The Conditional Mean 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

Goals

Define the conditional mean function

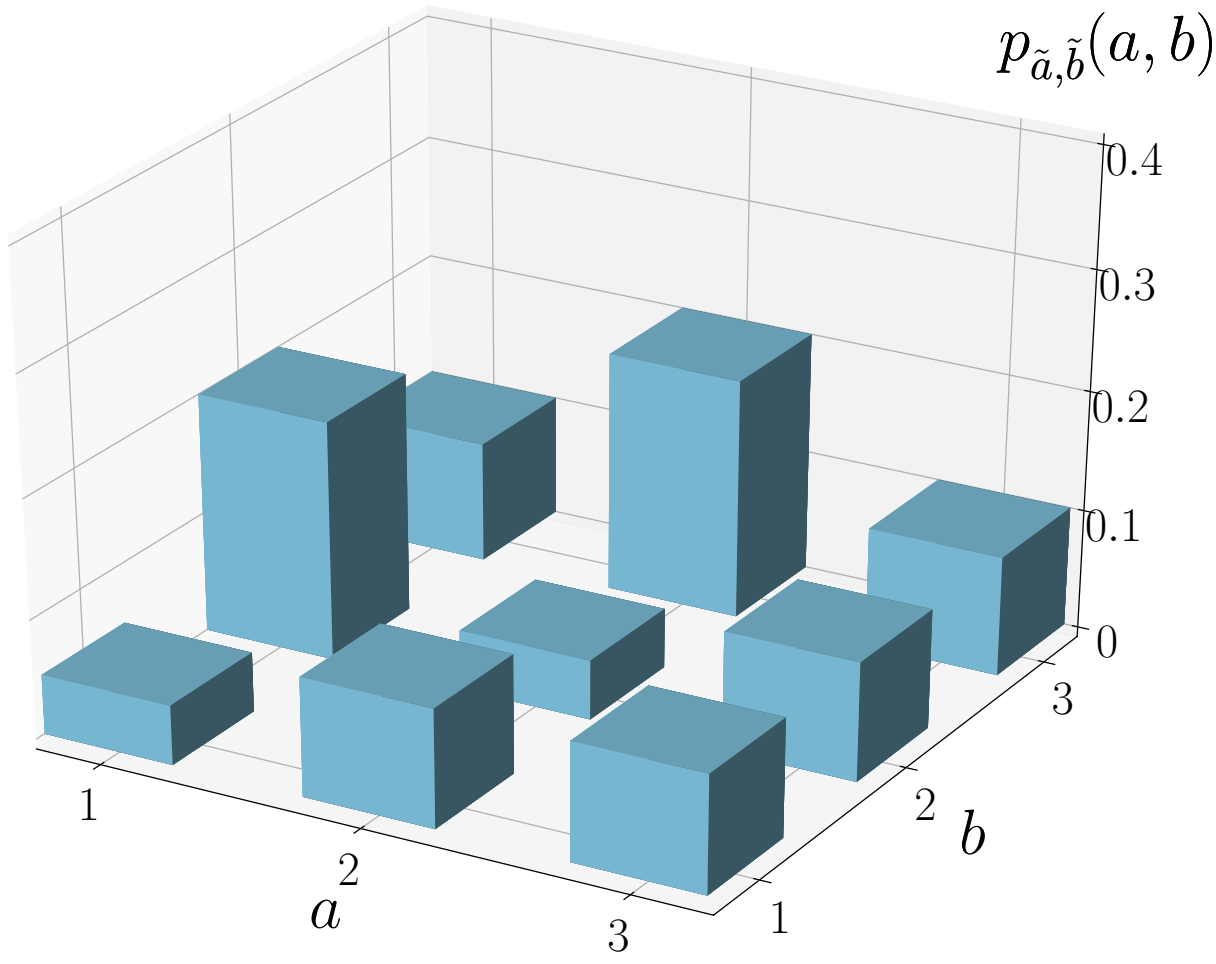
Explain how to estimate it from data

Conditional mean

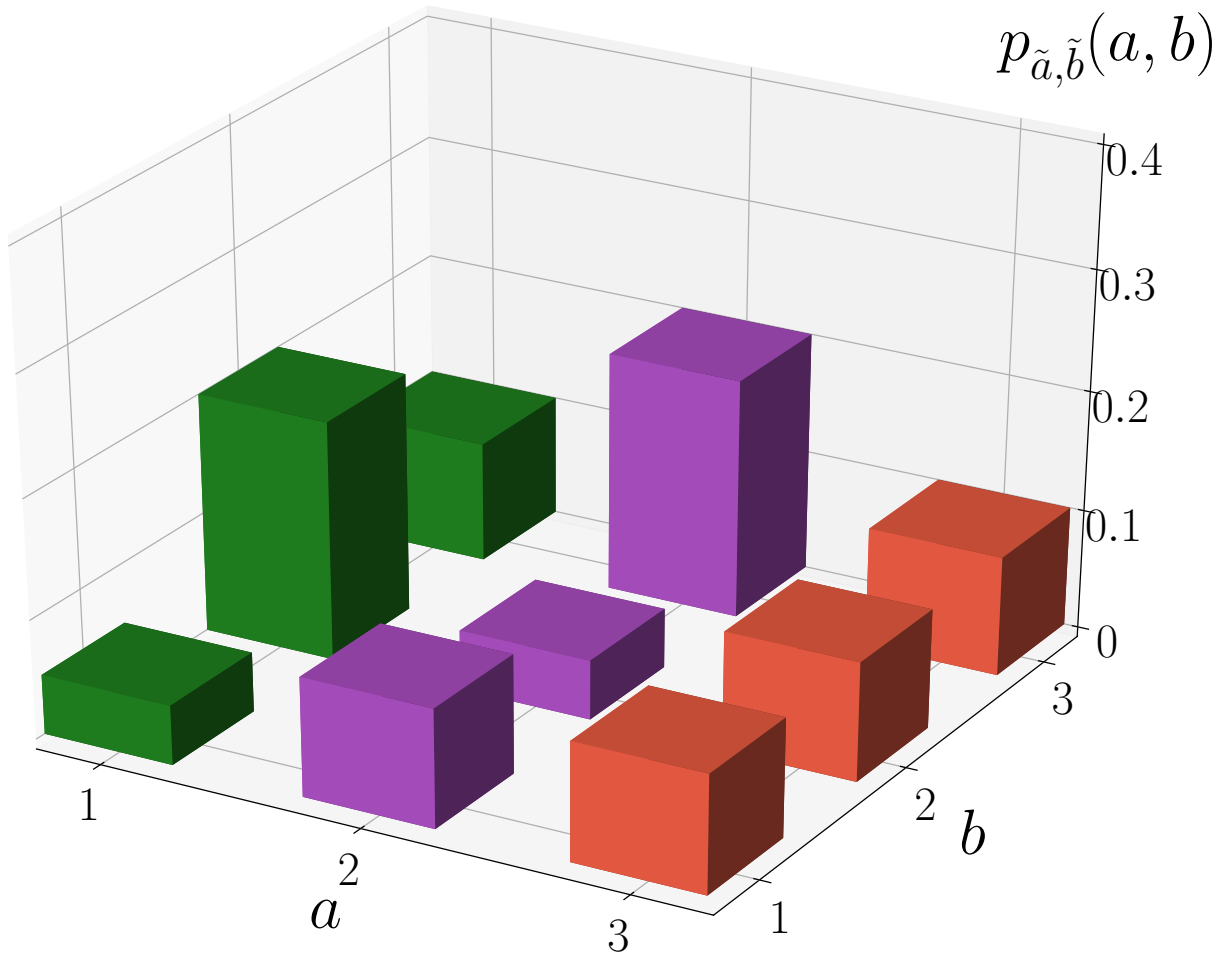
Random variables $\tilde{a}$ and $\tilde{b}$ belong to the same probability space

If $\tilde{a} = a$ what is the mean of $\tilde{b}$?

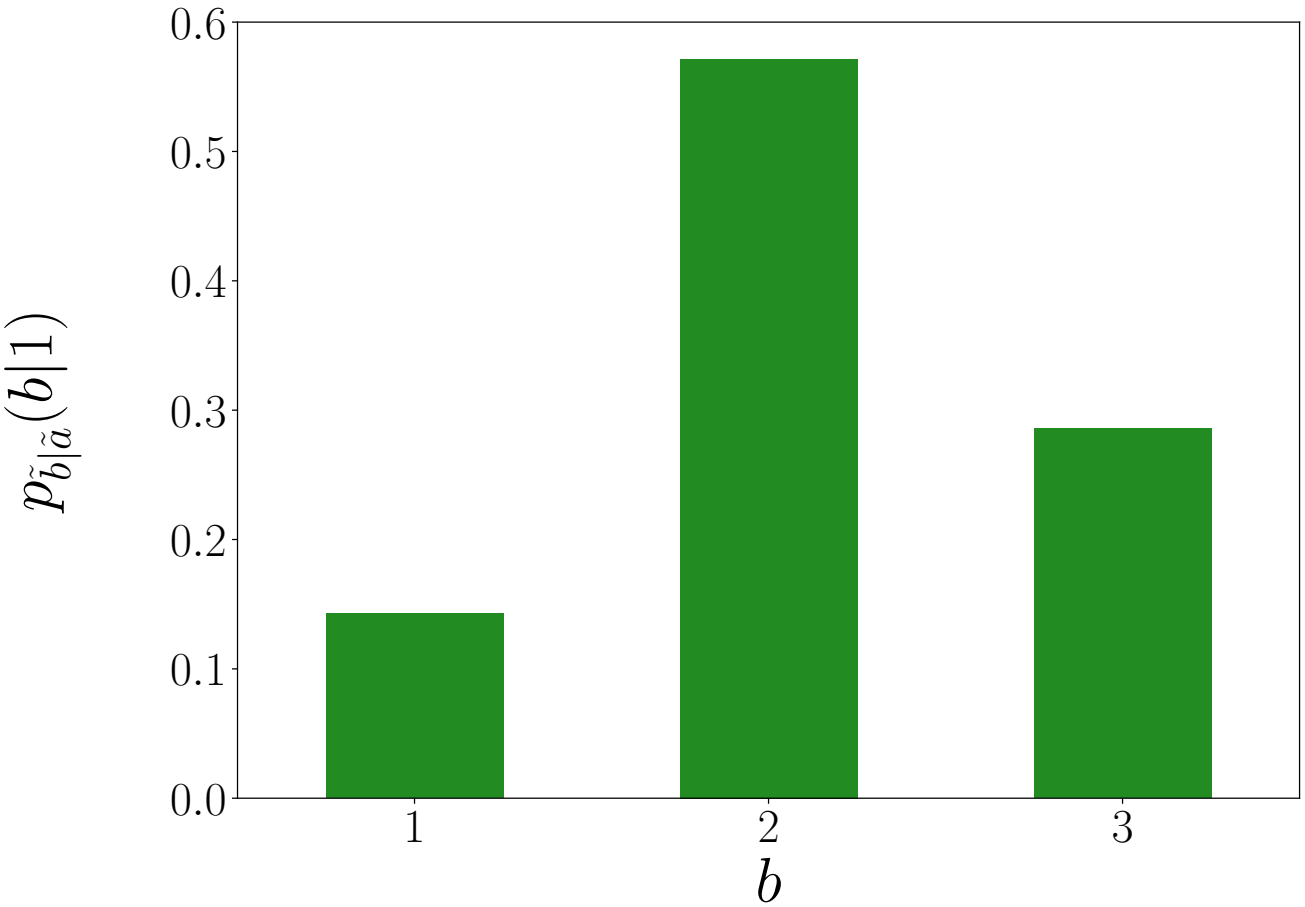
Joint pmf



Mean of $\tilde{b}$ if $\tilde{a}$ is known?



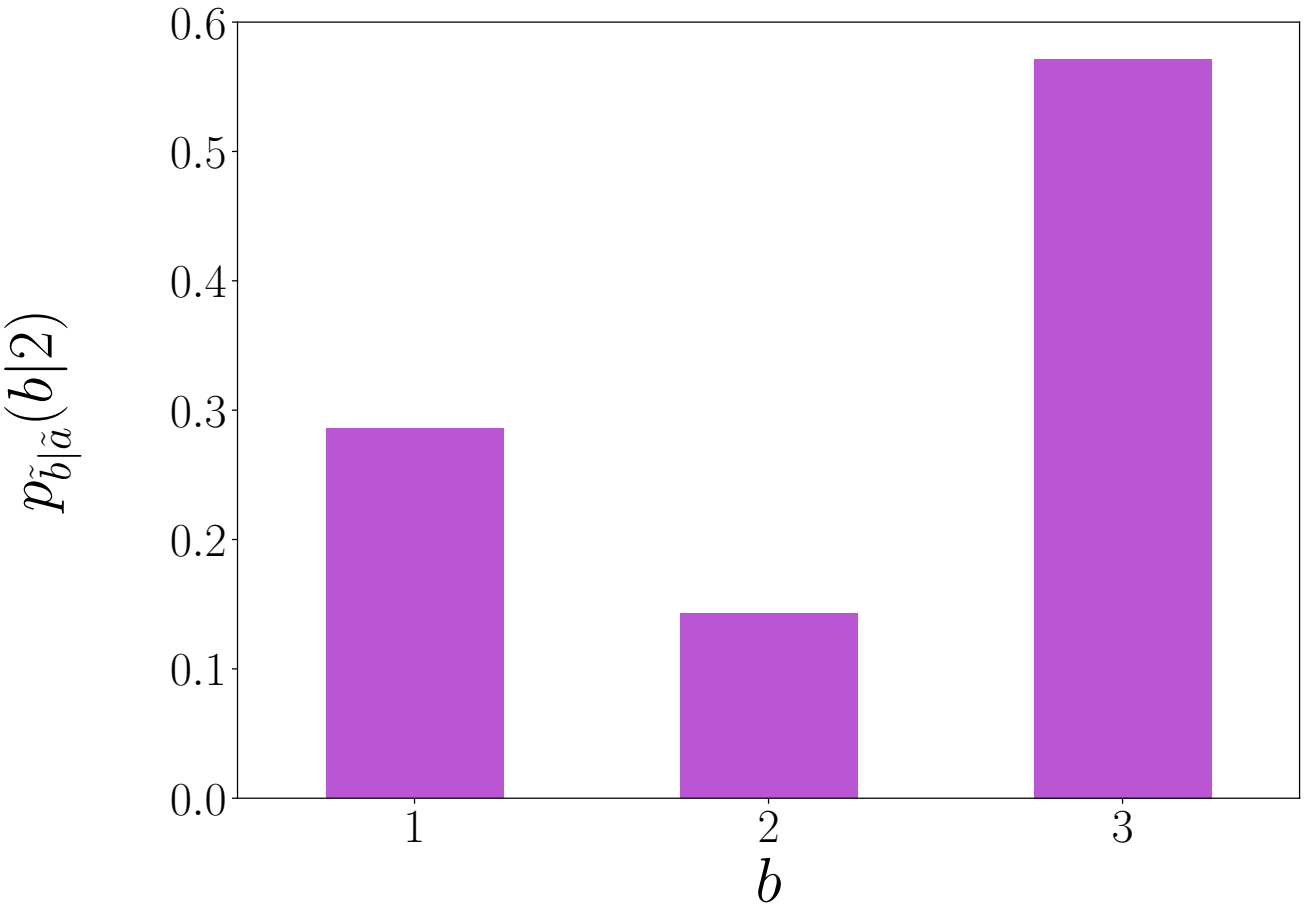
Mean of $\tilde{b}$ if $\tilde{a} = 1$



$$\sum_{b \in B} b p_{\tilde{b} \mid \tilde{a}}(b \mid 1) = 1 \cdot \frac{1}{7} + 2 \cdot \frac{4}{7} + 3 \cdot \frac{2}{7}$$

$$= \frac{15}{7}$$

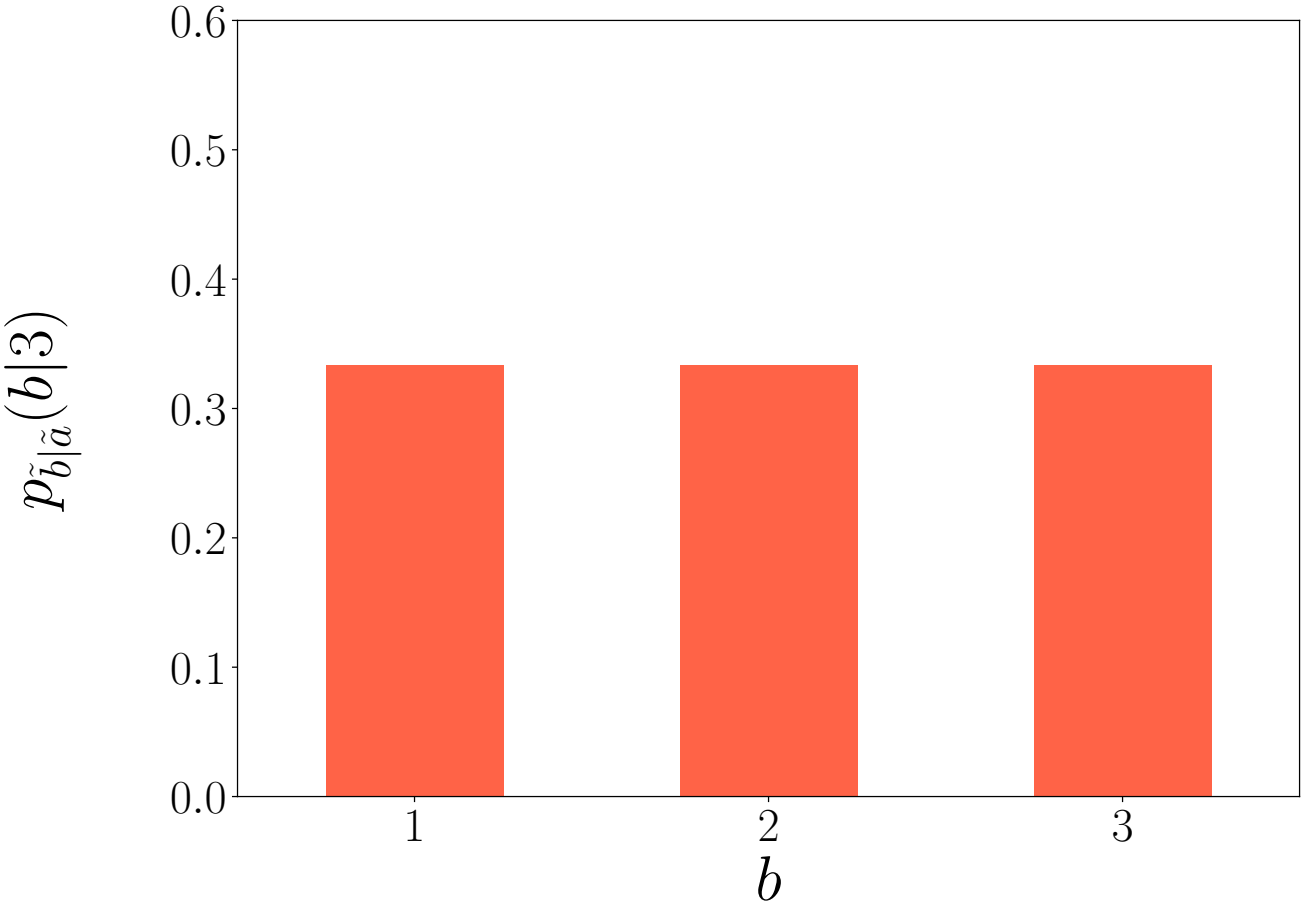
Mean of $\tilde{b}$ if $\tilde{a} = 2$



$$\sum_{b \in B} b p_{\tilde{b} \mid \tilde{a}}(b \mid 2) = 1 \cdot \frac{2}{7} + 2 \cdot \frac{1}{7} + 3 \cdot \frac{4}{7}$$

$$= \frac{16}{7}$$

Mean of $\tilde{b}$ if $\tilde{a} = 3$



$$\sum_{b \in B} b p_{\tilde{b} \,|\, \tilde{a}}(b \,|\, 3) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3}$$

$$= 2$$

Conditional mean function

The conditional mean function of a discrete random variable $\tilde{b}$ given $\tilde{a}$ is

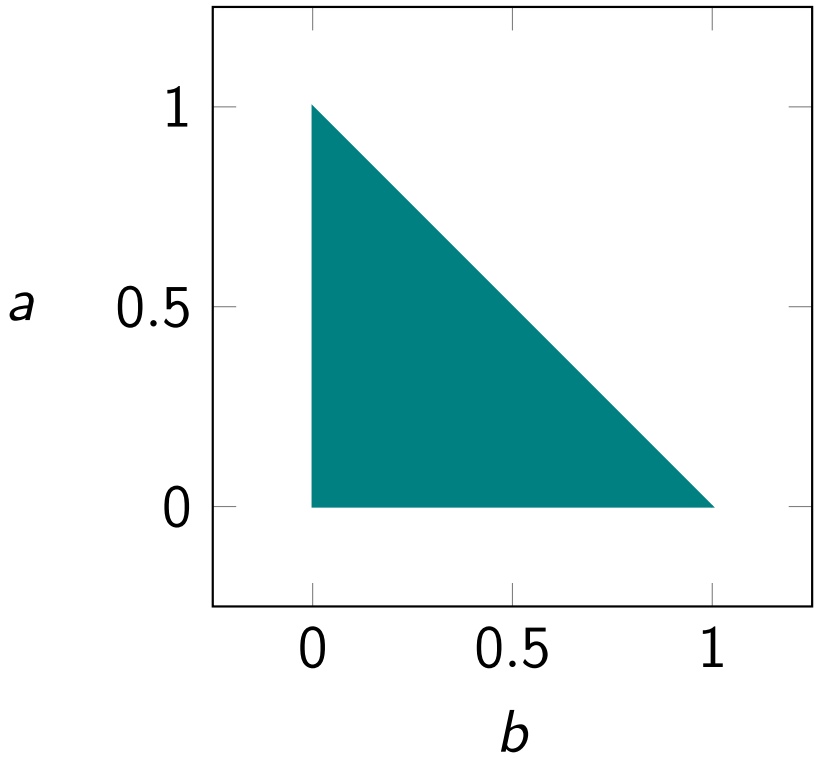
$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) := \sum_{b \in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a)$$

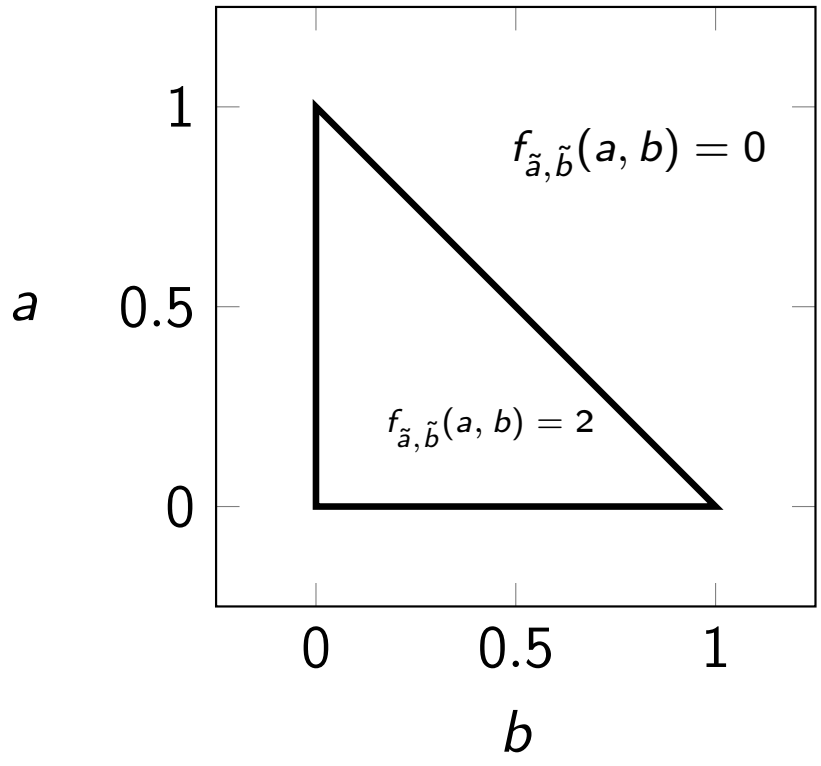
$$\mu_{\tilde{b}\,|\,\tilde{a}}(1) = \sum_{b \in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,1) = \frac{15}{7}$$

$$\mu_{\tilde{b}\,|\,\tilde{a}}(2) = \sum_{b \in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,2) = \frac{16}{7}$$

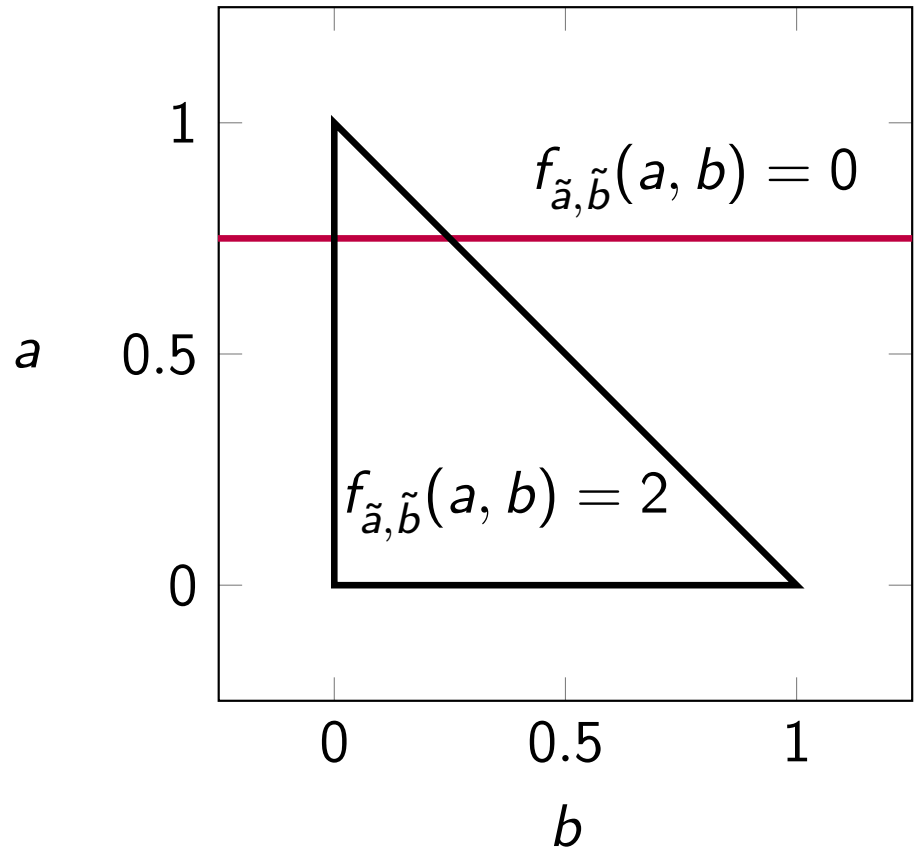
$$\mu_{\tilde{b}\,|\,\tilde{a}}(3) = \sum_{b \in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = 2$$

Triangle lake



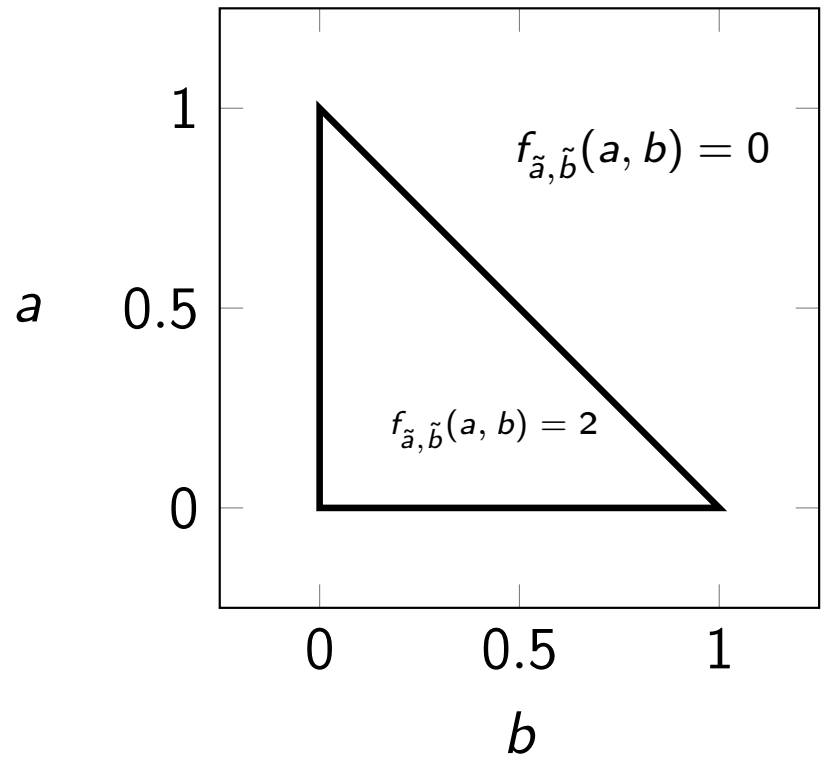


Mean of $\tilde{b}$ if $\tilde{a} = a$?



$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = \frac{f_{\tilde{a},\tilde{b}}(a,b)}{f_{\tilde{a}}(a)}$$

Marginal pdf



$$f_{\tilde{a}}(a) = \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\,\mathrm{d}b$$
$$= \int_{b=0}^{1-a} 2\,\mathrm{d}b = 2(1-a)$$

Mean of $\tilde{b}$ if $\tilde{a} = a$?

$$f_{\tilde{b} \mid \tilde{a}}(b \mid a) = \frac{f_{\tilde{a},\tilde{b}}(a, b)}{f_{\tilde{a}}(a)}$$
$$= \frac{2}{2(1-a)} = \frac{1}{1-a} \qquad b \in [0, 1-a]$$

$$\int_{b=-\infty}^{\infty} b f_{\tilde{b} \mid \tilde{a}}(b \mid a)\, \mathrm{d}b = \int_{b=0}^{1-a} \frac{b}{1-a}\, \mathrm{d}b$$
$$= \frac{(1-a)^2}{2(1-a)}$$
$$= \frac{1-a}{2}$$

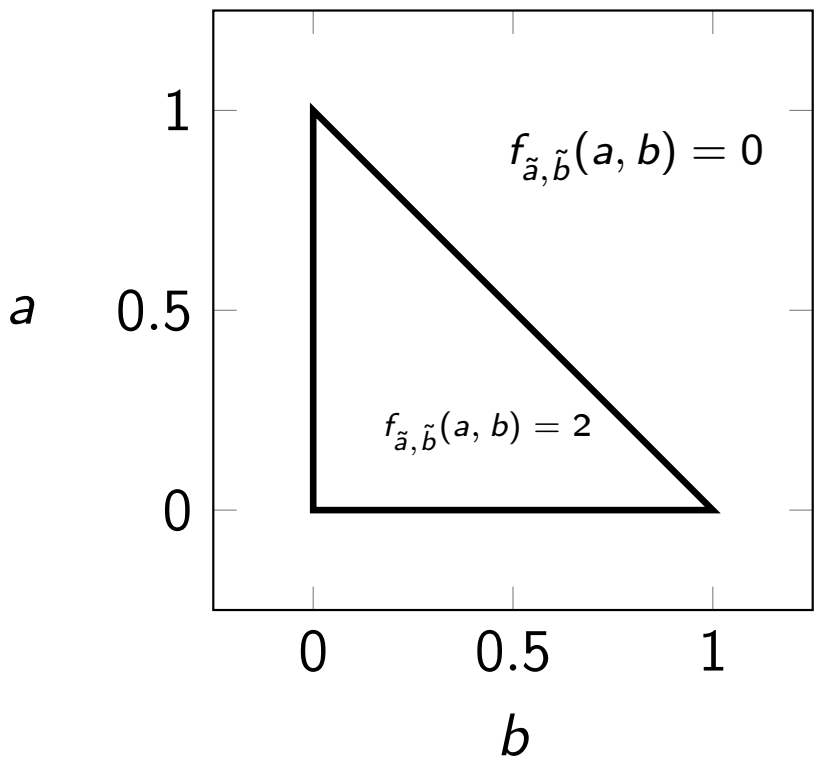
Conditional mean function

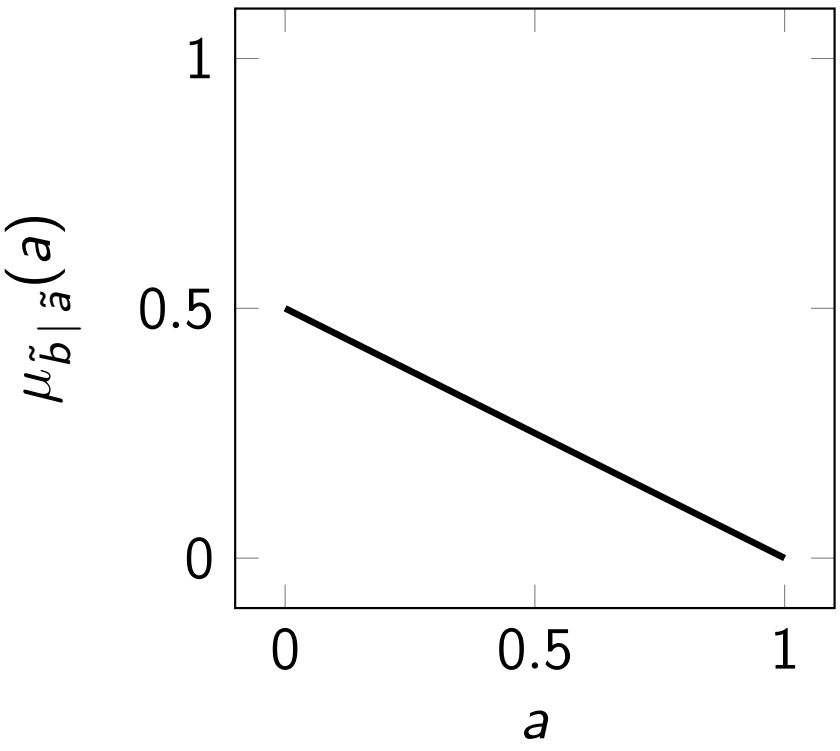
The conditional mean function of a continuous random variable $\tilde{b}$ given $\tilde{a}$ is

$$\mu_{\tilde{b} \mid \tilde{a}}(a) := \int_{b=-\infty}^{\infty} b \, f_{\tilde{b} \mid \tilde{a}}(b \mid a) \, \mathrm{d}b$$

Triangle lake

$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) = \frac{1-a}{2}$$





Sample conditional mean

Dataset $\mathcal{D}$: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), where $x_i \in A$

Data interpreted as samples from random variables $\tilde{a}$ (range A) and $\tilde{b}$

Estimate of $\mu_{\tilde{b} \mid \tilde{a}}$?

For any $a \in A$,

$$Y_a := \{y \mid (a, y) \in \mathcal{D}\}$$

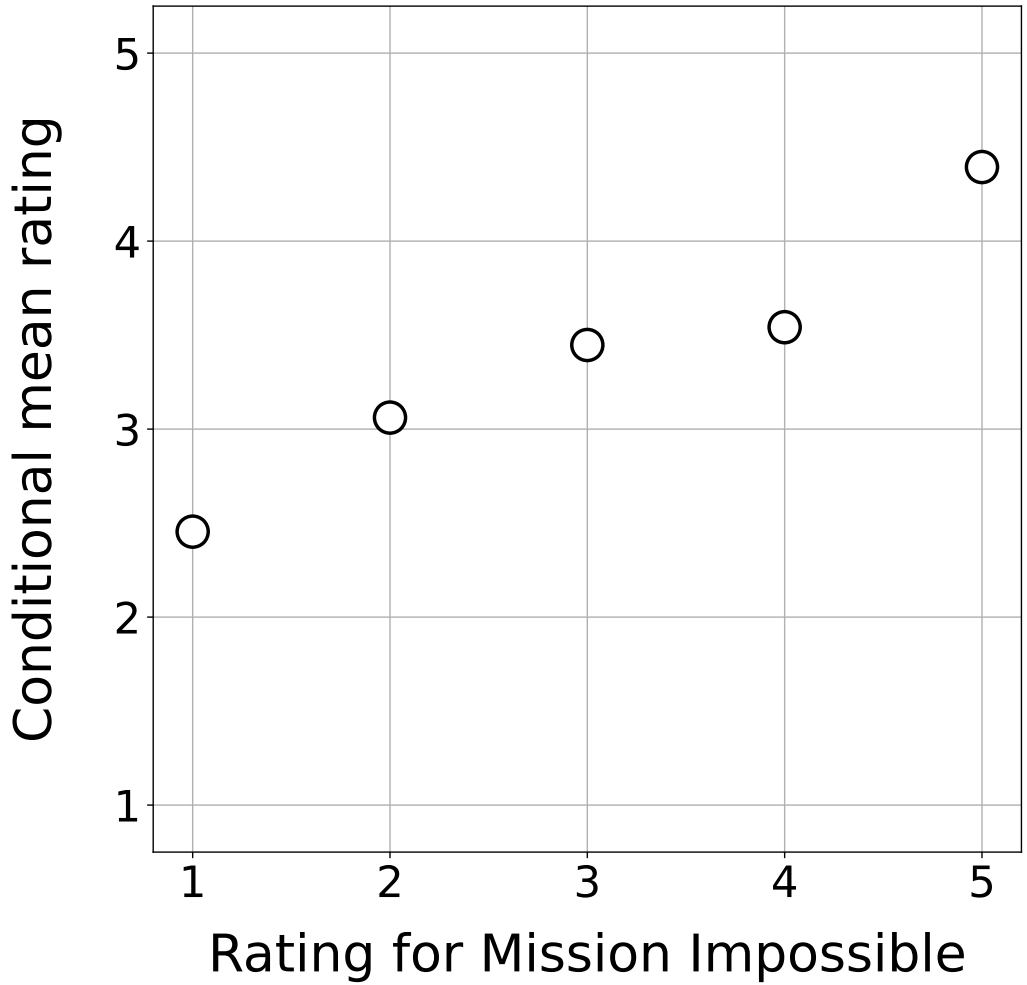
$$\widehat{m}_{\tilde{b} \mid \tilde{a}}(a) := \frac{1}{n_a} \sum_{y \in Y_a} y$$

n_a = number of elements of Y_a

Movie ratings

Mission Impossible \ Independence Day	**1**	**2**	**3**	**4**	**5**
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

Sample conditional mean function



Sample conditional mean function

Dataset $\mathcal{D}$: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)

Data interpreted as samples from random variables $\tilde{a}$ and $\tilde{b}$

If $\tilde{a}$ is continuous, estimate of $\mu_{\tilde{b} \mid \tilde{a}}$?

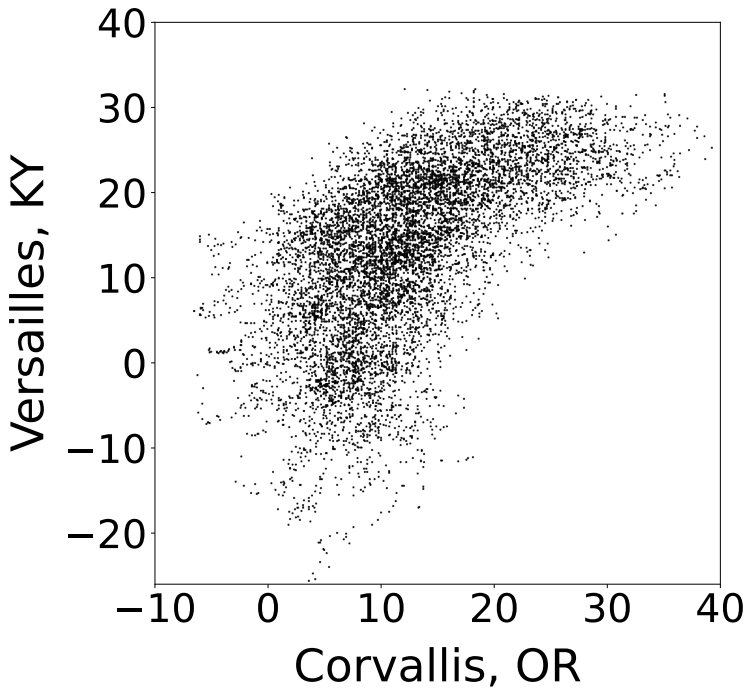
2 options:

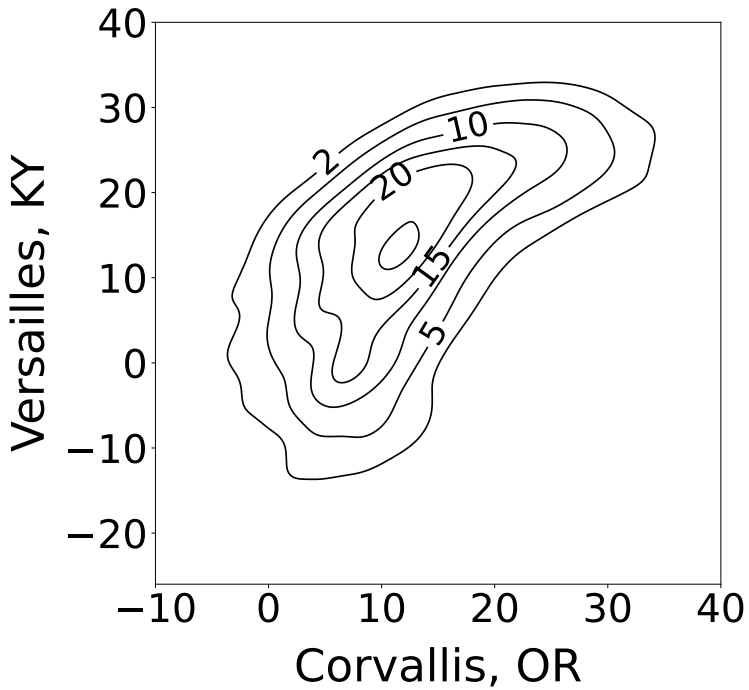
- Estimate $f_{\tilde{b} \mid \tilde{a}}$ using kernel density estimation and use it to approximate $\mu_{\tilde{b} \mid \tilde{a}}$
- For small ϵ,

$$Y_{a,\epsilon} := \{y \mid (x, y) \in \mathcal{D} \text{ for } |x - a| \leq \epsilon\}$$

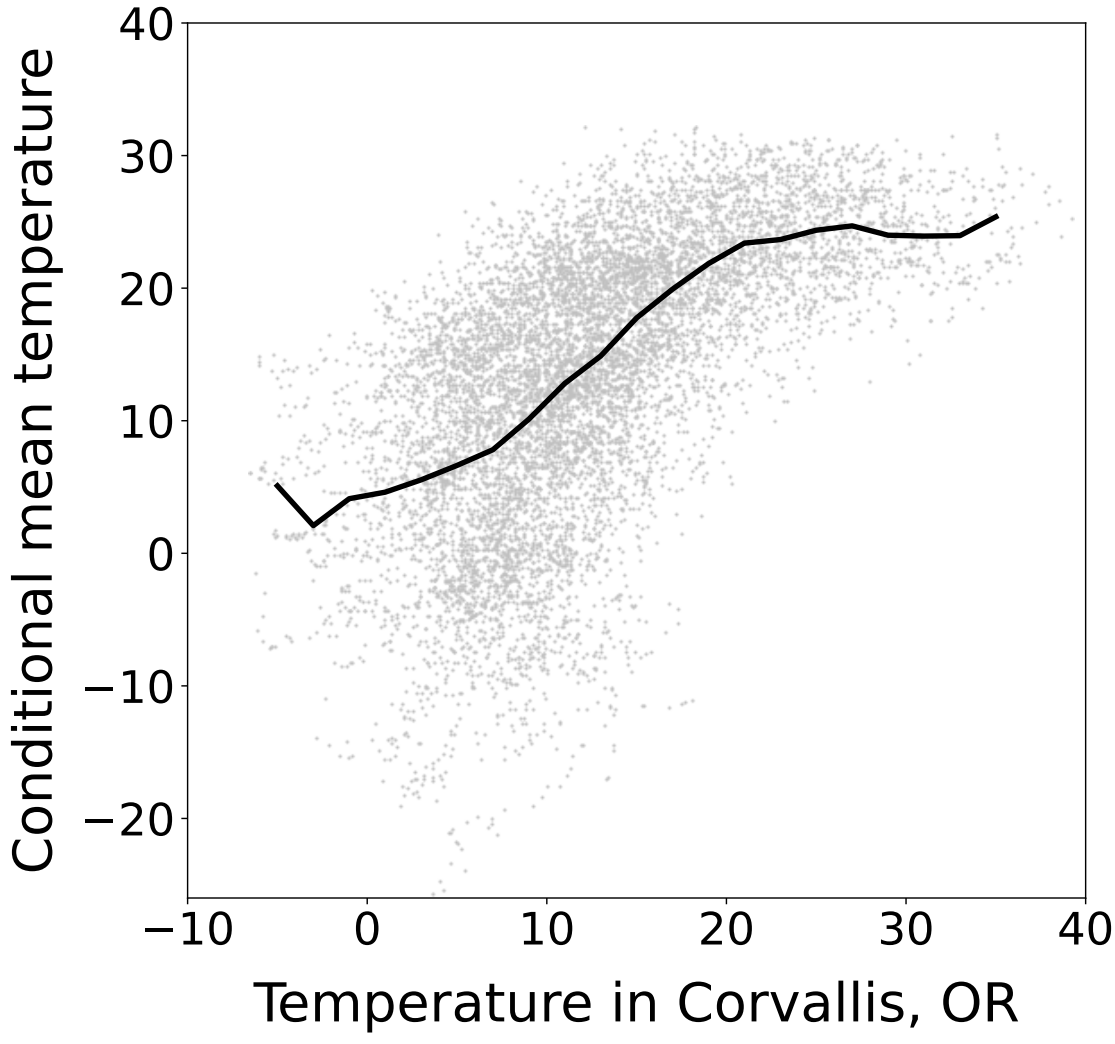
$$\widehat{m}_{\tilde{b} \mid \tilde{a}}(a) := \frac{1}{n_a} \sum_{y \in Y_{a,\epsilon}} y$$

Temperature in Corvallis and Versailles





Sample conditional mean function



What have we learned

Definition of conditional mean function

How to estimate it from data