) &= \mathrm{P}(\{\omega \mid \tilde{a}(\omega) \in S\}) \\ &= \mathrm{P}(\cup_{a_i \in S} A_i) \\ &= \sum_{a_i \in S} \mathrm{P}(A_i) \\ &= \sum_{a_i \in S} p_{\tilde{a}}(a) \end{aligned}$$

The pmf is all we need, we can forget about the probability space!

Any pmf must sum to one

$$\sum_{i=1,2,\ldots} p_{\tilde{a}}(a_i) = \mathrm{P}\left(\cup_i A_i\right)$$
$$= \mathrm{P}(\Omega) = 1$$

In practice

To model an uncertain quantity with values in a discrete set A using a discrete random variable $\tilde{a}$ we just estimate the pmf $p_{\tilde{a}}$

Mathematician: *How do we know there's an underlying probability space?*

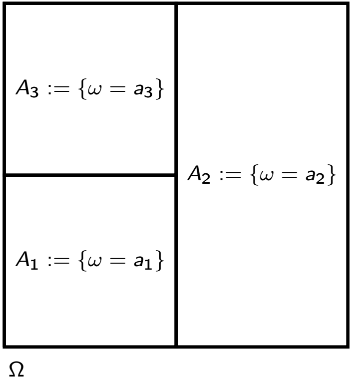
If $p_{\tilde{a}}$ is nonnegative and sums to one, we can build it:

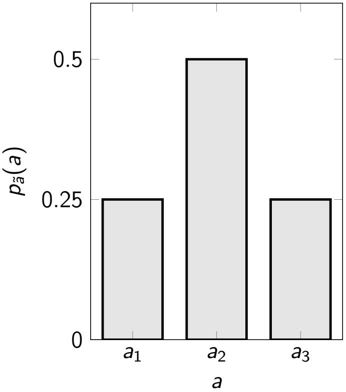
Sample space: A

Collection of events: Power set of A

Probability measure: $p_{\tilde{a}}$

Reverse-engineering the probability space





What have we learned?

Data scientist:

An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

That they are both right!