Overview of 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

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?

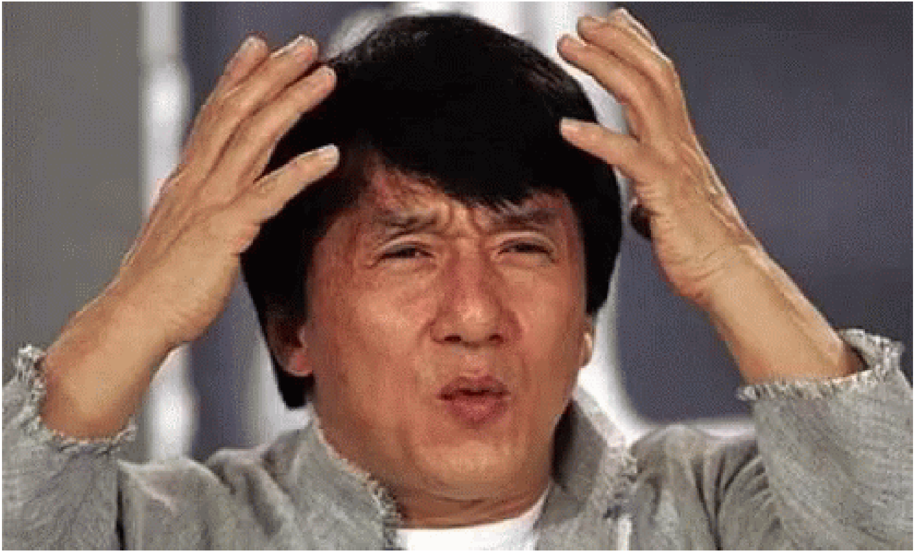
Data scientist:

An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

Me as a student



A random variable is like a car

Car motors are very complicated, but we *don't need to know about them to drive cars!*

We just use the steering wheel

Under the hood random variables are functions in probability spaces

But all we need to use them is their associated probabilities

Plan

- Mathematical definition of random variables
- The probability mass function
- Nonparametric modeling
- Parametric modeling

Rolling a die twice

Probability space representing two rolls of a six-sided die

Outcomes:

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

Quantity of interest: Result of first roll

Key insight: It can be represented as a function of the outcome

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

This is a random variable!

Probability mass function

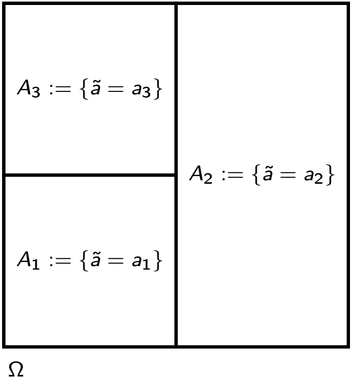
The probability mass function (pmf) $p_{\tilde{a}}$ of $\tilde{a}$ maps each value a to the probability that $\tilde{a} = a$

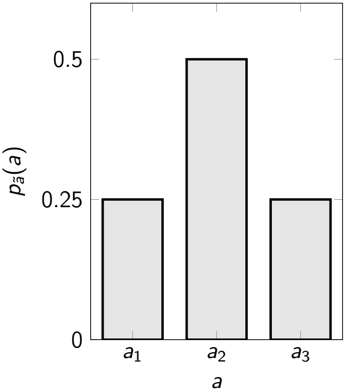
$$p_{\tilde{a}}(a) := \mathrm{P}(\{\omega \mid \tilde{a}(\omega) = a\})$$

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

Wait, *are we sure we can assign probabilities to these events?*

Probability mass function





Formal definition

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

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

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

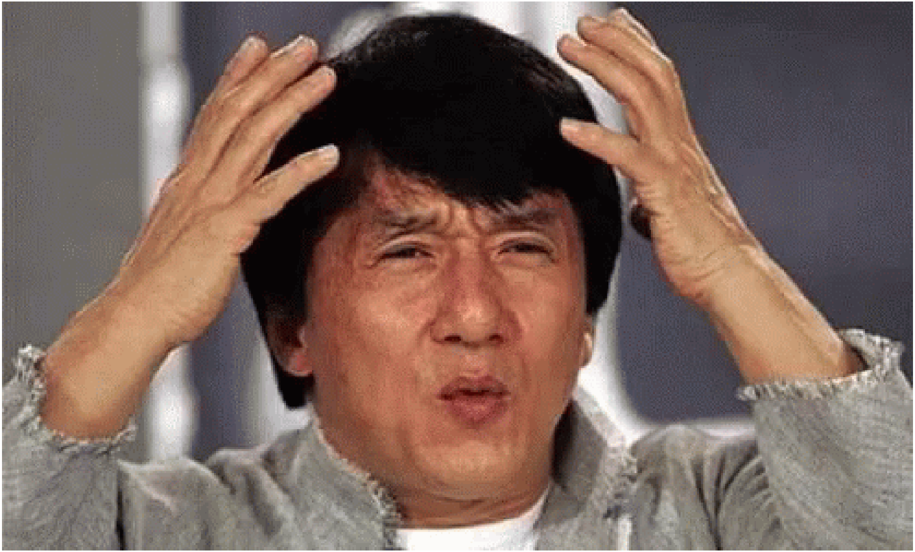
$$A_i := \{\omega \mid \tilde{a}(\omega) = a_i\} \qquad i = 1, 2, \ldots$$

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

Me as a student

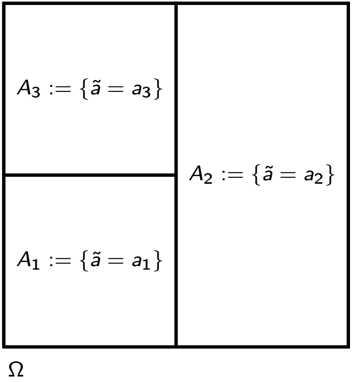


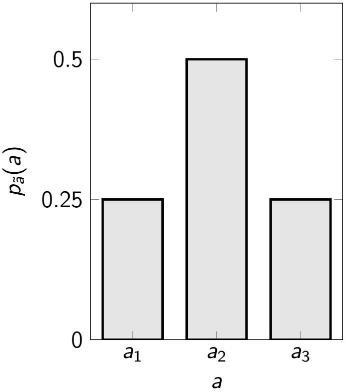
In practice

We never define random variables as functions of outcomes

Instead, we define them through their pmf

Probability mass function





Computing probabilities

Probability that $\tilde{a}$ is in any set S

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

The pmf is all we need, we can forget about the probability space!

In practice

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

Mathematician: *How do we know there's an underlying probability space?*

We can build a probability space (but we never do!)

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

This is a nonparametric estimator of the pmf

Free throws

Goal: Model streaks of consecutive free throws

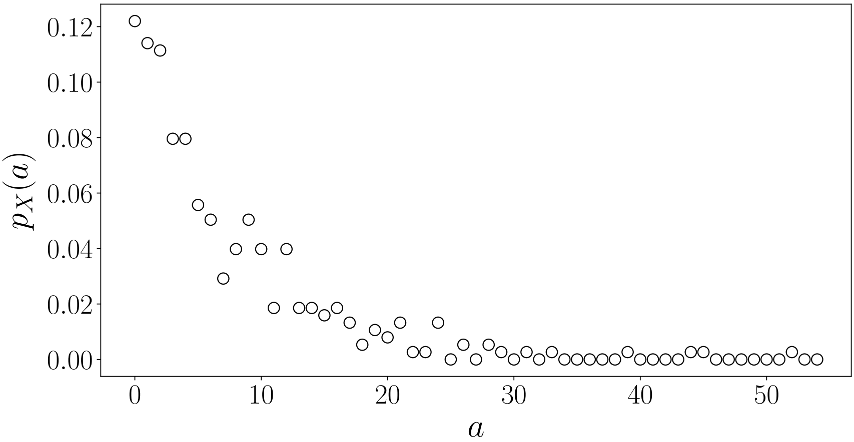
Data: 377 streaks from 3,015 free throws shot by Kevin Durant in the NBA

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

There are 42 streaks of length 2

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

Empirical pmf



Pretty noisy!

Possible solution: Use a parametric model

Discrete parametric distributions

- Bernoulli
- Binomial
- Geometric
- Poisson

Bernoulli distribution

Binary random variable $\tilde{a}$ equal to 1 with probability θ

$$p_{\tilde{a}}(1) = \theta \qquad\qquad p_{\tilde{a}}(0) = 1 - \theta$$

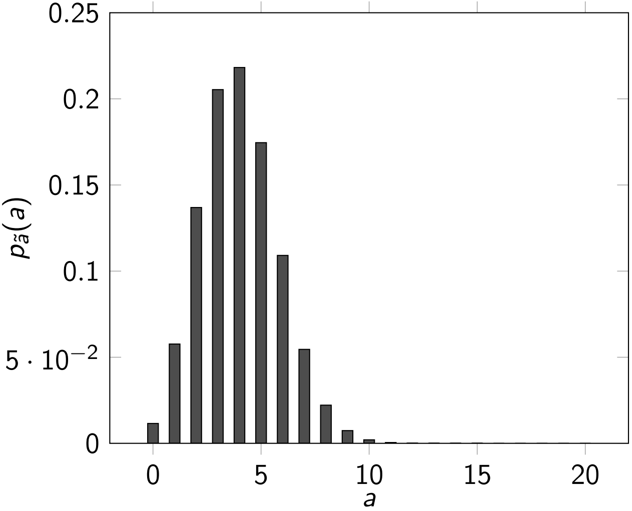
Binomial distribution

Flip n identical coins independently (probability of heads $= \theta$)

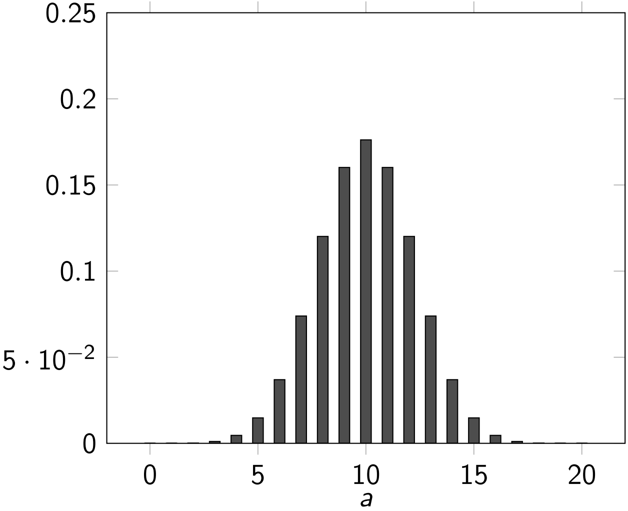
Number of heads is a binomial random variable with parameters n and θ

$$p_{\tilde{a}}(a) = \binom{n}{a} \theta^a (1-\theta)^{(n-a)} \qquad a = 0, 1, \ldots, n$$

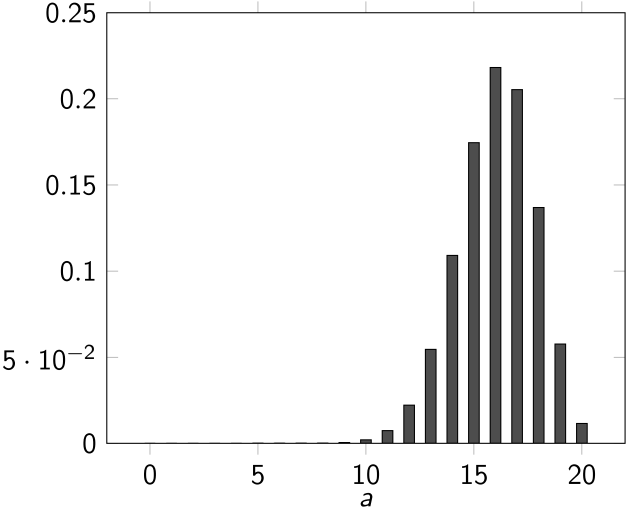
Binomial $n = 20$, $\theta = 0.2$



Binomial $n = 20$, $\theta = 0.5$



Binomial $n = 20$, $\theta = 0.8$



Geometric distribution

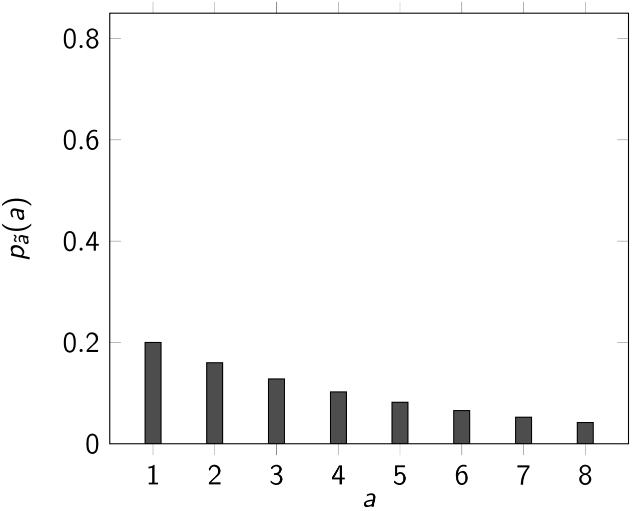
Flip coin independently until it lands on heads (probability of heads = θ)

Number of flips is a geometric random variable with parameter θ

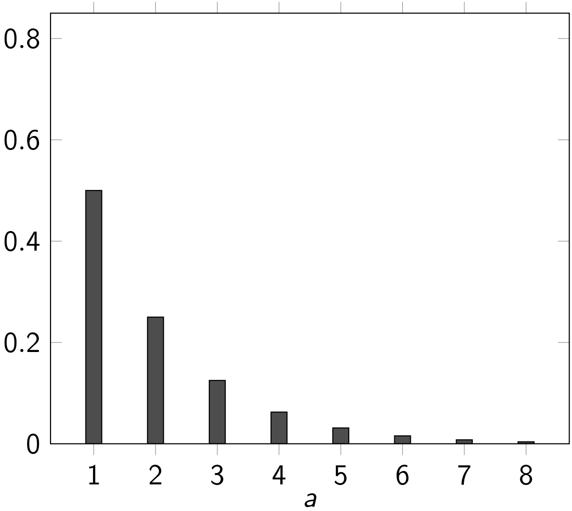
$$p_{\tilde{a}}(a) = (1-\theta)^{a-1}\,\theta \qquad a = 1, 2, \ldots$$

Can be used to model free-throw streaks

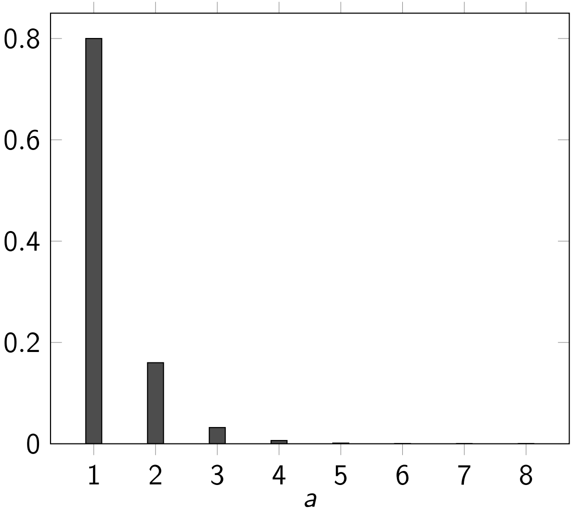
Geometric distribution $\theta = 0.2$



Geometric distribution $\alpha = 0.5$



Geometric distribution $\alpha = 0.8$



Modeling number of earthquakes

Assumptions:

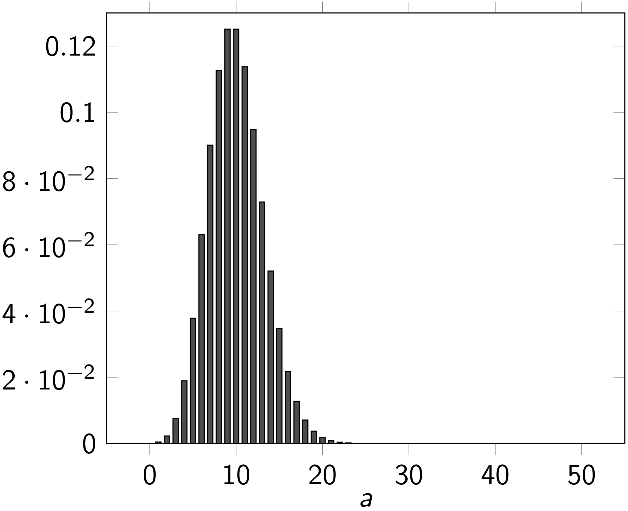
1. Earthquakes are independent
2. Probability of an earthquake in period of small length t is λt

Number of earthquakes is a Poisson random variable with parameter λ

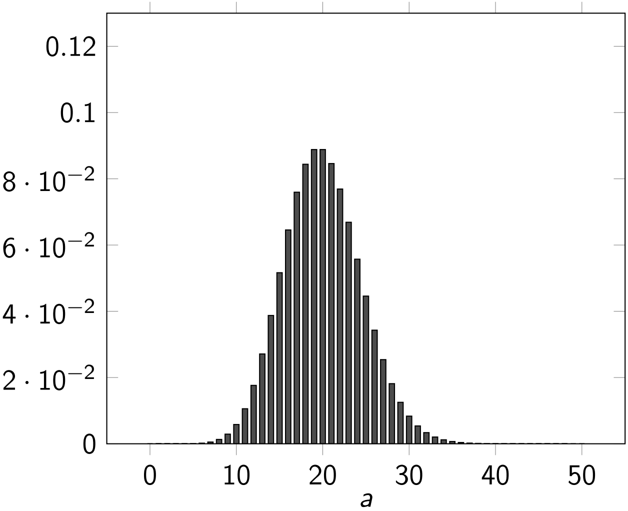
$$p_{\tilde{a}}(a) = \frac{\lambda^a e^{-\lambda}}{a!} \qquad a = 0, 1, 2, \ldots$$

Can be used to model calls, emails, particle decay...

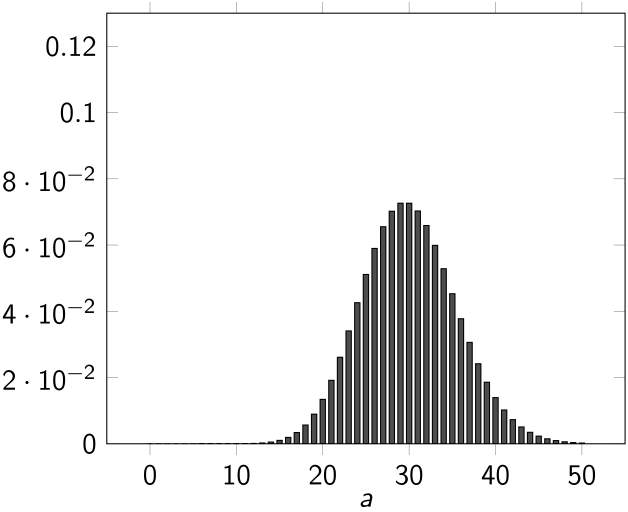
Poisson distribution $\lambda = 10$



Poisson distribution $\lambda = 20$



Poisson distribution $\lambda = 30$



How do we fit a parametric model to data?

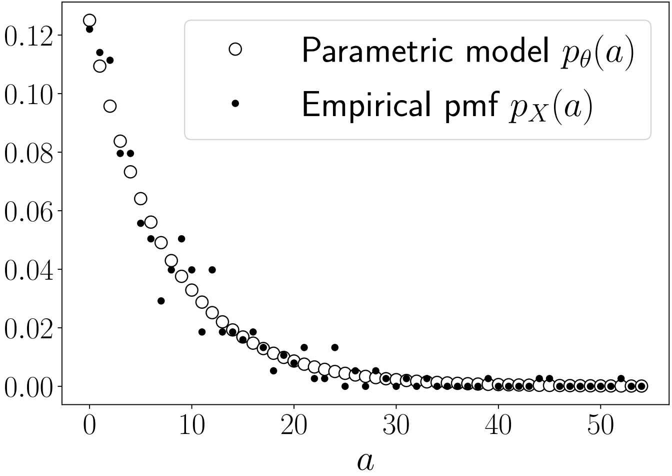
Derive probability of observing the data if the parametric model holds

Interpret probability as a function of the parameters

Choose parameters to make data as likely as possible

This is called *maximum-likelihood* estimation

Geometric model for free-throw streaks



What model is better?

Parametric models

Advantage: Can be fit robustly with very little data

Disadvantage: Require assumptions that are usually wrong

Nonparametric models

Advantage: Very flexible

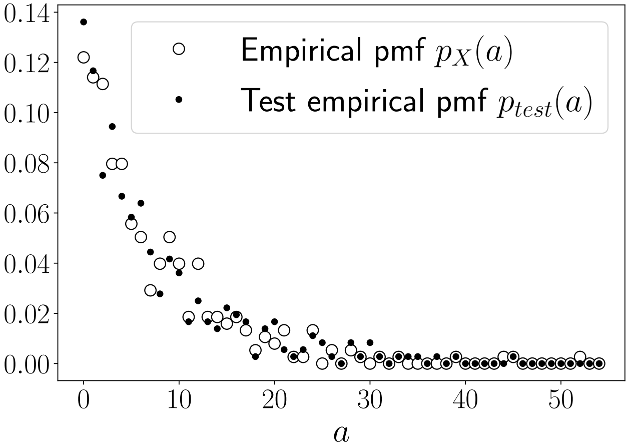
Disadvantage: Noisy unless we have a lot of data

Evaluation?

Use held-out test data!

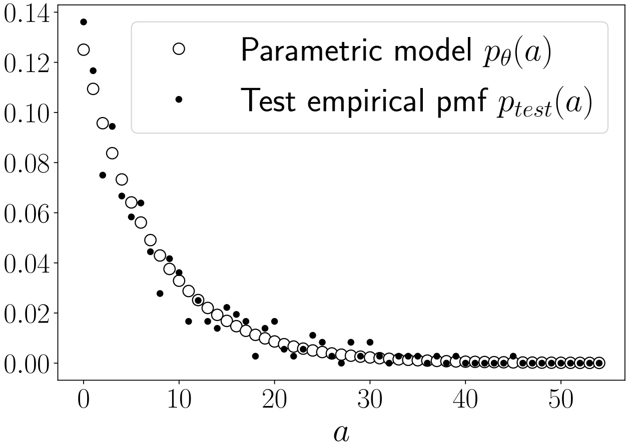
Nonparametric model

Test RMSE $= 7.67\,10^{-3}$



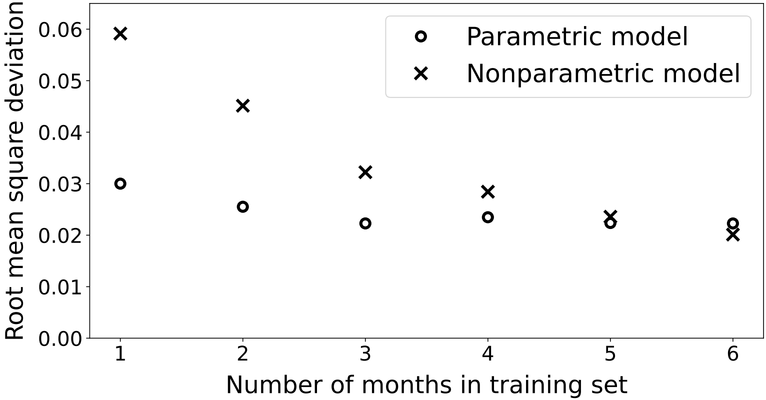
Parametric model

Test RMSE $= 5.61\,10^{-3}$



Nonparametric vs parametric Poisson model

Data: Number of calls arriving at a call center



What have we learned

- Mathematical definition of random variables
- Definition and properties of the probability mass function
- Nonparametric modeling
- Parametric modeling