Averaging

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

In data science we average all over the place

- To describe a quantity
- To describe the variation of a quantity
- To estimate a variable from another variable
- To estimate causal effects

Plan

- Average of a random variable
- The variance
- The conditional mean
- Causal inference

Average of a random variable?

Data: 3,4,3,4,4,3, ...

Interpreted as samples from random variable $\tilde{a}$ with range A

$$\frac{3+4+3+4+\cdots}{n}$$
$$= 3 \cdot \frac{\text{number of data} = 3}{n} + 4 \cdot \frac{\text{number of data} = 4}{n}$$

$$\approx \sum_{a \in A} a \, p_{\tilde{a}}(a)$$

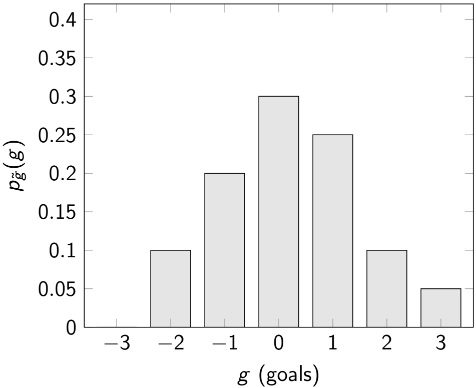
Mean of a discrete random variable

The mean, first moment or expected value of a discrete random variable $\tilde{a}$ with range A is

$$\mathrm{E}\left[\tilde{a}\right] := \sum_{a \in A} a \, p_{\tilde{a}}\left(a\right)$$

if the sum converges

Goal difference



$$
\begin{aligned}
\mathrm{E}[\tilde{g}] &= \sum_{g=-2}^{2} g \, p_{\tilde{g}}(g) \\
&= -2 \cdot 0.1 - 1 \cdot 0.2 + 0 \cdot 0.3 + 1 \cdot 0.25 + 2 \cdot 0.1 + 3 \cdot 0.05 \\
&= 0.2
\end{aligned}
$$

Average of function of a random variable?

Data: 3,4,3,4,4,3, ...

Interpreted as samples from random variable $\tilde{a}$ with range A

$$\frac{3^2 + 4^2 + 3^2 + 4^2 + \cdots}{n}$$

$$= 3^2 \cdot \frac{\text{number of data} = 3}{n} + 4^2 \cdot \frac{\text{number of data} = 4}{n}$$

$$\approx \sum_{a \in A} a^2 \, p_{\tilde{a}}(a)$$

Function of a random variable

The expected value of $h(\tilde{a})$, $h : \mathbb{R} \to \mathbb{R}$ is

$$\mathrm{E}\left[h(\tilde{a})\right] := \sum_{a \in A} h(a)\, p_{\tilde{a}}(a)$$

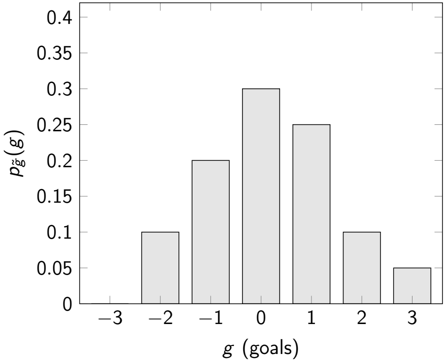
if $\tilde{a}$ is discrete and the sum converges

Converting goal difference to points

Points: $\tilde{x} := h(\tilde{g})$, where

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

Goal difference



$$
\begin{aligned}
\mathrm{E}[\tilde{x}] &= \mathrm{E}\left[h(\tilde{g})\right] \\
&= \sum_{g=-2}^{2} h(g) p_{\tilde{g}}(g) \\
&= 0 \cdot 0.1 + 0 \cdot 0.2 + 1 \cdot 0.3 + 3 \cdot 0.25 + 3 \cdot 0.1 + 3 \cdot 0.05 \\
&= 1.5
\end{aligned}
$$

Function of multiple random variables?

Data: $(3,1)$, $(4,2)$, $(4,1)$, $(3,2)$, $\ldots$, (x_n, y_n)

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

$$\frac{3\cdot 1 + 4\cdot 2 + 4\cdot 1 + 3\cdot 2 + \cdots}{n}$$
$$= 3\cdot 1\cdot \frac{\text{pairs} = (3{,}1)}{n} + 3\cdot 2\cdot \frac{\text{pairs} = (3{,}2)}{n} + \cdots$$

$$\approx \sum_{a\in A}\sum_{b\in B} a\cdot b\, p_{\tilde{a},\tilde{b}}(a, b)$$

Function of multiple random variables

If $\tilde{a}$ (range: A) and $\tilde{b}$ (range: B) are discrete, the expected value of $h(\tilde{a}, \tilde{b})$ is

$$\mathrm{E}[h(\tilde{a}, \tilde{b})] := \sum_{a \in A} \sum_{b \in B} h(a, b)\, p_{\tilde{a}, \tilde{b}}(a, b),$$

if the sum converges

Function of discrete random vector

If $\tilde{x}$ is a d-dimensional discrete random vector the expected value of $h(\tilde{x})$ of $\tilde{x}$ is

$$\mathrm{E}\left[h(\tilde{x})\right] := \sum_{x[1]\in X_1} \sum_{x[2]\in X_2} \cdots \sum_{x[d]\in X_d} h\left(x\right) p_{\tilde{x}}\left(x\right)$$

if the sum converges

Continuous random variable

The mean, first moment or expected value of a continuous random variable $\tilde{a}$ is

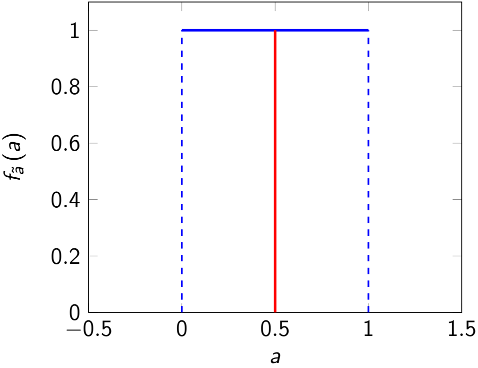
$$\mathrm{E}\left[\tilde{a}\right] := \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right)\, \mathrm{d}a$$

if the integral converges

Uniform random variable in $[a, b]$

$$\begin{aligned}
\mathrm{E}\left[\tilde{u}\right] &= \int_{u=-\infty}^{\infty} u f_{\tilde{a}}\left(u\right)\,\mathrm{d}u \\
&= \int_{u=a}^{b} \frac{u}{b-a}\,\mathrm{d}u \\
&= \frac{a+b}{2}
\end{aligned}$$

Uniform random variable in $[0, 1]$



Function of a random variable

The mean of $h(\tilde{a})$, $h : \mathbb{R} \to \mathbb{R}$ is

$$\mathrm{E}\left[h(\tilde{a})\right] := \int_{a=-\infty}^{\infty} h(a)\, f_{\tilde{a}}(a)\, \mathrm{d}a$$

if $\tilde{a}$ is continuous and the integral converges

Multiple random variables

If $\tilde{a}$, and $\tilde{b}$ are continuous, the expected value of $h(\tilde{a}, \tilde{b})$ is

$$\mathrm{E}[h(\tilde{a}, \tilde{b})] := \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} h\,(a, b)\, f_{\tilde{a},\tilde{b}}\,(a, b)\ \mathrm{d}a\, \mathrm{d}b$$

if the integral converges

Function of random vector

If $\tilde{x}$ is a d-dimensional continuous random vector the expected value of $h(\tilde{x})$ is

$$\mathrm{E}\left[h(\tilde{x})\right] := \int_{x \in \mathbb{R}^d} h\left(x\right) f_{\tilde{x}}\left(x\right) \, \mathrm{d}x$$

if the integral converges

Discrete and continuous quantities

If $\tilde{c}$ is continuous and $\tilde{d}$ is discrete with range D, the mean of $h(\tilde{c}, \tilde{d})$ is

$$\begin{aligned}
\mathrm{E}\left[h(\tilde{c}, \tilde{d})\right] &:= \int_{c=-\infty}^{\infty} \sum_{d \in D} h(c, d) f_{\tilde{c}}(c)\, p_{\tilde{d} \mid \tilde{c}}(d \mid c)\, \mathrm{d}c \\
&= \sum_{d \in D} \int_{c=-\infty}^{\infty} h(c, d) p_{\tilde{d}}(d)\, f_{\tilde{c} \mid \tilde{d}}(c \mid d)\, \mathrm{d}c,
\end{aligned}$$

if the sum and integral converge

Bayesian coin flip

We flip a coin but don't know the probability of heads $\tilde{\theta}$

We assume $\tilde{\theta}$ is uniform in [0,1]

Mean of the coin flip (heads = 1, tails = 0)?

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \int_{c=-\infty}^{\infty} \sum_{a=0}^{1} a f_{\tilde{\theta}}\left(\theta\right) p_{\tilde{a}\,|\,\tilde{\theta}}\left(a \,|\, \theta\right)\, \mathrm{d}\theta \\
&= \int_{0}^{1} \theta \,\mathrm{d}\theta \\
&= \frac{1}{2}
\end{aligned}$$

How do we estimate the mean from data?

We average

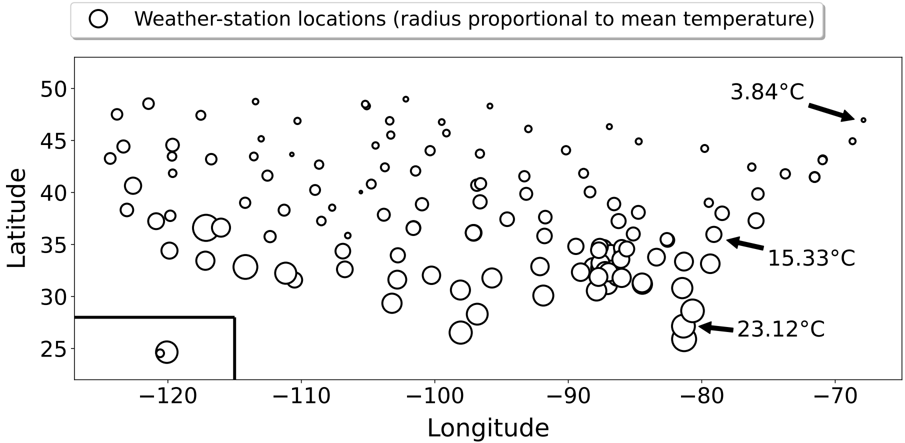
The sample mean of $X := \{x_1, x_2, \ldots, x_n\}$ is the arithmetic average

$$m(X) := \frac{\sum_{i=1}^{n} x_i}{n}$$

Same for discrete and continuous variables

Temperature dataset

Hourly temperatures at 134 weather stations in the US



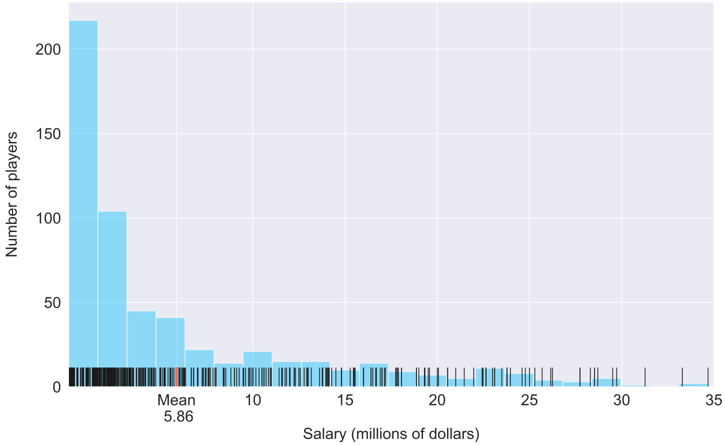
Method of moments

Distribution	Parameter	Maximum-likelihood estimator	Mean
Bernoulli	θ	$\frac{1}{n}\sum_{i=1}^{n} x_i = m(X)$	θ
Geometric	α	$\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1} = m(X)^{-1}$	α^{-1}
Poisson	λ	$\frac{1}{n}\sum_{i=1}^{n} x_i = m(X)$	λ
Exponential	λ	$\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1} = m(X)^{-1}$	λ^{-1}
Gaussian	μ	$\frac{1}{n}\sum_{i=1}^{n} x_i = m(X)$	μ

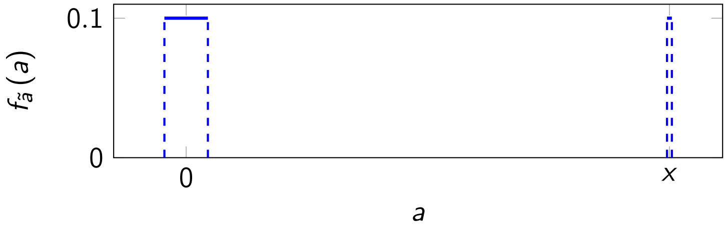
NBA salaries

How many earn more than mean?

Less than 1/3 of players (32.1%)

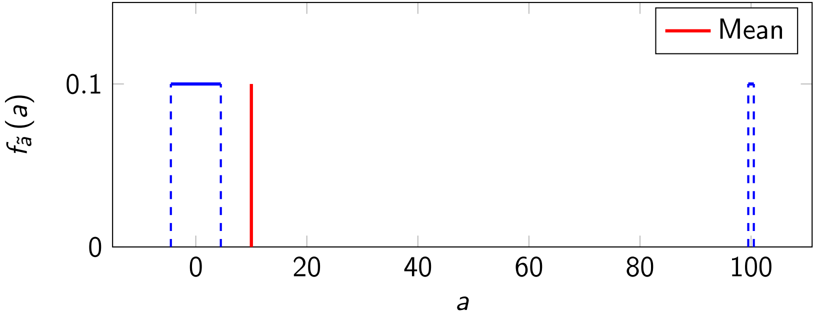


Extreme values

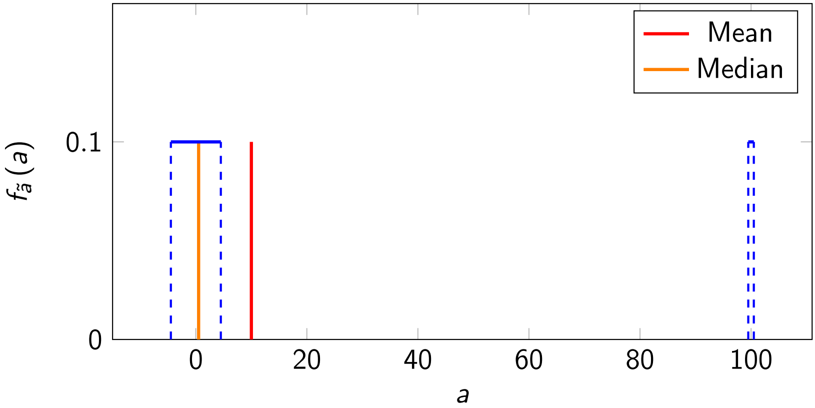


Random variable $\tilde{a}$ uniform in $[-4.5, 4.5]$ and $[x - 0.5, x + 0.5]$

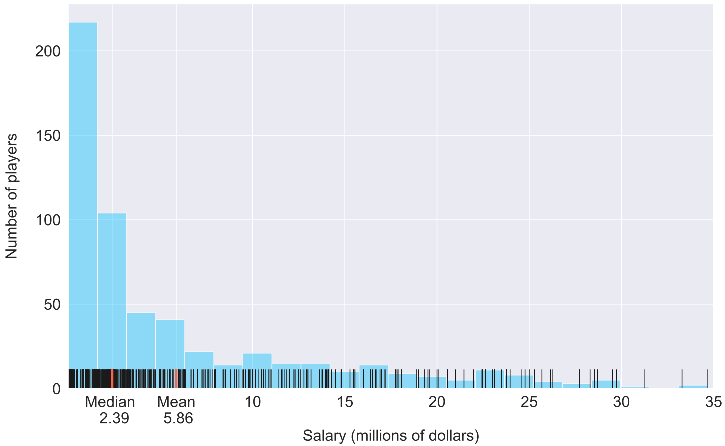
$$\mathrm{E}[\tilde{a}] = \frac{x}{10}$$



Median = 0.5



NBA salaries

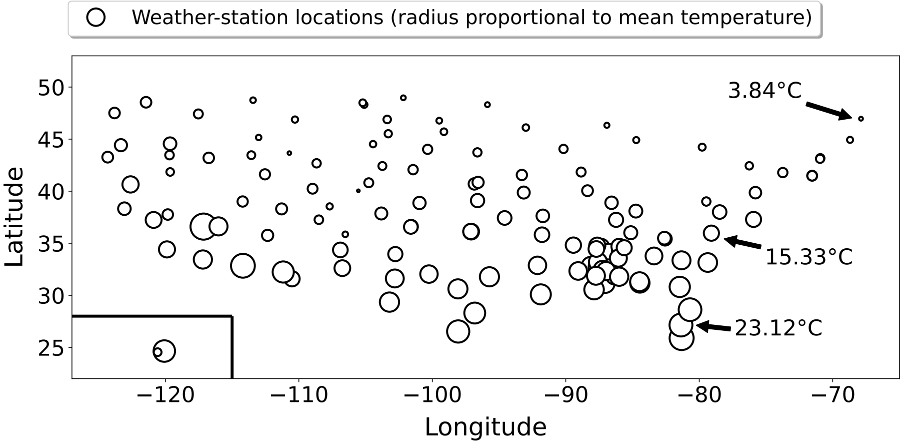


Two important properties

- Mean of linear combination is linear combination of means (always)
- Mean of product is product of means
  (only under independence)

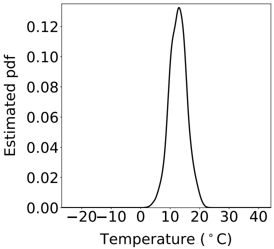
Temperature dataset

Hourly temperatures at 134 weather stations in the US

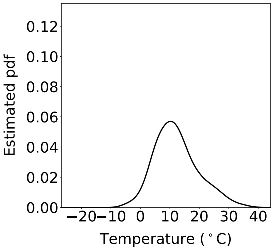


Same mean

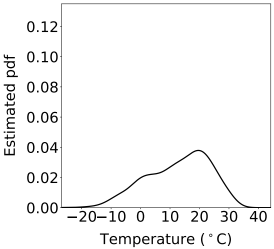
$m(X) = 12.7°\text{C}$



$m(X) = 12.3°\text{C}$



$m(X) = 12.7°\text{C}$



Challenge

Quantifying *magnitude* of deviation from the mean

Magnitude of a random variable?

Magnitude of real number a: $|a| = \sqrt{a^2}$

Euclidean length of vector x: $||x||_2 = \sqrt{\sum_{i=1}^{d} x[i]^2}$

Magnitude/energy of random variable $\tilde{a}$?

Mean square or second moment $\mathrm{E}\left[\tilde{a}^2\right]$

Variance

Mean squared distance of a random variable to its mean

$$\begin{aligned}\mathrm{Var}[\tilde{a}] &:= \mathrm{E}\left[(\tilde{a} - \mathrm{E}[\tilde{a}])^2\right] \\ &= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}[\tilde{a}]^2\end{aligned}$$

Standard deviation

The standard deviation $\sigma_{\tilde{a}}$ of $\tilde{a}$ is

$$\sigma_{\tilde{a}} := \sqrt{\mathrm{Var}\left[\tilde{a}\right]}$$

Sample variance

Dataset: $x_1, x_2, \ldots, x_n$

The sample variance is the average squared deviation from the sample mean

$$v(X) := \frac{\sum_{i=1}^{n} (x_i - m(X))^2}{n-1}$$

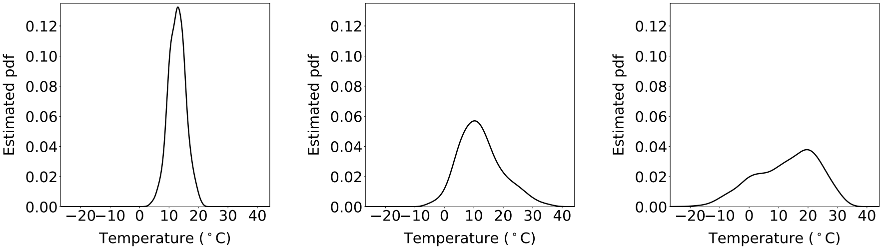
The sample standard deviation σ_X is the square root of the sample variance

Same mean

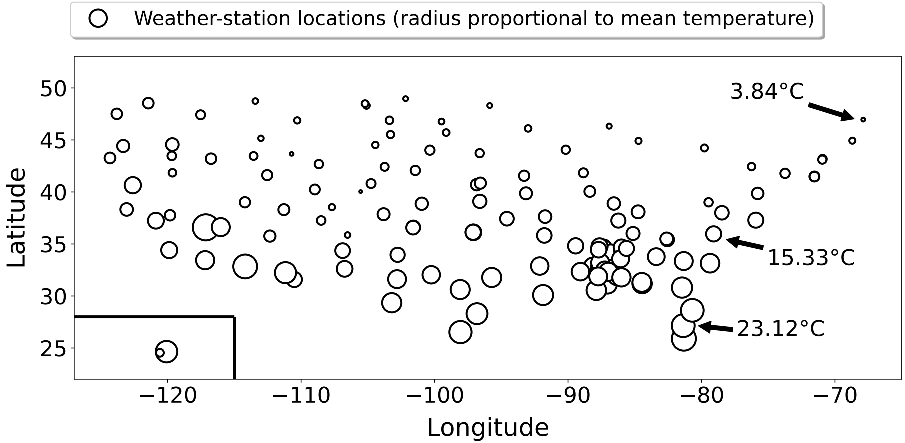
$m(X) = 12.7°\mathrm{C}$ $\sqrt{v(X)} = 2.9°\mathrm{C}$

$m(X) = 12.3°\mathrm{C}$ $\sqrt{v(X)} = 7.5°\mathrm{C}$

$m(X) = 12.7°\mathrm{C}$ $\sqrt{v(X)} = 10.6°\mathrm{C}$

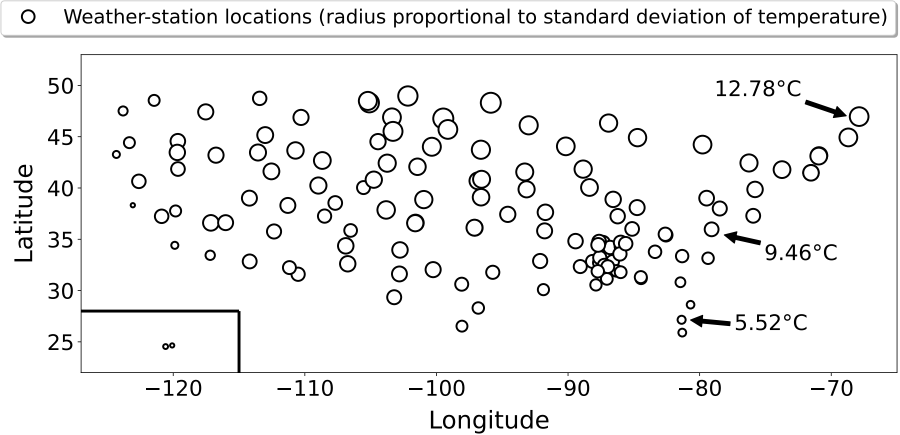


Means



Weather-station locations (radius proportional to mean temperature)
3.84°C
15.33°C
23.12°C
50
45
40
35
30
25
Latitude
−120
−110
−100
−90
−80
−70
Longitude

Standard deviations



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

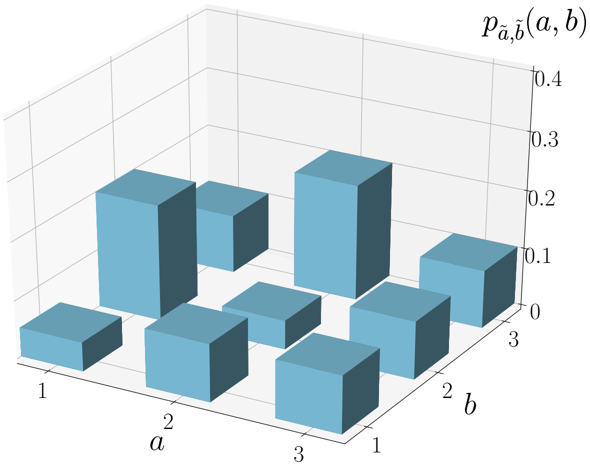
The conditional mean function of a discrete random variable $\tilde{b}$ given $\tilde{a}$ is

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

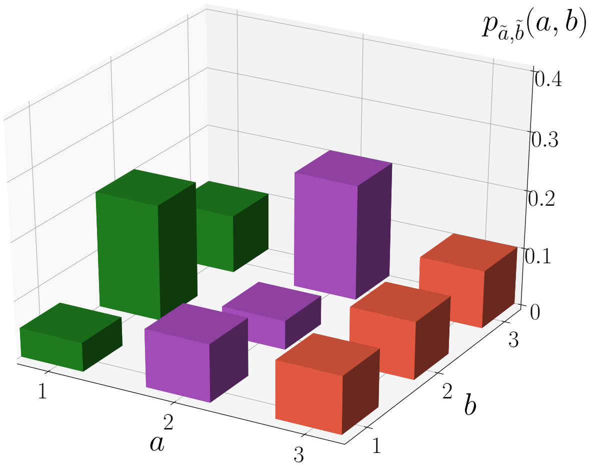
The conditional mean function of a continuous random variable $\tilde{b}$ given $\tilde{a}$ is

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

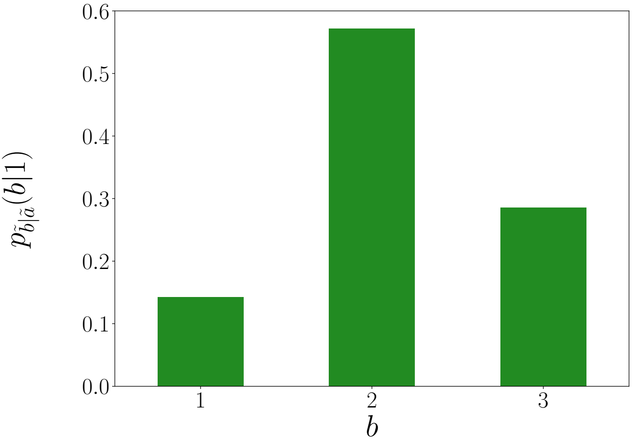
Joint pmf



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



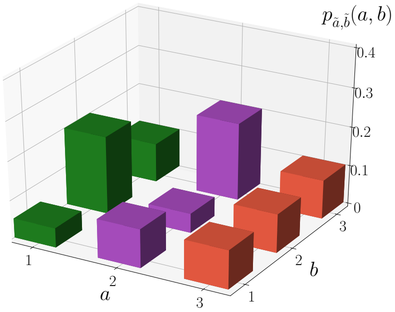
Mean of $\tilde{b}$ if $\tilde{a} = 1$



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

$$= 1 \cdot \frac{1}{7} + 2 \cdot \frac{4}{7} + 3 \cdot \frac{2}{7} = \frac{15}{7}$$

Conditional mean function

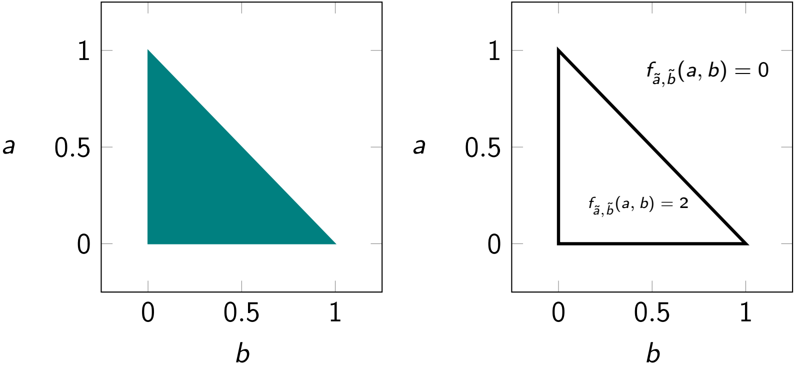


$$\mu_{\tilde{b}\,|\,\tilde{a}}(1) = \sum_{b \in B} b \, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,1) = \frac{15}{7}$$

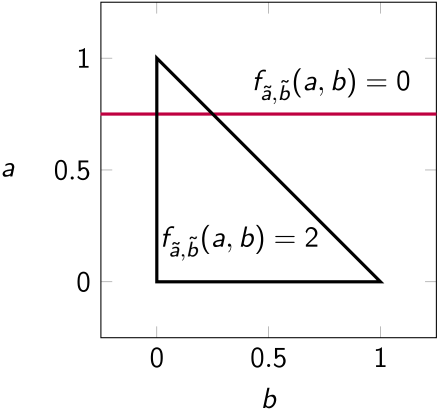
$$\mu_{\tilde{b}\,|\,\tilde{a}}(2) = \sum_{b \in B} b \, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,2) = \frac{16}{7}$$

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

Triangle lake



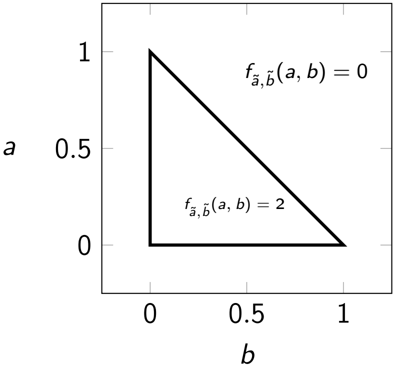
Mean of $\tilde{b}$ if $\tilde{a} = a$?

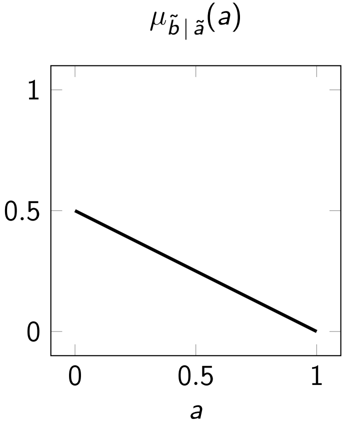


$$f_{\tilde{b} \mid \tilde{a}}(b \mid a) = \frac{1}{1-a} \qquad b \in [0, 1-a]$$

Triangle lake: Conditional mean function

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





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 $\mu_{\tilde{b}\,|\,\tilde{a}}$?

For any $a \in A$,

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

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

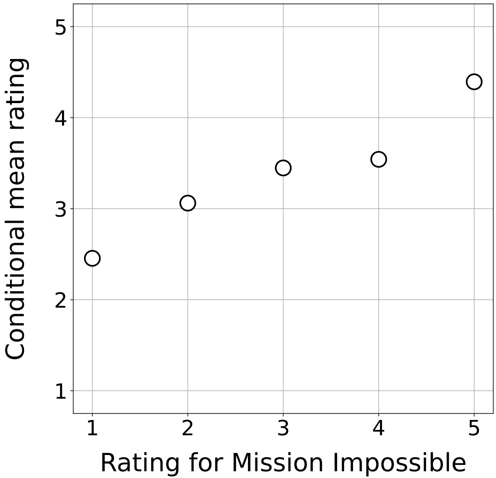
$n_a =$ number of elements of Y_a

Movie ratings

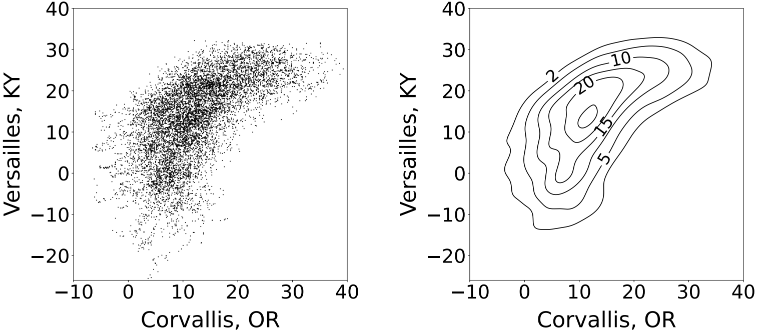
Independence Day

Mission Impossible	**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

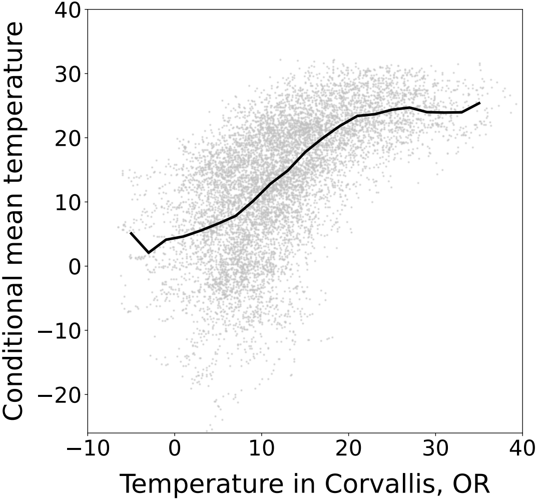
Sample conditional mean function



Temperature in Corvallis and Versailles



Sample conditional mean function



Iterated expectation

For any random variables $\tilde{a}$ and $\tilde{b}$ belonging to the same probability space

$$\mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right] = \mathrm{E}[\tilde{b}]$$

For any function $h : \mathbb{R}^2 \to \mathbb{R}$

$$\mathrm{E}[\mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(\tilde{a})] = \mathrm{E}\left[h(\tilde{a},\tilde{b})\right]$$

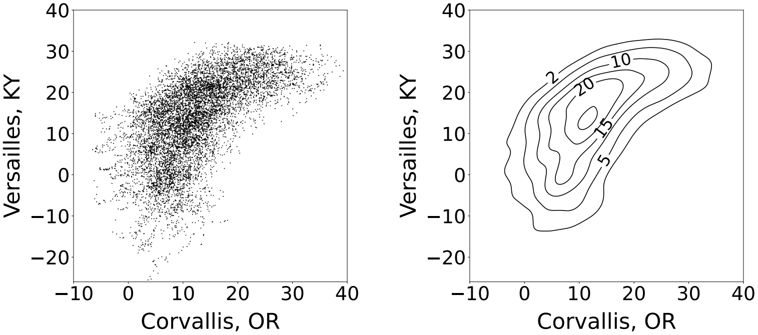
Regression

Independence Day

Mission Impossible	**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

Given rating for Mission Impossible, rating for Independence Day?

Regression



Given temperature in Corvallis, temperature in Versailles?

Regression

Goal: Find function h, such that $h(a)$ approximates $\tilde{b}$ when $\tilde{a} = a$

How do we evaluate the estimator?

Mean squared error (MSE)

$$\mathrm{E}\left[(\tilde{b} - h(\tilde{a}))^2\right] = \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} (b - h(a))^2 f_{\tilde{a},\tilde{b}}(a, b)\, \mathrm{d}b\, \mathrm{d}a$$

Minimum MSE constant estimate

Best constant estimate of $\tilde{a}$?

$$\arg\min_{c \in \mathbb{R}} \mathrm{E}\left[(c - \tilde{a})^2\right] = \mathrm{E}[\tilde{a}]$$

The mean $\mathrm{E}[\tilde{a}]$ is the minimum MSE constant estimate

Regression: Given $\tilde{a} = a$ how should we estimate $\tilde{b}$?

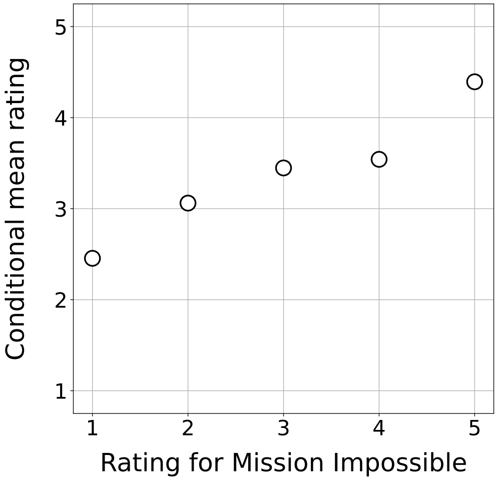
Conditional mean function of $\tilde{b}$ given $\tilde{a} = a$

MMSE estimator

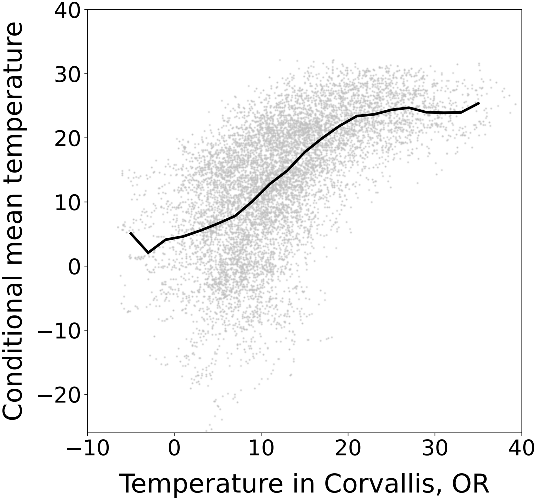
The conditional mean is the minimum MSE estimator

$$\mu_{\tilde{b} \mid \tilde{a}}(\tilde{a}) = \arg\min_{h(\tilde{a})} \mathrm{E}\left[\left(\tilde{b} - h(\tilde{a})\right)^2\right]$$

Movie ratings



Temperature in Corvallis and Versailles



Causal inference

Goal: Estimate causal effect of a *treatment* from data

All caps titles

Goal: Determine whether all caps titles cause YouTube videos to get more views

Treatment $\tilde{t}$: if title is all caps $\tilde{t} = 1$, if not $\tilde{t} = 0$

Data: Number of views $\tilde{y}$

Potential outcomes

$\widetilde{\text{po}}_0$: Views if all titles are proper case

$\widetilde{\text{po}}_1$: Views if all titles are all caps

Observed data:

$$\tilde{y} := \begin{cases} \widetilde{\text{po}}_0 & \text{if} \quad \tilde{t} = 0 \\ \\ \widetilde{\text{po}}_1 & \text{if} \quad \tilde{t} = 1 \end{cases}$$

Average treatment effect

$$\mathsf{ATE} := \mathrm{E}\left[\widetilde{\mathsf{po}}_1\right] - \mathrm{E}\left[\widetilde{\mathsf{po}}_0\right]$$

Challenge: We do not observe $\widetilde{\mathsf{po}}_0$ and $\widetilde{\mathsf{po}}_1$ directly

Observed data

Treatment $\tilde{t}$	Observed outcome $\tilde{y}$	Outcome if proper case $\widetilde{\text{po}}_0$	Outcome if all caps $\widetilde{\text{po}}_1$
✗	102	102	?
✗	45	45	?
✓	330	?	330
✓	121	?	121
✓	23	?	23

? are counterfactuals

Is $\mu_{\tilde{y}\,|\,\tilde{t}}(1) - \mu_{\tilde{y}\,|\,\tilde{t}}(0)$ a reasonable estimate for the ATE?

Estimating the ATE

$$
\begin{aligned}
\mu_{\tilde{y} \mid \tilde{t}}(1) &= \mu_{\widetilde{\mathsf{po}}_1 \mid \tilde{t}}(1) \\
&= \int_x x f_{\widetilde{\mathsf{po}}_1 \mid \tilde{t}}(x \mid 1) \, \mathrm{d}x \\
&= \int_x x f_{\widetilde{\mathsf{po}}_1}(x) \, \mathrm{d}x \qquad \text{if } \widetilde{\mathsf{po}}_1 \text{ and } t \text{ are independent} \\
&= \mathrm{E}\left[\widetilde{\mathsf{po}}_1\right]
\end{aligned}
$$

$$
\mu_{\tilde{y} \mid \tilde{t}}(0) = \mathrm{E}\left[\widetilde{\mathsf{po}}_0\right]
$$

$$
\mathsf{ATE} = \mu_{\tilde{y} \mid \tilde{t}}(1) - \mu_{\tilde{y} \mid \tilde{t}}(0)
$$

YouTube videos

All caps: 19

No all caps: 26

$$\begin{aligned} \text{ATE} &= \mu_{\tilde{y}\,|\,\tilde{t}}(1) - \mu_{\tilde{y}\,|\,\tilde{t}}(0) \\ &= 133 - 132 \approx 0 \end{aligned}$$

YouTube videos