Estimating Probabilities from Data

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

Plan

Explain how to estimate probabilities from data

Probability Theory and Statistics

Probability theory provides a mathematical framework to describe uncertainty, but does not care about connection to reality

Statistics aims to extract information from data

Intuitive definition of probability

$$\mathrm{P}(\text{event}) = \frac{\text{number of times event occurs}}{\text{total repetitions}}$$

Six-sided die

Data collection: We roll the die 60 times and observe 8 twos

Probability of event *Rolling a two*?

Empirical probability

Let A be an event in a sample space Ω

Let $X := \{x_1, x_2, \ldots, x_n\}$ be a set of data with values in Ω

The empirical probability of A is the statistical estimator

$$\mathrm{P}_X(A) := \frac{\sum_{i=1}^{n} 1_{x_i \in A}}{n}$$

where $1_{x_i \in A}$ is one if $x_i \in A$ and zero otherwise

Six-sided die

$\Omega := \{1, 2, 3, 4, 5, 6\}$

Collection: Power set of Ω

Probability measure: $P(\{i\}) = \theta_i$ for $1 \leq i \leq 6$

Data collection: We roll the die 60 times and obtain

$$x_1 := 10, \quad x_2 := 8, \quad x_3 := 18, \quad x_4 := 7, \quad x_5 := 7, \quad x_6 := 10$$

Empirical probability estimates

$$\mathrm{P}_X(\{1\}) = \frac{10}{60} \qquad \mathrm{P}_X(\{2\}) = \frac{8}{60} \qquad \mathrm{P}_X(\{3\}) = \frac{18}{60}$$

$$\mathrm{P}_X(\{4\}) = \frac{7}{60} \qquad \mathrm{P}_X(\{5\}) = \frac{7}{60} \qquad \mathrm{P}_X(\{6\}) = \frac{10}{60}$$

Coin flip

We simulate a fair coin flip twenty times

Heads (out of 20)	15	13	10	9	9	8	9	9	12	8
Empirical prob.	0.75	0.65	0.5	0.45	0.45	0.4	0.45	0.45	0.6	0.4

If we flip 21 times, no estimate can be exact!

House of Representatives 1984

		Duty-free exports	
		Yes	No
Budget	Yes	151	88
	No	21	140

Goal: Understand relationship between two issues

If representative votes Yes on Budget, are they more likely to vote Yes on Duty-free exports?

Probabilistic modeling

Interpret voting as a repeatable experiment and build a probability space

Outcomes? *Yes-Yes*, *Yes-No*, *No-Yes*, *No-No*

Events of interest: B (Yes on Budget), D (Yes on Duty-free)

Empirical probabilities

		Duty-free exports	
		Yes	No
Budget	Yes	151	88
	No	21	140

$$\mathrm{P}(B) = \frac{239}{400} = 0.598$$

$$\mathrm{P}(D) = \frac{172}{400} = 0.43$$

What about $\mathrm{P}(D \mid B)$?

Intuitive definition of conditional probability

$$\mathrm{P}(\text{event } B \,|\, \text{event } A) = \frac{\text{number of times } A \text{ and } B \text{ occur}}{\text{number of times } A \text{ occurs}}$$

Conditional probabilities

		Duty-free exports	
		Yes	No
Budget	Yes	151	88
	No	21	140

$$\mathrm{P}(D \mid B) = \frac{151}{239} = 0.632$$

$$\mathrm{P}(D \mid B^c) = \frac{21}{161} = 0.130$$

Empirical conditional probability

Let A and B be events in a sample space Ω

Let $X := \{x_1, x_2, \ldots, x_n\}$ be a set of data with values in Ω

The empirical conditional probability of B given A is

$$\mathrm{P}_X(B \mid A) := \frac{\sum_{i=1}^{n} 1_{x_i \in A \cap B}}{\sum_{i=1}^{n} 1_{x_i \in A}}$$

where $1_{x_i \in S}$ is one if $x_i \in S$ and zero otherwise

What have we learned?

How to estimate probabilities from data