Monte Carlo Simulation

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

Life is not a homework problem!

2021 Tokyo Olympics

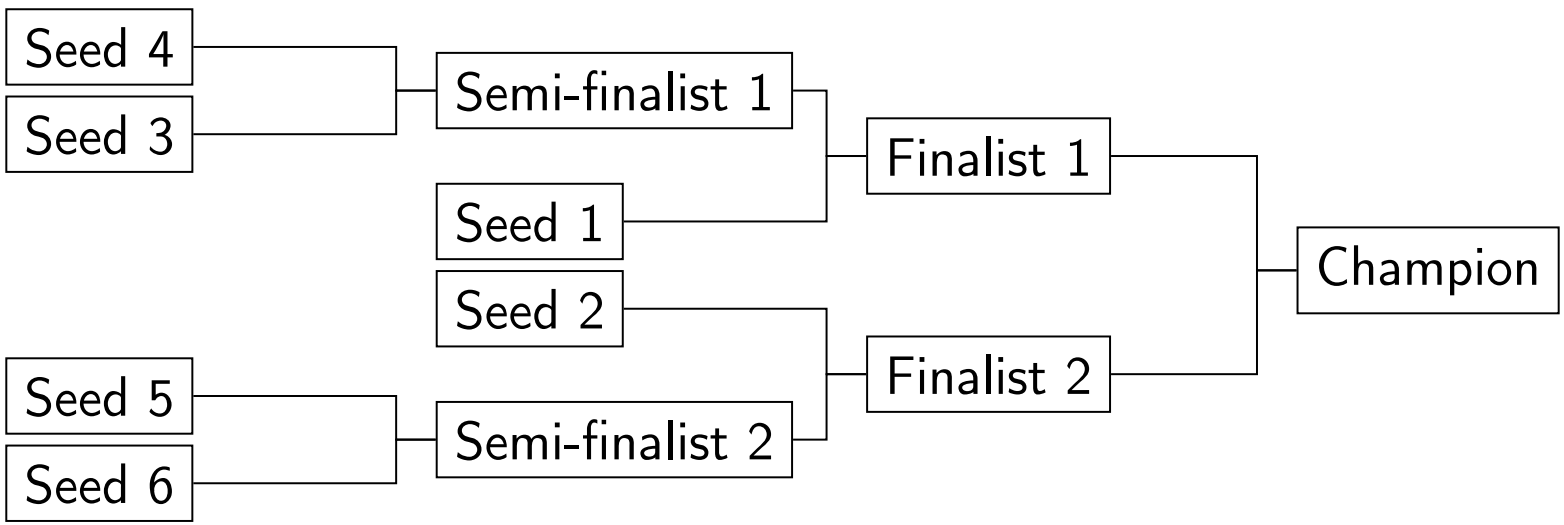
3x3 basketball tournament

Participants: Belgium, China, Japan, Latvia, the Netherlands, Poland, the Russian Olympic Committee (ROC), and Serbia

Goal: Estimate probability that each team wins gold, silver or bronze

Tournament

Group stage followed by bracket



Data

RANK		PLAYER	NATIONALITY	POINTS
1		**Dusan Bulut**	Serbia	**827,172**
2		**Nauris Miezis**	Latvia	**825,366**
3		**Aleksandar Ratkov**	Serbia	**788,975**
4		**Mihailo Vasic**	Serbia	**788,507**
5		**Stefan Kojic**	Serbia	**788,063**
6		**Karlis Lasmanis**	Latvia	**735,840**
7		**Edgars Krumins**	Latvia	**733,800**
8		**Dominique Jones**	United States	**689,922**

Assumption

Result of game only depends on total rating points of team members

$$\mathrm{P}(\text{A beats B}) = \frac{\text{Total points of A}}{\text{Total points of A} + \text{Total points of B}}$$

Points

Country	Ranking points
Serbia	2,997,304
Latvia	2,959,152
ROC	970,438
Netherlands	768,134
Belgium	664,381
Poland	654,908
China	356,522
Japan	334,018

$$\mathrm{P}(\text{Belgium beats Poland}) = 0.504$$

$$\mathrm{P}(\text{China beats Serbia}) = 0.106$$

$$\mathrm{P}(\text{Latvia beats Netherlands}) = 0.794$$

Probability that Belgium wins bronze?

Just compute the probability of all the ways that Belgium can win bronze and add them

Is this easy to do? Yes, just check all possible results

Wait, how many results do we need to check? $2^{40} > 10^{12}$!
(Total games: 40)

Monte Carlo method

Stanislaw Ulam had the same problem trying to model a card game in 1946

After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays

In our case: simulate tournament over and over, and compute the fraction of times Belgium wins bronze

Intuitive definition of probability

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

Monte Carlo method

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

Assume we can generate outcomes from Ω according to P

To approximate the probability of $A \in \mathcal{C}$, we

1. Generate n simulated outcomes: $s_1, s_2, \ldots, s_n \in \Omega$

2. Compute the fraction of the outcomes in A,

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

where $1_{x_i \in S}$ is one if $s_i \in S$ and zero otherwise for any event $S \in \mathcal{C}$

Estimated probabilities ($n = 10^4$)

Country	Ranking points	Probability of winning (%)			
		Gold	Silver	Bronze	Group
Serbia	2,997,304	43.2	27.1	19.6	43.3
Latvia	2,959,152	42.0	28.0	18.9	42.9
ROC	970,438	6.3	14.9	18.9	5.6
Netherlands	768,134	3.6	10.3	14.4	3.2
Belgium	664,381	2.2	8.5	11.4	2.4
Poland	654,908	2.2	7.7	11.3	2.1
China	356,522	0.3	1.7	3.1	0.4
Japan	334,018	0.2	1.7	2.5	0.2

Actual result? Gold: *Latvia* Silver: *ROC* Bronze: *Serbia*

What about conditional probabilities?

Intuitively,

$$\mathrm{P}(B \mid A) = \frac{\text{number of times } A \text{ and } B \text{ occurs}}{\text{number of times } A \text{ occurs}}$$

Conditional probabilities

To approximate the conditional probability of an event $B \in \mathcal{C}$ conditioned on A, we

1. Generate n simulated outcomes: $s_1, s_2, \ldots, s_n \in \Omega$

2. Compute fraction of outcomes in A that are also in B

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

Back to 3x3 tournament

Goal: Estimate probabilities of winning gold conditioned on Serbia being eliminated in group stage

Estimated conditional probabilities

Country	Probability of gold conditioned on the event *Serbia does not reach bracket* (%)		
	10^4 runs	10^6 runs	10^7 runs
Latvia	68.6	63.5	63.4
ROC	10.0	13.3	13.2
Netherlands	7.1	8.5	8.6
Belgium	10.0	6.5	6.3
Poland	4.3	6.2	6.1
China	0	1.2	1.3
Japan	0	0.8	1.1
Serbia	0	0	0
Runs where event occurs	70	5,539	55,719

What have we learned?

Life is not a homework problem

Probabilities can often not be computed exactly

Solution: Monte Carlo simulation

But be careful with rare events!