The Probability Mass Function

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

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

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, ...:

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

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

Properties

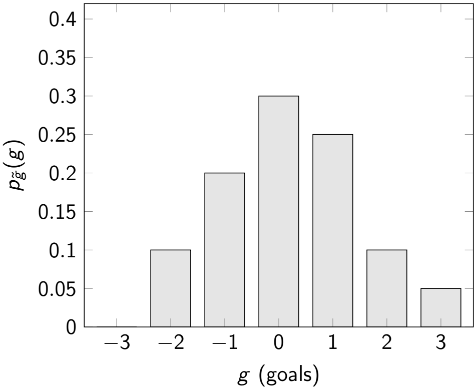
- Nonnegative
- Sums to one

$$\sum_{i=1,2,\ldots} p_{\tilde{a}}(a_i) = 1$$

- For any set $S \subseteq \{a_1, a_2, \ldots\}$

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

Goal difference



$$\mathrm{P}\left(\tilde{g} \in \{-2, 2\}\right) = p_{\tilde{g}}\left(-2\right) + p_{\tilde{g}}\left(2\right) = 0.2$$

$$\mathrm{P}\left(\tilde{g} > 1\right) = p_{\tilde{g}}\left(2\right) + p_{\tilde{g}}\left(3\right) = 0.15$$

Points

What is the distribution of the points gained by each team?

The points gained x equal

$$h(g) := \begin{cases} 0 & \text{if } g < 0 \\ 1 & \text{if } g = 0 \\ 3 & \text{if } g > 0 \end{cases}$$

Function of a random variable

Given a deterministic function h, is $\tilde{b} := h(\tilde{a})$ a random variable?

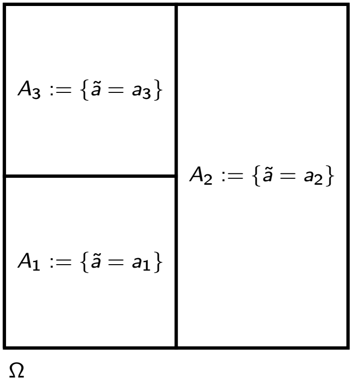
If $\tilde{a}$ takes values a_1, a_2, ... then $\tilde{b}$ takes values $h(a_1)$, $h(a_2)$, ...

$\tilde{b} := h \circ \tilde{a}$ is a function from Ω to $\{h(a_1), h(a_2), \ldots\}$

Mathematician:

Is it measurable?

Function of a random variable



Function of a random variable

We need $\mathrm{P}(\tilde{b} = b)$ to exist for all $b \in \{h(a_1), h(a_2), \ldots\}$

Probability measure of the probability space must assign a probability to

$$\begin{aligned} B_i &:= \left\{\omega \mid \tilde{b}\,(\omega) = b\right\} \\ &= \cup_{h(a_j)=b} \left\{\omega \mid \tilde{a}\,(\omega) = a_j\right\} \end{aligned}$$

It does because $A_i := \{\omega \mid \tilde{a}\,(\omega) = a_i\}$, $i = 1, 2, \ldots$, are assigned probabilities if $\tilde{a}$ is a random variable

Function of a random variable

How do we compute $p_{\tilde{b}}$ from $p_{\tilde{a}}$ when $\tilde{b} = h(\tilde{a})$?

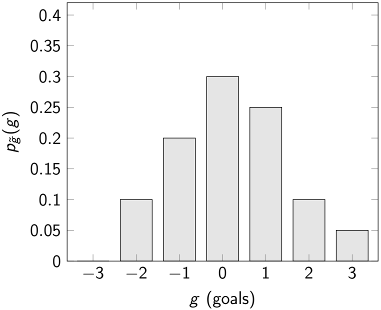
$$
\begin{aligned}
p_{\tilde{b}}(b) &= \mathrm{P}(\tilde{b} = b) \\
&= \mathrm{P}\left(h(\tilde{a}) = b\right) \\
&= \sum_{\{a \mid h(a) = b\}} p_{\tilde{a}}(a)
\end{aligned}
$$

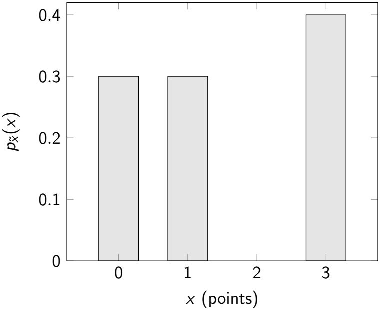
Converting goal difference to points

Probability mass function of $\tilde{x} := h(\tilde{g})$, where

$$h(g) := \begin{cases} 0 & \text{if } g < 0 \\ 1 & \text{if } g = 0 \\ 3 & \text{if } g > 0 \end{cases}$$

Converting goal difference to points





$$p_{\tilde{x}}(0) = 0.3 \qquad p_{\tilde{x}}(1) = 0.3 \qquad p_{\tilde{x}}(3) = 0.4$$

In practice

To model an uncertain quantity with a discrete random variable we only need to estimate the pmf

How to estimate a pmf from data

Observations: 1, 2, 1, 1, 2, 1

What is a reasonable estimate for $p_{\tilde{a}}(1)$?

Empirical pmf

Let $X := \{x_1, x_2, \ldots, x_n\}$ be data with values in discrete set A

The empirical probability mass function of the data is

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

where $1_{x_i=a}$ is one if $x_i = a$ and zero otherwise

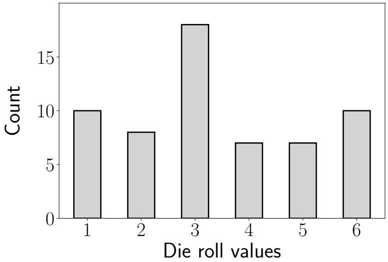
Is the empirical pmf a valid pmf?

Nonnegative? Yes

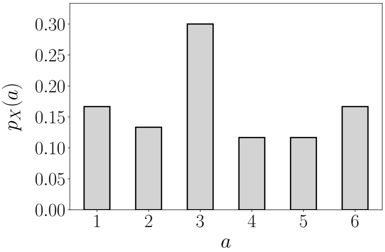
$$\begin{aligned}\sum_{a \in A} p_X(a) &= \sum_{a \in A} \frac{1}{n} \sum_{i=1}^{n} 1_{x_i = a} \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in A} 1_{x_i = a} \\ &= 1\end{aligned}$$

Die rolls

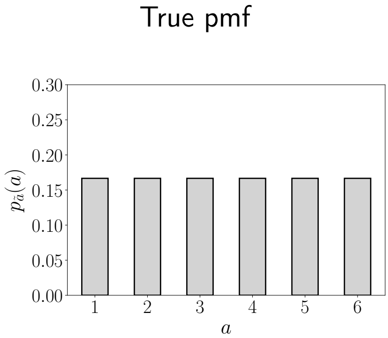
Histogram

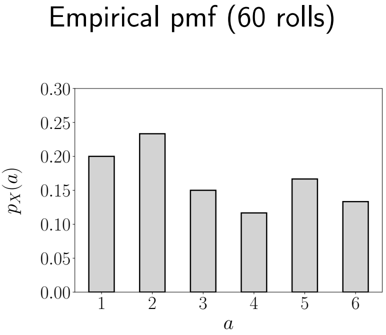


Empirical pmf



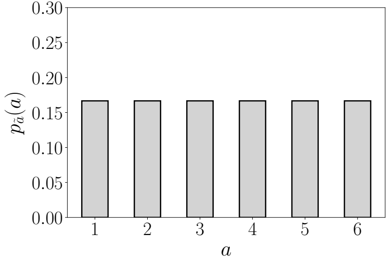
Empirical pmf from simulated fair rolls



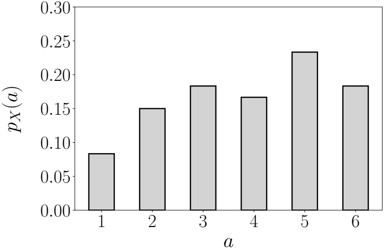


Empirical pmf from simulated fair rolls

True pmf

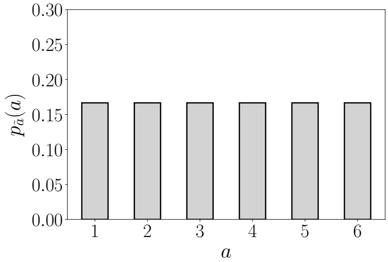


Empirical pmf (60 rolls)

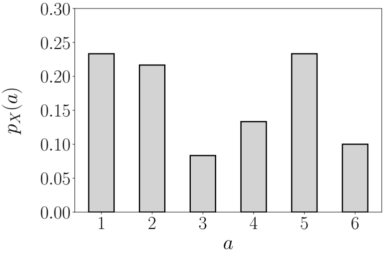


Empirical pmf from simulated fair rolls

True pmf



Empirical pmf (60 rolls)



Free throws

Goal: Model streaks of consecutive free throws

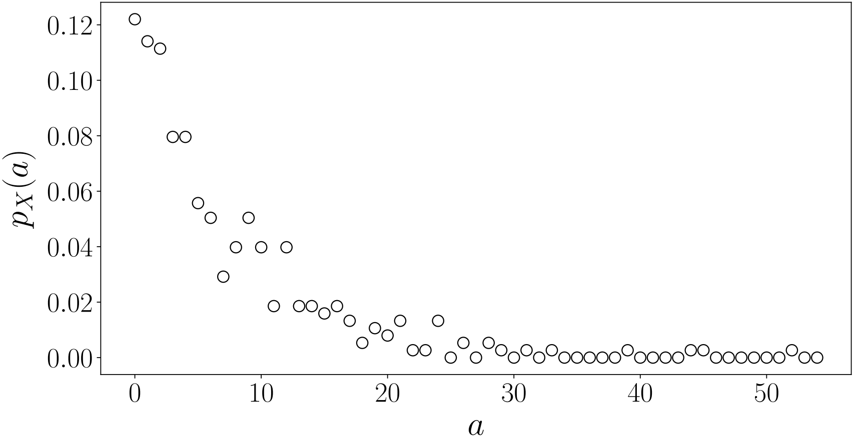
Data: 377 streaks from 3,015 free throws shot by Kevin Durant in the NBA

$X := \{2, 4, 17, 3, 2, \ldots\}$

There are 42 streaks of length 2

$$p_X(2) = \frac{42}{377} = 0.114$$

Empirical pmf



What have we learned?

Properties of the pmf

How to derive the pmf of a function of a random variable

How to estimate the pmf from data