Mathematical Definition of Continuous Random Variables

Probability and Statistics for Data Science

Carlos Fernandez-Granda

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

Motivation

Physical quantities such as length, mass, or time are usually modeled as being continuous

Goal: Define continuous random variables to represent uncertain continuous quantities

Notation

Deterministic variables: a, b, x, y

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

What is a random variable?

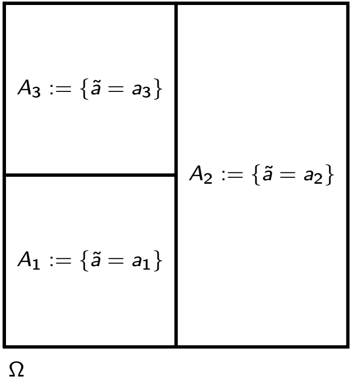
Data scientist:

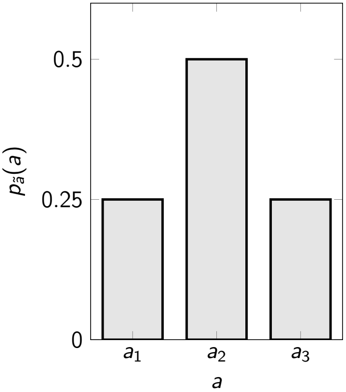
An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

Discrete random variables





Discrete random variables

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

Key question

Can we describe an uncertain continuous quantity $\tilde{a}$ through probabilities of the form

$$\mathrm{P}(\tilde{a} = a)?$$

No!

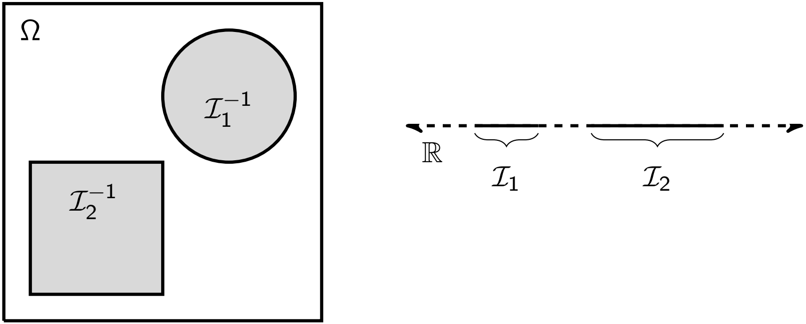
Intuitive reason: Individual points should have zero probability

Mathematical reason: If we assign nonzero probability to an uncountable set of points, the probability of the set explodes to ∞

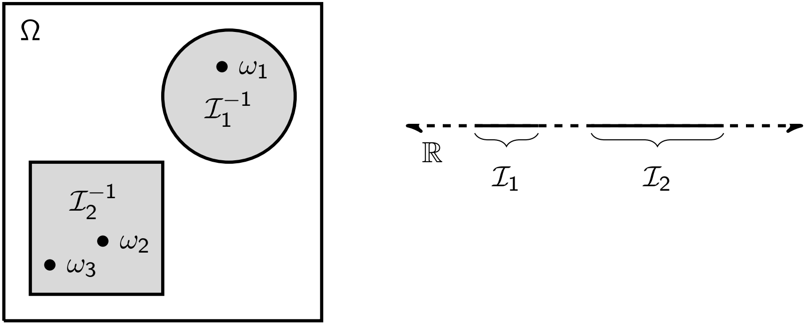
Strategy

Describe continuous random variables using the probability that they belong to intervals

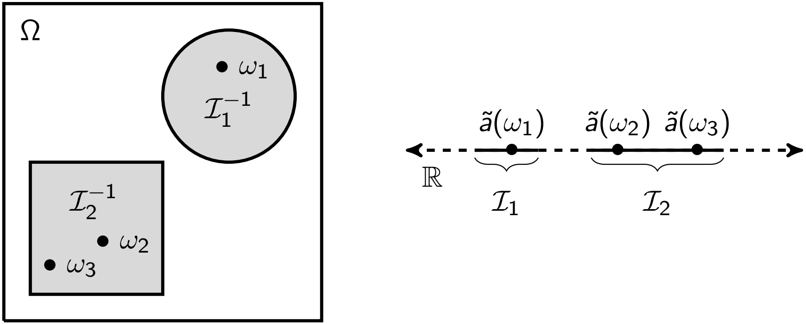
Continuous random variables



Continuous random variables



Continuous random variables



Continuous random variable

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

Function $\tilde{a} : \Omega \to \mathbb{R}$

The function $\tilde{a}$ is a valid random variable if for any interval $\mathcal{I} := [a, b] \subseteq \mathbb{R}$, $a \leq b$

$$\mathcal{I}^{-1} := \{\omega \mid \tilde{a}(\omega) \in \mathcal{I}\}$$

is in the collection $\mathcal{C}$, so

$$\mathrm{P}(\tilde{a} \in \mathcal{I}) = \mathrm{P}(\mathcal{I}^{-1}) \quad \text{is well defined}$$

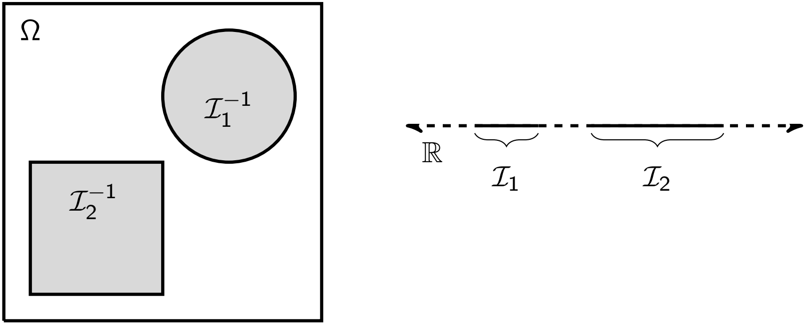
Such functions are called measurable

Continuous random variables

We say that a random variable $\tilde{a}$ is continuous if for any individual real value $a \in \mathbb{R}$

$$\mathrm{P}(\tilde{a} = a) = 0$$

$$\mathrm{P}(\tilde{a} \in \mathcal{I}_1 \cup \mathcal{I}_2) = \mathrm{P}(\tilde{a} \in \mathcal{I}_1) + \mathrm{P}(\tilde{a} \in \mathcal{I}_2)?$$



Unions of intervals

Let $\mathcal{I}_1$, $\mathcal{I}_2$, ..., $\mathcal{I}_n$ be disjoint intervals of $\mathbb{R}$

$$\begin{aligned}
\mathrm{P}(\tilde{a} \in \cup_{i=1}^{n} \mathcal{I}_i) &= \mathrm{P}(\{\omega \mid \tilde{a}(\omega) \in \cup_{i=1}^{n} \mathcal{I}_i\}) \\
&= \mathrm{P}\left(\left\{\omega \mid \omega \in (\cup_{i=1}^{n} \mathcal{I}_i)^{-1}\right\}\right) \\
&= \mathrm{P}\left(\left\{\omega \mid \omega \in \cup_{i=1}^{n} \mathcal{I}_i^{-1}\right\}\right) \\
&= \sum_{i=1}^{n} \mathrm{P}\left(\left\{\omega \mid \omega \in \mathcal{I}_i^{-1}\right\}\right) \\
&= \sum_{i=1}^{n} \mathrm{P}(\tilde{a} \in \mathcal{I}_i)
\end{aligned}$$

Intervals

For any $[a, b] \subseteq \mathbb{R}$, $a \leq b$,

$$\begin{aligned} \mathrm{P}(\tilde{a} \in [a, b]) &= \mathrm{P}(\tilde{a} = a) + \mathrm{P}(\tilde{a} \in (a, b)) + \mathrm{P}(\tilde{a} = b) \\ &= \mathrm{P}(\tilde{a} \in (a, b)) \end{aligned}$$

Borel sets

Technically, the probability that $\tilde{a} \in S$ is only well defined if S is a union of countable intervals

These sets are called Borel sets

Non-Borel sets exist!

Do we care? No

Conclusion

We describe continuous random variables in terms of the probability that they belong to any interval

How do we encode this information?

Using the cumulative distribution function or the probability density function