7

Averaging

Overview

Averaging is a fundamental operation in probability and statistics. In Section 7.1 we define an averaging operation for random variables known as the mean. Section 7.2 shows that the mean is linear, and Section 7.3 that the mean of the product of independent random variables is equal to the product of their means. In Section 7.4 we derive the mean of several popular distributions. In Section 7.5 we show that the mean can be distorted by extreme values, unlike the median. Sections 7.6 defines the mean square, which represents the average squared value of a random variable. Section 7.7 defines the variance, which quantifies the average variation of a random variable. In Section 7.8 we define the conditional mean of a random variable given another random variable, discussing its connection to regression. Finally, in Section 7.9 we discuss the estimation of causal effects using conditional averages.

7.1 The Mean

In this section we define an averaging operation for random quantities, known as the mean. Sections 7.1.1 and 7.1.2 define the mean for discrete and continuous random variables respectively. In Section 7.1.3 we explain how to compute the mean of quantities that depend on both discrete and continuous random variables. Finally, Section 7.1.4 describes how to estimate the mean from data.

7.1.1 Discrete Random Variables

In order to define an averaging operation that is appropriate for discrete random variables, we consider a dataset $x_1, \ldots, x_n$ with values in a discrete set A. The arithmetic average of the data is

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{a \in A} a \cdot \frac{\text{number of data equal to } a}{n}. \tag{7.1}$$

Let us interpret the data as samples from a random variable $\tilde{a}$. According to our intuitive definition of probability in (1.1), for large n, the probability that $\tilde{a} = a$

for any $a \in A$ represents the fraction of data that equal a:

$$\mathrm{P}(\tilde{a} = a) \approx \frac{\text{number of data equal to } a}{n}. \tag{7.2}$$

Plugging this into (7.1) yields a formula for the average in terms of the entries of the pmf of $\tilde{a}$:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \approx \sum_{a \in A} a\mathrm{P}(\tilde{a} = a) \tag{7.3}$$

$$= \sum_{a \in A} a p_{\tilde{a}}(a). \tag{7.4}$$

We call this average the mean or expected value of $\tilde{a}$. This averaging operator, which is sometimes called the *expectation* operator, maps a random variable to a single number.

Definition 7.1 (Mean of a discrete random variable)**.** *The mean, expected value or first moment of a discrete random variable $\tilde{a}$ with range A and pmf $p_{\tilde{a}}$ is*

$$\mathrm{E}\left[\tilde{a}\right] := \sum_{a \in A} a p_{\tilde{a}}\left(a\right), \tag{7.5}$$

if the sum converges.

It is possible for a random variable to have an infinite mean, if the sum in Definition 7.1 tends to infinity, or to have no mean at all, when the sum is not well defined. We provide examples of such random variables and discuss the implications in Section 9.6.

Example 7.2 (Expected goals in soccer game)**.** In Example 2.7 we consider a discrete random variable $\tilde{g}$ that represents the goal difference in a soccer game between Barcelona and Atlético de Madrid. The pmf of $\tilde{g}$ equals $p_{\tilde{g}}\left(-2\right) := 0.1$, $p_{\tilde{g}}\left(-1\right) := 0.2$, $p_{\tilde{g}}\left(0\right) := 0.3$, $p_{\tilde{g}}\left(1\right) := 0.25$, $p_{\tilde{g}}\left(2\right) := 0.1$, $p_{\tilde{g}}\left(3\right) := 0.05$. The mean of $\tilde{g}$ is

$$\mathrm{E}[\tilde{g}] := \sum_{g=-2}^{3} g p_{\tilde{g}}\left(g\right) \tag{7.6}$$

$$= -2 \cdot 0.1 - 1 \cdot 0.2 + 0 \cdot 0.3 + 1 \cdot 0.25 + 2 \cdot 0.1 + 3 \cdot 0.05 \tag{7.7}$$

$$= 0.2. \tag{7.8}$$

In probabilistic modeling we are often interested in the behavior of functions of the available data. Imagine that we want to compute the mean of $h(\tilde{a})$, where h is a deterministic function and $\tilde{a}$ is a discrete random variable. If we apply h to a dataset $x_1, \ldots, x_n$ for very large n, then by (7.2), the arithmetic average

should be well approximated by the sum of the possible values of $h(a)$ weighted by the probability of a:

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i) = \sum_{a\in A} h(a)\cdot\frac{\text{number of data equal to } a}{n} \tag{7.9}$$

$$\approx \sum_{a\in A} h(a)p_{\tilde{a}}(a), \tag{7.10}$$

where A is the range of $\tilde{a}$. This motivates the following generalization of Definition 7.1.

Definition 7.3 (Mean of a function of a discrete random variable)**.** *Let $\tilde{a}$ be a discrete random variable with range A and pmf $p_{\tilde{a}}$ and let h be a deterministic function. Then,*

$$\mathrm{E}\left[h(\tilde{a})\right] := \sum_{a\in A} h(a)p_{\tilde{a}}(a). \tag{7.11}$$

The mean of a function of a discrete random variable can also be computed by first deriving the pmf of the transformed variable, and then applying Definition 7.1. Exercise 7.2 shows that this yields the same result.

Example 7.4 (Expected points in soccer game)**.** In Example 2.9 we derive the distribution of points $\tilde{x} := h(\tilde{g})$ corresponding to the goal difference $\tilde{g}$ from Example 7.2. A win is worth three points, a draw one point, and a loss zero points, so

$$h(g) := \begin{cases} 0 & \text{if } g < 0, \\ 1 & \text{if } g = 0, \\ 3 & \text{if } g > 0. \end{cases} \tag{7.12}$$

By Definition 7.3, we can derive the mean of the points without deriving the pmf of $\tilde{g}$:

$$\mathrm{E}[\tilde{x}] = \mathrm{E}\left[h(\tilde{g})\right] \tag{7.13}$$

$$= \sum_{g=-2}^{3} h(g)p_{\tilde{g}}\left(g\right) \tag{7.14}$$

$$= 0\cdot 0.1 + 0\cdot 0.2 + 1\cdot 0.3 + 3\cdot 0.25 + 3\cdot 0.1 + 3\cdot 0.05 \tag{7.15}$$

$$= 1.5. \tag{7.16}$$

We can also define the mean of functions of multiple discrete random variables. For instance, let us consider a bivariate function $h : \mathbb{R}^2 \to \mathbb{R}$ applied to a dataset (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). If we interpret the data as joint realizations of the random variables $\tilde{a}$ and $\tilde{b}$ taking values in the discrete sets A and B, respectively,

Table 7.1 ***Cats and dogs.*** *Joint pmf $p_{\tilde{c},\tilde{d}}(c,d)$ of the random variables $\tilde{c}$ and $\tilde{d}$ in Example 7.6.*

		Cats (c)			
		0	**1**	**2**	**3**
Dogs (d)	**0**	0.35	0.15	0.1	0.05
	1	0.2	0.05	0.03	0
	2	0.05	0.02	0	0

then for large n,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i) &= \sum_{a\in A}\sum_{b\in B} h(a,b)\cdot\frac{\text{number of pairs } (x,y) \text{ for which } x=a \text{ and } y=b}{n}\\ &\approx \sum_{a\in A}\sum_{b\in B} h(a,b)p_{\tilde{a},\tilde{b}}(a,b). \end{aligned} \tag{7.17}$$

The same logic applies to more than two random variables, or the entries of a d-dimensional random vector.

Definition 7.5 (Mean of a function of multiple discrete random variables)**.** *If $\tilde{a}$ and $\tilde{b}$ are discrete random variables with ranges A and B, the mean or expected value of a function $h(\tilde{a},\tilde{b})$ of $\tilde{a}$ and $\tilde{b}$, $h:\mathbb{R}^2\to\mathbb{R}$, is*

$$\mathrm{E}[h(\tilde{a},\tilde{b})] := \sum_{a\in A}\sum_{b\in B} h(a,b)\, p_{\tilde{a},\tilde{b}}(a,b), \tag{7.18}$$

if the sum converges.

Let $\tilde{x}$ be a d-dimensional discrete random vector, where the ith entry $\tilde{x}[i]$, $1\le i\le n$, is a random variable with range X_i. The mean or expected value of a function $h(\tilde{x})$ of $\tilde{x}$, $h:\mathbb{R}^d\to\mathbb{R}$, is

$$\mathrm{E}\left[h(\tilde{x})\right] := \sum_{x[1]\in X_1}\sum_{x[2]\in X_2}\cdots\sum_{x[d]\in X_d} h(x)\, p_{\tilde{x}}(x), \tag{7.19}$$

if the sum converges.

Example 7.6 (Cats and dogs)**.** A producer of pet food wants to compute the expected total number of cats and dogs per household in a certain city. Table 7.1 shows the joint pmf of the cats and dogs, represented by the random variables $\tilde{c}$ and $\tilde{d}$. The expected number of pets is the mean of the sum $\tilde{c}+\tilde{d}$. By Definition 7.5, it equals

$$\mathrm{E}[\tilde{c}+\tilde{d}] = \sum_{c=0}^{3}\sum_{d=0}^{2}(c+d)p_{\tilde{c},\tilde{d}}(c,d) \tag{7.20}$$

$$= 0.15 + 2\cdot 0.1 + 3\cdot 0.05 + 0.2 + 2\cdot 0.05 + 3\cdot 0.03 + 2\cdot 0.05 + 3\cdot 0.02$$

$$= 1.05. \tag{7.21}$$

7.1.2 Continuous Random Variables

In this section we define an averaging operation that can be applied to continuous random variables. To motivate the definition, let us partition the real line using a grid with step size equal to ϵ, $a_m := m\epsilon$, where $m \in \mathbb{Z}$. Imagine that we want to average a very large real-valued dataset $x_1, \ldots, x_n$, which we interpret as samples from a continuous random variable $\tilde{a}$. For small ϵ, as n tends to infinity, the average is approximately equal to

$$\frac{1}{n}\sum_{i=1}^{n} x_i \approx \sum_{m\in\mathbb{Z}} \frac{a_m \cdot \text{number of data between } a_m - \epsilon \text{ and } a_m}{n} \tag{7.22}$$

$$\approx \sum_{m\in\mathbb{Z}} a_m \mathrm{P}\left(a_m - \epsilon < \tilde{a} \leq a_m\right) \tag{7.23}$$

$$\approx \sum_{m\in\mathbb{Z}} a_m f_{\tilde{a}}(a_m)\epsilon, \tag{7.24}$$

because for small ϵ, $f_{\tilde{a}}(a_m)\epsilon$ approximates the probability that the random variable is between $a_m - \epsilon$ and a_m, as explained in Section 3.3. If we take the limit $\epsilon \to 0$ so that the grid becomes arbitrarily fine, then the sum in (7.24) converges to the integral $\int_{a\in\mathbb{R}} a f_{\tilde{a}}(a)\ \mathrm{d}a$. The same argument can be applied to deduce that the average of samples from a random variable $\tilde{a}$ transformed by a deterministic random variable h should be well approximated by $\int_{a\in\mathbb{R}} h(a) f_{\tilde{a}}(a)\ \mathrm{d}a$. This motivates our definition of the mean of a continuous random variable.

Definition 7.7 (Mean of a continuous random variable)**.** *The mean, expected value or first moment of a continuous random variable $\tilde{a} : \Omega \to \mathbb{R}$ is defined as*

$$\mathrm{E}\left[\tilde{a}\right] := \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right)\ \mathrm{d}a, \tag{7.25}$$

if the integral converges.

Let $\tilde{a}$ be a continuous random variable with pdf $f_{\tilde{a}}$ and let $h : \mathbb{R} \to \mathbb{R}$ be a deterministic function, such that $h(\tilde{a})$ is a valid random variable. Then,

$$\mathrm{E}\left[h(\tilde{a})\right] := \int_{a\in\mathbb{R}} h(a) f_{\tilde{a}}(a)\ \mathrm{d}a, \tag{7.26}$$

if the integral converges.

As in the discrete case, it is possible for a continuous random variable to have an infinite mean, if the integral in Definition 7.7 tends to infinity. Similarly, a continuous random variable does not have a mean if the integral is not well defined. We provide examples and discuss the implications in Section 9.6.

The mean of a uniform random variable is the midpoint of the interval where its density is nonzero, so it is equal to its median. For most distributions, this is

not the case. The mean and median of a random variable can be very different, as we discuss in Section 7.5.

Lemma 7.8 (Mean of a uniform random variable)**.** *The mean of random variable $\tilde{u}$ that is uniformly distributed in the interval $[a, b]$, $b > a$, equals*

$$\mathrm{E}[\tilde{u}] = \frac{a+b}{2}. \tag{7.27}$$

Proof

$$\mathrm{E}\left[\tilde{u}\right] = \int_{u=-\infty}^{\infty} u f_{\tilde{a}}\left(u\right) \, \mathrm{d}u \tag{7.28}$$

$$= \int_{u=a}^{b} \frac{u}{b-a} \, \mathrm{d}u \tag{7.29}$$

$$= \frac{b^2 - a^2}{2\left(b-a\right)} \tag{7.30}$$

$$= \frac{a+b}{2}. \tag{7.31}$$

■

Example 7.9 (Expected power)**.** In Example 3.22 we explain how to derive the pdf of the power $\tilde{w}$ dissipated in a resistor as a function of the pdf of the current $\tilde{c}$ passing through it. If we are only interested in the mean power, we can compute it directly from $f_{\tilde{c}}$ by applying (7.26), because the power is a deterministic function of the current: $\tilde{w} = r\tilde{c}^2$, where r denotes the resistance of the resistor. For instance, if the current is uniformly distributed between -1 and 1 amperes,

$$\mathrm{E}[\tilde{w}] = \mathrm{E}\left[r\tilde{c}^2\right] \tag{7.32}$$

$$= \int_{c=-1}^{1} \frac{rc^2}{2} \, \mathrm{d}c \tag{7.33}$$

$$= \frac{r}{3}. \tag{7.34}$$

We now generalize Definition 7.7 to define the expected value of a function of multiple continuous random variables. The definition can be justified following a similar argument to (7.24).

Definition 7.10 (Mean of a function of multiple continuous random variables)**.** *If $\tilde{a}$ and $\tilde{b}$ are continuous random variables with joint pdf $f_{\tilde{a},\tilde{b}}$, the mean or expected value of a function $h(\tilde{a}, \tilde{b})$, $h : \mathbb{R}^2 \to \mathbb{R}$, of $\tilde{a}$ and $\tilde{b}$ is*

$$\mathrm{E}[h(\tilde{a}, \tilde{b})] := \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} h\left(a, b\right) f_{\tilde{a},\tilde{b}}\left(a, b\right) \, \mathrm{d}b \, \mathrm{d}a \tag{7.35}$$

if the integral converges.

If $\tilde{x}$ is a d-dimensional continuous random vector with entries $\tilde{x}[i]$, the expected value of a function $h(\tilde{x})$, $h : \mathbb{R}^d \to \mathbb{R}$, of $\tilde{x}$ is

$$\mathrm{E}\left[h(\tilde{x})\right] := \int_{x[1]\in\mathbb{R}} \int_{x[2]\in\mathbb{R}} \cdots \int_{x[d]\in\mathbb{R}} h\left(x\right) f_{\tilde{x}}\left(x\right) \, \mathrm{d}x, \tag{7.36}$$

if the integral converges.

Example 7.11 (Sugar)**.** You put your hand in a bag that has 1 kg of sugar and grab an amount of sugar that is uniformly distributed between 0 and 1 kg. While transferring the sugar to another bag, you spill an amount that is uniformly distributed between 0 and the amount that you grabbed. What is the expected amount of spilled sugar?

We represent the grabbed sugar by a random variable $\tilde{g}$ and the spilled sugar by a random variable $\tilde{s}$. From the information above, we can deduce the joint distribution of the two random variables. Setting $h(\tilde{g}, \tilde{s}) := \tilde{s}$ in Definition 7.10,

$$\begin{aligned}
\mathrm{E}[\tilde{s}] &= \int_g \int_s s f_{\tilde{g},\tilde{s}}(g,s) \, \mathrm{d}s \, \mathrm{d}g && (7.37)\\
&= \int_g \int_s s f_{\tilde{g}}(g) f_{\tilde{s}\,|\,\tilde{g}}(s\,|\,g) \, \mathrm{d}s \, \mathrm{d}g && (7.38)\\
&= \int_{g=0}^{1} \int_{s=0}^{g} \frac{s}{g} \, \mathrm{d}s \, \mathrm{d}g && (7.39)\\
&= \int_{g=0}^{1} \frac{g}{2} \, \mathrm{d}g && (7.40)\\
&= \frac{1}{4}. && (7.41)
\end{aligned}$$

7.1.3 Discrete And Continuous Random Variables

To compute the mean of a quantity that depends on a discrete and continuous random variable, we use their marginal and conditional distributions. The definition can be justified following the same logic used to motivate Definitions 7.5 and 7.7.

Definition 7.12 (Mean of a discrete and a continuous random variable)**.** *If $\tilde{c}$ is a continuous random variable and $\tilde{d}$ a discrete random variable with range D defined on the same probability space, the mean or expected value of a function $h(\tilde{c}, \tilde{d})$ of $\tilde{c}$ and $\tilde{d}$ is*

$$\begin{aligned}
\mathrm{E}\left[h(\tilde{c},\tilde{d})\right] &:= \int_{c=-\infty}^{\infty} \sum_{d\in D} h\left(c,d\right) f_{\tilde{c}}\left(c\right) p_{\tilde{d}|\tilde{c}}\left(d\,|\,c\right) \, \mathrm{d}c && (7.42)\\
&= \sum_{d\in D} \int_{c=-\infty}^{\infty} h\left(c,d\right) p_{\tilde{d}}\left(d\right) f_{\tilde{c}|\tilde{d}}\left(c|d\right) \, \mathrm{d}c, && (7.43)
\end{aligned}$$

if the sum and integral converge.

Example 7.13 (Mean of a Bayesian coin flip)**.** In Example 6.16 we model the result of a coin flip $\tilde{x}$ as a Bernoulli random variable ($\tilde{x} = 1$ is heads and $\tilde{x} = 0$ is tails) with a random parameter $\tilde{\theta}$. Let us compute the mean of $\tilde{x}$ under the two different priors for $\tilde{\theta}$ described in Example 6.16. If $\tilde{\theta}$ is uniformly distributed in $[0, 1]$, by Definition 7.12

$$\mathrm{E}\left[\tilde{x}\right] = \int_{\theta=-\infty}^{\infty} \sum_{x=0}^{1} x f_{\tilde{\theta}}\left(\theta\right) p_{\tilde{x}\,|\,\tilde{\theta}}\left(x \,|\, \theta\right) \,\mathrm{d}\theta \tag{7.44}$$

$$= \int_{0}^{1} \theta \,\mathrm{d}\theta \tag{7.45}$$

$$= \frac{1}{2}. \tag{7.46}$$

If $\tilde{\theta}$ has a triangular pdf $f_{\tilde{\theta}}\left(\theta\right) = 2\theta$, $\theta \in [0, 1]$, then

$$\mathrm{E}\left[\tilde{x}\right] = \int_{\theta=-\infty}^{\infty} \sum_{x=0}^{1} x f_{\tilde{\theta}}\left(\theta\right) p_{\tilde{x}\,|\,\tilde{\theta}}\left(x \,|\, \theta\right) \,\mathrm{d}\theta \tag{7.47}$$

$$= \int_{0}^{1} 2\theta^2 \,\mathrm{d}\theta \tag{7.48}$$

$$= \frac{2}{3}. \tag{7.49}$$

7.1.4 The Sample Mean

As explained in Sections 7.1.1, 7.1.2 and 7.1.3, the mean of a random variable can be interpreted as an average of the uncertain quantity represented by the random variable. Consequently, if we want to approximate the mean of a random variable using data, it makes sense to estimate it via averaging. In probability and statistics, this estimate is known as the sample mean, because it is a mean computed from samples.

Definition 7.14 (Sample mean)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The sample mean is the arithmetic average*

$$m(X) := \frac{\sum_{i=1}^{n} x_i}{n}. \tag{7.50}$$

Note that although the mathematical formula for the mean changes depending on whether the random variables involved are discrete or continuous, *the sample mean is always the same*. We just average the data. In Section 9.5 we show that the sample mean provides an accurate estimate of the mean for distributions with finite variance, as long as it is computed from independent samples.

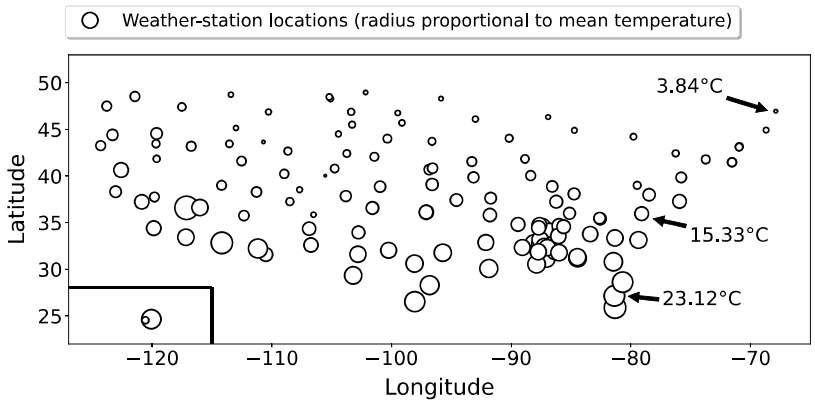

Figure 7.1 Mean temperature in the United States. The graph provides a visualization of the mean temperature at 134 weather stations in the United States in 2015. The bottom left corner shows the two stations in Hawaii (one is in Mauna Loa at high altitude, so its temperature is lower). The radius of each circular marker is proportional to the mean temperature. The mean temperature increases from north to south. The mean temperatures at Durham (North Carolina), Limestone (Maine) and Sebring (Florida) are included for reference.

Example 7.15 (Mean temperature)**.** We consider hourly temperature data measured at 134 weather stations in the United States in 2015, extracted from Dataset 9. Figure 7.1 shows the sample means of the temperature at each station. The means provide an effective summary; we clearly see that it is colder at higher latitudes, and warmer at lower latitudes.

7.2 Linearity Of Expectation

The mean is a sum or an integral weighted by probabilities or densities of probability, respectively. As a result, it is linear, as illustrated by the following example.

Example 7.16 (Cost of a latte)**.** The owner of a cafe is interested in estimating the mean cost of a latte next year. She models the price of one kilogram of coffee as a random variable $\tilde{c}$ with expected value 2.5 dollars, and the price of a gallon of milk as a random variable $\tilde{m}$ with expected value 3.5 dollars. Since a latte has

0.02 kg of coffee and 0.1 gallons of milk,

$$\begin{aligned}
\mathrm{E}[\tilde{\ell}] &= \mathrm{E}[0.02\tilde{c} + 0.1\tilde{m}] && (7.51)\\
&= \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} (0.02c + 0.1m) f_{\tilde{c},\tilde{m}}(c,m)\ \mathrm{d}c\ \mathrm{d}m && (7.52)\\
&= 0.02 \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} c f_{\tilde{c},\tilde{m}}(c,m)\ \mathrm{d}c\ \mathrm{d}m + 0.1 \int_{c\in\mathbb{R}} \int_{m\in\mathbb{R}} m f_{\tilde{c},\tilde{m}}(c,m)\ \mathrm{d}c\ \mathrm{d}m\\
&= 0.02 \int_{c\in\mathbb{R}} c f_{\tilde{c}}(c)\ \mathrm{d}c + 0.1 \int_{m\in\mathbb{R}} m f_{\tilde{m}}(m)\ \mathrm{d}m && (7.53)\\
&= 0.02\,\mathrm{E}[\tilde{c}] + 0.1\,\mathrm{E}[\tilde{m}] && (7.54)\\
&= 0.4. && (7.55)
\end{aligned}$$

The expected cost is 40 cents.

In Example 7.16, we show that $\mathrm{E}[0.02\tilde{c} + 0.1\tilde{m}] = 0.02\,\mathrm{E}[\tilde{c}] + 0.1\,\mathrm{E}[\tilde{m}]$, i.e. that the mean of the linear combination of the random variables is equal to the linear combination of their respective means. This is a fundamental property, which we use all the time when computing and manipulating the mean.

Theorem 7.17 (Linearity of expectation)**.** *For any constants $c_1, c_2 \in \mathbb{R}$, any functions $h_1, h_2 : \mathbb{R}^n \to \mathbb{R}$ and any continuous or discrete random variables $\tilde{a}$ and $\tilde{b}$,*

$$\mathrm{E}\left[c_1\, h_1(\tilde{a},\tilde{b}) + c_2\, h_2(\tilde{a},\tilde{b})\right] = c_1\,\mathrm{E}\left[h_1(\tilde{a},\tilde{b})\right] + c_2\,\mathrm{E}\left[h_2(\tilde{a},\tilde{b})\right]. \tag{7.56}$$

Proof The theorem follows immediately from the linearity of sums and integrals, just as in Example 7.16. ■

Linearity of expectation makes it possible to compute the mean of linear functions of random variables, without having to derive their joint pdf or pmf, which is usually much more complicated. We illustrate this by deriving the mean of a binomial random variable.

Lemma 7.18 (Mean of a binomial random variable)**.** *The mean of a binomial random variable $\tilde{a}$ with parameters n and θ equals $\mathrm{E}[\tilde{a}] = n\theta$.*

Proof The binomial random variable $\tilde{a}$ can be represented as a function of n independent Bernoulli random variables $\tilde{b}_1, \ldots, \tilde{b}_n$ with parameter θ, as shown in Example 2.16 (each coin flip is a Bernoulli random variable). Specifically, $\tilde{a}$ is equal to the number of Bernoulli variables that equal one, or equivalently to their sum. By linearity of expectation,

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \mathrm{E}\left[\sum_{k=1}^{n} \tilde{b}_k\right] && (7.57)\\
&= \sum_{k=1}^{n} \mathrm{E}[\tilde{b}_k] && (7.58)\\
&= n\theta. && (7.59)
\end{aligned}$$

Note that this holds even if the Bernoulli random variables are not independent. ■

7.3 Independent Random Variables

If two random variables are independent, then the mean of their product equals the product of their means.

Theorem 7.19 (Mean of the product of independent random variables)**.** *If $\tilde{a}$ and $\tilde{b}$ are independent random variables defined on the same probability space, and $g, h : \mathbb{R} \to \mathbb{R}$ are deterministic real-valued functions, then*

$$\mathrm{E}\left[g\left(\tilde{a}\right) h(\tilde{b})\right] = \mathrm{E}\left[g\left(\tilde{a}\right)\right] \mathrm{E}[h(\tilde{b})], \tag{7.60}$$

as long as all the expected values in the expression are well defined.

Proof If $\tilde{a}$ and $\tilde{b}$ are continuous, then the joint pdf factorizes into the product of the marginals (see Definition 5.15), so

$$\begin{aligned}
\mathrm{E}\left[g\left(\tilde{a}\right) h(\tilde{b})\right] &= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} g\left(a\right) h\left(b\right) f_{\tilde{a},\tilde{b}}\left(a, b\right) \, \mathrm{d}a \, \mathrm{d}b && (7.61)\\
&= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} g\left(a\right) h\left(b\right) f_{\tilde{a}}\left(a\right) f_{\tilde{b}}\left(b\right) \, \mathrm{d}a \, \mathrm{d}b && (7.62)\\
&= \mathrm{E}\left[g\left(\tilde{a}\right)\right] \mathrm{E}[h(\tilde{b})]. && (7.63)
\end{aligned}$$

The proof is the same for discrete random variables. In that case the joint pmf factorizes into the product of the marginal pmfs (see Definition 4.16). If $\tilde{a}$ is discrete and $\tilde{b}$ is continuous, then by independence the conditional pdf of $\tilde{b}$ given $\tilde{a}$ is equal to the marginal pdf of $\tilde{b}$ (see Definition 6.8), so

$$\begin{aligned}
\mathrm{E}\left[g\left(\tilde{a}\right) h(\tilde{b})\right] &= \sum_{a \in A} \int_{b=-\infty}^{\infty} g\left(a\right) h\left(b\right) p_{\tilde{a}}\left(a\right) f_{\tilde{b} \,|\, \tilde{a}}\left(b \,|\, a\right) \, \mathrm{d}b && (7.64)\\
&= \sum_{a \in A} g\left(a\right) p_{\tilde{a}}\left(a\right) \int_{b=-\infty}^{\infty} h\left(b\right) f_{\tilde{b}}\left(b\right) \, \mathrm{d}a \, \mathrm{d}b && (7.65)\\
&= \mathrm{E}\left[g\left(\tilde{a}\right)\right] \mathrm{E}[h(\tilde{b})]. && (7.66)
\end{aligned}$$

■

The following example shows that the independence assumption is crucial in Theorem 7.19.

Example 7.20 (Restaurant)**.** The owner of a restaurant wants to estimate the expected revenue per night. Looking at past data, she determines that the mean number of customers per night is 50 and the mean amount spent per customer is 40 dollars. She concludes that the expected revenue is 2,000 dollars. However, looking at the actual data, she realizes that this is not the case! This is because the number of customers and the money spent by each customer are not independent.

It turns out that a good model for the customers in the restaurant is that each

night is busy or calm with probability 1/2. In busy nights, there are exactly 80 customers and each spend 60 dollars motivated by the lively atmosphere. In calm nights, there are exactly 20 customers, and each spend 20 dollars. Consequently, the mean of the random variable $\tilde{c}$ representing the number of customers is 50, and the mean of the random variable representing the amount spent per customer $\tilde{a}$ is 40 dollars. However, the mean of the product equals

$$\begin{aligned} \mathrm{E}[\tilde{c}\tilde{a}] &= \frac{80 \cdot 60}{2} + \frac{20 \cdot 20}{2} && (7.67) \\ &= 2600. && (7.68) \end{aligned}$$

Due to the dependence between the random variables, the mean of their product is greater than the product of their means.

7.4 Mean Of Parametric Distributions

In this section, we derive the mean of different parametric distributions as a function of their parameters.

In the case of the Bernoulli distribution, the mean is just equal to the parameter. We can therefore never expect a Bernoulli random variable to equal its expected value (unless the parameter is zero or one).

Lemma 7.21 (Mean of a Bernoulli random variable)**.** *The mean of a Bernoulli random variable $\tilde{a}$ with parameter θ equals* $\mathrm{E}[\tilde{a}] = \theta$.

Proof

$$\begin{aligned} \mathrm{E}\left[\tilde{a}\right] &= 0 \cdot p_{\tilde{a}}\left(0\right) + 1 \cdot p_{\tilde{a}}\left(1\right) && (7.69) \\ &= \theta. && (7.70) \end{aligned}$$

■

In the case of a geometric random variable with parameter α, the mean is $1/\alpha$. For example, the expected number of times that we need to flip a fair coin until it lands on heads is two.

Lemma 7.22 (Mean of a geometric random variable)**.** *The mean of a geometric random variable $\tilde{a}$ with parameter α equals* $\mathrm{E}[\tilde{a}] = \frac{1}{\alpha}$.

Proof We apply the geometric series identity $\sum_{k=1}^{\infty} k r^{k-1} = (1-r)^{-2}$ for $0 < r < 1$. Setting $r = 1 - \alpha$,

$$\begin{aligned} \mathrm{E}\left[\tilde{a}\right] &= \sum_{a=1}^{\infty} a \, p_{\tilde{a}}\left(a\right) && (7.71) \\ &= \sum_{a=1}^{\infty} a \, \alpha \left(1-\alpha\right)^{a-1} && (7.72) \\ &= \frac{1}{\alpha}. && (7.73) \end{aligned}$$

■

The mean of a Poisson random variable is equal to its parameter λ. This justifies our interpretation of λ in Example 2.21 as *the total earthquakes per year that we expect to occur.*

Lemma 7.23 (Mean of a Poisson random variable)**.** *The mean of a Poisson random variable $\tilde{a}$ with parameter λ equals*

$$\mathrm{E}[\tilde{a}] = \lambda. \tag{7.74}$$

Proof By the Taylor series expansion of the exponential function, we have

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}, \tag{7.75}$$

which implies

$$\begin{aligned} \mathrm{E}\left[\tilde{a}\right] &= \sum_{a=0}^{\infty} a \, p_{\tilde{a}}\left(a\right) && (7.76)\\ &= \sum_{a=1}^{\infty} \frac{\lambda^a e^{-\lambda}}{(a-1)!} && (7.77)\\ &= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} && (7.78)\\ &= \lambda. && (7.79) \end{aligned}$$

■

The mean of an exponential random variable is the inverse of its parameter λ.

Lemma 7.24 (Mean of an exponential random variable)**.** *The mean of an exponential random variable $\tilde{a}$ with parameter λ equals $1/\lambda$.*

Proof Applying integration by parts,

$$\begin{aligned} \mathrm{E}\left[\tilde{a}\right] &= \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right) \, \mathrm{d}a && (7.80)\\ &= \int_{a=0}^{\infty} a\lambda e^{-\lambda a} \, \mathrm{d}a && (7.81)\\ &= -ae^{-\lambda a}\big]_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} \lambda e^{-\lambda a} \, \mathrm{d}a && (7.82)\\ &= \frac{1}{\lambda}. && (7.83) \end{aligned}$$

■

Reassuringly, the mean of a Gaussian random variable is equal to its mean parameter μ.

Table 7.2 ***Mean of parametric models and the corresponding maximum-likelihood estimator.*** *Parameters of several popular parametric distributions as a function of their mean. For these distributions, maximum likelihood estimation is equivalent to the method of moments, which produces a parameter estimate by plugging in the sample mean $m(X)$ into these functions.*

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

Lemma 7.25 (Mean of a Gaussian random variable)**.** *The mean of a Gaussian random variable $\tilde{a}$ with parameters μ and σ equals μ.*

Proof We apply the change of variables $t = (a - \mu)/\sigma$:

$$\mathrm{E}[\tilde{a}] = \int_{a=-\infty}^{\infty} a f_{\tilde{a}}(a)\, \mathrm{d}a \tag{7.84}$$

$$= \int_{a=-\infty}^{\infty} \frac{a}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}\, \mathrm{d}a \tag{7.85}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} t e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{\mu}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} e^{-\frac{t^2}{2}}\, \mathrm{d}t \tag{7.86}$$

$$= \mu, \tag{7.87}$$

where the last step follows from the fact that the integral of a bounded odd function over a symmetric interval is zero. ■

The parameters of all the parametric distributions that we consider in this section can be expressed as a function of their mean. This suggests a very simple approach to estimate the parameters from data: compute the sample mean and use it to replace the true mean. For the Bernoulli, Poisson and Gaussian distributions, the parameter estimate is equal to the sample mean. For the geometric and exponential distributions, the parameter estimate is the inverse of the sample mean. This parameter-estimation strategy is called the *method of moments*. For the distributions in this section, it yields the same estimate as maximum likelihood estimation (see Table 7.2).

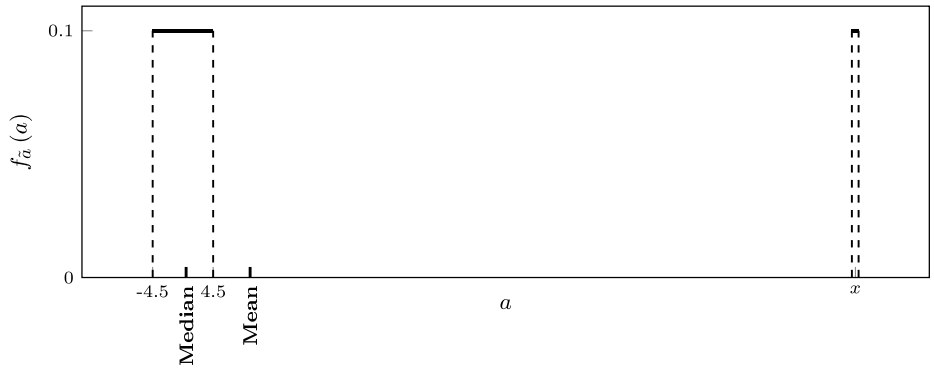

Figure 7.2 The mean can be far from the median. Distribution of the random variable in Example 7.26. The pdf is uniform in $[-4.5, 4.5] \cup [x - 0.5, x + 0.5]$. The mean is $x/10$, so for large x it is very different to the median, which always equals 0.5.

7.5 Sensitivity Of The Mean To Extreme Values

The mean is often interpreted as representing a *typical* value taken by a random variable. However, it can be severely distorted by a small subset of extreme values. In such cases, the median is a good alternative, due to its robustness to outliers.

Example 7.26 (Mean vs. median)**.** Consider a random variable $\tilde{a}$ that is uniformly distributed in $[-4.5, 4.5]$, but can also belong to a small interval of extreme values $[x - 0.5, x + 0.5]$, where x is a large constant. The mean of $\tilde{a}$ equals

$$\mathrm{E}[\tilde{a}] = \int_{a=-4.5}^{4.5} a f_{\tilde{a}}(a)\, \mathrm{d}a + \int_{a=x-0.5}^{x+0.5} a f_{\tilde{a}}(a)\, \mathrm{d}a \tag{7.88}$$

$$= \frac{1}{10} \frac{(x + 0.5)^2 - (x - 0.5)^2}{2} \tag{7.89}$$

$$= \frac{x}{10}. \tag{7.90}$$

The mean scales linearly with x. It is completely driven by the most extreme values that $\tilde{a}$ can take, even though they only represent 10% of its range. In contrast, the median ignores these values. The cdf of $\tilde{a}$ between -4.5 and 4.5 is equal to

$$F_{\tilde{a}}(q) = \int_{-4.5}^{q} f_{\tilde{a}}(a)\, \mathrm{d}a \tag{7.91}$$

$$= \frac{q + 4.5}{10}. \tag{7.92}$$

Setting this equal to 1/2 shows that the median is equal to 0.5, with no dependence on x. Figure 7.2 shows the pdf of $\tilde{a}$ and the location of the median and the

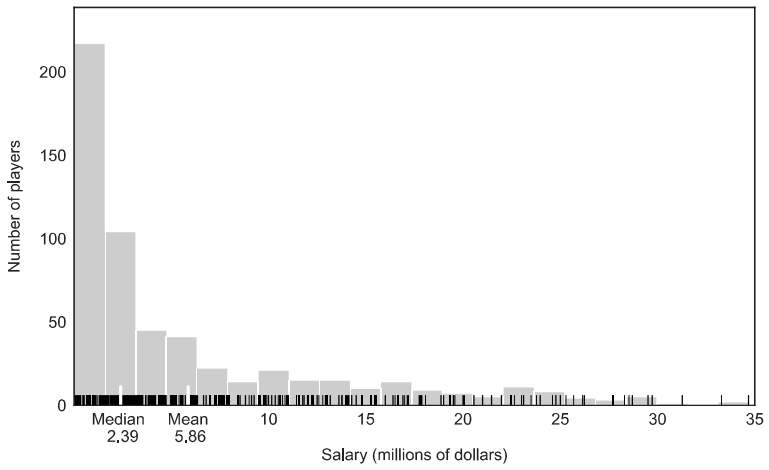

Figure 7.3 Salaries of NBA players in the 2017/2018 season. The plot shows a rug plot and a histogram of the salaries, as well as their sample mean and their median (computed using the empirical pmf). Less than one third of the players (32.1%) earn more than the mean salary.

mean. The median clearly provides a more realistic measure of the center of the distribution.

Example 7.27 (NBA salaries)**.** Figure 7.3 shows a real-world example where the mean is distorted by extreme values. The data are salaries of NBA players in the 2017/2018 season (Dataset 14). The median salary (2.39 million) is less than half the mean (5.86 million). Less than one third of the players (32.1%) earn more than the mean. The mean is therefore not a good characterization of a *typical* salary; it is inflated by the massive salaries of the best-paid players.

7.6 The Mean Square

The mean square provides a measure of the *magnitude* of a random variable. We often use squaring to quantify the magnitude of deterministic quantities. The magnitude of a scalar is the square root of its square. The length or Euclidean norm of a vector is the square root of the sum of its squared entries. Analogously, we use the mean of the square of a random variable to quantify its magnitude or energy. This mean aggregates all the possible squared values of the random variable, preventing cancellations.

Definition 7.28 (Mean square)**.** *The mean square or second moment of a random variable $\tilde{a}$ is the expected value of $\tilde{a}^2$:* $\mathrm{E}\left[\tilde{a}^2\right]$.

The mean square can be applied to quantify the difference between two random variables. In particular, we can use it as a metric to evaluate estimators. This metric is called the mean squared error.

Definition 7.29 (Mean squared error)**.** *The mean squared error (MSE) between an estimator $\tilde{e}$ and a random variable $\tilde{a}$ is* $\mathrm{E}[(\tilde{a} - \tilde{e})^2]$.

An interesting property of the mean of a random variable is that it is the constant that best approximates the random variable in terms of MSE.

Theorem 7.30 (Constant minimum MSE estimator)**.** *For any random variable $\tilde{a}$ with mean* $\mathrm{E}[\tilde{a}]$*,*

$$\mathrm{E}\left[\tilde{a}\right] = \arg\min_{c \in \mathbb{R}} \mathrm{E}\left[(c - \tilde{a})^2\right]. \tag{7.93}$$

Proof Let $\mathrm{MSE}(c) := \mathrm{E}\left[(c - \tilde{a})^2\right] = c^2 - 2c\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}^2\right]$ denote the mean squared error as a function of the constant estimate c. The derivatives of the MSE with respect to c equal

$$\frac{\mathrm{d}\,\mathrm{MSE}(c)}{\mathrm{d}c} = 2(c - \mathrm{E}[\tilde{a}]), \tag{7.94}$$

$$\frac{\mathrm{d}^2\,\mathrm{MSE}(c)}{\mathrm{d}c^2} = 2. \tag{7.95}$$

The error is strictly convex because the second derivative is positive. Its minimum is attained at $c = \mathrm{E}[\tilde{a}]$, where the first derivative is zero. ■

When performing estimation of uncertain quantities from data, we typically approximate the mean squared error using the average of the squared errors. Minimizing this cost function is known as *least-squares* estimation. The following lemma is the finite-data counterpart to Theorem 7.30; it establishes that the sample mean is the best constant least-squares estimator.

Theorem 7.31 (Constant least-squares estimator)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a real-valued dataset. The sample mean of the dataset is the constant estimate that minimizes the residual sum of squares:*

$$m(X) = \arg\min_{c} \sum_{i=1}^{n} (x_i - c)^2. \tag{7.96}$$

Proof We follow the same steps as in the proof of Theorem 7.30. As a function of the constant estimate c the residual sum of squares equals

$$\mathrm{RSS}(c) := \sum_{i=1}^{n} (x_i - c)^2 \tag{7.97}$$

$$= nc^2 - 2c\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2. \tag{7.98}$$

Its derivatives are

$$\frac{\mathrm{d}\,\mathrm{RSS}(c)}{\mathrm{d}c} = 2\left(nc - \sum_{i=1}^{n} x_i\right), \tag{7.99}$$

$$\frac{\mathrm{d}^2\,\mathrm{RSS}(c)}{\mathrm{d}c^2} = 2n. \tag{7.100}$$

The error is strictly convex, so the minimum is at $c = m(X)$, where the first derivative is zero. ■

7.7 The Variance

7.7.1 Definition

The mean square of the difference between a random variable and its mean is called the *variance* of the random variable. It quantifies the variation of the random variable around its mean. The square root of this quantity can be interpreted as an average deviation from the mean; it is called the *standard deviation* of the random variable.

Definition 7.32 (Variance and standard deviation)**.** *The variance of a random variable $\tilde{a}$ is the mean square deviation from the mean,*

$$\mathrm{Var}\left[\tilde{a}\right] := \mathrm{E}\left[\left(\tilde{a} - \mathrm{E}\left[\tilde{a}\right]\right)^2\right]. \tag{7.101}$$

It is sometimes referred to as the second central moment of $\tilde{a}$. The standard deviation $\sigma_{\tilde{a}}$ of $\tilde{a}$ is the square root of the variance,

$$\sigma_{\tilde{a}} := \sqrt{\mathrm{Var}\left[\tilde{a}\right]}. \tag{7.102}$$

Just as in the case of the mean, it is possible for a random variable to have infinite variance, or to not have a well-defined variance.

The following lemma provides a convenient way to compute the variance: subtracting the squared mean from the mean square.

Lemma 7.33. *For any random variable $\tilde{a}$ with finite variance,*

$$\mathrm{Var}\left[\tilde{a}\right] = \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2. \tag{7.103}$$

Proof By linearity of expectation,

$$\mathrm{Var}\left[\tilde{a}\right] := \mathrm{E}\left[\left(\tilde{a} - \mathrm{E}\left[\tilde{a}\right]\right)^2\right] \tag{7.104}$$

$$= \mathrm{E}\left[\tilde{a}^2 - 2\tilde{a}\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}\right]\right]^2 \tag{7.105}$$

$$= \mathrm{E}\left[\tilde{a}^2\right] - 2\mathrm{E}\left[\tilde{a}\right]\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}\right]^2 \tag{7.106}$$

$$= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2. \tag{7.107}$$

■

The following lemma derives the variance of a uniform random variable in the unit interval. The standard deviation turns out to be slightly more than one fourth ($1/\sqrt{12} \approx 0.289$) of the length of the interval.

Lemma 7.34 (Variance of a uniform random variable)**.** *The variance of a random variable $\tilde{u}$ that is uniformly distributed in the interval $[a, b]$, $b > a$, equals*

$$\operatorname{Var}[\tilde{u}] = \frac{(b-a)^2}{12}. \tag{7.108}$$

Proof The mean square equals

$$\begin{aligned}
\mathrm{E}\left[\tilde{u}^2\right] &= \int_{u=-\infty}^{\infty} u^2 f_{\tilde{a}}(u)\, \mathrm{d}u && (7.109)\\
&= \int_{u=a}^{b} \frac{u^2}{b-a}\, \mathrm{d}u && (7.110)\\
&= \frac{b^3 - a^3}{3(b-a)} && (7.111)\\
&= \frac{a^2 + ab + b^2}{3}. && (7.112)
\end{aligned}$$

Subtracting $\mathrm{E}[\tilde{u}]^2 = \left(\frac{a+b}{2}\right)^2$ (see Lemma 7.8) from the mean square yields the result by Lemma 7.33. ■

Unlike the mean, the variance is not linear. The following lemma shows what happens to the variance of a random variable when we scale and shift it.

Lemma 7.35 (Variance of scaled, shifted random variable)**.** *For any constants c_1 and c_2*

$$\operatorname{Var}\left[c_1\tilde{a} + c_2\right] = c_1^2 \operatorname{Var}\left[\tilde{a}\right]. \tag{7.113}$$

Proof

$$\begin{aligned}
\operatorname{Var}\left[c_1\tilde{a} + c_2\right] &= \mathrm{E}\left[\left(c_1\tilde{a} + c_2 - \mathrm{E}\left[c_1\tilde{a} + c_2\right]\right)^2\right] && (7.114)\\
&= \mathrm{E}\left[\left(c_1\tilde{a} + c_2 - c_1\mathrm{E}\left[\tilde{a}\right] - c_2\right)^2\right] && (7.115)\\
&= c_1^2\, \mathrm{E}\left[\left(\tilde{a} - \mathrm{E}\left[\tilde{a}\right]\right)^2\right] && (7.116)\\
&= c_1^2 \operatorname{Var}\left[\tilde{a}\right]. && (7.117)
\end{aligned}$$

■

The result is very intuitive. If we shift the random variable by adding a constant, then the variance is not affected because the variance only measures the deviation from the mean. If we multiply a random variable by a constant, the standard deviation is scaled accordingly.

7.7.2 The Sample Variance

In order to estimate the variance from data, we compute the average squared deviation from the sample mean. The resulting estimator is called the sample variance.

Definition 7.36 (The sample variance and sample standard deviation)**.** *Let* $X := \{x_1, x_2, \ldots, x_n\}$ *denote a real-valued dataset. The sample variance is the average squared deviation from the sample mean,*

$$v(X) := \frac{\sum_{i=1}^{n} (x_i - m(X))^2}{n-1}. \tag{7.118}$$

The square root of the sample variance is called the sample standard deviation.

You may have noticed that we divide by $n-1$ when computing the average, instead of by n. This is to ensure that the estimator is unbiased, when we compute it from independent samples, as shown in Theorem 9.8. In practice, it does not make much of a difference.

Example 7.37 (Variance of temperature)**.** Figure 7.4 shows the sample variance of the temperature at three weather stations in the United States in 2015, extracted from Dataset 9. The mean temperature at the three stations is essentially the same, but their probability densities look very different. This is captured by the respective variances. A higher variance indicates that the pdf is more spread out, because the temperature is farther from its mean on average.

Figure 7.5 shows the sample standard deviation of the temperature in weather stations all over the United States, complementing the map of the means in Figure 7.1. This provides a holistic summary of temperature deviations, showing that coastal locations have very small variance, and that the variance increases as we move towards the interior, especially in the north.

7.7.3 Variance Of Parametric Distributions

In this section we derive the variance of several popular parametric distributions. Figure 7.6 provides a visual summary of these results by displaying the range of values that fall within one standard deviation of the mean for each of the distributions.

Lemma 7.38 (Variance of a Bernoulli random variable)**.** *The variance of a Bernoulli random variable* $\tilde{a}$ *with parameter* θ *equals* $\mathrm{Var}[\tilde{a}] = \theta(1-\theta)$*.*

Proof The mean square equals

$$\mathrm{E}\left[\tilde{a}^2\right] = 0 \cdot p_{\tilde{a}}(0) + 1 \cdot p_{\tilde{a}}(1) \tag{7.119}$$

$$= \theta. \tag{7.120}$$

The result then follows from Lemmas 7.33 and 7.21. ■

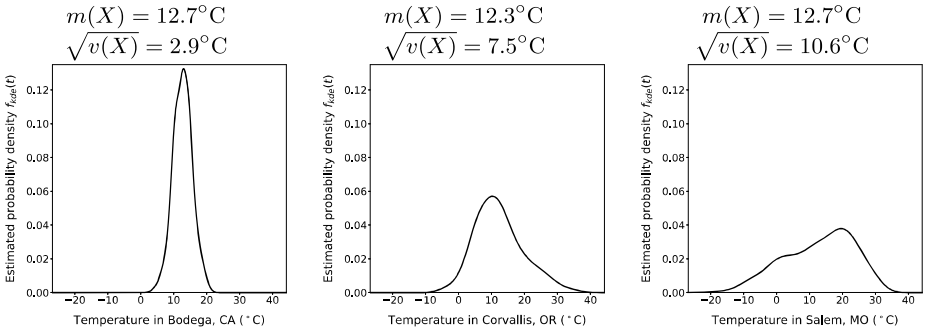

Figure 7.4 Distributions with similar mean but different variances. The plots show the pdfs of the temperature in Bodega (California), Corvallis (Oregon), and Salem (Missouri) in 2015, estimated via kernel density estimation (see Section 3.5.2). The mean temperatures at the three stations are almost the same, but their standard deviations are very different, which is reflected in their pdfs.

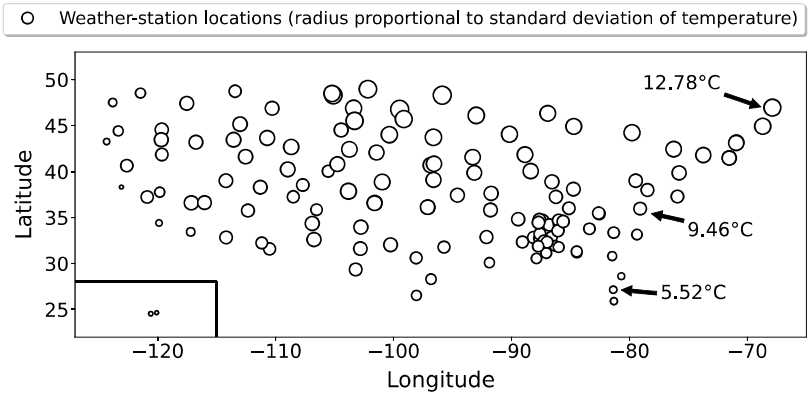

Figure 7.5 Standard deviation of the temperature in the United States. The graph provides a visualization of the standard deviation of the temperature at 134 weather stations in the United States in 2015. The bottom left corner shows the two stations in Hawaii. The radius of each circular marker is proportional to the standard deviation. The standard deviation is lower at stations situated on the coast, especially on the West Coast, Hawaii and Florida. The standard deviations of the temperatures at Durham (North Carolina), Limestone (Maine) and Sebring (Florida) are included for reference.

The variance of a Bernoulli random variable is maximized when $\theta = 0.5$ (you can verify this by taking derivatives with respect to θ). This makes sense, as it maximizes the variability of the outcomes.

The standard deviation of a geometric random variable is close to its mean.

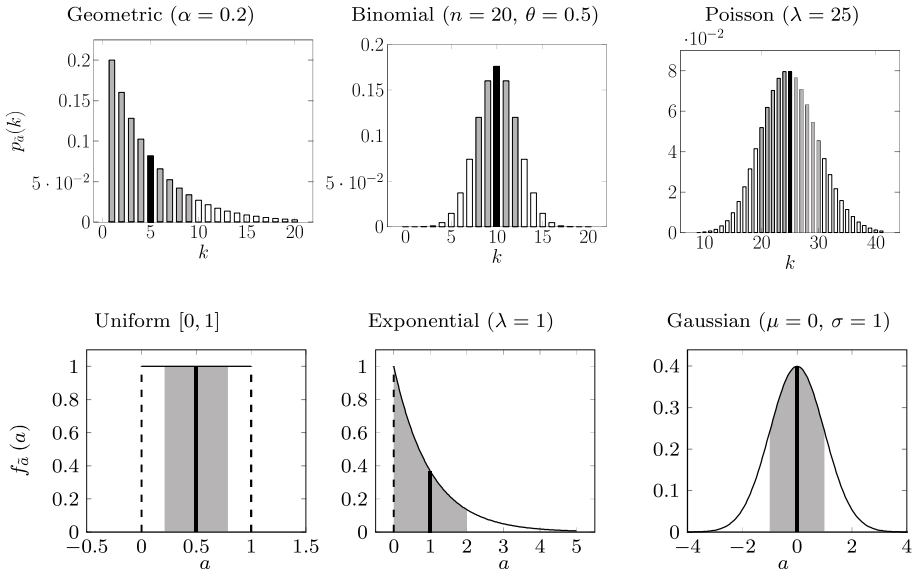

Figure 7.6 Mean and standard deviation of popular distributions. The graphs show the pmfs of several discrete distributions (top row) and the pdfs of several continuous distributions (bottom row). The mean of the random variable is highlighted in black. Values within one standard deviation of the mean are shaded in gray. The means and variances of these distributions are derived in Sections 7.4 and 7.7.3, except for the mean and variance of the binomial, which are derived in Lemmas 7.18 and 9.12.

Lemma 7.39 (Variance of a geometric random variable)**.** *The variance of a geometric random variable $\tilde{a}$ with parameter α equals*

$$\text{Var}[\tilde{a}] = \frac{1-\alpha}{\alpha^2}. \tag{7.121}$$

Proof To derive the second moment we apply the geometric series identity $\sum_{k=1}^{\infty} k^2 r^k = r(1+r)(1-r)^{-3}$ for $0 < r < 1$. Setting $r = 1 - \alpha$, we have

$$\begin{aligned} \text{E}\left[\tilde{a}^2\right] &= \sum_{a=1}^{\infty} a^2 \, p_{\tilde{a}}(a) && (7.122)\\ &= \sum_{a=1}^{\infty} a^2 \, \alpha \, (1-\alpha)^{a-1} && (7.123)\\ &= \frac{\alpha}{1-\alpha} \sum_{a=1}^{\infty} a^2 \, (1-\alpha)^{a} && (7.124)\\ &= \frac{2-\alpha}{\alpha^2}. && (7.125) \end{aligned}$$

The result then follows from Lemmas 7.33 and 7.22. ■

An interesting property of the Poisson distribution is that the mean is equal to the variance.

Lemma 7.40 (Variance of a Poisson random variable)**.** *The variance of a Poisson random variable $\tilde{a}$ with parameter λ equals λ.*

Proof The second moment equals

$$\mathrm{E}\left[\tilde{a}^2\right] = \sum_{a=1}^{\infty} a^2 \, p_{\tilde{a}}\left(a\right) \tag{7.126}$$

$$= \sum_{a=1}^{\infty} \frac{a\lambda^a e^{-\lambda}}{(a-1)!} \tag{7.127}$$

$$= e^{-\lambda}\left(\sum_{a=2}^{\infty} \frac{(a-1)\,\lambda^a}{(a-1)!} + \sum_{a=1}^{\infty} \frac{\lambda^a}{(a-1)!}\right) \tag{7.128}$$

$$= e^{-\lambda}\left(\sum_{m=0}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!}\right) \tag{7.129}$$

$$= \lambda^2 + \lambda. \tag{7.130}$$

The result then follows from Lemmas 7.33 and 7.23. ■

In the case of the exponential random variable, the standard deviation is equal to the mean.

Lemma 7.41 (Variance of an exponential random variable)**.** *The variance of an exponential random variable $\tilde{a}$ with parameter λ equals $1/\lambda^2$.*

Proof Applying integration by parts,

$$\mathrm{E}\left[\tilde{a}^2\right] = \int_{a=-\infty}^{\infty} a^2 f_{\tilde{a}}\left(a\right) \, \mathrm{d}a \tag{7.131}$$

$$= \int_{a=0}^{\infty} a^2 \lambda e^{-\lambda a} \, \mathrm{d}a \tag{7.132}$$

$$= -a^2 e^{-\lambda a}\big]_0^{\infty} + 2\frac{1}{\lambda}\int_0^{\infty} a\lambda e^{-\lambda a} \, \mathrm{d}a \tag{7.133}$$

$$= \frac{2}{\lambda^2}. \tag{7.134}$$

The result then follows from Lemmas 7.33 and 7.24. ■

Example 7.42 (Variance of call center data)**.** According to Lemma 7.41, quantities modeled as exponential random variables should have a standard deviation that is similar to their mean. In Example 3.29, we fit an exponential random variable to data consisting of inter-arrival times of calls at a call center (Dataset 3). For calls arriving at the bank call center on weekdays from 9 am to 10 am, the sample mean equals 30.8 and the sample standard deviation is 33.6, which is indeed quite close.

The variance of a Gaussian random variable is equal to its variance parameter. It is therefore not very surprising that the maximum-likelihood estimator of this parameter is almost equal to the sample variance (compare Theorem 3.36 and Definition 7.36). The only difference is that the sum is normalized by n instead of $n-1$, which has a negligible effect as long as n is not very small.

Lemma 7.43 (Variance of a Gaussian random variable)**.** *The variance of a Gaussian random variable $\tilde{a}$ with parameters μ and σ equals σ^2.*

Proof We apply the change of variables $t = (a-\mu)/\sigma$ and integrate by parts,

$$\mathrm{E}\left[\tilde{a}^2\right] = \int_{a=-\infty}^{\infty} a^2 f_{\tilde{a}}(a)\, \mathrm{d}a \tag{7.135}$$

$$= \int_{a=-\infty}^{\infty} \frac{a^2}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}\, \mathrm{d}a \tag{7.136}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} t e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{\mu^2}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \left(-t e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{t=-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \right) + \mu^2 \tag{7.137}$$

$$= \sigma^2 + \mu^2. \tag{7.138}$$

The result then follows from Lemmas 7.33 and 7.25. ■

7.8 The Conditional Mean

The conditional mean of a random variable represents how its average value varies depending on the value of other random variables. In Section 7.8.1 we define the conditional mean formally and explain how to estimate it from data. Section 7.8.2 explains how to analyze the probabilistic behavior of the conditional mean, motivated by iterated expectation, an identity that allows us to compute the mean from the conditional mean. Section 7.8.3 introduces the problem of regression, and shows that the conditional mean is an optimal estimator in terms of mean squared error.

7.8.1 The Conditional Mean Function

The conditional mean function of a random variable $\tilde{b}$ given another random variable $\tilde{a}$ is equal to the mean of $\tilde{b}$ conditioned on the event $\tilde{a} = a$. It can be computed using the conditional distribution of $\tilde{b}$ given $\tilde{a}$.

Definition 7.44 (The conditional mean function)**.** *Let $\tilde{a}$ be a random variable or a random vector, and let $\tilde{b}$ be another random variable belonging to the same probability space. The conditional mean function $\mu_{\tilde{b}\,|\,\tilde{a}}(a)$ of $\tilde{b}$ given $\tilde{a} = a$ is the mean of $\tilde{b}$ computed according to the conditional distribution of $\tilde{b}$ given $\tilde{a} = a$,*

where a is in the range of $\tilde{a}$. If $\tilde{b}$ is a discrete random variable, then

$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) := \sum_{b\in B} b\, p_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)\,, \tag{7.139}$$

where B is the range of $\tilde{b}$ conditioned on $\tilde{a} = a$. If $\tilde{b}$ is a continuous random variable with conditional pdf $f_{\tilde{b}\,|\,\tilde{a}}$, then

$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) := \int_{b=-\infty}^{\infty} b\, f_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)\;\mathrm{d}b. \tag{7.140}$$

The conditional mean function $\mu_{\tilde{b}\,|\,\tilde{a}}(a)$ is only defined for values of a for which $\tilde{a}$ has nonzero probability (if it is discrete) or nonzero density (if it is continuous).

In addition, for any deterministic function $h : \mathbb{R} \to \mathbb{R}$, we define the conditional mean function of $h(\tilde{a}, \tilde{b})$ given $\tilde{a} = a$ as

$$\mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(a) := \sum_{b\in B} h(a,b)\, p_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)\,, \tag{7.141}$$

if $\tilde{b}$ is discrete, and

$$\mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(a) := \int_{b=-\infty}^{\infty} h(a,b)\, f_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)\;\mathrm{d}b, \tag{7.142}$$

if $\tilde{b}$ is continuous.

The conditional mean $\mu_{\tilde{b}\,|\,\tilde{a}}(a)$ is often denoted by $\mathrm{E}[\tilde{b}\,|\,\tilde{a} = a]$ in the literature. We have chosen our notation to emphasize that $\mu_{\tilde{b}\,|\,\tilde{a}}(a)$ is a *deterministic function* of a.

Example 7.45 (Simple example: Conditional mean function)**.** Consider the random variables $\tilde{a}$ and $\tilde{b}$ in Example 4.6. In order to compute the conditional mean function of $\tilde{b}$ given $\tilde{a} = a$, we first need to derive the conditional pmf of $\tilde{b}$ given $\tilde{a}$. As in Example 4.15, we divide the joint pmf by the marginal pmf of $\tilde{a}$ to obtain

$$p_{\tilde{b}\,|\,\tilde{a}}(1\,|\,1) = \frac{1}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(2\,|\,1) = \frac{4}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(3\,|\,1) = \frac{2}{7} \tag{7.143}$$

$$p_{\tilde{b}\,|\,\tilde{a}}(1\,|\,2) = \frac{2}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(2\,|\,2) = \frac{1}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(3\,|\,2) = \frac{4}{7} \tag{7.144}$$

$$p_{\tilde{b}\,|\,\tilde{a}}(1\,|\,3) = \frac{1}{3}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(2\,|\,3) = \frac{1}{3}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(3\,|\,3) = \frac{1}{3}. \tag{7.145}$$

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

$$\mu_{\tilde{b}\,|\,\tilde{a}}(1) = \sum_{b\in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,1) = \frac{15}{7}, \tag{7.146}$$

$$\mu_{\tilde{b}\,|\,\tilde{a}}(2) = \sum_{b\in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,2) = \frac{16}{7}, \tag{7.147}$$

$$\mu_{\tilde{b}\,|\,\tilde{a}}(3) = \sum_{b\in B} b\, p_{\tilde{b}\,|\,\tilde{a}}(b\,|\,3) = 2. \tag{7.148}$$

Example 7.46 (Triangle lake: Conditional mean function)**.** In Example 5.5 the position of an otter is modeled as being uniformly distributed in a triangular lake. In Example 5.13 we derive the conditional pdf of the vertical coordinate $\tilde{b}$ given the horizontal coordinate $\tilde{a}$,

$$f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a) = \frac{1}{1-a} \qquad 0 \leq b \leq 1-a. \tag{7.149}$$

The conditional mean function of $\tilde{b}$ given $\tilde{a} = a \in [0,1]$ is equal to

$$\mu_{\tilde{b}\,|\,\tilde{a}}(a) = \int_{b\in\mathbb{R}} b f_{\tilde{b}\,|\,\tilde{a}}(b\,|\,a)\ \mathrm{d}b \tag{7.150}$$

$$= \int_{b=0}^{1-a} \frac{b}{1-a}\ \mathrm{d}b \tag{7.151}$$

$$= \frac{(1-a)^2}{2(1-a)} \tag{7.152}$$

$$= \frac{1-a}{2}. \tag{7.153}$$

The function is shown in the left plot of Figure 7.8.

In order to estimate the conditional mean function from real data, we use the sample conditional mean, which is a conditional average.

Definition 7.47 (Sample conditional mean function of discrete data)**.** *Consider a dataset $\mathcal{D}$ formed by pairs of the form (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where x_i belongs to a discrete set A for $1 \leq i \leq n$. We interpret the samples as realizations from a pair of random variables $\tilde{a}$ and $\tilde{b}$ belonging to the same probability space. For any $a \in A$, let*

$$Y_a := \{y : (a, y) \in \mathcal{D}\} \tag{7.154}$$

denote the bag of values of y for all pairs of data points (x, y) for which $x = a$. The sample conditional mean function is the average of Y_a as a function of a,*

$$\widehat{m}_{\tilde{b}\,|\,\tilde{a}}(a) := \frac{1}{n_a} \sum_{y\in Y_a} y, \tag{7.155}$$

where n_a denotes the number of elements of Y_a.

Example 7.48 (Movie ratings: Sample conditional mean function)**.** The conditional mean function provides a summary of the dependence between two quantities. The left plot in Figure 7.7 shows the sample conditional mean function of the ratings for the movie Independence Day given the rating for the movie Mission Impossible. The function is computed following Definition 7.47 using the movie-rating data in Example 4.8 extracted from Dataset 8. In the plot, we can see that people that like Mission Impossible rate Independence Day more highly on average than those who don't.

*Y_a does not satisfy the mathematical definition of a set, because it can contain duplicate values.

When approximating the conditional mean of a random variable $\tilde{b}$ given the value of a *continuous* random variable $\tilde{a}$, we encounter a challenge. The sample conditional mean given $\tilde{a} = a$ is the average of the data points for which $\tilde{a}$ equals a (the set Y_a in (7.154)). The problem is that, for continuous random variables, a takes uncountably infinite possible values, but the available data are finite. Consequently, Y_a is empty for most values of a! To overcome this issue, we have two options. The first is to approximate the conditional pdf or pmf of $\tilde{b}$ (for instance, via kernel density estimation, as in Example 5.14) and use it to compute the conditional mean. The second is to average over data points for which $\tilde{a}$ is close to a instead of exactly equal to a, assuming that the conditional mean function is relatively smooth. This smoothness assumption is governed by a parameter ϵ, which has an analogous role to the histogram bin width or the kernel bandwidth in density estimation (see Section 3.5).

Definition 7.49 (Sample conditional mean of continuous data)**.** *Consider a dataset $\mathcal{D}$ formed by pairs of the form (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where x_i and y_i, $1 \leq i \leq n$, are real values. We interpret the samples as realizations from a pair of random variables $\tilde{a}$ and $\tilde{b}$ belonging to the same probability space. For any $a \in \mathbb{R}$ and a fixed ϵ, let*

$$Y_{a,\epsilon} := \{y : (x, y) \in \mathcal{D} \text{ for } |x - a| \leq \epsilon\} \tag{7.156}$$

denote the values taken by the second variable when the first one is close to a (at a distance of less than ϵ). The sample conditional mean is the average of $Y_{a,\epsilon}$ as a function of a,

$$\widehat{m}_{\tilde{b}\,|\,\tilde{a}}(a) := \frac{1}{n_a} \sum_{y \in Y_{a,\epsilon}} y, \tag{7.157}$$

where n_a denotes the number of elements of $Y_{a,\epsilon}$.

Example 7.50 (Sample conditional mean temperature)**.** The right plot in Figure 7.7 shows the sample conditional mean function of the temperature in Versailles (Kansas) given the temperature in Corvallis (Oregon), computed using hourly temperature measurements in 2015 extracted from Dataset 9. We set the ϵ parameter in Definition 7.49 to equal one degree. Between 0° and 20°, the average temperature in Versailles grows proportionally to the temperature in Corvallis. However, beyond 20°, the sample conditional mean is essentially constant. Below 0°, the sample conditional mean is quite noisy because of the small number of data in that region.

7.8.2 The Conditional Mean And Iterated Expectation

In some situations, the conditional mean function $\mu_{\tilde{b}\,|\,\tilde{a}}$ of a random variable $\tilde{b}$ given another random variable $\tilde{a}$ is much easier to derive than the mean of $\tilde{b}$ (see for instance Examples 7.57 and 7.58 below). It is therefore very useful to

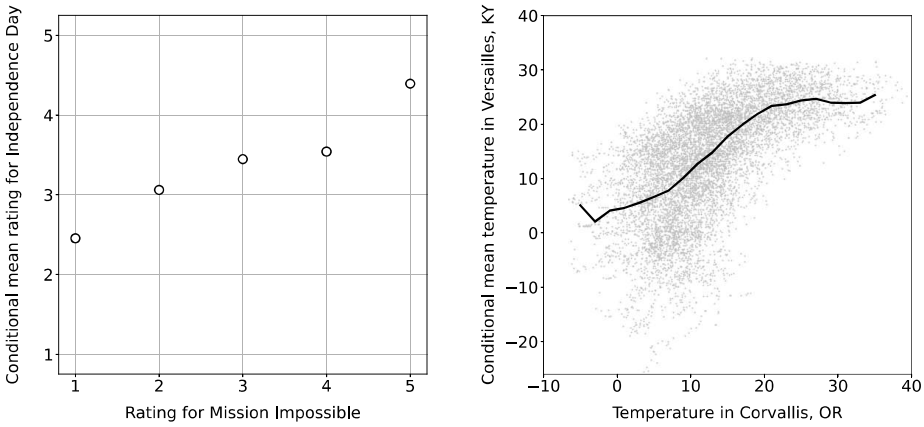

Figure 7.7 Estimating the conditional mean function from data. The left plot shows the sample conditional mean of the rating for Independence Day given the rating for Mission Impossible, computed following Definition 7.47. The right plot shows the sample conditional mean of the temperature in Versailles (Kansas) given the temperature in Corvallis (Oregon), computed following Definition 7.49 with $\epsilon := 1°\text{C}$.

be able to compute the mean of $\tilde{b}$ from $\mu_{\tilde{b} \mid \tilde{a}}$. This can be achieved via iterated expectation.

To motivate iterated expectation, we begin with an intuitive derivation. Consider a dataset (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), where we interpret the data as joint realizations of the random variables $\tilde{a}$ and $\tilde{b}$. As explained in Section 7.1, the mean of a random variable is defined to represent the average of a large number of realizations of the random variable. Consequently, the conditional mean function of $\tilde{b}$ given $\tilde{a} = a$ represents the average of the second entry of our data, when the first entry equals a:

$$\mu_{\tilde{b} \mid \tilde{a}}(a) \approx \frac{1}{n_a} \sum_{y \in Y_a} y, \tag{7.158}$$

where Y_a is the bag of values of y for all pairs of data points (x, y) for which $x = a$, as in Definition 7.47, and n_a is the number of elements of Y_a.

Now, let us express the mean of $\tilde{b}$ in terms of the conditional mean function. Assuming that $\tilde{a}$ is discrete with range A, notice that each y_i, $1 \leq i \leq n$, belongs to exactly one of the bags Y_a (the one for which $x_i = a$), so

$$\text{E}[\tilde{b}] \approx \frac{1}{n} \sum_{i=1}^{n} y_i \tag{7.159}$$

$$= \frac{1}{n} \sum_{a \in A} \sum_{y \in Y_a} y. \tag{7.160}$$

According to our intuitive definition of probability in (1.1), for large n, the probability $p_{\tilde{a}}(a)$ that $\tilde{a} = a$ equals n_a/n (the fraction of first entries equal to a). Therefore,

$$\mathrm{E}[\tilde{b}] = \sum_{a \in A} \frac{n_A}{n} \frac{1}{n_A} \sum_{y \in Y_a} y \tag{7.161}$$

$$\approx \sum_{a \in A} \frac{n_A}{n} \mu_{\tilde{b} \,|\, \tilde{a}}(a) \tag{7.162}$$

$$\approx \sum_{a \in A} p_{\tilde{a}}(a) \mu_{\tilde{b} \,|\, \tilde{a}}(a). \tag{7.163}$$

The mean of $\tilde{b}$ is equal to the sum of the conditional mean function at every possible value of $\tilde{a}$, weighted by the corresponding probability. This is known as iterated expectation, because we are first computing the mean or expected value of $\tilde{b}$ conditioned on $\tilde{a}$, and then we are computing the expected value of the conditional mean.

If $\tilde{a}$ is continuous, a similar argument (involving a grid and limits as in the intuitive derivation of the mean of a continuous random variable in Section 7.1.2), yields an expression for the mean of $\tilde{b}$ as a weighted integral of the conditional mean function,

$$\mathrm{E}[\tilde{b}] \approx \int_{a=-\infty}^{\infty} f_{\tilde{a}}(a) \mu_{\tilde{b} \,|\, \tilde{a}}(a) \, \mathrm{d}a. \tag{7.164}$$

The conditional mean function $\mu_{\tilde{b} \,|\, \tilde{a}}(a)$ is a deterministic function of a, as explained in Section 7.8.1. Consequently, setting $h := \mu_{\tilde{b} \,|\, \tilde{a}}$ in Definitions 7.3 and 7.7, we can interpret (7.163) and (7.164) as the mean of the random variable $\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})$, obtained by plugging $\tilde{a}$ into the conditional mean function. This random variable is known as the conditional mean of $\tilde{b}$ given $\tilde{a}$. It describes the conditional average behavior of $\tilde{b}$ given $\tilde{a}$ when $\tilde{a}$ is *unknown*.

Definition 7.51 (The conditional mean)**.** *Let $\tilde{a}$ be a random variable or a random vector, and let $\tilde{b}$ be another random variable belonging to the same probability space. The conditional mean of $\tilde{b}$ given $\tilde{a}$ is the random variable $\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})$ obtained by plugging $\tilde{a}$ into the conditional mean function from Definition 7.44.*

A common notation for $\mu_{\tilde{b} \,|\, \tilde{a}}(\tilde{a})$ in the literature is $\mathrm{E}[\tilde{b} \,|\, \tilde{a}]$. We have chosen our notation to emphasize that the conditional mean is a random variable, *not an expected value*. Consequently, we describe its behavior using its pmf if it is discrete, or its cdf or pdf if it is continuous.

Example 7.52 (Simple example: Conditional mean)**.** In Example 7.45 the conditional mean function only takes three different values: 15/7, 16/7 and 2, depending on whether $\tilde{a}$ is equal to 1, 2 or 3, respectively. It is straightforward to

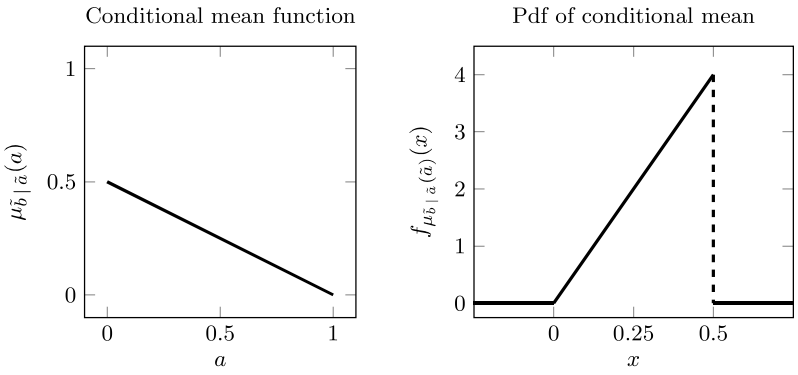

Figure 7.8 Conditional mean of the vertical coordinate in the triangle lake. The black line in the left plot is the conditional mean function of the vertical coordinate of the otter in Example 5.5 given the horizontal coordinate, as derived in Example 7.46. The pdf of the conditional mean, computed in Example 7.53, is shown on the right.

compute the pmf of $\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})$ from the marginal pmf of $\tilde{a}$ derived in Example 4.11:

$$p_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}\left(\frac{15}{7}\right) = \mathrm{P}\left(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) = \frac{15}{7}\right) \tag{7.165}$$

$$= \mathrm{P}\left(\tilde{a} = 1\right) \tag{7.166}$$

$$= 0.35. \tag{7.167}$$

Similarly,

$$p_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}\left(\frac{16}{7}\right) = 0.35, \tag{7.168}$$

$$p_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}\left(2\right) = 0.3. \tag{7.169}$$

.....

Example 7.53 (Triangle lake: Conditional mean)**.** To obtain the conditional mean of the otter's vertical coordinate $\tilde{b}$ given the horizontal coordinate $\tilde{a}$, we plug $\tilde{a}$ into the conditional mean function derived in Example 7.46:

$$\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) = \frac{1-\tilde{a}}{2}. \tag{7.170}$$

This is a continuous random variable. Since the density of $\tilde{a}$ is nonzero only between 0 and 1, we only need to consider values between 0 and 0.5. Its cdf and

pdf can be computed using the marginal pdf of $\tilde{a}$ derived in Example 5.9:

$$
\begin{aligned}
F_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}(x) &= \mathrm{P}\left(\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a}) \leq x\right) && (7.171)\\
&= \mathrm{P}\left(\frac{1-\tilde{a}}{2} \leq x\right) && (7.172)\\
&= \mathrm{P}\left(\tilde{a} \geq 1-2x\right) && (7.173)\\
&= \int_{1-2x}^{1} p_{\tilde{a}}(t)\ \mathrm{d}t && (7.174)\\
&= \int_{1-2x}^{1} 2(1-t)\ \mathrm{d}t && (7.175)\\
&= 4x^2 && (7.176)
\end{aligned}
$$

for $x \in [0, 0.5]$. The pdf, shown in the right plot of Figure 7.8, equals

$$
f_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}(x) = 8x, \tag{7.177}
$$

for $x \in [0, 0.5]$ and zero otherwise.

We can now formally prove iterated expectation, establishing that the mean of a random variable is indeed equal to the mean of its conditional mean.

Theorem 7.54 (Iterated expectation)**.** *For any random variables $\tilde{a}$ and $\tilde{b}$ belonging to the same probability space,*

$$
\mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right] = \mathrm{E}\left[\tilde{b}\right]. \tag{7.178}
$$

Similarly, for any function $h : \mathbb{R}^2 \to \mathbb{R}$,

$$
\mathrm{E}[\mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(\tilde{a})] = \mathrm{E}\left[h(\tilde{a},\tilde{b})\right]. \tag{7.179}
$$

Proof We prove the result for continuous random variables. The proof for discrete random variables and for functions that depend on both continuous and discrete random variables is essentially the same. By the definition of the conditional mean and the chain rule,

$$
\begin{aligned}
\mathrm{E}[\mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(\tilde{a})] &= \int_{a=-\infty}^{\infty} f_{\tilde{a}}\left(a\right) \mu_{h(\tilde{a},\tilde{b})\,|\,\tilde{a}}(a)\ \mathrm{d}a && (7.180)\\
&= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} f_{\tilde{a}}\left(a\right) f_{\tilde{b}\,|\,\tilde{a}}\left(b\,|\,a\right) h\left(a,b\right)\ \mathrm{d}b\ \mathrm{d}a && (7.181)\\
&= \int_{a=-\infty}^{\infty} \int_{b=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}\left(a,b\right) h\left(a,b\right)\ \mathrm{d}b\ \mathrm{d}a && (7.182)\\
&= \mathrm{E}[h(\tilde{a},\tilde{b})]. && (7.183)
\end{aligned}
$$

Setting $h(\tilde{a},\tilde{b}) := \tilde{b}$ establishes (7.178). ■

The following two simple examples illustrate the iterated-expectation identity.

Example 7.55 (Simple example: Iterated expectation)**.** The mean of the conditional mean $\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})$ derived in Example 7.52 equals

$$\begin{aligned} \mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right] &= \sum_{x\in\{2,15/7,16/7\}} x p_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}(x) && (7.184)\\ &= 2\cdot 0.3 + \frac{15}{7}\cdot 0.35 + \frac{16}{7}\cdot 0.35 && (7.185)\\ &= 2.15. && (7.186) \end{aligned}$$

We verify that this indeed equals the mean of $\tilde{b}$ using the marginal pmf derived in Example 4.11,

$$\begin{aligned} \mathrm{E}[\tilde{b}] &= \sum_{b=1}^{3} b p_{\tilde{b}}(b) && (7.187)\\ &= 1\cdot 0.25 + 2\cdot 0.35 + 3\cdot 0.4 && (7.188)\\ &= 2.15 = \mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right]. && (7.189) \end{aligned}$$

Example 7.56 (Triangle lake: Iterated expectation)**.** The mean of the conditional mean $\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})$ derived in Example 7.53 is

$$\begin{aligned} \mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right] &= \int_{x=-\infty}^{\infty} x f_{\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})}(x)\,\mathrm{d}x && (7.190)\\ &= \int_{x=0}^{\frac{1}{2}} 8x^2\,\mathrm{d}x = \frac{1}{3}. && (7.191) \end{aligned}$$

The marginal pdf of $\tilde{b}$ (see Example 5.9) equals

$$f_{\tilde{b}}(b) = \int_{a=-\infty}^{\infty} f_{\tilde{a},\tilde{b}}(a,b)\,\mathrm{d}a = 2(1-b) \tag{7.192}$$

for $0 \leq b \leq 1$ and zero otherwise. We confirm that the mean of $\tilde{b}$ is equal to the mean of the conditional mean,

$$\begin{aligned} \mathrm{E}[\tilde{b}] &= \int_{b=-\infty}^{\infty} b f_{\tilde{b}}(b)\,\mathrm{d}b && (7.193)\\ &= \int_{b=0}^{1} 2(1-b)b\,\mathrm{d}b && (7.194)\\ &= \frac{1}{3} = \mathrm{E}\left[\mu_{\tilde{b}\,|\,\tilde{a}}(\tilde{a})\right]. && (7.195) \end{aligned}$$

Iterated expectation allows us to obtain the mean of quantities that depend on several quantities very easily, as long as we have access to the marginal and conditional distributions.

Example 7.57 (Computer)**.** We want to model the time until a computer breaks down. We know that the time depends on how often the computer is turned off, and whether the owner is careful. We model the time $\tilde{t}$ as an exponential random variable with a parameter $\tilde{\lambda}$ that is itself a random variable defined as

$$\tilde{\lambda} := \frac{1}{\tilde{o} + \tilde{c}}. \tag{7.196}$$

The random variable $\tilde{o}$ represents the fraction of time the computer is off; it is uniformly distributed in $[0, 1]$. The random variable $\tilde{c}$ represents how careful the owner is; it is also uniformly distributed in $[0, 1]$. In order to compute the mean of $\tilde{t}$ we could derive the corresponding pdf, but this would be quite complicated. Fortunately, iterated expectation comes to the rescue. Conditioned on $\tilde{o} = o$ and $\tilde{c} = c$, $\tilde{t}$ is exponential with parameter $\lambda := \frac{1}{o+c}$, so by Lemma 7.24 the mean equals

$$\mu_{\tilde{t} \mid \tilde{o}, \tilde{c}}(o, c) = \frac{1}{\lambda} \tag{7.197}$$

$$= o + c. \tag{7.198}$$

By Lemma 7.8 the means of $\tilde{o}$ and $\tilde{c}$ equal 0.5, so by iterated expectation and linearity of expectation,

$$\mathrm{E}\left[\tilde{t}\right] = \mathrm{E}\left[\mu_{\tilde{t} \mid \tilde{o}, \tilde{c}}(o, c)\right] \tag{7.199}$$

$$= \mathrm{E}\left[\tilde{o} + \tilde{c}\right] \tag{7.200}$$

$$= 0.5 + 0.5 \tag{7.201}$$

$$= 1. \tag{7.202}$$

Example 7.58 (Mean of a mixture model)**.** In Example 6.3 we use a Gaussian mixture model to model height in a population. In the model, height is represented by a continuous random variable $\tilde{h}$ and sex as a discrete random variable $\tilde{s}$. The sample mean of height equals 163 cm for the women ($\tilde{s} = 0$) and 176 cm for the men ($\tilde{s} = 1$). This coincides with our estimate of the mean parameters of the corresponding Gaussian distributions, because the maximum-likelihood estimate for the mean parameter is the sample mean by Theorem 3.36. The conditional mean function therefore equals

$$\mu_{\tilde{h} \mid \tilde{s}}(0) = 163, \tag{7.203}$$

$$\mu_{\tilde{h} \mid \tilde{s}}(1) = 176. \tag{7.204}$$

The sex $\tilde{s}$ is Bernoulli with parameter 0.67. Iterated expectation allows us to compute the mean of $\tilde{h}$ without having to derive its marginal pdf,

$$\mathrm{E}[\tilde{h}] = \mathrm{E}[\mu_{\tilde{h} \mid \tilde{s}}(\tilde{s})] \tag{7.205}$$

$$= p_{\tilde{s}}(0)\mu_{\tilde{h} \mid \tilde{s}}(0) + p_{\tilde{s}}(1)\mu_{\tilde{h} \mid \tilde{s}}(1) \tag{7.206}$$

$$= 171.7 \text{ cm}. \tag{7.207}$$

7.8.3 Regression Via Conditional Averaging

Regression is the problem of estimating a certain quantity of interest, called the *response*, from observed variables called *features*. From a probabilistic viewpoint, the goal is to approximate a random variable $\tilde{y}$, representing the response, as a deterministic function h of a random vector $\tilde{x}$, representing the features. In this section, we study the regression problem under the assumption that we know the joint distribution of the features and the response.

A popular metric to evaluate a regression estimate is the mean squared error (MSE), introduced in Definition 7.29, between the estimate and the response. Recall that, by Theorem 7.30, the mean is the best constant estimate of a random variable in terms of MSE. Inspired by this, if we observe that the features equal a certain value, $\tilde{x} = x$, we can estimate the response $\tilde{y}$ using its conditional mean given $\tilde{x} = x$, provided by the conditional mean function $\mu_{\tilde{y}\,|\,\tilde{x}}(x)$ defined in Section 7.8.1. This estimation strategy is optimal: it achieves the minimum possible MSE.

Theorem 7.59 (Minimum MSE estimator)**.** *Let $\tilde{x}$ and $\tilde{y}$ be a random vector and a random variable, representing the features and response in a regression problem. Among all estimators of the response $\tilde{y}$ that only depend on the features $\tilde{x}$, the conditional mean $\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})$ achieves the minimum MSE,*

$$\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) = \arg\min_{h(\tilde{x})} \mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right]. \tag{7.208}$$

As a result, the conditional mean is often referred to as the minimum mean-squared-error (MMSE) estimator.

Proof Let h be an arbitrary real-valued function, which provides an estimate $h(\tilde{x})$ of $\tilde{y}$ when we plug in $\tilde{x}$. Our goal is to show that this estimate cannot achieve a smaller MSE than the conditional mean. By linearity of expectation,

$$\mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right] \tag{7.209}$$

$$= \mathrm{E}\left[\left(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) + \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) - h(\tilde{x})\right)^2\right] \tag{7.210}$$

$$= \mathrm{E}\left[(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}))^2\right] + \mathrm{E}\left[(\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) - h(\tilde{x}))^2\right] + 2\mathrm{E}\left[(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}))(\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) - h(\tilde{x}))\right].$$

The proof is completed by showing that the third term equals zero,

$$\mathrm{E}\left[(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}))(\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) - h(\tilde{x}))\right] \tag{7.211}$$

$$= \mathrm{E}\left[\tilde{y}\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})\right] - \mathrm{E}[\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})^2] + \mathrm{E}\left[\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})h(\tilde{x})\right] - \mathrm{E}\left[\tilde{y}h(\tilde{x})\right] = 0, \tag{7.212}$$

which implies

$$\mathrm{E}\left[(\tilde{y} - h(\tilde{x}))^2\right] = \mathrm{E}\left[(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}))^2\right] + \mathrm{E}\left[(\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}) - h(\tilde{x}))^2\right] \tag{7.213}$$

$$\geq \mathrm{E}\left[(\tilde{y} - \mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x}))^2\right], \tag{7.214}$$

Table 7.3 ***Conditional pmf of cats given dogs.*** *Conditional pmf* $p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,d)$ *of* $\tilde{c}$ *given* $\tilde{d}$ *in Example 7.6.*

Dogs (d) \ Cats (c)	**0**	**1**	**2**	**3**
0	0.54	0.23	0.15	0.08
1	0.71	0.18	0.11	0
2	0.71	0.29	0	0

because the second term in (7.213) is nonnegative (it is an integral or sum of a nonnegative quantity).

All that remains is to prove that (7.212) holds. Given $\tilde{x} = x$, $h(\tilde{x})$ is just a constant equal to $h(x)$, so the conditional mean function of $h(\tilde{x})\tilde{y}$ given $\tilde{x} = x$ is equal to $h(x)$ multiplied by the conditional mean function of $\tilde{y}$ given $\tilde{x} = x$. Assuming that $\tilde{x}$ is continuous∗,

$$\mu_{h(\tilde{x})\tilde{y}\,|\,\tilde{x}}(x) = \int_{y=-\infty}^{\infty} h(x) y f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x)\,\mathrm{d}y \tag{7.215}$$

$$= h(x) \int_{y=-\infty}^{\infty} y f_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x)\,\mathrm{d}y \tag{7.216}$$

$$= h(x)\mu_{\tilde{y}\,|\,\tilde{x}}(x). \tag{7.217}$$

By iterated expectation,

$$\mathrm{E}\left[\tilde{y}h(\tilde{x})\right] = \mathrm{E}\left[\mu_{\tilde{y}h(\tilde{x})\,|\,\tilde{x}}(\tilde{x})\right] \tag{7.218}$$

$$= \mathrm{E}\left[\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})h(\tilde{x})\right]. \tag{7.219}$$

Since the conditional mean function $\mu_{\tilde{y}\,|\,\tilde{x}}(x)$ is a deterministic function of x, the same argument (setting $h := \mu_{\tilde{y}\,|\,\tilde{x}}$) implies that

$$\mathrm{E}\left[\tilde{y}\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})\right] = \mathrm{E}[\mu_{\tilde{y}\,|\,\tilde{x}}(\tilde{x})^2]. \tag{7.220}$$

We conclude that (7.212) holds and the proof is complete. ■

Example 7.60 (Cats and dogs: MMSE estimator)**.** Example 7.6 provides the joint pmf of the number of cats and dogs per household, represented by the random variables $\tilde{c}$ and $\tilde{d}$, in a certain city according to a fictitious pet-food producer. We consider the problem of estimating the number of cats in a household from the number of dogs. The conditional pmf of $\tilde{c}$ given $\tilde{d}$, computed as described in Section 4.3, is shown in Table 7.3. The corresponding conditional mean function

∗The same argument holds if $\tilde{x}$ is discrete, replacing the integral by a sum and the conditional pdf by the conditional pmf.

equals

$$\mu_{\tilde{c}\,|\,\tilde{d}}(0) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,0) = 0.77, \tag{7.221}$$

$$\mu_{\tilde{c}\,|\,\tilde{d}}(1) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,1) = 0.39, \tag{7.222}$$

$$\mu_{\tilde{c}\,|\,\tilde{d}}(2) = \sum_{c=0}^{3} c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,2) = 0.29. \tag{7.223}$$

By Theorem 7.59, this is the optimal estimator of $\tilde{c}$ given $\tilde{d}$ in terms of MSE. The corresponding MSE equals

$$\mathrm{E}\left[\left(\tilde{c} - \mu_{\tilde{c}\,|\,\tilde{d}}(\tilde{d})\right)^2\right] = \sum_{c=0}^{3}\sum_{d=0}^{2} p_{\tilde{c},\tilde{d}}(c,d)\left(c - \mu_{\tilde{c}\,|\,\tilde{d}}(d)\right)^2 \tag{7.224}$$

$$= 0.756. \tag{7.225}$$

Notice that our estimator is optimal in terms of MSE, but it is abysmal in terms of probability of error: nobody can own 0.77, 0.4 or 0.29 cats, so we are always wrong! By Theorem 4.30, the best estimate in terms of probability of error is the MAP estimator of $\tilde{c}$ given $\tilde{d} = d$,

$$\mathrm{MAP}(0) = \arg\max_c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,0) = 0, \tag{7.226}$$

$$\mathrm{MAP}(1) = \arg\max_c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,1) = 0, \tag{7.227}$$

$$\mathrm{MAP}(2) = \arg\max_c\, p_{\tilde{c}\,|\,\tilde{d}}(c\,|\,2) = 0. \tag{7.228}$$

The MAP estimate $c\ell\left(\tilde{c}\,|\,\tilde{d}\right)$ is that the number of cats is zero, no matter what number of dogs are observed. The corresponding MSE is

$$\mathrm{E}\left[\left(\tilde{c} - \mathrm{MAP}(\tilde{d})\right)^2\right] = \sum_{c=0}^{3}\sum_{d=0}^{2} p_{\tilde{c},\tilde{d}}(c,d)c^2 \tag{7.229}$$

$$= 1.19, \tag{7.230}$$

which is substantially larger than the MSE of the MMSE estimator.

Theorem 7.59 makes regression sound easy, but it is actually usually very challenging! The conditional mean is indeed an optimal estimator, but it is *intractable to compute* unless the number of features is very small. The reason is that the number of possible conditional distributions of the response given the features explodes exponentially. For instance, consider the regression problem of estimating the rating for the movie Independence Day from the rating for Mission Impossible, in Example 7.48. The MMSE estimator can be easily approximated via the sample conditional mean function, shown in the left plot of Figure 7.7, because there is only one feature. However, what if we want to estimate the rating for

Independence Day from the ratings of *100 other movies*? Then, to approximate the minimum MSE estimator, we need to compute the sample conditional mean given every possible rating sequence of length 100. However, there are $5^{100} > 10^{68}$ such sequences, so this is impossible! In fact, when we encounter a new user, they are likely to have a unique set of features that we have not observed before. This is a manifestation of the curse of dimensionality, described in Section 4.7. In Chapter 12 we describe several strategies to address this challenge when performing regression in practice.

7.9 The Average Treatment Effect

In Section 4.6 we explain how to estimate the causal effect of a treatment on a binary outcome of interest, with two possible values (e.g. whether a patient recovers or not). In this section, we study causal effects associated to outcomes with multiple possible values.∗ As a motivating example, consider the problem of determining whether the title capitalization in YouTube videos has an effect on the number of views. We interpret title capitalization as a binary treatment $\tilde{t}$: if $\tilde{t} = 1$, the title is all caps; if $\tilde{t} = 0$, the title is proper case.

In order to study the causal effect of the treatment, we define the potential outcomes associated to it, as in Section 4.6.1. In our example, the potential outcome $\widetilde{\text{po}}_0$ represents the number of views that we would observe in a hypothetical situation where all titles would be in proper case. Conversely, the potential outcome $\widetilde{\text{po}}_1$ represents the number of views if all titles were in all caps. The average treatment effect is the difference between the means of these two potential outcomes.

Definition 7.61 (Average treatment effect)**.** *Given two potential outcomes* $\widetilde{\text{po}}_0$ *and* $\widetilde{\text{po}}_1$ *associated to a treatment* $\tilde{t}$*, the average treatment effect is*

$$\text{ATE} := \text{E}\left[\widetilde{\text{po}}_1\right] - \text{E}\left[\widetilde{\text{po}}_0\right]. \tag{7.231}$$

The challenge in estimating the ATE is that our observations of the potential outcomes are incomplete. The observed outcome $\tilde{y}$ is equal to $\widetilde{\text{po}}_0$ only when $\tilde{t} = 0$ and to $\widetilde{\text{po}}_1$ only when $\tilde{t} = 1$,

$$\tilde{y} := \begin{cases} \widetilde{\text{po}}_0 & \text{if} \quad \tilde{t} = 0, \\ \widetilde{\text{po}}_1 & \text{if} \quad \tilde{t} = 1. \end{cases} \tag{7.232}$$

In order to determine the ATE, we need to estimate the mean of the potential outcomes $\text{E}\left[\widetilde{\text{po}}_0\right]$ and $\text{E}\left[\widetilde{\text{po}}_1\right]$ from $\tilde{y}$. An option is to use the conditional mean of the observed outcome given the treatment, $\mu_{\tilde{y}\,|\,\tilde{t}}(0)$ and $\mu_{\tilde{y}\,|\,\tilde{t}}(1)$. However, this is *not* necessarily a good estimate. In the case of the YouTube videos, the conditional

∗Recall that in causal inference, the term *outcome* does not refer to an element in a sample space, as in Section 1.2. Instead, it is a random variable representing a quantity of interest that may (or may not) be affected by the treatment.

mean of the observed views given all-caps titles is

$$\mu_{\tilde{y}\,|\,\tilde{t}}(1) = \sum_{y=0}^{\infty} y\, p_{\tilde{y}\,|\,\tilde{t}}(y\,|\,1) \tag{7.233}$$

$$= \sum_{y=0}^{\infty} y\, p_{\widetilde{\mathrm{po}}_1\,|\,\tilde{t}}(y\,|\,1), \tag{7.234}$$

whereas the mean of $\widetilde{\mathrm{po}}_1$ equals

$$\mathrm{E}\left[\widetilde{\mathrm{po}}_1\right] = \sum_{y=0}^{\infty} y\, p_{\widetilde{\mathrm{po}}_1}(y). \tag{7.235}$$

The conditional mean of the observed outcome only equals the mean of the potential outcome if the treatment and the potential outcome are independent. In that case, the conditional pmf of $\widetilde{\mathrm{po}}_1$ given that a video is selected to have all-caps titles, $p_{\widetilde{\mathrm{po}}_1\,|\,\tilde{t}}(\cdot\,|\,1)$, is the same as the marginal pmf $p_{\widetilde{\mathrm{po}}_1}$. This means that the selected videos are not more (or less prone) to have more views if their titles are set to all caps. Consequently, we can use the conditional mean of the observed views to estimate the ATE. An important consequence is that the ATE can be estimated reliably from randomized experiments, such as randomized controlled trials, which guarantee that the independence condition holds.

Theorem 7.62 (Estimation of the average treatment effect)**.** *Let $\tilde{y}$ be a random variable following the definition in (7.232), where $\widetilde{\mathrm{po}}_0$ and $\widetilde{\mathrm{po}}_1$ denote the two potential outcomes associated to a treatment $\tilde{t}$. If $\tilde{t}$ and $\widetilde{\mathrm{po}}_0$ are independent, and $\tilde{t}$ and $\widetilde{\mathrm{po}}_1$ are also independent, then*

$$\mathrm{ATE} = \mu_{\tilde{y}\,|\,\tilde{t}}(1) - \mu_{\tilde{y}\,|\,\tilde{t}}(0). \tag{7.236}$$

Proof We assume that the potential outcomes are continuous, the same argument holds for discrete outcomes replacing the corresponding integrals by sums and the pdfs by pmfs. If $\tilde{t} = 1$, then $\tilde{y} = \widetilde{\mathrm{po}}_1$, so $\mu_{\tilde{y}\,|\,\tilde{t}}(1) = \mu_{\widetilde{\mathrm{po}}_1\,|\,\tilde{t}}(1)$. By the independence assumption, the conditional pdf of $\widetilde{\mathrm{po}}_1$ given $\tilde{t}$ is equal to the marginal pdf, so

$$\mu_{\widetilde{\mathrm{po}}_1\,|\,\tilde{t}}(1) = \int_{y=-\infty}^{\infty} y f_{\widetilde{\mathrm{po}}_1\,|\,\tilde{t}}(y\,|\,1)\ \mathrm{d}y \tag{7.237}$$

$$= \int_{y=-\infty}^{\infty} y f_{\widetilde{\mathrm{po}}_1}(y)\ \mathrm{d}y \tag{7.238}$$

$$= \mathrm{E}\left[\widetilde{\mathrm{po}}_1\right]. \tag{7.239}$$

By the same argument, $\mu_{\tilde{y}\,|\,\tilde{t}}(0) = \mathrm{E}\left[\widetilde{\mathrm{po}}_0\right]$, which completes the proof. ■

Example 7.63 (All caps titles in YouTube videos)**.** In order to evaluate the effect of capitalization on YouTube videos, we randomized the title capitalization of 46 videos on probability and statistics for data science. Each time a video was posted, the title was set to all caps with probability 1/2 and to proper case with

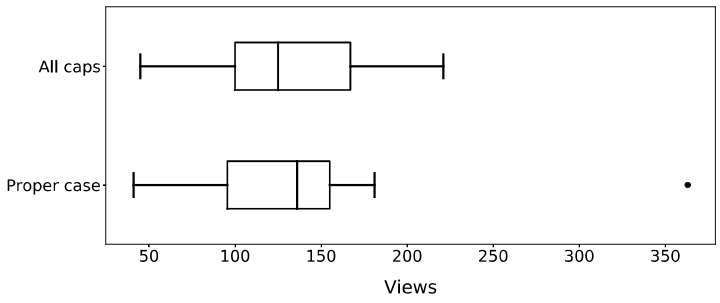

Figure 7.9 Effect of title capitalization in YouTube videos. Boxplots of the number of views of the videos with all-caps and proper-case titles in Example 7.63. The distribution of views is very similar for both types of title. Since the choice of capitalization was randomized, this is strong evidence that capitalization does not affect the number of views.

probability 1/2. The mean number of views for the 19 videos with all-caps titles was 133. The mean number of views for the 27 videos with proper-case titles was 132. By Theorem 7.62, the ATE is equal to the difference, which is very small (133-132=1). This suggests that capitalization makes no difference in the number of views. Figure 7.9 shows the distribution of the number of views for both types of title. The median number of views for the proper-case titles is higher than that of the all-caps titles.

The following example shows that in observational studies, where we have no control over the treatment, the ATE is *not* necessarily equal to the difference between the conditional mean of the observed outcome $\tilde{y}$ given $\tilde{t} = 1$ and given $\tilde{t} = 0$. The reason is that confounders can completely distort the conditional means. This is the same phenomenon described in Example 4.22 and Section 4.6.2.

Example 7.64 (Private lessons)**.** Our goal in this example is to study the causal effect of receiving private math lessons on the performance of Portuguese high-school students, using data extracted from Dataset 15. The observed outcome of interest is the grade obtained in a math class, which we denote by $\tilde{y}$. The private lessons are represented by a random variable $\tilde{t}$, since we interpret them as a *treatment*: $\tilde{t} = 1$ indicates that the student received private lessons and belongs to the treatment group; $\tilde{t} = 0$ indicates that they did not receive private lessons and therefore belong to the control group.

The conditional mean function of $\tilde{y}$ given $\tilde{t}$ equals,

$$\mu_{\tilde{y}\,|\,\tilde{t}}(1) = 10.92, \qquad \mu_{\tilde{y}\,|\,\tilde{t}}(0) = 9.99. \tag{7.240}$$

We define the *observed* ATE as the difference between the two values of the

conditional mean,

$$\text{observed ATE} := \mu_{\tilde{y} \mid \tilde{t}}(1) - \mu_{\tilde{y} \mid \tilde{t}}(0) = 0.93. \tag{7.241}$$

From this value, it seems that private lessons increase a student grade by one point (out of 20) on average, which is quite substantial. In a randomized study, the observed ATE is equal to the true ATE by Theorem 7.62, so we might be tempted to conclude that private lessons are effective. However, our data are *not* randomized, so we should determine whether there are any systematic differences between the treatment and control groups, which might account for the observed ATE.

The data also report what students had previously failed the course, which is a possible confounding factor. We represent it using a random variable $\tilde{c}$: $\tilde{c} = 1$ indicates that the student previously failed the course, and $\tilde{c} = 0$ that they didn't. The conditional mean function of the grade $\tilde{y}$ given $\tilde{c}$ and $\tilde{t}$ for students that had previously failed is

$$\mu_{\tilde{y} \mid \tilde{c},\tilde{t}}(1,1) = 8.95, \qquad \mu_{\tilde{y} \mid \tilde{c},\tilde{t}}(1,0) = 6.66. \tag{7.242}$$

For these students, it looks like the private lessons may be helpful: the conditional average grade is more than two points higher given $\tilde{t} = 1$. The conditional mean function for students who had not failed is

$$\mu_{\tilde{y} \mid \tilde{c},\tilde{t}}(0,1) = 11.20, \qquad \mu_{\tilde{y} \mid \tilde{c},\tilde{t}}(0,0) = 11.31. \tag{7.243}$$

These students have better grades than the *previously failed* group, whether they receive private lessons or not. Those that receive private lessons have a slightly worse grade on average than those who don't.

A key consideration when analyzing a possible confounding factor is whether it is independent from the treatment or not. Our confounder $\tilde{c}$ is definitely not independent from the treatment, because

$$p_{\tilde{c} \mid \tilde{t}}(1 \mid 1) = 0.122 \quad \text{and} \quad p_{\tilde{c} \mid \tilde{t}}(1 \mid 0) = 0.285 \tag{7.244}$$

are very different. Students in the control group are more than twice as likely to have failed than those in the treatment group. This is very problematic for our analysis. Students who previously failed have lower grades than the rest of the students. Consequently, the treatment group may have a higher average grade just because it contains less such students, *not because the private lessons are useful*.

To analyze the effect of the confounder quantitatively, we derive an expression of the conditional mean function of the observed grades $\tilde{y}$ given $\tilde{t}$ as a function

of the conditional mean function of $\tilde{y}$ given $\tilde{t}$ and also $\tilde{c}$,

$$\mu_{\tilde{y}\,|\,\tilde{t}}(t) = \int_{y=-\infty}^{\infty} f_{\tilde{y}\,|\,\tilde{t}}(y\,|\,t) y \, \mathrm{d}y \tag{7.245}$$

$$= \sum_{c=0}^{1} p_{\tilde{c}\,|\,\tilde{t}}(c\,|\,t) \int_{y=-\infty}^{\infty} f_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(y\,|\,c,t) y \, \mathrm{d}y \tag{7.246}$$

$$= \sum_{c=0}^{1} p_{\tilde{c}\,|\,\tilde{t}}(c\,|\,t) \mu_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(c,t), \tag{7.247}$$

where we have used the fact that, by the chain rule for discrete and continuous random variables (Theorem 6.6),

$$f_{\tilde{y}\,|\,\tilde{t}}(y\,|\,t) = \sum_{c=0}^{1} p_{\tilde{c}\,|\,\tilde{t}}(c\,|\,t) f_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(y\,|\,c,t). \tag{7.248}$$

We can use (7.247) to separate the contributions of the *previously failed* and *not previously failed* groups:

$$\mu_{\tilde{y}\,|\,\tilde{t}}(1) = p_{\tilde{c}\,|\,\tilde{t}}(0\,|\,1) \mu_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(0,1) + p_{\tilde{c}\,|\,\tilde{t}}(1\,|\,1) \mu_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(1,1) \tag{7.249}$$

$$= 0.878 \cdot 11.20 + 0.122 \cdot 8.95 \tag{7.250}$$

$$= \underset{\uparrow \tilde{c}=0}{9.83} + \underset{\uparrow \tilde{c}=1}{1.09} = 10.92. \tag{7.251}$$

Similarly,

$$\mu_{\tilde{y}\,|\,\tilde{t}}(0) = p_{\tilde{c}\,|\,\tilde{t}}(0\,|\,0) \mu_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(0,0) + p_{\tilde{c}\,|\,\tilde{t}}(1\,|\,0) \mu_{\tilde{y}\,|\,\tilde{c},\tilde{t}}(1,0) \tag{7.252}$$

$$= 0.715 \cdot 11.31 + 0.285 \cdot 6.66 \tag{7.253}$$

$$= \underset{\uparrow \tilde{c}=0}{8.09} + \underset{\uparrow \tilde{c}=1}{1.90} = 9.99. \tag{7.254}$$

It turns out that the contribution of the *not previously failed* group drives the observed ATE up, even though for that group private lessons seem to make no difference. In contrast, the contribution of the *previously failed* group drives the observed ATE down, even though private lessons do seem to result in higher grades for that group. As we had feared, the observed ATE is severely distorted by the dependence between the confounder and the treatment, just like in Simpson's paradox (see Section 4.6.2).

In Section 4.6.4 we explain how to adjust for confounders in the case of binary outcomes. The same procedure can be used to control for confounders when estimating the ATE. Crucially we again require that the potential outcomes be conditionally independent from the treatment given the confounder. In that case, the conditional means are not distorted by additional confounders, and we can correct the ATE by reweighting the conditional means of the outcome given the treatment and the confounder.

Theorem 7.65 (Adjusting the ATE to control for confounders)**.** *Let $\tilde{y}$ be a random variable following the definition in (7.232), where $\widetilde{\text{po}}_0$ and $\widetilde{\text{po}}_1$ denote the two potential outcomes associated to a treatment $\tilde{t}$, and let $\tilde{c}$ be a confounder, represented by a discrete random variable with range C. If the treatment $\tilde{t}$ and the potential outcomes $\widetilde{\text{po}}_0$ and $\widetilde{\text{po}}_1$ are conditionally independent given $\tilde{c}$, then*

$$\text{ATE} = \sum_{c \in C} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 1) - \sum_{c \in C} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 0). \tag{7.255}$$

Proof If $\tilde{t} = 1$, then $\tilde{y} = \widetilde{\text{po}}_1$, so $\mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 1) = \mu_{\widetilde{\text{po}}_1 \mid \tilde{c}, \tilde{t}}(c, 1)$. By the conditional independence assumption,

$$\begin{aligned} \mu_{\widetilde{\text{po}}_1 \mid \tilde{c}, \tilde{t}}(c, 1) &= \int_x x f_{\widetilde{\text{po}}_1 \mid \tilde{c}, \tilde{t}}(x \mid c, 1) \, \mathrm{d}x & (7.256) \\ &= \int_x x f_{\widetilde{\text{po}}_1 \mid \tilde{c}}(x \mid c) \, \mathrm{d}x & (7.257) \\ &= \mu_{\widetilde{\text{po}}_1 \mid \tilde{c}}(c). & (7.258) \end{aligned}$$

By the same argument,

$$\mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 0) = \mu_{\widetilde{\text{po}}_0 \mid \tilde{c}}(c). \tag{7.259}$$

We have assumed that the potential outcomes are continuous, the same argument holds for discrete outcomes replacing the corresponding integrals by sums and the pdfs by pmfs.

The result then follows from aggregating these conditional means and applying iterated expectation (Theorem 7.54),

$$\begin{aligned} &\sum_{c \in C} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 1) - \sum_{c \in C} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 0) & (7.260) \\ &= \sum_{c \in C} p_{\tilde{c}}(c) \mu_{\widetilde{\text{po}}_1 \mid \tilde{c}}(c) - \sum_{c \in C} p_{\tilde{c}}(c) \mu_{\widetilde{\text{po}}_0 \mid \tilde{c}}(c) & (7.261) \\ &= \text{E}\left[\mu_{\widetilde{\text{po}}_1 \mid \tilde{c}}(\tilde{c})\right] - \text{E}\left[\mu_{\widetilde{\text{po}}_0 \mid \tilde{c}}(\tilde{c})\right] & (7.262) \\ &= \text{E}\left[\widetilde{\text{po}}_1\right] - \text{E}\left[\widetilde{\text{po}}_0\right] = \text{ATE}\,. & (7.263) \end{aligned}$$

■

Example 7.66 (Private lessons: Adjusting for previous failure)**.** Example 7.64 shows that the different rates at which students in the treatment and control groups had previously failed distorts the ATE. In order to correct for this, we compute the fraction of students that previously failed from the data, which equals 0.21. By Theorem 7.65, the adjusted ATE is

$$\begin{aligned} \text{adjusted ATE} &:= \sum_{c=0}^{1} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 1) - \sum_{c=0}^{1} p_{\tilde{c}}(c) \mu_{\tilde{y} \mid \tilde{c}, \tilde{t}}(c, 0) & (7.264) \\ &= (0.79 \cdot 11.20 + 0.21 \cdot 8.95) - (0.79 \cdot 11.31 + 0.21 \cdot 6.66) & (7.265) \\ &= 0.39. & (7.266) \end{aligned}$$

After the correction, the ATE is reduced to less than half, from 0.93 to 0.39.

Should we conclude that private lessons are not very useful? Definitely not! The conditional independence assumption in Theorem 7.65 implies that within the *not previously failed* group and the *previously failed* group, the students that take private lessons are not systematically different from the students that do not take private lessons. It is difficult to believe that this is the case: students often take private lessons *because they need them*. If the students receiving private lessons are systematically weaker, then this will artificially lower our adjusted ATE.

A counterargument is that students often take private lessons *because they can afford them*. It is therefore possible that students taking private lessons have more resources at their disposal, and would do better anyway. In conclusion, more analysis is needed in order to control for additional confounders such as academic ability and socioeconomic status. This example reflects the shortcomings of observational studies. However, although randomizing the treatment automatically controls for all possible confounders, it is often problematic to implement in practice: imagine the reactions of some parents if their kid were assigned to a control group without private lessons...