Multivariate Discrete 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`

Goal

Learn to model interactions between multiple uncertain discrete quantities

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

Plan

Mathematical characterization of multiple random variables

Practical description of multiple random variables via probabilities

Estimation from data

Rolling a die twice

Probability space representing two rolls of a six-sided die

Outcomes:

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

Random variables

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

$$\tilde{b}(\omega) := \omega_2$$

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

The outcome fixes the values of all random variables simultaneously

If $\omega = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$ $\quad \tilde{a}(\omega) = 3 \quad \tilde{b}(\omega) = 1 \quad \tilde{c}(\omega) = 4$

Definition of random variable

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

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

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

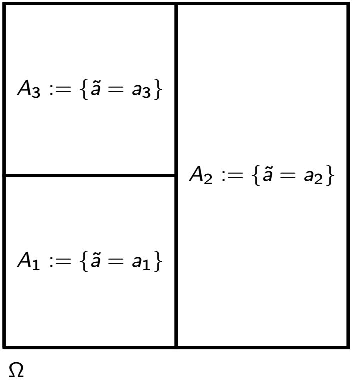
$$A_i := \{\omega \mid \tilde{a}(\omega) = a_i\} \qquad i = 1, 2, \ldots$$

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

Probability mass function

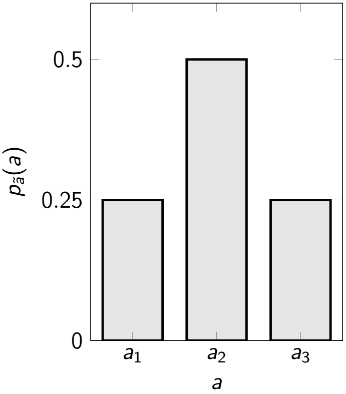


$A_3 := \{\tilde{a} = a_3\}$

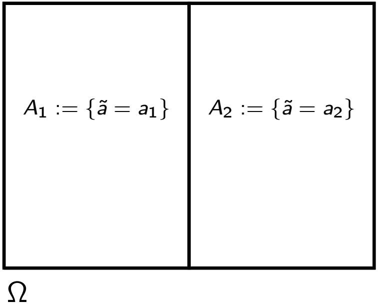
$A_2 := \{\tilde{a} = a_2\}$

$A_1 := \{\tilde{a} = a_1\}$

Ω



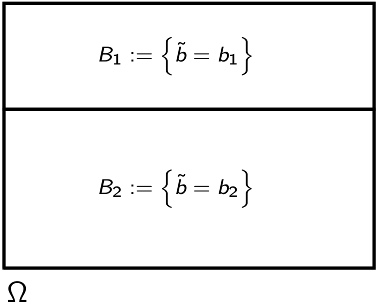
Sample space



$A_1 := \{\tilde{a} = a_1\}$

$A_2 := \{\tilde{a} = a_2\}$

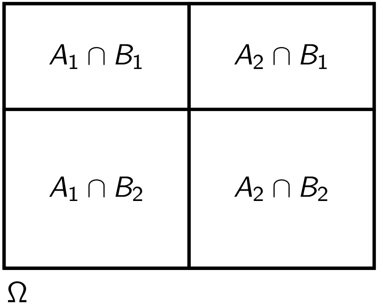
Ω



$B_1 := \left\{\tilde{b} = b_1\right\}$

$B_2 := \left\{\tilde{b} = b_2\right\}$

Ω



$A_1 \cap B_1$

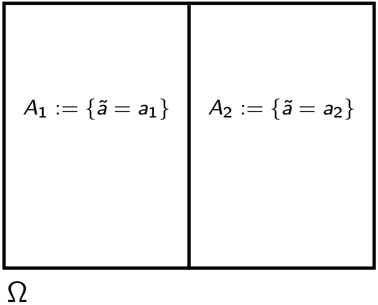
$A_2 \cap B_1$

$A_1 \cap B_2$

$A_2 \cap B_2$

Ω

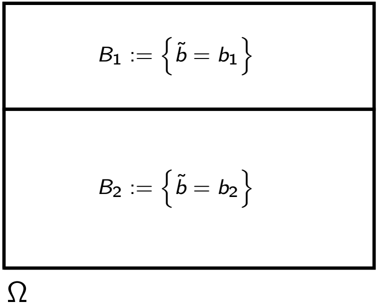
Sample space



$A_1 := \{\tilde{a} = a_1\}$

$A_2 := \{\tilde{a} = a_2\}$

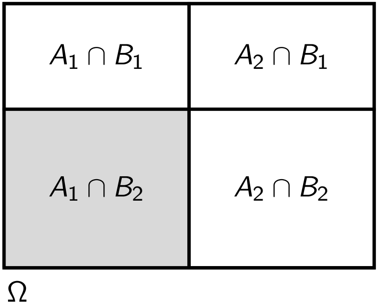
Ω



$B_1 := \left\{\tilde{b} = b_1\right\}$

$B_2 := \left\{\tilde{b} = b_2\right\}$

Ω



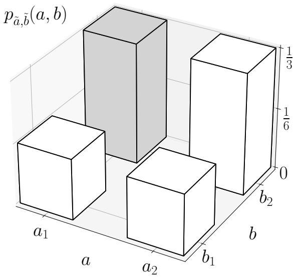
$A_1 \cap B_1$

$A_2 \cap B_1$

$A_1 \cap B_2$

$A_2 \cap B_2$

Ω

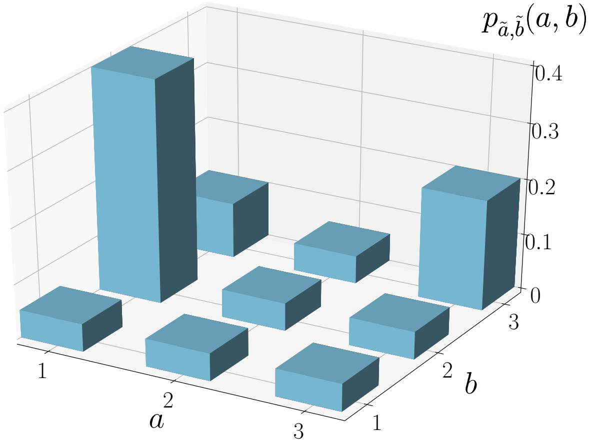


Joint probability mass function

The joint pmf of $\tilde{a} : \Omega \to A$ and $\tilde{b} : \Omega \to B$ is defined as

$$p_{\tilde{a},\tilde{b}}(a, b) := \mathrm{P}\left(\tilde{a} = a, \tilde{b} = b\right)$$

Joint pmf



Random vector

Each entry $\tilde{x}[i]$ is a random variable in the same probability space

$$\tilde{x} := \begin{bmatrix} \tilde{x}[1] \\ \tilde{x}[2] \\ \cdots \\ \tilde{x}[d] \end{bmatrix}$$

Joint probability mass function

The joint pmf of a discrete random vector $\tilde{x}$ is

$$p_{\tilde{x}}(x) := \mathrm{P}\left(\tilde{x}[1] = x[1], \tilde{x}[2] = x[2], \ldots, \tilde{x}[d] = x[d]\right)$$

Computing probabilities

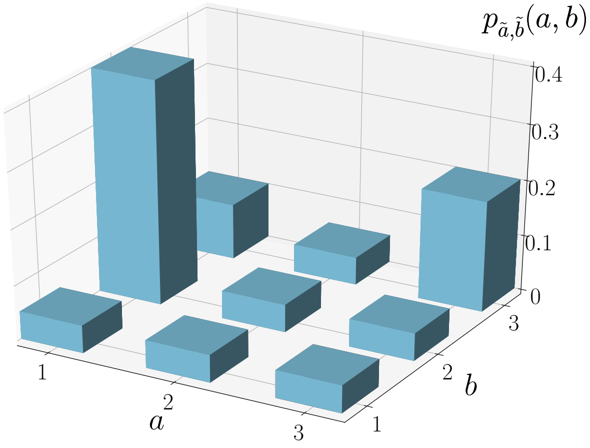
For any set S

$$\begin{aligned}
\mathrm{P}\left((\tilde{a},\tilde{b}) \in S\right) &= \mathrm{P}\left(\cup_{(a,b)\in S}\left\{\tilde{a}=a,\tilde{b}=b\right\}\right) \\
&= \sum_{(a,b)\in S} \mathrm{P}\left(\tilde{a}=a,\tilde{b}=b\right) \\
&= \sum_{(a,b)\in S} p_{\tilde{a},\tilde{b}}\left(a,b\right)
\end{aligned}$$

Similarly, for a d-dimensional random vector

$$\mathrm{P}\left(\tilde{x}\in S\right) = \sum_{x\in S} p_{\tilde{x}}\left(x\right)$$

Computing probabilities



$$\mathrm{P}(\{\tilde{a} < 2, \tilde{b} > 1\}) = p_{\tilde{a},\tilde{b}}(1,2) + p_{\tilde{a},\tilde{b}}(1,3) = 0.5$$

Properties

Joint pmfs are nonnegative (they are probabilities)

$$\sum_{a \in A} \sum_{b \in B} p_{\tilde{a},\tilde{b}}\left(a,b\right) = \mathrm{P}\left(\left\{\tilde{a} \in A\right\} \cap \left\{\tilde{b} \in B\right\}\right) = 1$$

$$\sum_{x[1] \in R_1} \sum_{x[2] \in R_2} \cdots \sum_{x[d] \in R_d} p_{\tilde{x}}\left(x\right) = 1$$

Any function with these properties is a valid joint pmf

Estimating a joint pmf from data

If data equal

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

How would you estimate $p_{\tilde{a},\tilde{b}}\left(\left[\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right]\right)$?

Empirical joint pmf

Data: $X := \{x_1, x_2, \ldots, x_n\}$

The empirical joint pmf is

$$p_X(v) := \frac{\sum_{i=1}^{n} 1_{x_i = v}}{n},$$

where $1_{x_i = v}$ equals one if $x_i = v$ and zero otherwise

Movie ratings

Movielens dataset

Users give 1-5 ratings to movies

Goal: Model ratings for Independence Day and Mission Impossible

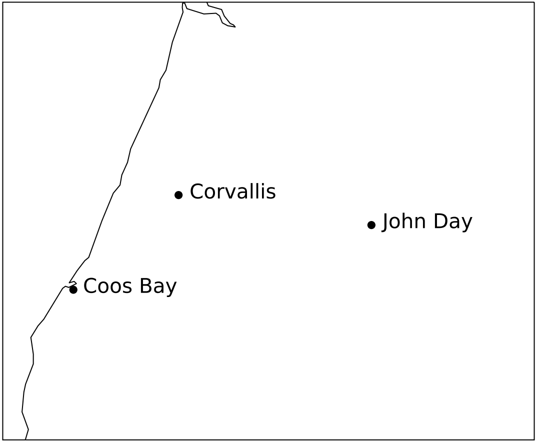
Movie ratings

		Independence Day				
		1	**2**	**3**	**4**	**5**
Mission Impossible	**1**	2	3	5	1	0
	2	3	12	18	11	5
	3	5	14	37	41	17
	4	6	15	20	47	19
	5	0	0	4	12	17

Empirical joint pmf (%)

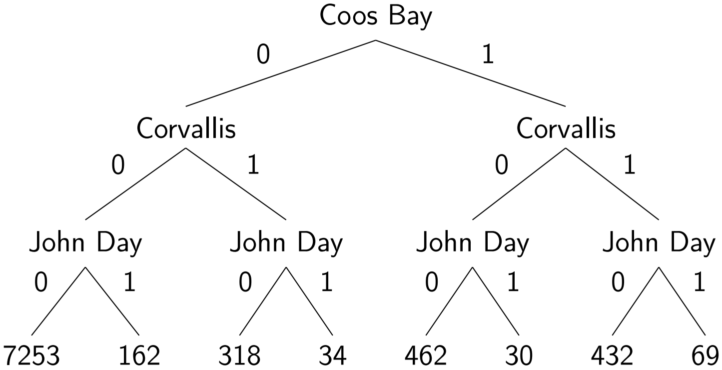
				Independence Day		
		1	**2**	**3**	**4**	**5**
Mission Impossible	**1**	0.6	1	1.6	0.3	0
	2	1	3.8	5.7	3.5	1.6
	3	1.6	4.5	11.8	13.1	5.4
	4	1.9	4.8	6.4	15	6.1
	5	0	0	1.3	3.8	5.4

Precipitation in Oregon



Goal: Model precipitation in Coos Bay, Corvallis, John Day

Precipitation in Oregon



Coos Bay

0 1

Corvallis Corvallis

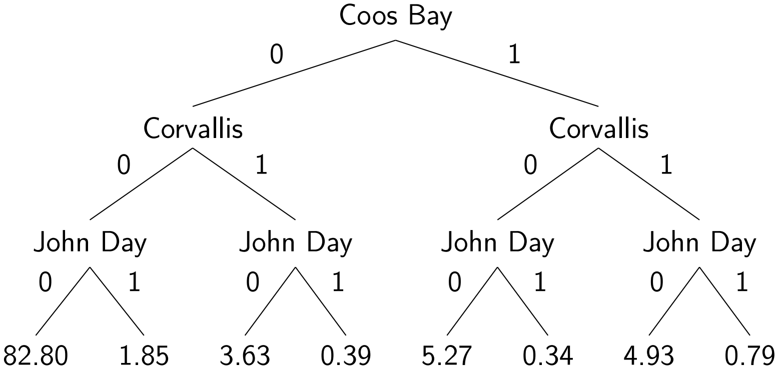
0 1 0 1

John Day John Day John Day John Day

0 1 0 1 0 1 0 1

7253 162 318 34 462 30 432 69

Empirical joint pmf (%)



Coos Bay

0 1

Corvallis Corvallis

0 1 0 1

John Day John Day John Day John Day

0 1 0 1 0 1 0 1

82.80 1.85 3.63 0.39 5.27 0.34 4.93 0.79

What have we learned?

Mathematical definition of multivariate discrete random variables

Definition and properties of the joint pmf

How to estimate the joint pmf from data