Joint Distribution of Discrete and 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

Goal

Manipulate discrete and continuous quantities in the same probabilistic model

Notation

Deterministic variables: a, b, x, y

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

What is a random variable?

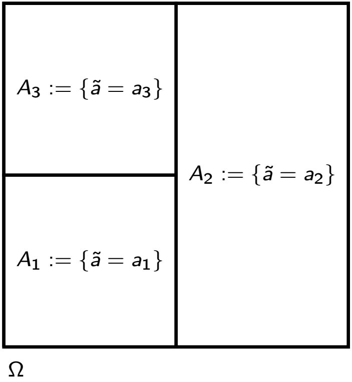
Data scientist:

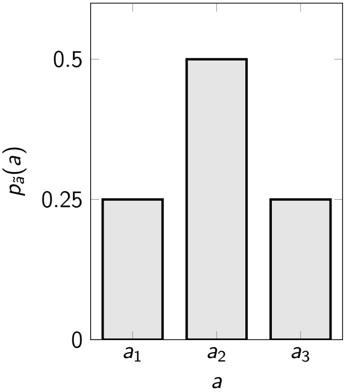
An uncertain variable described by probabilities estimated from data

Mathematician:

A function mapping outcomes in a probability space to real numbers

Discrete random variable



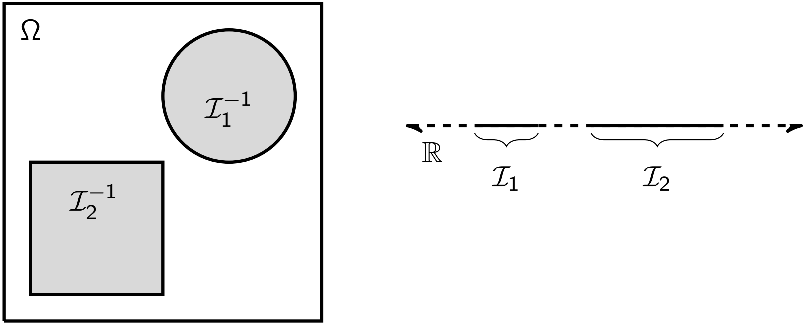


Probability mass function

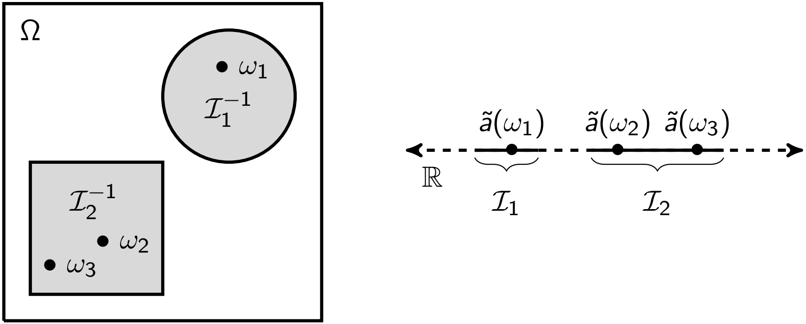
The probability mass function (pmf) of $\tilde{a}$ is the probability that $\tilde{a}$ equals each of its possible values a_1, a_2, ...:

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

Continuous random variables



Continuous random variables



User interface

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 $a \in \mathbb{R}$

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

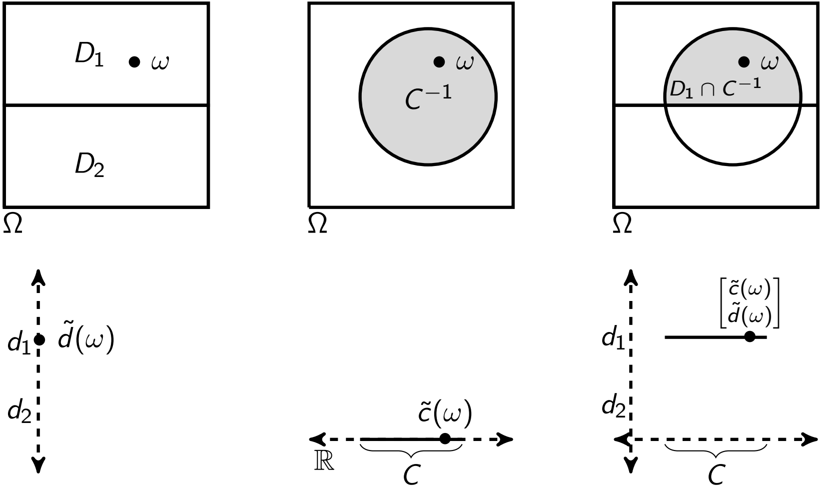
$$f_{\tilde{a}}(a) := \frac{\mathrm{d}F_{\tilde{a}}(a)}{\mathrm{d}a}$$

Discrete and continuous variables

How can we jointly model discrete and continuous quantities?

We represent them as random variables in the same probability space

Discrete and continuous variables



User interface

Joint pmf? ✗

Joint pdf? ✗

Joint cdf? ✓ but ☹

How about marginal pmf and conditional cdf/pdf?

Pmf + conditional cdf / pdf

Discrete random variable $\tilde{d}$ and continuous random variable $\tilde{c}$

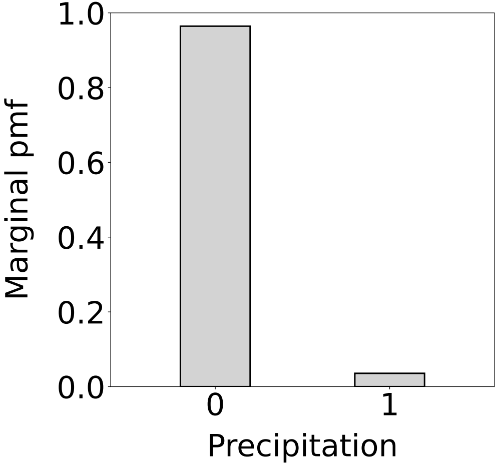
$$
\begin{aligned}
\mathrm{P}\left(\tilde{d} = d, \tilde{c} \leq c\right) &= \mathrm{P}\left(\tilde{d} = d\right) \mathrm{P}\left(\tilde{c} \leq c \,|\, \tilde{d} = d\right) \\
&= p_{\tilde{d}}(d) \, F_{\tilde{c} \,|\, \tilde{d}}(c \,|\, d)
\end{aligned}
$$

$$
\begin{aligned}
f_{\tilde{c} \,|\, \tilde{d}}(c \,|\, d) &:= \lim_{\epsilon \to 0} \frac{\mathrm{P}(c - \epsilon < \tilde{c} \leq c \,|\, \tilde{d} = d)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{F_{\tilde{c} \,|\, \tilde{d}}(c \,|\, d) - F_{\tilde{c} \,|\, \tilde{d}}(c - \epsilon \,|\, d)}{\epsilon} \\
&= \frac{dF_{\tilde{c} \,|\, \tilde{d}}(c \,|\, d)}{dc}
\end{aligned}
$$

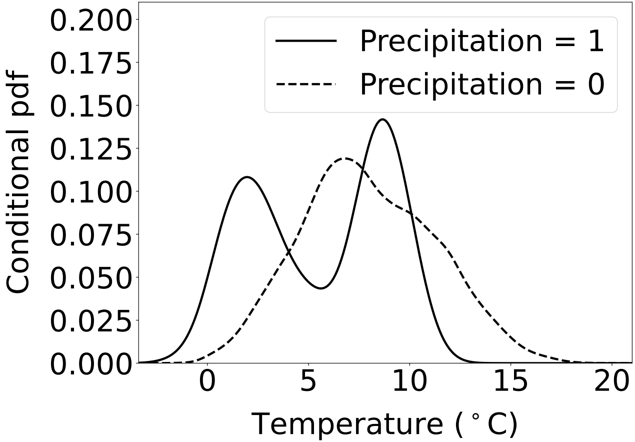
Mauna Loa

Temperature ($\tilde{c}$) and precipitation ($\tilde{d}$)

Marginal pmf of precipitation



Conditional pdf of temperature given precipitation



Marginal distribution of $\tilde{c}$

We know $p_{\tilde{d}}$ and $f_{\tilde{c}\,|\,\tilde{d}}(\cdot\,|\,d)$ for all d

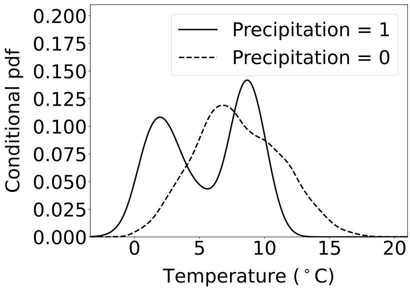
Marginal distribution of $\tilde{c}$?

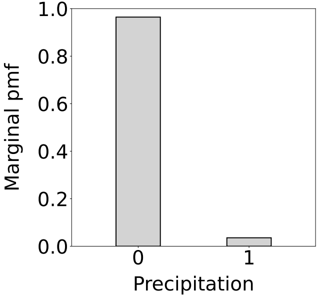
$$\begin{aligned} F_{\tilde{c}}(c) &= \mathrm{P}(\tilde{c} \leq c) \\ &= \sum_{d \in D} \mathrm{P}\left(\tilde{d} = d\right) \mathrm{P}(\tilde{c} \leq c \,|\, d) \\ &= \sum_{d \in D} p_{\tilde{d}}(d)\, F_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d) \end{aligned}$$

$$f_{\tilde{c}}(c) = \sum_{d \in D} p_{\tilde{d}}(d)\, f_{\tilde{c}\,|\,\tilde{d}}(c \,|\, d)$$

Mauna Loa

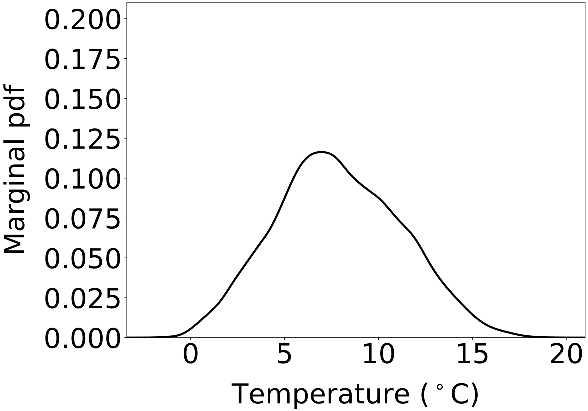
Temperature ($\tilde{c}$) and precipitation (d)





Mauna Loa

$$f_{\tilde{c}}(c) = p_{\tilde{d}}(0)\, f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,0) + p_{\tilde{d}}(1)\, f_{\tilde{c}\,|\,\tilde{d}}(c\,|\,1)$$

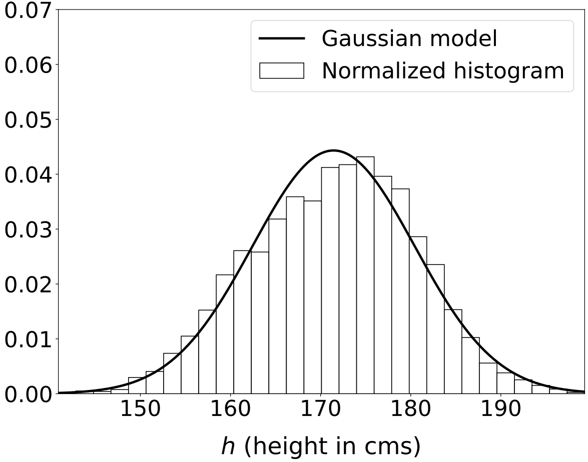


Height data

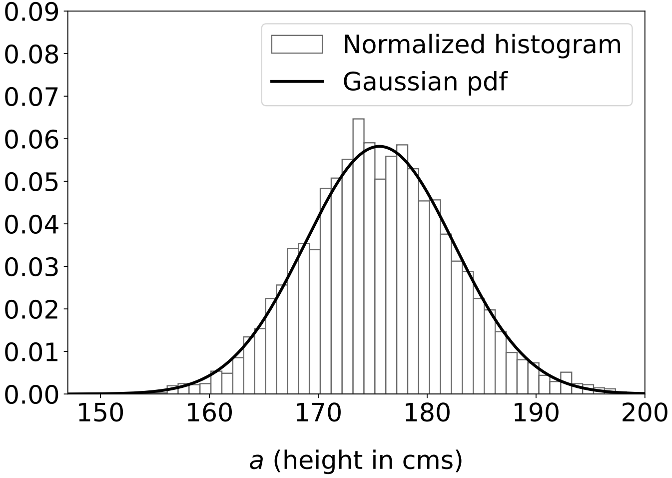
4,082 men and 1,986 women in the United States army

Goal: Design parametric model

Gaussian model



Just the men



Normalized histogram
Gaussian pdf
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
150
160
170
180
190
200
a (height in cms)

Gaussian mixture model

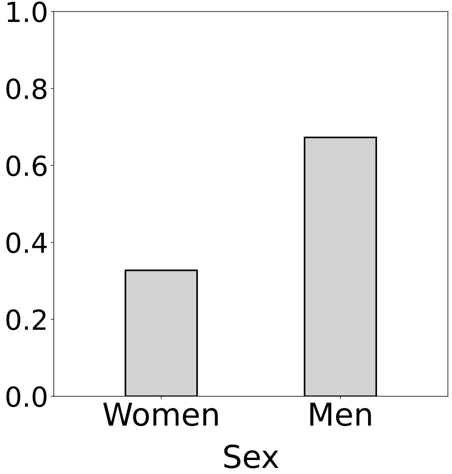
Height: Continuous random variable $\tilde{h}$

Sex: Discrete random variable $\tilde{s}$

Conditional distribution of $\tilde{h}$ given $\tilde{s}$ is Gaussian

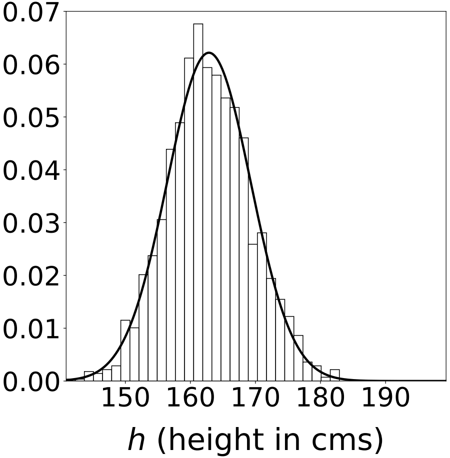
Distribution of $\tilde{s}$?

1,986 women and 4,082 men



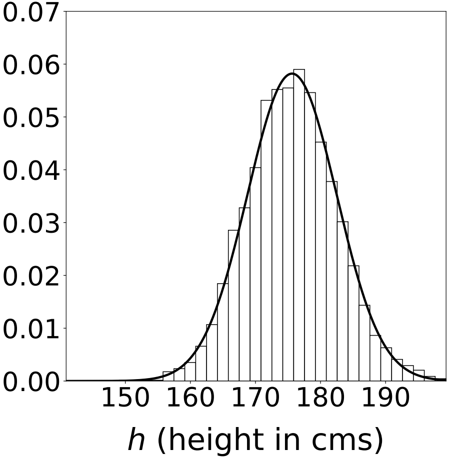
Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = woman?

Gaussian with $\mu_{\text{women}} = 163$ cm and $\sigma_{\text{women}} = 6.4$ cm



Conditional distribution of $\tilde{h}$ given $\tilde{s}$ = man?

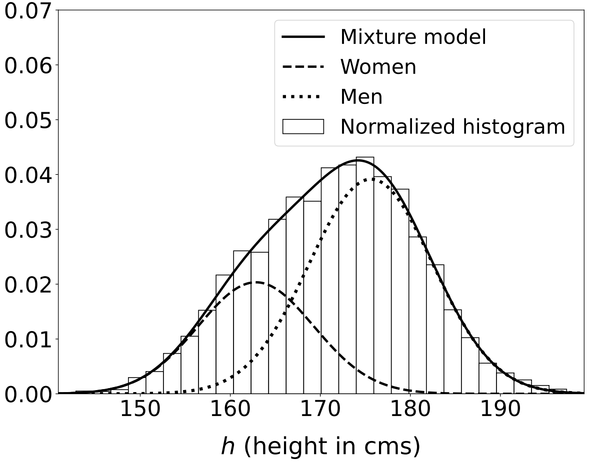
Gaussian with $\mu_{\text{men}} = 176$ cm and $\sigma_{\text{men}} = 6.9$ cm



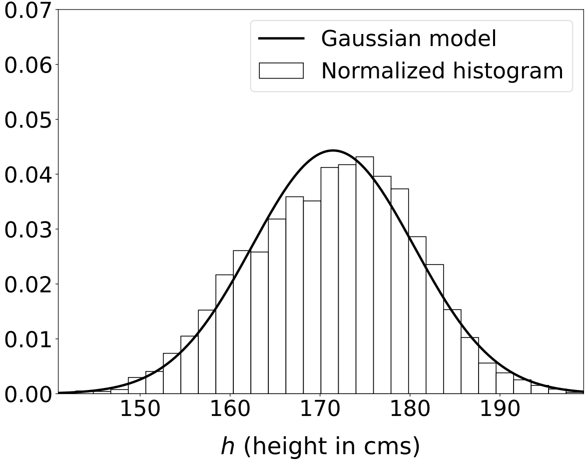
Marginal distribution of $\tilde{h}$?

$$
\begin{aligned}
f_{\tilde{h}}(h) &= \sum_{s=0}^{1} p_{\tilde{s}}(s)\, f_{\tilde{h}\,|\,\tilde{s}}(h\,|\,s) \\
&= \frac{\pi_{\text{women}}}{\sqrt{2\pi}\sigma_{\text{women}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{women}}}{\sigma_{\text{women}}}\right)^2\right) \\
&\quad + \frac{\pi_{\text{men}}}{\sqrt{2\pi}\sigma_{\text{men}}} \exp\left(-\frac{1}{2}\left(\frac{h-\mu_{\text{men}}}{\sigma_{\text{men}}}\right)^2\right)
\end{aligned}
$$

Gaussian mixture model



Gaussian model



What have we learned?

How to jointly model discrete and continuous variables

Gaussian mixture models