Overview of Continuous 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

Motivation

Physical quantities such as length, mass, or time are usually interpreted as continuous

Goal: Define continuous random variables to model uncertain continuous quantities

Notation

Deterministic variables: a, b, x, y

Random variables: $\tilde{a}$, $\tilde{b}$, $\tilde{x}$, $\tilde{y}$

What is a random variable?

Data scientist:

An uncertain variable described by probabilities estimated from data

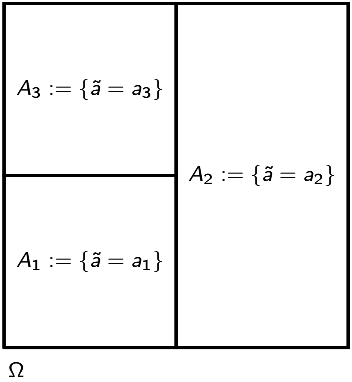
Mathematician:

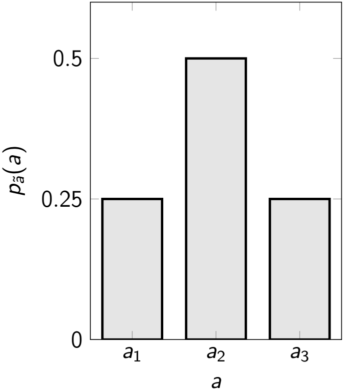
A function mapping outcomes in a probability space to real numbers

Plan

- Mathematical definition of continuous random variables
- The cumulative distribution function
- Probability density
- Nonparametric modeling
- Parametric modeling

Discrete random variables





Key question

Can we describe an uncertain continuous quantity $\tilde{a}$ through probabilities of the form

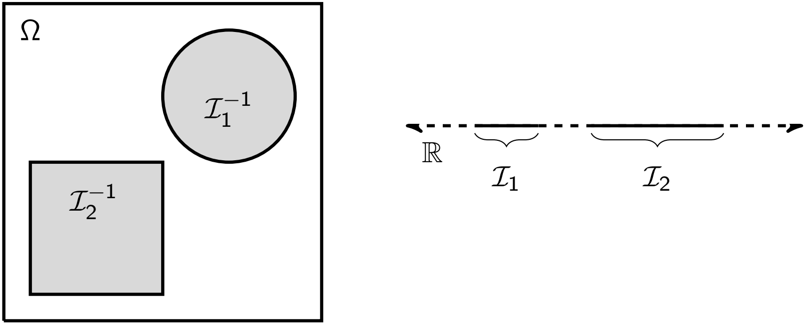
$$\mathrm{P}\left(\tilde{a} = a\right)?$$

No!

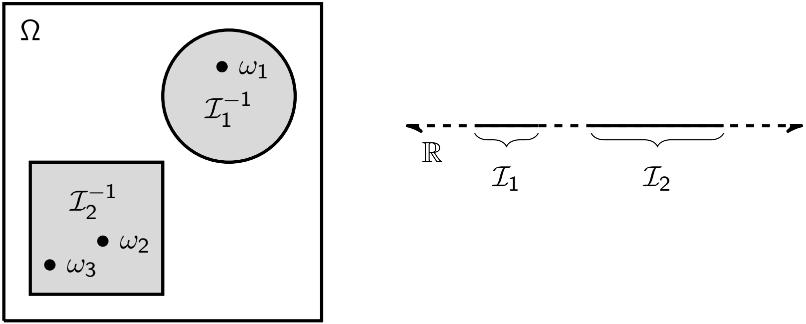
Individual points should have zero probability

Instead, we use the probability that $\tilde{a}$ belongs to different intervals

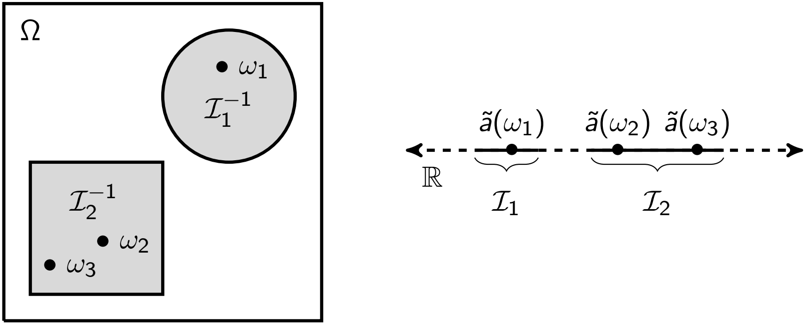
Continuous random variables



Continuous random variables



Continuous random variables



Continuous random variable

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

Function $\tilde{a} : \Omega \to \mathbb{R}$

The function $\tilde{a}$ is a valid random variable if for any interval $\mathcal{I} := [a, b] \subseteq \mathbb{R}$, $a \leq b$

$$\mathcal{I}^{-1} := \{\omega \mid \tilde{a}(\omega) \in \mathcal{I}\}$$

is in the collection $\mathcal{C}$, so

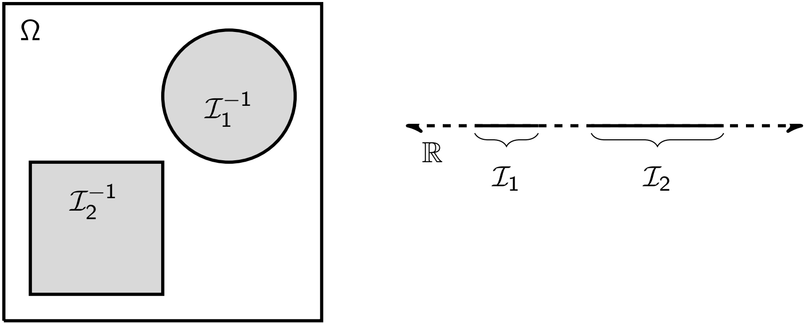
$$\mathrm{P}(\tilde{a} \in \mathcal{I}) = \mathrm{P}(\mathcal{I}^{-1}) \quad \text{is well defined}$$

Continuous random variables

We say that a random variable $\tilde{a}$ is continuous if for any individual real value $a \in \mathbb{R}$

$$\mathrm{P}(\tilde{a} = a) = 0$$

$$\mathrm{P}(\tilde{a} \in \mathcal{I}_1 \cup \mathcal{I}_2) = \mathrm{P}(\tilde{a} \in \mathcal{I}_1) + \mathrm{P}(\tilde{a} \in \mathcal{I}_2)?$$



Unions of intervals

Let $\mathcal{I}_1$, $\mathcal{I}_2$, ..., $\mathcal{I}_n$ be disjoint intervals of $\mathbb{R}$

$$\begin{aligned}
\mathrm{P}\left(\tilde{a} \in \cup_{i=1}^{n} \mathcal{I}_i\right) &= \mathrm{P}\left(\left\{\omega \mid \tilde{a}\left(\omega\right) \in \cup_{i=1}^{n} \mathcal{I}_i\right\}\right) \\
&= \sum_{i=1}^{n} \mathrm{P}\left(\tilde{a} \in \mathcal{I}_i\right)
\end{aligned}$$

Conclusion

We describe continuous random variables in terms of the probability that they belong to any interval

How do we encode this information?

Using the cumulative distribution function or the probability density function

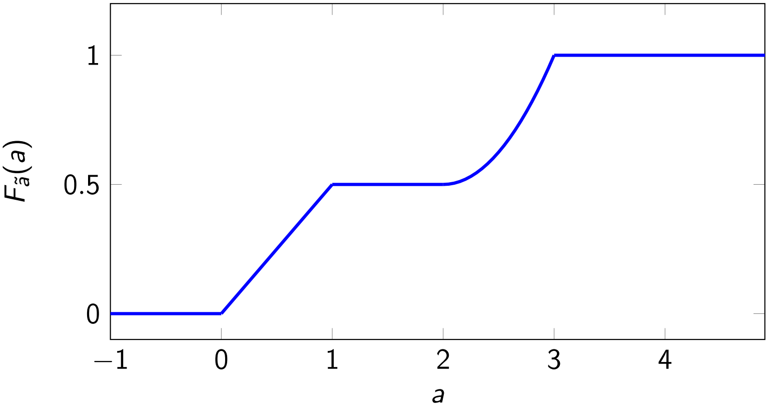
Cumulative distribution function

The cumulative distribution function (cdf) of a random variable $\tilde{a}$ is

$$F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a)$$

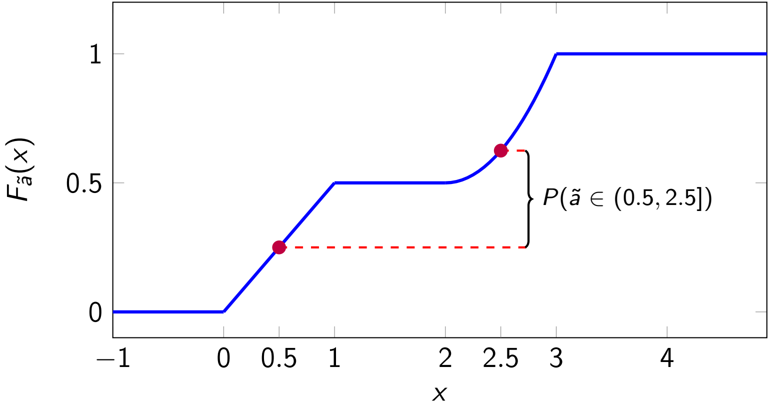
Probability that $\tilde{a}$ is less than or equal to a, for all $a \in \mathbb{R}$

Cumulative distribution function

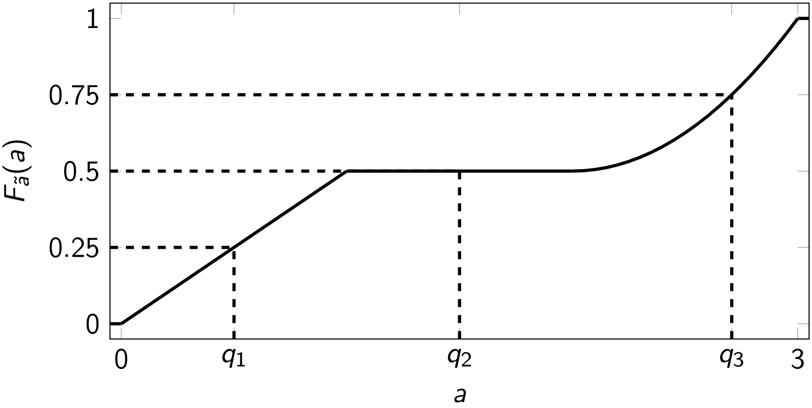


Probability of an interval

$$\mathrm{P}(a < \tilde{a} \leq b) = F_{\tilde{a}}(b) - F_{\tilde{a}}(a)$$



Quantiles



Quantiles

The n-quantiles of $\tilde{a}$ are $n-1$ points q_1, q_2, ..., q_n such that

$$\mathrm{P}(\tilde{a} \leq q_1) = \mathrm{P}(q_1 \leq \tilde{a} \leq q_2) = \cdots = \mathrm{P}(\tilde{a} \geq q_{n-1})$$

or equivalently

$$F_{\tilde{a}}(q_i) = \mathrm{P}(\tilde{a} \leq q_i) = \frac{i}{n} \qquad i = 1, 2, \ldots, n-1$$

4-quantiles are called quartiles: q_1, q_2, q_3

Median

The median q_2 of a continuous random variable $\tilde{a}$ satisfies

$$\mathrm{P}(\tilde{a} \leq q_2) = \mathrm{P}(\tilde{a} > q_2) = \frac{1}{2}$$

or equivalently

$$F_{\tilde{a}}(q_2) = \frac{1}{2}$$

Estimating the cdf from data

For any a, $F_{\tilde{a}}(a) := \mathrm{P}(\tilde{a} \leq a)$ is a probability

Use empirical probability!

Empirical cdf

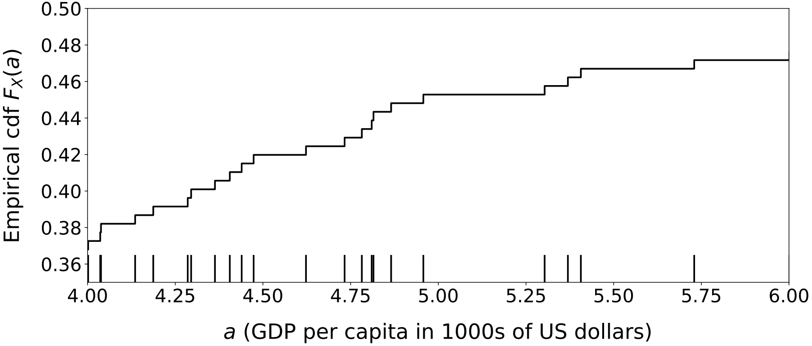
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The empirical cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ equals

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

where $1_{x_i \leq a}$ equals one if $x_i \leq a$ and zero otherwise

Empirical cdf



Quantile estimation

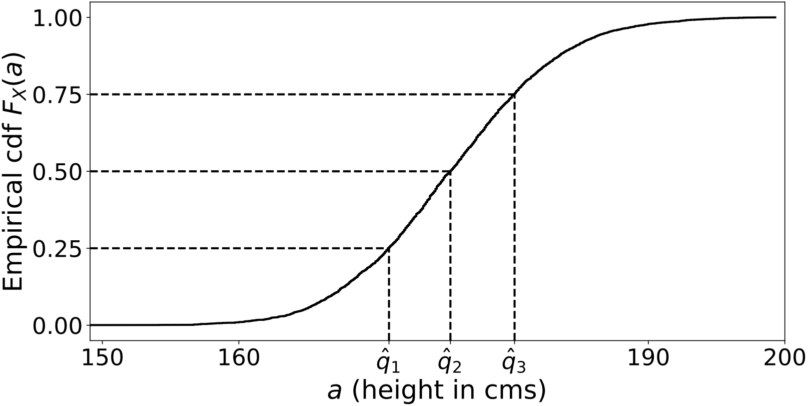
Dataset $X := \{x_1, x_2, \ldots, x_n\}$

The n-quantiles of the data are $n-1$ points $\hat{q}_1$, $\hat{q}_2$, ..., $\hat{q}_n$ such that

$$\mathrm{P}_X(\tilde{a} \leq \hat{q}_1) = \mathrm{P}_X(\hat{q}_1 \leq \tilde{a} \leq \hat{q}_2) = \cdots = \mathrm{P}_X(\tilde{a} \geq \hat{q}_{n-1})$$

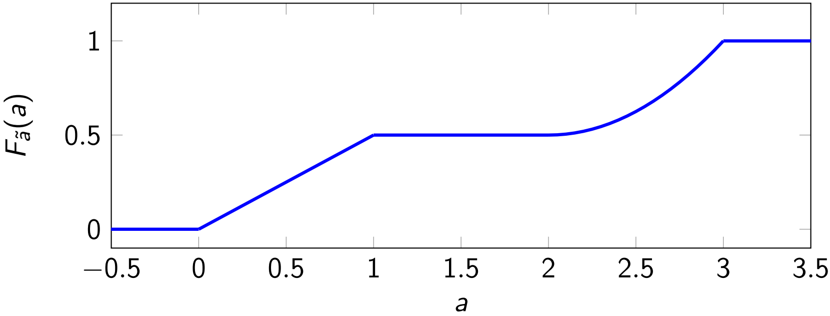
where P_X is the empirical probability of the data

Height in US army



Probability density

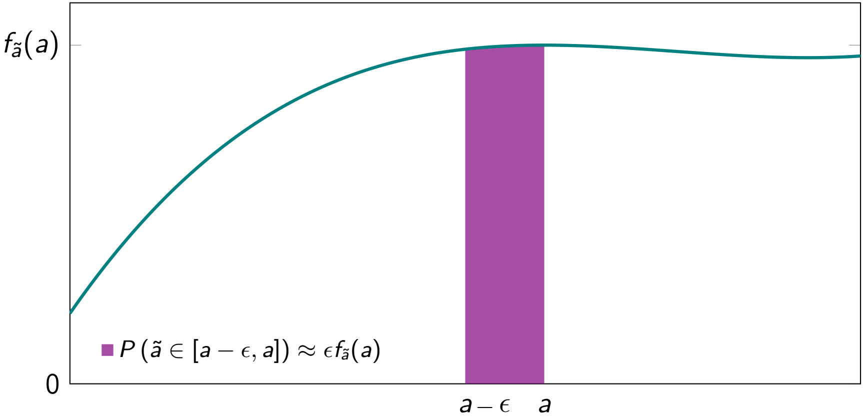
The cdf is a global quantity



How can we characterize local behavior?

Use density!

Probability density



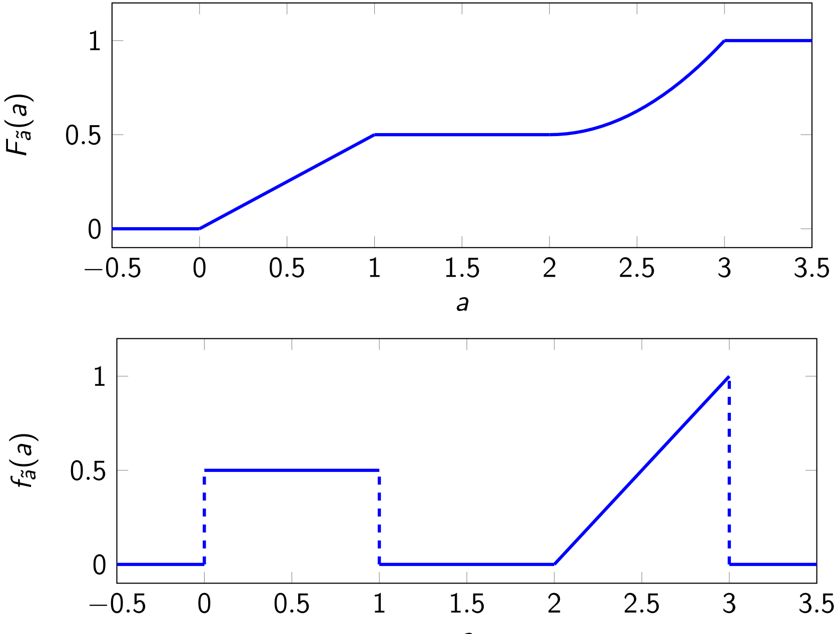
Probability density function

Let $\tilde{a} : \Omega \to \mathbb{R}$ be a random variable with cdf $F_{\tilde{a}}$

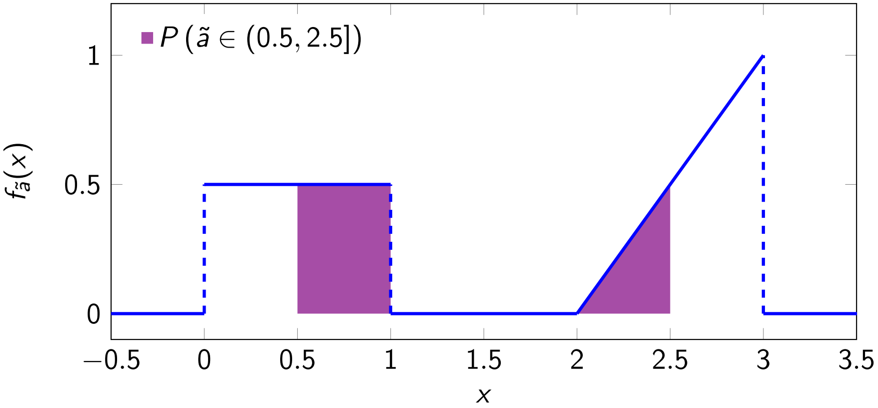
If $F_{\tilde{a}}$ is differentiable, the probability density function (pdf) of $\tilde{a}$ is

$$f_{\tilde{a}}(a) := \frac{\mathrm{d}F_{\tilde{a}}(a)}{\mathrm{d}a}$$

The pdf is the derivative of the cdf



Using pdf to compute probabilities



How to estimate a pdf from data?

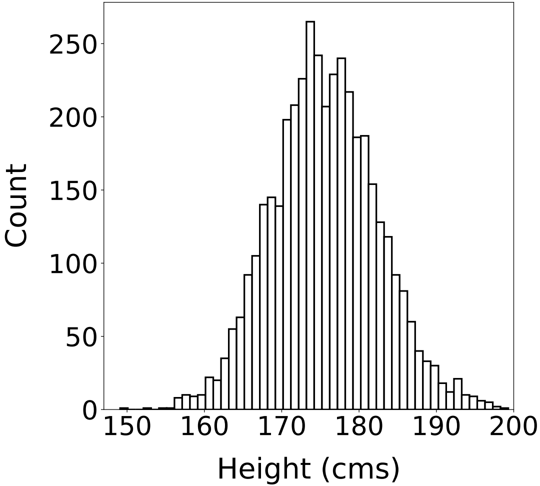
Nonparametric estimate: Normalized histogram / kernel density estimation

Parametric estimate fit using maximum likelihood

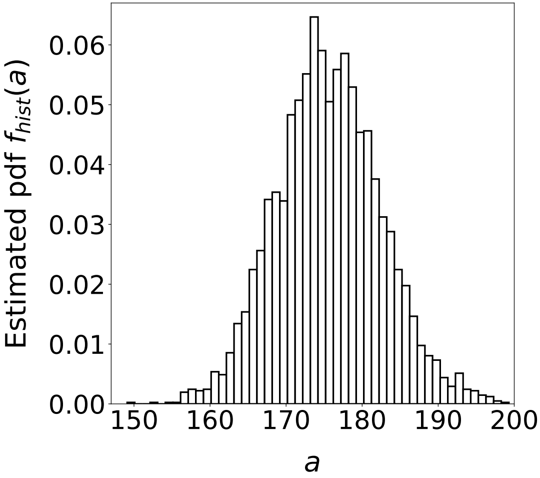
Nonparametric models

Assumption: Density is locally constant / smooth

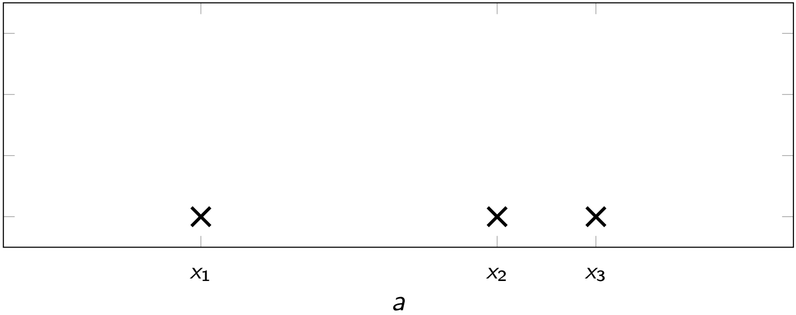
Histogram



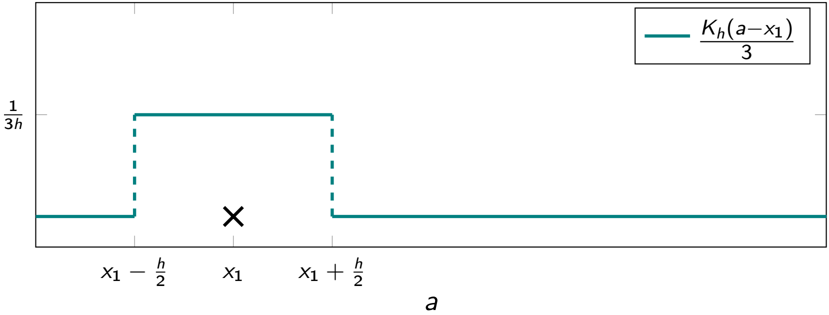
Normalized histogram



Kernel density estimation

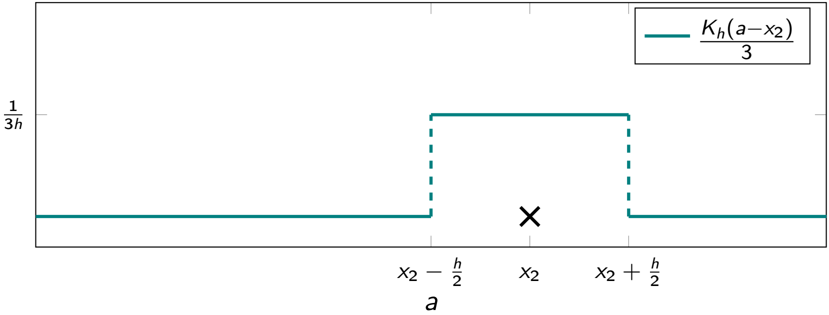


Kernel density estimation



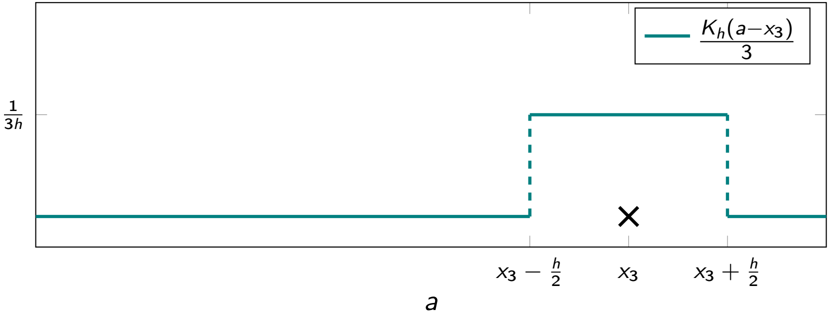
$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right)$$

Kernel density estimation



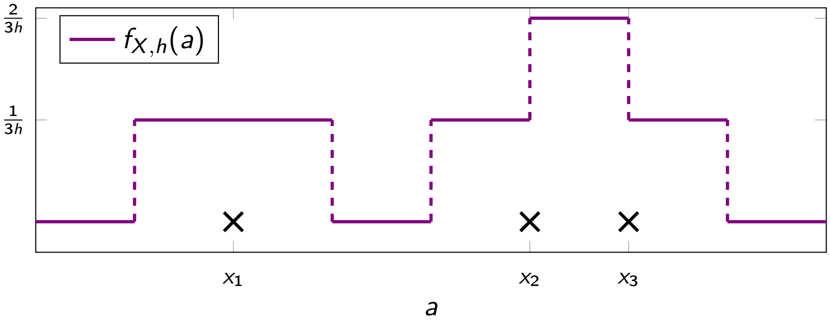
$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right)$$

Kernel density estimation



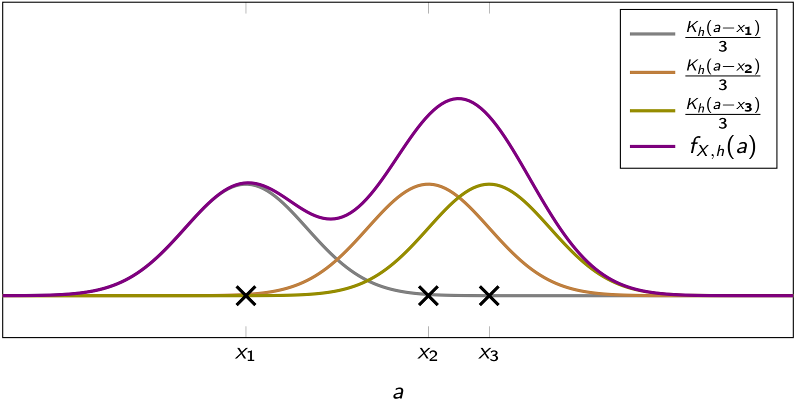
$$f_{X,h}(a) := \frac{1}{3h}K\left(\frac{a-x_1}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_2}{h}\right) + \frac{1}{3h}K\left(\frac{a-x_3}{h}\right)$$

Kernel density estimation



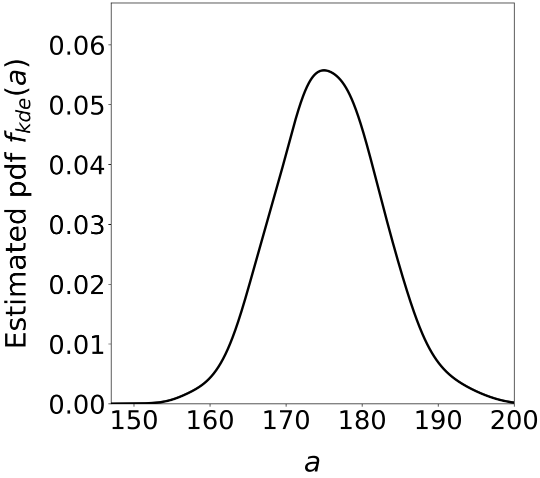
$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_3}{h}\right)$$

Kernel density estimation



$$f_{X,h}(a) := \frac{1}{3h} K\left(\frac{a - x_1}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_2}{h}\right) + \frac{1}{3h} K\left(\frac{a - x_3}{h}\right)$$

Kernel density estimation



Parametric models

Advantage: Can be fit robustly with very little data

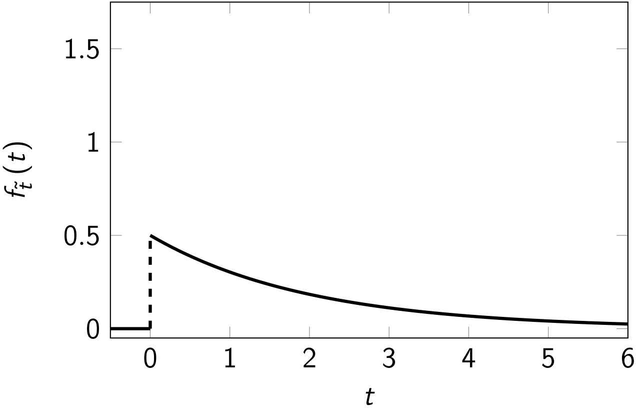
Disadvantage: Require assumptions that are usually wrong

Exponential distribution

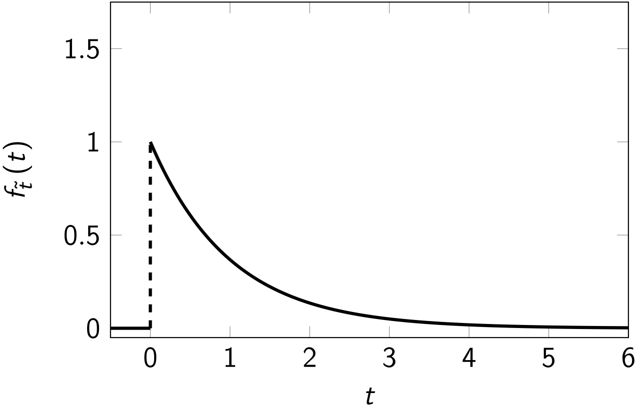
The pdf of an exponential random variable $\tilde{t}$ with parameter λ is

$$f_{\tilde{t}}(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

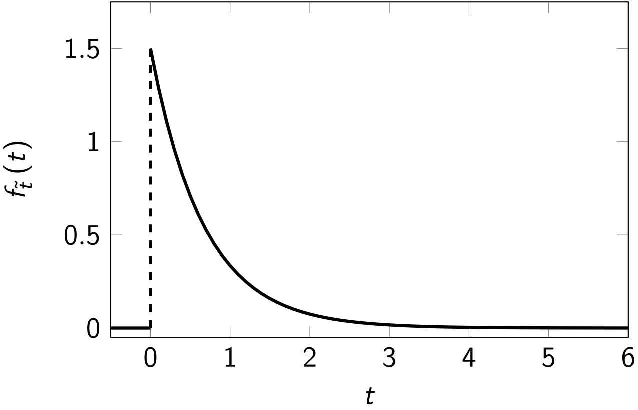
$\lambda = 0.5$



$\lambda = 1$



$\lambda = 1.5$



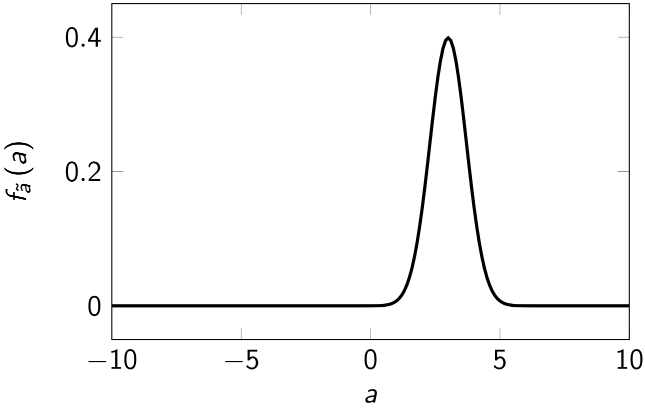
Gaussian distribution

Motivation: Sum of independent quantities is approximately Gaussian

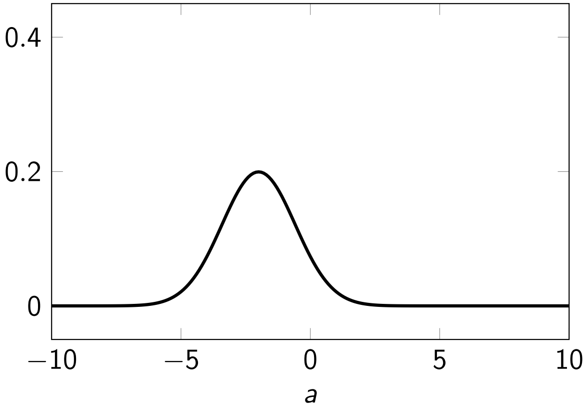
The Gaussian or normal parametric pdf with mean μ and standard deviation σ is

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

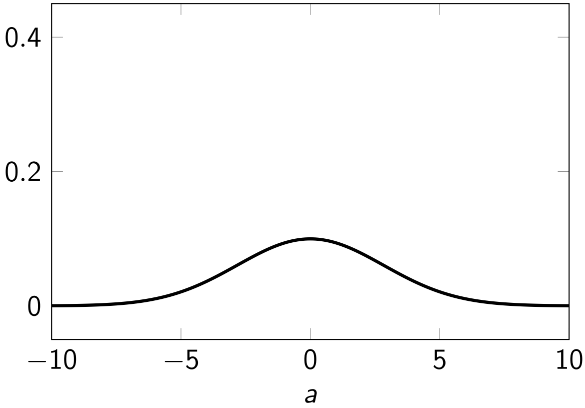
$\mu = 3,\ \sigma = 1$



$\mu = -2$, $\sigma = 2$



$\mu = 0,\ \sigma = 4$

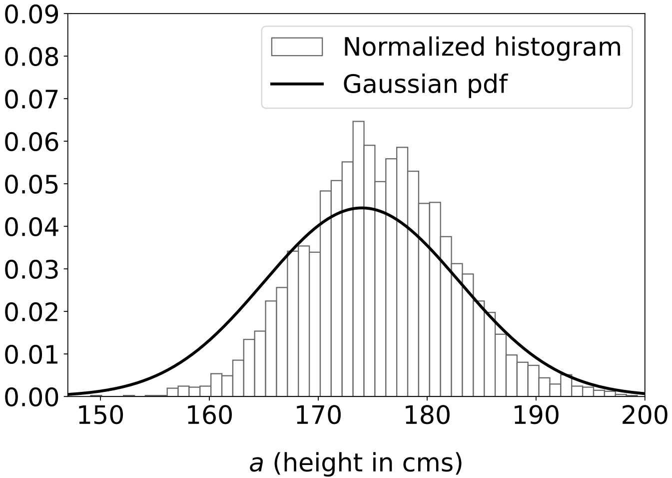


Parametric modeling

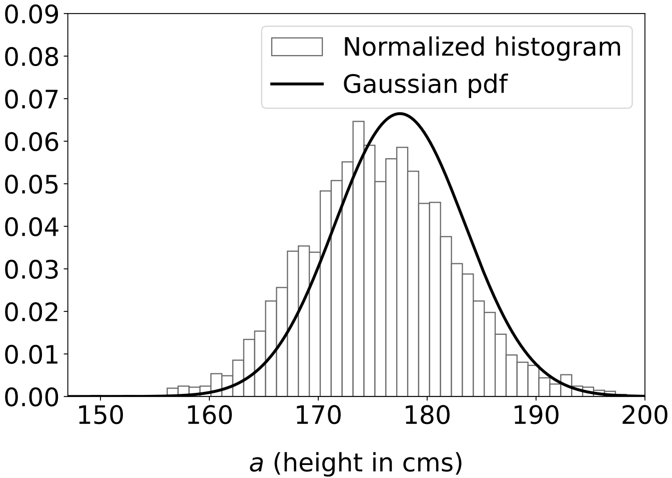
Choose an appropriate parametric model

Estimate parameters from data

$\mu_1 := 174,\ \sigma_1 := 9$



$\mu_2 := 177,\ \sigma_2 := 6$



Maximum likelihood

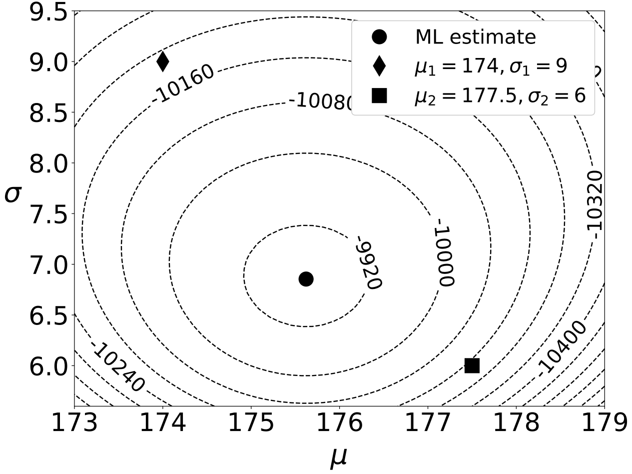
Derive probability density at the data if the parametric model holds

Interpret density as a function of the parameters

Choose parameters to make density as high as possible

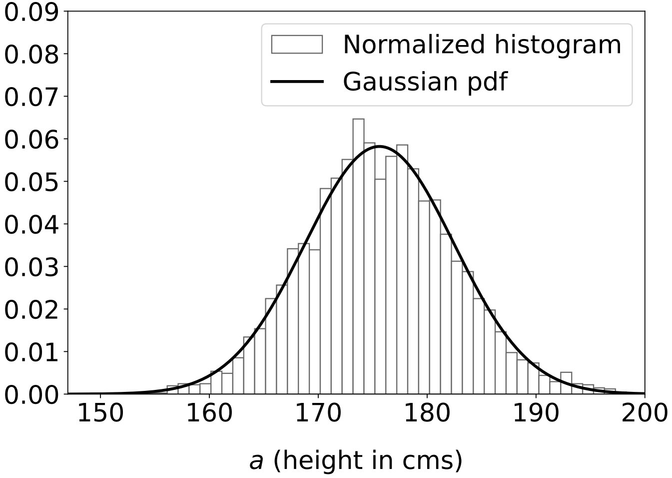
This is called *maximum-likelihood* estimation

Log likelihood (height data)



Maximum-likelihood estimate

$\mu_{\text{ML}} := 177$, $\sigma_2 := 6$



What have we learned?

- Mathematical definition of continuous random variables
- The cumulative distribution function
- Probability density
- Nonparametric modeling
- Parametric modeling