Expectation

Last modified: 2026-11-20 00:10:36 (PST)

Definition 0.1 (Expectation, expected value, population mean)**.** The **expectation**, **expected value**, or **population mean** of a *discrete* random variable X, denoted $\mathbb{E}[X]$, $\mu(X)$, or μ_X, is the mean of X's possible values [1], weighted by the probability mass function of those values:

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \sum_{x \in \mathcal{R}(X)} x \cdot \mathbb{P}(X = x)$$

The **expectation** of a *continuous* random variable X with probability density function $\mathbb{p}(X = x)$ is the mean of X's possible values, weighted by that density:

$$\mathbb{E}[X] \stackrel{\text{def}}{=} \int_{x \in \mathcal{R}(X)} x \cdot \mathbb{p}(X = x)\, dx$$

In both cases, the expectation is defined only when the sum or integral converges absolutely, that is, when it is finite with $|x|$ in place of x.

Proposition 0.1 (The mean of numbers is the expectation of one chosen at random)**.** *Let $a_1, \ldots, a_n$ be distinct real numbers, and let Y be one of them chosen at random. Then the expectation of Y is the mean of the numbers:*

$$\mathbb{E}[Y] = \frac{1}{n} \sum_{i=1}^{n} a_i$$

Proof

Proof.

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{i=1}^{n} a_i \cdot \mathbb{P}(Y = a_i) && \text{(definition of expectation; the range of } Y \text{ is } \{a_1, \dots, a_n\}) \\
&= \sum_{i=1}^{n} a_i \cdot \frac{1}{n} && \text{(each of the } n \text{ numbers is chosen with probability } 1/n) \\
&= \frac{1}{n} \sum_{i=1}^{n} a_i && \text{(definition of the mean)}
\end{aligned}$$

□

Example 0.1 (Expectation of one of four numbers chosen at random)**.** Let Y be one of the four distinct numbers $2, 3, 5, 10$, chosen at random. By Proposition 0.1, the expectation of Y is the mean of the four numbers:

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{1}{4} \cdot (2 + 3 + 5 + 10) && \text{(the mean of four numbers)} \\
&= \frac{20}{4} && \text{(add)} \\
&= 5 && \text{(divide)}
\end{aligned}$$

Remark (Expectation is a probability-weighted average)

Remark 0.1 (Expectation is a probability-weighted average). The expectation of a random variable X is a weighted average of the values of X, and the probabilities are the weights.

- A discrete random variable has a sum over its possible values. Each value x is multiplied by its weight $\mathbb{P}(X = x)$, and the products are added. The weights are non-negative and add to 1.
- A continuous random variable has an integral over its possible values. Each value x is multiplied by its weight, the density $\mathbb{p}(X = x)$, and the products are accumulated. The density is non-negative and integrates to 1.

As Proposition 0.1 shows, the mean of n numbers is the special case with every weight equal to $1/n$.
A value with a larger probability counts for more in the expectation. The expectation is the balance point of the possible values: the point where a ruler would balance if each value x carried a weight proportional to $\mathbb{P}(X = x)$.

Example 0.2 (Expectation of a fair die roll)**.** Let Y be the result of a fair die roll, so the range is $\mathcal{R}(Y) = \{1,2,3,4,5,6\}$ and $\mathbb{P}(Y = y) = 1/6$ for each y in the range. Every value has the same weight, so the expectation is the ordinary mean of $1, \dots, 6$:

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{y \in \{1,2,3,4,5,6\}} y \cdot \mathbb{P}(Y = y) && \text{(definition of expectation for discrete r.v.)} \\
&= (1+2+3+4+5+6) \cdot \frac{1}{6} && \text{(every weight is } 1/6\text{)} \\
&= \frac{21}{6} && \text{(add)} \\
&= 3.5 && \text{(divide)}
\end{aligned}$$

Now load the die so that the result 6 has probability $1/2$, and each of $1, \dots, 5$ has probability $1/10$. The weights still add to 1, and the expectation moves toward 6:

$$\begin{aligned}
\mathbb{E}[Y] &= (1+2+3+4+5) \cdot \frac{1}{10} + 6 \cdot \frac{1}{2} && \text{(weights } 1/10 \text{ and } 1/2\text{)} \\
&= 1.5 + 3 && \text{(multiply)} \\
&= 4.5 && \text{(add)}
\end{aligned}$$

Example 0.3 (The standard Cauchy distribution has no expectation)**.** Let X have the standard Cauchy distribution, with density $\mathbb{p}(X = x) = \frac{1}{\pi(1+x^2)}$ for every real x. For $x \geq 0$, $\frac{1}{2\pi}\log(1+x^2)$ is an antiderivative of $|x| \cdot \frac{1}{\pi(1+x^2)} = \frac{x}{\pi(1+x^2)}$, and for $x \leq 0$, $-\frac{1}{2\pi}\log(1+x^2)$ is an antiderivative of $|x| \cdot \frac{1}{\pi(1+x^2)} = \frac{-x}{\pi(1+x^2)}$. So, integrating over each half-line:

$$\begin{aligned}
\int_0^\infty |x| \cdot \frac{1}{\pi(1+x^2)}\,dx &= \lim_{b\to\infty}\left[\frac{1}{2\pi}\log(1+x^2)\right]_0^b && \text{(antiderivative on } [0,\infty)\text{)} \\
&= \lim_{b\to\infty} \frac{1}{2\pi}\log(1+b^2) && \text{(evaluate at the bounds; } \log 1 = 0\text{)} \\
&= \infty && \text{(} \log(1+b^2) \to \infty \text{ as } b \to \infty\text{)}
\end{aligned}$$

$$\begin{aligned}
\int_{-\infty}^0 |x| \cdot \frac{1}{\pi(1+x^2)}\,dx &= \lim_{a\to-\infty}\left[-\frac{1}{2\pi}\log(1+x^2)\right]_a^0 && \text{(antiderivative on } (-\infty,0]\text{)} \\
&= \lim_{a\to-\infty} \frac{1}{2\pi}\log(1+a^2) && \text{(evaluate at the bounds; } \log 1 = 0\text{)} \\
&= \infty && \text{(} \log(1+a^2) \to \infty \text{ as } a \to -\infty\text{)}
\end{aligned}$$

Both halves diverge, so $\int_{-\infty}^{\infty} |x| \cdot \frac{1}{\pi(1+x^2)}\,dx$ diverges, and by Definition 0.1, X has no expectation.

Theorem 0.1 (Expectation of the Bernoulli distribution)**.** *If* $X \sim \text{Ber}(\pi)$ *(Bernoulli distribution), that is,* $\mathbb{P}(X = 1) = \pi$ *and* $\mathbb{P}(X = 0) = 1 - \pi$*, then:*

$$\mathbb{E}[X] = \pi$$

Proof

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathcal{R}(X)} x \cdot \mathbb{P}(X = x) && \text{(definition of expectation for discrete r.v.)} \\
&= \sum_{x \in \{0,1\}} x \cdot \mathbb{P}(X = x) && \text{(range of Bernoulli r.v. is } \{0, 1\}) \\
&= (0 \cdot \mathbb{P}(X = 0)) + (1 \cdot \mathbb{P}(X = 1)) && \text{(expand sum over } x = 0 \text{ and } x = 1) \\
&= (0 \cdot (1 - \pi)) + (1 \cdot \pi) && \text{(substitute Bernoulli PMF values)} \\
&= 0 + \pi && \text{(multiplication by 0 and 1)} \\
&= \pi && \text{(addition of 0)}
\end{aligned}$$

□

Theorem 0.2 (Expectation of time-to-event variables)**.** *If* T *is a non-negative random variable with survival function* $\mathbb{S}(t)$ *and a defined expectation, then:*

$$\mathbb{E}[T] = \int_{t=0}^{\infty} \mathbb{S}(t)\, dt$$

Proof

Proof. We prove the continuous case, in which T has a density $\mathbb{f}$. The integrand $\mathbb{f}(u) \cdot \mathbb{1}(0 \le t \le u)$ is non-negative on $[0, \infty) \times [0, \infty)$, so Tonelli's theorem (the non-negative case of the Fubini–Tonelli theorem; Billingsley [2], Theorem 18.3) lets us exchange the order of integration:

$$
\begin{aligned}
\mathbb{E}[T] &= \int_{u=0}^{\infty} u\,\mathbb{f}(u)\,du && \text{(definition of expectation; } T \geq 0\text{)} \\
&= \int_{u=0}^{\infty} \left(\int_{t=0}^{u} 1\,dt \right) \mathbb{f}(u)\,du && \left(u = \textstyle\int_0^u 1\,dt\right) \\
&= \int_{u=0}^{\infty} \int_{t=0}^{u} \mathbb{f}(u)\,dt\,du && \text{(move } \mathbb{f}(u) \text{ inside the inner integral)} \\
&= \int_{t=0}^{\infty} \int_{u=t}^{\infty} \mathbb{f}(u)\,du\,dt && \text{(Tonelli: exchange the order over } 0 \leq t \leq u\text{)} \\
&= \int_{t=0}^{\infty} \Pr(T > t)\,dt && \text{(integrate the density over } (t, \infty)\text{)} \\
&= \int_{t=0}^{\infty} \mathbb{S}(t)\,dt && \text{(definition of the survival function)}
\end{aligned}
$$

Every step also holds when $\int_{u=0}^{\infty} u\,\mathbb{f}(u)\,du$ is infinite, because Tonelli's theorem needs only a non-negative integrand; so $\int_{t=0}^{\infty} \mathbb{S}(t)\,dt$ is finite exactly when $\mathbb{E}[T]$ is defined.
See [3] for an alternative presentation of this proof. □

Example 0.4 (Mean of an exponential random variable via survival function)**.** Let T be exponential with rate $\lambda > 0$, so $\mathbb{S}(t) = \mathrm{e}^{-\lambda t}$ for $t \geq 0$, as computed on the random variables page. By Theorem 0.2:

$$
\begin{aligned}
\mathbb{E}[T] &= \int_{0}^{\infty} \mathbb{S}(t)\,dt && \text{(expectation via the survival function)} \\
&= \int_{0}^{\infty} \mathrm{e}^{-\lambda t}\,dt && \text{(substitute the survival function)} \\
&= \left[-\frac{1}{\lambda} \mathrm{e}^{-\lambda t} \right]_{0}^{\infty} && \text{(antiderivative)} \\
&= 0 - \left(-\frac{1}{\lambda} \right) && \text{(evaluate at the bounds)} \\
&= \frac{1}{\lambda} && \text{(simplify)}
\end{aligned}
$$

Definition 0.2 (Expectation of a random matrix)**.** For a random matrix $\mathbf{A}$ of size $m \times n$ with (i, j)-th element A_{ij}, the **expectation** $\mathbb{E}\mathbf{A}$ is the $m \times n$ matrix whose (i, j)-th element is $\mathbb{E}\big[A_{ij}\big]$:

$$\mathbb{E}\mathbf{A} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbb{E}[A_{11}] & \mathbb{E}[A_{12}] & \cdots & \mathbb{E}[A_{1n}] \\ \mathbb{E}[A_{21}] & \mathbb{E}[A_{22}] & \cdots & \mathbb{E}[A_{2n}] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[A_{m1}] & \mathbb{E}[A_{m2}] & \cdots & \mathbb{E}[A_{mn}] \end{pmatrix}$$

Remark

Remark. In other words, expectation is applied **element-wise** to a random matrix.

Example 0.5 (Expectation of a random vector of coin flips)**.** For the two fair coin flips in the independence page's example, the random vector $\underset{\sim}{X} = (X_1, X_2)^\top$ is a 2×1 random matrix, and:

$$\mathbb{E}\underset{\sim}{X} = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}$$

since each X_i is Bernoulli with $\pi = 1/2$ (Theorem 0.1).

Theorem 0.3 (Law of the Unconscious Statistician (LOTUS))**.** ***Discrete case.*** *For any function g of a* discrete *random variable X [4]:*

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathbb{P}(X = x)$$

Continuous case. *For any function g of a* continuous *random variable X with density $\mathbb{p}(X = x)$:*

$$\mathbb{E}[g(X)] = \int_{x \in \mathcal{R}(X)} g(x) \cdot \mathbb{p}(X = x)\, dx$$

Proof

Proof. We prove the case of a discrete random variable first. Let $Y = g(X)$. For each $y \in \mathcal{R}(Y)$, the event $\{Y = y\}$ is the disjoint union of the events $\{X = x\}$ over the values x with $g(x) = y$, and each $x \in \mathcal{R}(X)$ belongs to exactly one such group.

$$
\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}[Y] && \text{(substitution } Y = g(X)\text{)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y) && \text{(definition of expectation for discrete r.v.)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{\substack{x \in \mathcal{R}(X) \\ g(x) = y}} \mathbb{P}(X = x) && \text{(countable additivity over the disjoint events } \{X = x\}\text{)} \\
&= \sum_{y \in \mathcal{R}(Y)} \sum_{\substack{x \in \mathcal{R}(X) \\ g(x) = y}} y \cdot \mathbb{P}(X = x) && \text{(distribute } y \text{ into the inner sum)} \\
&= \sum_{y \in \mathcal{R}(Y)} \sum_{\substack{x \in \mathcal{R}(X) \\ g(x) = y}} g(x) \cdot \mathbb{P}(X = x) && (y = g(x) \text{ for every term of the inner sum}) \\
&= \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathbb{P}(X = x) && \text{(the groups partition } \mathcal{R}(X)\text{)}
\end{aligned}
$$

The last step regroups a possibly infinite sum, which is valid because $\mathbb{E}[g(X)]$ is defined only when the sum converges absolutely.
We prove the case of a continuous random variable X and a function g that is strictly increasing and continuously differentiable on $\mathcal{R}(X)$, with inverse function $h = (g)^{-1}$. Let $Y = g(X)$. Because g is strictly increasing, $\{Y \le y\} = \{X \le h(y)\}$, so the CDF of Y is $\Pr(Y \le y) = \Pr(X \le h(y))$. Differentiating with the chain rule gives the density of Y: $\mathbb{p}(Y = y) = \mathbb{p}(X = h(y)) \cdot h'(y)$.

$$
\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}[Y] && \text{(substitution } Y = g(X)\text{)} \\
&= \int y \cdot \mathbb{p}(Y = y)\, dy && \text{(definition of expectation for continuous r.v.)} \\
&= \int y \cdot \mathbb{p}(X = h(y)) \cdot h'(y)\, dy && \text{(density of } Y\text{)} \\
&= \int g(x) \cdot \mathbb{p}(X = x)\, dx && \text{(substitute } y = g(x)\text{, so } h(y) = x \text{ and } h'(y)\, dy = dx\text{)}
\end{aligned}
$$

Each integral runs over the range of its variable, and g carries $\mathcal{R}(X)$ onto $\mathcal{R}(Y)$. For a strictly decreasing g, the same calculation holds with $\Pr(Y \le y) = \Pr(X \ge h(y))$ and $|h'(y)|$ in place of $h'(y)$. □

Remark

Remark. LOTUS says that to compute $\mathbb{E}[g(X)]$, we do not need to first find the distribution of $g(X)$; we can compute the expectation directly using the distribution of X. The proof for a continuous random variable follows the same plan, weighted by its density.

The proof above covers a strictly monotone, continuously differentiable g; a fully general proof needs the change-of-variables theorem for integrals against a pushforward measure — approximating g by simple functions and passing to the limit — which is beyond this page's scope [2], [5], [6], [7].

Example 0.6 (Expected value of X^2 for a Bernoulli variable)**.** Let $X \sim \text{Ber}(\pi)$. By LOTUS (Theorem 0.3, discrete case):

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{x \in \{0,1\}} x^2 \cdot \mathbb{P}(X = x) && \text{(LOTUS, discrete case)} \\
&= 0^2 \cdot \mathbb{P}(X = 0) + 1^2 \cdot \mathbb{P}(X = 1) && \text{(expand sum over } x = 0 \text{ and } x = 1) \\
&= 0^2 \cdot (1 - \pi) + 1^2 \cdot \pi && \text{(substitute Bernoulli PMF values)} \\
&= 0 + \pi && \text{(simplify each term)} \\
&= \pi && \text{(add)}
\end{aligned}$$

Example 0.7 (Expected value of X^2 for a Uniform(0,1) variable)**.** Let $X \sim \text{Uniform}(0, 1)$ (uniform distribution), so $\mathbb{p}(X = x) = 1$ for $x \in [0, 1]$. By LOTUS (Theorem 0.3, continuous case):

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^1 x^2 \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS, continuous case)} \\
&= \int_0^1 x^2 \cdot 1\, dx && (\mathbb{p}(X = x) = 1 \text{ on } [0, 1]) \\
&= \left[\frac{x^3}{3}\right]_0^1 && \text{(antiderivative of } x^2) \\
&= \frac{1}{3}. && \text{(evaluate at the bounds)}
\end{aligned}$$

Theorem 0.4 (Linearity of expectation)**.** *For random variables X and Y with defined expectations, and constants a, b, and c:*

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

Proof

Proof. We prove the discrete case, applying Theorem 0.3 to the pair (X, Y), which is a single discrete random object with values (x, y), and $g(x, y) = ax + by + c$:

$$
\begin{aligned}
\mathbb{E}[aX+bY+c] &= \sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} (ax+by+c)\cdot\mathbb{P}(X=x,Y=y) \\
&= a\sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} x\cdot\mathbb{P}(X=x,Y=y) + b\sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} y\cdot\mathbb{P}(X=x,Y=y) + c\sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)}\mathbb{P} \\
&= a\sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} x\cdot\mathbb{P}(X=x,Y=y) + b\sum_{y\in\mathcal{R}(Y)}\sum_{x\in\mathcal{R}(X)} y\cdot\mathbb{P}(X=x,Y=y) + c\sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)}\mathbb{P} \\
&= a\sum_{x\in\mathcal{R}(X)} x\sum_{y\in\mathcal{R}(Y)} \mathbb{P}(X=x,Y=y) + b\sum_{y\in\mathcal{R}(Y)} y\sum_{x\in\mathcal{R}(X)} \mathbb{P}(X=x,Y=y) + c\cdot 1 \\
&= a\sum_{x\in\mathcal{R}(X)} x\cdot\mathbb{P}(X=x) + b\sum_{y\in\mathcal{R}(Y)} y\cdot\mathbb{P}(Y=y) + c \\
&= a\mathbb{E}[X] + b\mathbb{E}[Y] + c
\end{aligned}
$$

Splitting and exchanging the sums is valid because the sums converge absolutely, since $|ax+by+c| \le |a||x| + |b||y| + |c|$ and $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are defined.
Beyond the discrete case, this proof does not carry over by replacing sums with integrals. That replacement would need a joint density for (X,Y), and some pairs have none: if X is continuous, the pair (X,X) has no joint density. In general, expectation is an integral against the probability measure, and the theorem is the linearity of that (Lebesgue) integral [2], which is beyond these notes' scope. □

Example 0.8 (Expected total of two dice)**.** Let X and Y be the results of two six-sided die rolls. Each has expectation

$$
\begin{aligned}
\sum_{x=1}^{6} x\cdot\tfrac{1}{6} &= \tfrac{21}{6} \\
&= 3.5,
\end{aligned}
$$

so:

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \mathbb{E}[X]+\mathbb{E}[Y] && \text{(linearity of expectation with } a=1,b=1,c=0) \\
&= 3.5+3.5 && \text{(substitute)} \\
&= 7 && \text{(add)}
\end{aligned}
$$

This calculation does not need the rolls to be independent.

Corollary 0.1 (Expectation of a linear function of one random variable)**.** *For a random variable X with a defined expectation, and constants a and c:*

$$\mathbb{E}[aX+c] = a\mathbb{E}[X] + c$$

Proof

Proof. For discrete X, this identity is Theorem 0.4 with $b = 0$. For continuous X with density $\mathbb{p}(X = x)$, apply Theorem 0.3 with $g(x) = ax + c$:

$$\begin{aligned}
\mathbb{E}[aX + c] &= \int_{x \in \mathcal{R}(X)} (ax + c) \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS, continuous case)} \\
&= a \int_{x \in \mathcal{R}(X)} x \cdot \mathbb{p}(X = x)\, dx + c \int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x)\, dx && \text{(linearity of the integral)} \\
&= a\mathbb{E}[X] + c \int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x)\, dx && \text{(definition of expectation)} \\
&= a\mathbb{E}[X] + c \cdot 1 && \text{(a density integrates to 1)} \\
&= a\mathbb{E}[X] + c && \text{(simplify)}
\end{aligned}$$

Splitting the integral is valid because both pieces converge absolutely: the first because $\mathbb{E}[X]$ is defined, and the second because its integrand is a non-negative density. □

Example 0.9 (Expectation of a rescaled uniform variable)**.** Let $X \sim \text{Uniform}(0, 1)$, so

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 x\, dx \\
&= \tfrac{1}{2}.
\end{aligned}$$

By Corollary 0.1 with $a = 2$ and $c = 1$:

$$\begin{aligned}
\mathbb{E}[2X + 1] &= 2\mathbb{E}[X] + 1 && \text{(expectation of a linear function)} \\
&= 2 \cdot \tfrac{1}{2} + 1 && \text{(substitute } \mathbb{E}[X] = \tfrac{1}{2}\text{)} \\
&= 2 && \text{(evaluate)}
\end{aligned}$$

Exercise 0.1 (Expected value of a density)**.** Let X be a continuous random variable with density $\mathbb{p}(X = x)$. The density evaluated at X is itself a random variable, $\mathbb{p}(X = X)$. Use Theorem 0.3 to write $\mathbb{E}[\mathbb{p}(X = X)]$ as an integral of the density.

Solution

Solution 0.1. Apply Theorem 0.3 to a continuous random variable, with $g(x) = \mathbb{p}(X = x)$:

$$\begin{aligned}
\mathbb{E}[\mathbb{p}(X = X)] &= \int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x) \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS, continuous case, with } g(x) = \mathbb{p}(X = x)) \\
&= \int_{x \in \mathcal{R}(X)} [\mathbb{p}(X = x)]^2\, dx. && \text{(combine the two factors)}
\end{aligned}$$

The expectation of the density is the integral of the squared density. It is large when the density is concentrated, because then X tends to land where the density is high.

Exercise 0.2 (Expected value of a uniform density)**.** By Solution 0.1, the expectation of the density of a continuous random variable X is $\int_{x \in \mathcal{R}(X)} [\mathbb{p}(X = x)]^2\, dx$.
Evaluate $\mathbb{E}[\mathbb{p}(X = X)]$ when $X \sim \text{Uniform}(a, b)$ (uniform distribution).

Solution

Solution 0.2. The density is $\mathbb{p}(X = x) = \frac{1}{b-a}$ on $[a, b]$ and 0 elsewhere:

$$\begin{aligned}
\mathbb{E}[\mathbb{p}(X = X)] &= \int_a^b \left(\frac{1}{b-a}\right)^2 dx && \text{(integral of the squared density; the density is 0 outside } [a, b]) \\
&= \frac{1}{(b-a)^2} \cdot (b-a) && \text{(the integrand does not depend on } x\text{, so the integral is the integrand time} \\
&= \frac{1}{b-a}. && \text{(cancel one factor of } b-a)
\end{aligned}$$

The density takes the single value $\frac{1}{b-a}$ on $[a, b]$, so its expectation is that same value.

Exercise 0.3 (Expected value of a normal density)**.** By Solution 0.1, the expectation of the density of a continuous random variable X is $\int_{x \in \mathcal{R}(X)} [\mathbb{p}(X = x)]^2\, dx$.
Evaluate $\mathbb{E}[\mathbb{p}(X = X)]$ when $X \sim \text{N}(\mu, \sigma^2)$ (normal distribution), using the integral $\int_{-\infty}^{\infty} \exp\{-u^2\}\, du = \sqrt{\pi}$.

Solution

Solution 0.3. The density is $\mathbb{p}(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$:

$$
\begin{aligned}
\mathbb{E}[\mathbb{p}(X = X)] &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(x-\mu)^2}{\sigma^2}\right\} dx && \text{(integral of the squared density; square the density)} \\
&= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\{-u^2\}\, \sigma\, du && \text{(substitute } u = (x-\mu)/\sigma\text{, so } dx = \sigma\, du\text{)} \\
&= \frac{\sigma\sqrt{\pi}}{2\pi\sigma^2} && \text{(the given integral)} \\
&= \frac{1}{2\sigma\sqrt{\pi}}. && \text{(simplify)}
\end{aligned}
$$

A larger standard deviation σ spreads the density out and lowers its expected value.

Lemma 0.1 (Linearity of expectation for finite sums)**.** *For random variables $X_1, \ldots, X_n$ with defined expectations, and constants $a_1, \ldots, a_n$:*

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

Proof

Proof. Prove this by induction on n, applying linearity of expectation for two variables in each step.
For $n = 1$:

$$
\begin{aligned}
\mathbb{E}[a_1 X_1] &= \mathbb{E}[a_1 X_1 + 0 \cdot X_1] && (0 \cdot X_1 = 0) \\
&= \mathbb{E}[a_1 X_1 + 0 \cdot X_1 + 0] && (v + 0 = v) \\
&= a_1 \mathbb{E}[X_1] + 0 \cdot \mathbb{E}[X_1] + 0 && \text{(linearity of expectation, } a = a_1, b = 0, c = 0\text{)} \\
&= a_1 \mathbb{E}[X_1] + 0 && (0 \cdot v = 0) \\
&= a_1 \mathbb{E}[X_1] && (v + 0 = v)
\end{aligned}
$$

For the inductive step, assume the lemma holds for $n-1$ variables, and let $W \stackrel{\text{def}}{=} \sum_{i=1}^{n-1} a_i X_i$:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] &= \mathbb{E}[W + a_n X_n] && \text{(split off the last term)} \\
&= \mathbb{E}[1 \cdot W + a_n X_n] && (1 \cdot W = W) \\
&= \mathbb{E}[1 \cdot W + a_n X_n + 0] && (v + 0 = v) \\
&= 1 \cdot \mathbb{E}[W] + a_n \mathbb{E}[X_n] + 0 && \text{(linearity of expectation, } a = 1, b = a_n, c = 0) \\
&= \mathbb{E}[W] + a_n \mathbb{E}[X_n] + 0 && (1 \cdot v = v) \\
&= \mathbb{E}[W] + a_n \mathbb{E}[X_n] && (v + 0 = v) \\
&= \sum_{i=1}^{n-1} a_i \mathbb{E}[X_i] + a_n \mathbb{E}[X_n] && \text{(inductive hypothesis)} \\
&= \sum_{i=1}^{n} a_i \mathbb{E}[X_i] && \text{(include the last term in the sum)}
\end{aligned}
$$

□

Example 0.10 (Expected total of five dice)**.** Let $X_1, \ldots, X_5$ be the outcomes of five rolls of a fair die. Each roll has expectation $\mathbb{E}[X_i] = 3.5$. Applying Lemma 0.1 with every $a_i = 1$ gives the expected total:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{5} X_i\right] &= \sum_{i=1}^{5} \mathbb{E}[X_i] && \text{(linearity of expectation for finite sums)} \\
&= 5 \cdot 3.5 && \text{(each roll has expectation 3.5)} \\
&= 17.5 && \text{(multiply)}
\end{aligned}
$$

Exercise 0.4 (Why the sample mean is a random variable)**.** Let $X_1, \ldots, X_n$ be independent draws from the same distribution, each with population mean μ and variance σ^2, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their sample mean.
Use Theorem 0.4 to show that $\mathbb{E}[\bar{X}] = \mu$, and say in one sentence what that does *not* tell us.

Solution

Solution 0.4. Pull the constant $1/n$ out and distribute the sum over the terms, both by Theorem 0.4:

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \frac{1}{n}(n\mu) \\ &= \mu\end{aligned}$$

So the sample mean is right *on average*. Independence was never used: Theorem 0.4 holds regardless, so this much would be true even for draws that influence one another. What it does not tell us is how far any *particular* sample mean falls from μ — that is a statement about variance, not expectation, and it is where independence does the work. The distinction matters because an estimator or fitted model is a function of one particular sample. Knowing that the procedure is right on average says nothing about the estimate actually in front of you, and separating those two statements is foundational to understanding bias and variance.

Theorem 0.5 (The expectation of a sample mean)**.** *Let $X_1, \ldots, X_n$ be random variables, with $n \geq 1$, that each have finite population mean μ, and let $\bar{X} \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. The X_i need not be independent. Then*

$$\mathbb{E}[\bar{X}] = \mu$$

Proof

Proof. This is the solution to Exercise 0.4, worked in Solution 0.4. It follows from Theorem 0.4, which needs no independence and applies because each $\mathbb{E}[X_i] = \mu$ is finite. □

Conditional distributions and expectations

Definition 0.1 (Conditional probability mass function)**.** Let X and Y be jointly distributed discrete random variables. The **conditional probability mass function** of Y given $X = x$ (for values of x with $\mathbb{P}(X = x) > 0$) is:

$$\mathbb{P}(Y = y \mid X = x) \overset{\text{def}}{=} \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)}$$

Example 0.1 (Conditional PMF from a vaccine trial)**.** The `vaccine` dataset in the `dobson` package (Dobson and Barnett [8], Table 9.6) records the responses in a flu vaccine trial: X is treatment group (placebo or vaccine) and Y is the level of response to treatment (small, moderate, or large).

```
response_levels <- c("small", "moderate", "large")
vaccine_tab <-
  dobson::vaccine |>
  dplyr::mutate(response = factor(response, levels = response_levels)) |>
  xtabs(frequency ~ treatment + response, data = _)
pander::pander(as.data.frame.matrix(vaccine_tab))
```

	small	moderate	large
placebo	25	8	5
vaccine	6	18	11

```
n_vaccine <- sum(vaccine_tab)
n_placebo <- vaccine_tab["placebo", ]
```

Choose one of the $n = 73$ trial participants at random (a random choice), and let (X, Y) be that participant's treatment and response. The joint PMF of (X, Y) is then the table of frequencies divided by the total n trial participants. By the marginal PMF theorem, the marginal probability $\mathbb{P}(X = \text{placebo})$ is:

$$\begin{aligned}\mathbb{P}(X = \text{placebo}) &= \mathbb{P}(X = \text{placebo}, Y = \text{small}) + \mathbb{P}(X = \text{placebo}, Y = \text{moderate}) + \mathbb{P}(X = \text{placebo}, Y = \text{large}) \\ &= \tfrac{25}{73} + \tfrac{8}{73} + \tfrac{5}{73} \\ &= \tfrac{38}{73}\end{aligned}$$

By Definition 0.1, the conditional PMF of Y given $X = \text{placebo}$ is:

$$\begin{aligned}\mathbb{P}(Y = \text{small} \mid X = \text{placebo}) &= \frac{\mathbb{P}(X = \text{placebo}, Y = \text{small})}{\mathbb{P}(X = \text{placebo})} && \text{(definition of conditional PMF)} \\ &= \frac{25/73}{38/73} && \text{(substitute)} \\ &= \frac{25}{38} && \text{(cancel 73)}\end{aligned}$$

Definition 0.2 (Conditional probability density function)**.** Let X and Y be jointly distributed continuous random variables with joint density $\mathbb{p}(X = x, Y = y)$ and marginal density $\mathbb{p}(X = x)$. The **conditional probability density function** of Y given $X = x$ (for values of x with $\mathbb{p}(X = x) > 0$) is:

$$\mathbb{p}(Y = y \mid X = x) \stackrel{\text{def}}{=} \frac{\mathbb{p}(X = x, Y = y)}{\mathbb{p}(X = x)}$$

Example 0.2 (Conditional PDF from a bivariate normal model of birthweight data)**.** The `birthweight` dataset in the `dobson` package (Dobson and Barnett [8], Table 2.3) records gestational age (weeks) and birthweight (grams) for 12 boys and 12 girls. Let X be gestational age and Y be birthweight, pooling both sexes into $n = 24$ observations.

```
birthweight <- dobson::birthweight
ga <- c(
  birthweight[["boys gestational age"]],
  birthweight[["girls gestational age"]]
)
wt <- c(birthweight[["boys weight"]], birthweight[["girls weight"]])
mu_x <- mean(ga)
sigma_x <- sd(ga)
mu_y <- mean(wt)
sigma_y <- sd(wt)
rho <- cor(ga, wt)
beta <- cov(ga, wt) / var(ga)
intercept <- mu_y - beta * mu_x
x0 <- 40
marg_dens_x0 <- dnorm(x0, mean = mu_x, sd = sigma_x)
cond_mean <- intercept + beta * x0
cond_sd <- sigma_y * sqrt(1 - rho^2)
tibble::tibble(
  parameter = c("mu_x", "sigma_x", "mu_y", "sigma_y", "rho"),
  value = round(c(mu_x, sigma_x, mu_y, sigma_y, rho), 4)
) |>
  pander::pander()
```

parameter	value
mu_x	38.54
sigma_x	1.817
mu_y	2968

sigma_y	282.1
rho	0.7443

Modeling (X, Y) as bivariate normal with parameters set equal to these values computed from the data (they stand in for the parameters here): μ_X and μ_Y are the means of X and Y, σ_X^2 and σ_Y^2 their variances, and ρ their correlation (the next page defines the last two). The joint density is (a standard result; e.g. Casella and Berger [6]):

$$\mathbb{p}(X = x,\, Y = y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]}$$

A further standard fact about the bivariate normal (Casella and Berger [6]) is that the marginal distribution of X is normal, $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$. At $x = 40$ weeks (a full-term pregnancy), $\mu_X = 38.5417$ and $\sigma_X = 1.8173$, so:

$$\begin{aligned}
\mathbb{p}(X = 40) &= \frac{1}{\sigma_X\sqrt{2\pi}} \mathrm{e}^{-\frac{(40-\mu_X)^2}{2\sigma_X^2}} && \text{(normal density at } x = 40) \\
&= \frac{1}{1.8173\sqrt{2\pi}} \mathrm{e}^{-\frac{(40-38.5417)^2}{2(1.8173)^2}} && \text{(substitute } \mu_X \text{ and } \sigma_X) \\
&\approx 0.1591 && \text{(evaluate)}
\end{aligned}$$

By Definition 0.2, dividing the joint density by this marginal density and simplifying the exponent (completing the square in y; Casella and Berger [6]) gives the conditional PDF of Y given $X = 40$, which is itself normal with mean shifted along the regression line and variance reduced by a factor of $1 - \rho^2$:

$$\begin{aligned}
\mathbb{p}(Y = y \mid X = 40) &= \frac{\mathbb{p}(X = 40,\, Y = y)}{\mathbb{p}(X = 40)} && \text{(definition of the c} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{(40-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(40-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right] + \frac{(40-\mu_X)^2}{2\sigma_X^2}} && \text{(substitute both d} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-\frac{\rho^2(40-\mu_X)^2}{2\sigma_X^2(1-\rho^2)} + \frac{\rho(40-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y(1-\rho^2)} - \frac{(y-\mu_Y)^2}{2\sigma_Y^2(1-\rho^2)}} && \text{(combine the (40 -} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[\rho^2\frac{\sigma_Y^2}{\sigma_X^2}(40-\mu_X)^2 - 2\rho\frac{\sigma_Y}{\sigma_X}(40-\mu_X)(y-\mu_Y) + (y-\mu_Y)^2\right]} && \text{(factor } -\tfrac{1}{2\sigma_Y^2(1-\rho^2)} \\
&= \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{-\frac{\left(y - \left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(40-\mu_X)\right]\right)^2}{2\sigma_Y^2(1-\rho^2)}} && \text{(the bracket is a p}
\end{aligned}$$

Multiplying out each of the last two exponents reproduces the one before it.
So $Y \mid X = 40 \sim \mathrm{N}(3136.15,\, 188.37^2)$: the conditional mean, 3136.15 g, is exactly the fitted regression line's prediction at $x = 40$ ($-1484.9846 + 115.5283 \times 40 = 3136.15$), matching R's `lm(wt ~ ga)` fit directly:

```
pander::pander(coef(lm(wt ~ ga)))
```

(Intercept)	ga
-1485	115.5

Definition 0.3 (Conditional expectation)**. Discrete case.** Let X and Y be jointly distributed discrete random variables. The **conditional expectation** of Y given $X = x$, using Definition 0.1, is [9]:

$$\mathbb{E}[Y \mid X = x] \stackrel{\text{def}}{=} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x)$$

Continuous case. Let X and Y be jointly distributed continuous random variables. The **conditional expectation** of Y given $X = x$, using Definition 0.2, is:

$$\mathbb{E}[Y \mid X = x] \stackrel{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x)\, dy$$

Example 0.3 (Conditional expectation from real trial and birthweight data)**. Discrete case.** Continuing Example 0.1, score the vaccine trial's response levels small = 1, moderate = 2, large = 3. The conditional PMF of Y given $X =$ placebo (from Example 0.1) is:

$$\begin{aligned}
\mathbb{P}(Y = \text{small} \mid X = \text{placebo}) &= \tfrac{25}{38} \\
\mathbb{P}(Y = \text{moderate} \mid X = \text{placebo}) &= \tfrac{8}{38} \\
\mathbb{P}(Y = \text{large} \mid X = \text{placebo}) &= \tfrac{5}{38}
\end{aligned}$$

so:

$$\begin{aligned}
\mathbb{E}[Y \mid X = \text{placebo}] &= 1 \cdot \tfrac{25}{38} + 2 \cdot \tfrac{8}{38} + 3 \cdot \tfrac{5}{38} && \text{(definition of conditional expectation)} \\
&= \frac{56}{38} && \text{(common denominator)} \\
&\approx 1.47 && \text{(divide)}
\end{aligned}$$

Continuous case. Continuing Example 0.2, $Y \mid X = 40 \sim \mathrm{N}(3136.15, 188.37^2)$. The mean of a normal distribution is its location parameter (Casella and Berger [6]), so integrating y against this conditional density gives:

$$\begin{aligned}
\mathbb{E}[Y \mid X = 40] &= \int_{-\infty}^{\infty} y \cdot \mathbb{p}(Y = y \mid X = 40)\, dy && \text{(definition of conditional expectation)} \\
&= 3136.15 \text{ g} && \text{(the mean of a normal distribution is its location parameter)}
\end{aligned}$$

matching the fitted regression line's prediction at $x = 40$ weeks exactly, as expected since the conditional mean of a bivariate normal *is* the linear regression of Y on X.

Definition 0.4 (Conditional expectation: mixed case)**.** Suppose exactly one of X, Y is discrete and the other is continuous, with joint density-mass function $\mathbb{p}(X = x,\, Y = y)$.

X **discrete,** Y **continuous.** Here $\mathbb{p}(X = x,\, Y = y)$ is, for each fixed x, a probability density in y, with $\int_{y \in \mathcal{R}(Y)} \mathbb{p}(X = x,\, Y = y)\, dy = \mathbb{P}(X = x)$. The conditional PDF of Y given $X = x$ (for values of x with $\mathbb{P}(X = x) > 0$) is:

$$\mathbb{p}(Y = y \mid X = x) \stackrel{\text{def}}{=} \frac{\mathbb{p}(X = x,\, Y = y)}{\mathbb{P}(X = x)}$$

and the conditional expectation of Y given $X = x$ is:

$$\mathbb{E}[Y \mid X = x] \stackrel{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x)\, dy$$

X **continuous,** Y **discrete.** Here $\mathbb{p}(X = x,\, Y = y)$ is, for each fixed y, a probability density in x, and $\sum_{y \in \mathcal{R}(Y)} \mathbb{p}(X = x,\, Y = y) = \mathbb{p}(X = x)$ is a density of X. The conditional PMF of Y given $X = x$ (for values of x with $\mathbb{p}(X = x) > 0$) is:

$$\mathbb{P}(Y = y \mid X = x) \stackrel{\text{def}}{=} \frac{\mathbb{p}(X = x,\, Y = y)}{\mathbb{p}(X = x)}$$

and the conditional expectation of Y given $X = x$ is:

$$\mathbb{E}[Y \mid X = x] \stackrel{\text{def}}{=} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x)$$

Example 0.4 (Conditional expectation, one discrete variable and one continuous variable)**.** X **discrete,** Y **continuous.** The `plasma` dataset in the `dobson` package (Dobson and Barnett [8], Table 6.25) records plasma inorganic phosphate levels (mg/dL) one hour after a glucose tolerance test, for hyperinsulinemic obese (`H-O`) and control (`C`) participants. Let X be group and Y be phosphate level.

```r
plasma_summary <-
  dobson::plasma |>
  dplyr::filter(Group %in% c("H-O", "C")) |>
  dplyr::summarise(
    .by = Group,
    n = dplyr::n(),
    mean = mean(phosphate),
    sd = sd(phosphate)
  )
group_stat <- function(group, stat) {
  plasma_summary[[stat]][plasma_summary[["Group"]] == group]
}
n_ho <- group_stat("H-O", "n")
n_c <- group_stat("C", "n")
mean_ho <- group_stat("H-O", "mean")
sd_ho <- group_stat("H-O", "sd")
mean_c <- group_stat("C", "mean")
sd_c <- group_stat("C", "sd")
pander::pander(plasma_summary)
```

Group	n	mean	sd
H-O	11	3.945	0.7776
C	12	2.783	0.4086

Modeling phosphate as approximately normal within each group, with parameters set equal to each group's sample mean and SD, and $\mathbb{P}(X = \text{H-O}) = 11/23$:

$$Y \mid X = \text{H-O} \sim \text{N}(3.95,\, 0.78^2)$$

$$Y \mid X = \text{C} \sim \text{N}(2.78,\, 0.41^2)$$

By Definition 0.4, since the mean of a normal distribution is its location parameter:

$$\mathbb{E}[Y \mid X = \text{H-O}] = 3.95 \text{ mg/dL}, \qquad \mathbb{E}[Y \mid X = \text{C}] = 2.78 \text{ mg/dL}$$

X continuous, Y discrete. The `senility` dataset in the `dobson` package (Dobson and Barnett [8], Table 7.8) records, for 54 elderly people, a WAIS (Wechsler Adult Intelligence Scale) score and whether symptoms of senility were present. Let X be WAIS score and Y indicate senility symptoms.

```
senility_fit <- glm(s ~ x, data = dobson::senility, family = binomial)
b0 <- coef(senility_fit)[["(Intercept)"]]
b1 <- coef(senility_fit)[["x"]]
pander::pander(coef(senility_fit))
```

(Intercept)	x
2.404	-0.3235

Modeling $\mathbb{P}(Y = 1 \mid X = x)$ with the fitted logistic regression:

$$\operatorname{logit} \mathbb{P}(Y = 1 \mid X = x) = 2.4040 - 0.3235\, x$$

By Definition 0.4, since $Y \mid X = x$ is Bernoulli:

$$\begin{aligned}
\mathbb{E}[Y \mid X = x] &= 0 \cdot \mathbb{P}(Y = 0 \mid X = x) + 1 \cdot \mathbb{P}(Y = 1 \mid X = x) && \text{(definition of conditional expectation, mixed ca} \\
&= \mathbb{P}(Y = 1 \mid X = x) && \text{(simplify)} \\
&= \frac{1}{1 + \mathrm{e}^{-(2.4040 - 0.3235\, x)}} && \text{(invert the logit)}
\end{aligned}$$

```
senility_p10 <- predict(
  senility_fit,
  newdata = tibble::tibble(x = 10),
  type = "response"
)
senility_p15 <- predict(
  senility_fit,
  newdata = tibble::tibble(x = 15),
  type = "response"
)
```

At $x = 10$: $\mathbb{E}[Y \mid X = 10] = 0.3$. At $x = 15$: $\mathbb{E}[Y \mid X = 15] = 0.08$ — a lower WAIS score is associated with higher predicted probability of senility symptoms.

Definition 0.5 (Conditional expectation function)**.** The **conditional expectation function** $\mathbb{E}[Y \mid X]$ is the function (and hence random variable) of X obtained by evaluating $\mathbb{E}[Y \mid X = x]$ at X; specifically, $\mathbb{E}[Y \mid X] = g(X)$ where $g(x) \stackrel{\text{def}}{=} \mathbb{E}[Y \mid X = x]$.

Example 0.5 (The conditional expectation function of the birthweight model)**.** Continuing Example 0.2 and Example 0.3, (X, Y) (gestational age, birthweight) is modeled as bivariate normal. The general form of the conditional mean derived in Example 0.2, evaluated at an arbitrary x instead of just $x = 40$, gives the conditional expectation function directly:

$$\begin{aligned}
g(x) &= \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \\
&= -1484.9846 + 115.5283\, x
\end{aligned}$$

which is exactly the fitted regression line (the general algebraic identity $\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x -$

$\mu_X) = \left(\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right) + \rho\frac{\sigma_Y}{\sigma_X}x$ is what makes the bivariate-normal conditional mean linear in x). As a check, evaluating at $x = 40$ recovers Example 0.3's result:

$$g(40) = -1484.9846 + 115.5283 \times 40 = 3136.15 \text{ g}$$

Definition 0.6 (Conditional expectation of a function of X and Y)**.** Let X and Y be jointly distributed random variables, h a function of two arguments, and x a value at which the conditional PMF or PDF of Y given $X = x$ is defined (Definition 0.1, Definition 0.2, or Definition 0.4). The **conditional expectation** of $h(X, Y)$ given $X = x$ is, for discrete Y:

$$\mathbb{E}[h(X,Y) \mid X = x] \stackrel{\text{def}}{=} \sum_{y \in \mathcal{R}(Y)} h(x,y) \cdot \mathbb{P}(Y = y \mid X = x)$$

and for continuous Y:

$$\mathbb{E}[h(X,Y) \mid X = x] \stackrel{\text{def}}{=} \int_{y \in \mathcal{R}(Y)} h(x,y) \cdot \mathbb{p}(Y = y \mid X = x)\, dy$$

when the sum or integral converges absolutely. Evaluating this function of x at X gives the random variable $\mathbb{E}[h(X, Y) \mid X]$.

Example 0.6 (The second flip, given the first)**.** In the joint PMF of two coin flips, write X for the first flip and Y for the total number of heads. Given $X = 1$, Y is 1 or 2 with conditional probability 1/2 each, and $Y - X$ is the second flip. With $h(x, y) = (y - x)^2$:

$$\begin{aligned}
\mathbb{E}[(Y-X)^2 \mid X = 1] &= (1-1)^2 \cdot \tfrac{1}{2} + (2-1)^2 \cdot \tfrac{1}{2} && \text{(definition, with } x = 1\text{)} \\
&= 0 + \tfrac{1}{2} && \text{(evaluate each term)} \\
&= \tfrac{1}{2} && \text{(add)}
\end{aligned}$$

Corollary 0.1 (With $h(x, y) = y$, the general definition gives $\mathbb{E}[Y \mid X = x]$)**.** *For jointly distributed X and Y and $h(x, y) = y$, Definition 0.6 gives the conditional expectation of Y given $X = x$ from Definition 0.3 (when X and Y are both discrete or both continuous) or from Definition 0.4 (when one is discrete and the other continuous):*

$$\mathbb{E}[h(X,Y) \mid X = x] = \mathbb{E}[Y \mid X = x]$$

Proof

Proof. For discrete Y:

$$\begin{aligned}
\mathbb{E}[h(X,Y) \mid X = x] &= \sum_{y \in \mathcal{R}(Y)} h(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && \text{(conditional expectation of a function, discrete } Y\text{)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x) && (h(x,y) = y) \\
&= \mathbb{E}[Y \mid X = x] && \text{(conditional expectation, discrete } Y\text{)}
\end{aligned}$$

The last step is Definition 0.3 when X is discrete and Definition 0.4 when X is continuous. For continuous Y, the same steps hold with integrals over y in place of the sums and $\mathbb{p}(Y = y \mid X = x)$ in place of $\mathbb{P}(Y = y \mid X = x)$. □

Theorem 0.1 (Conditional LOTUS, discrete case)**.** *Let X and Y be jointly distributed discrete random variables, h a function of two arguments, $W \stackrel{def}{=} h(X,Y)$, and x a value with $\mathbb{P}(X = x) > 0$. Then W is discrete, and $\mathbb{E}[h(X,Y) \mid X = x]$ from Definition 0.6 equals $\mathbb{E}[W \mid X = x]$ from Definition 0.3, with one sum converging absolutely exactly when the other does.*

Proof

Proof. W is discrete because its range is contained in $\{h(u,y) : u \in \mathcal{R}(X),\, y \in \mathcal{R}(Y)\}$, which is countable. For each $w \in \mathcal{R}(W)$, the event $\{X = x,\, W = w\}$ is the disjoint union of the events $\{X = x,\, Y = y\}$ over the values $y \in \mathcal{R}(Y)$ with $h(x,y) = w$, so:

$$\begin{aligned}
\mathbb{P}(W = w \mid X = x) &= \frac{\mathbb{P}(X = x,\, W = w)}{\mathbb{P}(X = x)} && \text{(conditional PMF of } W\text{)} \\
&= \frac{1}{\mathbb{P}(X = x)} \sum_{\substack{y \in \mathcal{R}(Y) \\ h(x,y) = w}} \mathbb{P}(X = x,\, Y = y) && \text{(countable additivity over the disjoint events } \{X \\
&= \sum_{\substack{y \in \mathcal{R}(Y) \\ h(x,y) = w}} \mathbb{P}(Y = y \mid X = x) && \text{(divide each term; conditional PMF of } Y\text{)}
\end{aligned}$$

Every $y \in \mathcal{R}(Y)$ with $\mathbb{P}(Y = y \mid X = x) > 0$ has $h(x,y) \in \mathcal{R}(W)$, because some outcome has $X = x$ and $Y = y$, and $W = h(x,y)$ there; the other values of y contribute terms equal to 0. So the groups $\{y \in \mathcal{R}(Y) : h(x,y) = w\}$, over $w \in \mathcal{R}(W)$, contain every y with a non-zero term, each exactly once:

$$
\begin{aligned}
\mathbb{E}[W \mid X = x] &= \sum_{w \in \mathcal{R}(W)} w \cdot \mathbb{P}(W = w \mid X = x) && \text{(conditional expectation of } W\text{)} \\
&= \sum_{w \in \mathcal{R}(W)} w \cdot \sum_{\substack{y \in \mathcal{R}(Y) \\ h(x,y)=w}} \mathbb{P}(Y = y \mid X = x) && \text{(the display above)} \\
&= \sum_{w \in \mathcal{R}(W)} \sum_{\substack{y \in \mathcal{R}(Y) \\ h(x,y)=w}} w \cdot \mathbb{P}(Y = y \mid X = x) && \text{(distribute } w \text{ into the inner sum)} \\
&= \sum_{w \in \mathcal{R}(W)} \sum_{\substack{y \in \mathcal{R}(Y) \\ h(x,y)=w}} h(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && (w = h(x,y) \text{ for every term of the inner sum}) \\
&= \sum_{y \in \mathcal{R}(Y)} h(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && \text{(the groups contain every non-zero term once)} \\
&= \mathbb{E}[h(X,Y) \mid X = x] && \text{(conditional expectation of a function of } X \text{ and } Y
\end{aligned}
$$

The same steps with $|w|$ and $|h(x,y)|$ in place of w and $h(x,y)$ regroup a sum of non-negative terms, which is valid whether or not it converges, so one sum converges absolutely exactly when the other does, and then the regrouping above is valid too. □

i Remark

Remark. Definition 0.6 is the conditional version of LOTUS: for discrete X, it is LOTUS applied under the probability measure $\Pr(\cdot \mid X = x)$, and by Theorem 0.1, results about $\mathbb{E}[W \mid X]$ apply to it. Y can also be a pair (Y, Z), with the sum or integral taken over both.

Theorem 0.2 (Linearity of conditional expectation)**.** *For jointly distributed X and Y, functions h_1 and h_2 whose conditional expectations given $X = x$ are defined, and constants a and b:*

$$\mathbb{E}[ah_1(X,Y) + bh_2(X,Y) \mid X = x] = a\mathbb{E}[h_1(X,Y) \mid X = x] + b\mathbb{E}[h_2(X,Y) \mid X = x]$$

Evaluating both sides at X gives $\mathbb{E}[ah_1(X,Y) + bh_2(X,Y) \mid X] = a\mathbb{E}[h_1(X,Y) \mid X] + b\mathbb{E}[h_2(X,Y) \mid X]$.

i Proof

Proof. For discrete Y:

$$
\begin{aligned}
&\mathbb{E}[ah_1(X,Y) + bh_2(X,Y) \mid X = x] \\
&= \sum_{y \in \mathcal{R}(Y)} (ah_1(x,y) + bh_2(x,y)) \cdot \mathbb{P}(Y = y \mid X = x) && \text{(definition of conditional expec} \\
&= a \sum_{y \in \mathcal{R}(Y)} h_1(x,y) \cdot \mathbb{P}(Y = y \mid X = x) + b \sum_{y \in \mathcal{R}(Y)} h_2(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && \text{(distribute the sum over the te} \\
&= a\mathbb{E}[h_1(X,Y) \mid X = x] + b\mathbb{E}[h_2(X,Y) \mid X = x] && \text{(definition of conditional expec}
\end{aligned}
$$

Splitting the sum is valid because both sums converge absolutely. For continuous Y, the same steps hold with integrals over y in place of the sums, by linearity of the integral: unlike Theorem 0.4, only one variable, y, is integrated, against one conditional density. □

Example 0.7 (Linearity, given the first flip)**.** Continuing Example 0.6, given $X = 1$, Y is 1 or 2 with probability $1/2$ each, so $\mathbb{E}[Y \mid X = 1] = 3/2$ and

$$
\begin{aligned}
\mathbb{E}[Y^2 \mid X = 1] &= (1 + 4)/2 \\
&= 5/2.
\end{aligned}
$$

By Theorem 0.2 with

$$
\begin{aligned}
a &= b \\
&= 1 :
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[Y + Y^2 \mid X = 1] &= \mathbb{E}[Y \mid X = 1] + \mathbb{E}[Y^2 \mid X = 1] && \text{(linearity of conditional expectation)} \\
&= \tfrac{3}{2} + \tfrac{5}{2} && \text{(substitute)} \\
&= 4 && \text{(add)}
\end{aligned}
$$

As a check, directly:

$$
\begin{aligned}
(1 + 1) \cdot \tfrac{1}{2} + (2 + 4) \cdot \tfrac{1}{2} &= 1 + 3 \\
&= 4.
\end{aligned}
$$

Theorem 0.3 (A function of X factors out of a conditional expectation given X)**.** *For jointly distributed X and Y, a function g of one argument, and a function h of two arguments whose conditional expectation given $X = x$ is defined:*

$$
\mathbb{E}[g(X)\, h(X,Y) \mid X = x] = g(x) \cdot \mathbb{E}[h(X,Y) \mid X = x]
$$

Evaluating both sides at X gives $\mathbb{E}[g(X)\, h(X,Y) \mid X] = g(X) \cdot \mathbb{E}[h(X,Y) \mid X]$. In partic-

ular, taking $h = 1$ gives $\mathbb{E}[g(X) \mid X] = g(X)$.

Proof

Proof. For discrete Y:

$$\begin{aligned}
\mathbb{E}[g(X)\,h(X,Y) \mid X = x] &= \sum_{y \in \mathcal{R}(Y)} g(x)\,h(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && \text{(definition of conditional expectation)} \\
&= g(x) \sum_{y \in \mathcal{R}(Y)} h(x,y) \cdot \mathbb{P}(Y = y \mid X = x) && (g(x) \text{ does not depend on } y) \\
&= g(x) \cdot \mathbb{E}[h(X,Y) \mid X = x] && \text{(definition of conditional expectation)}
\end{aligned}$$

For $h = 1$, the remaining sum is $\sum_{y \in \mathcal{R}(Y)} \mathbb{P}(Y = y \mid X = x) = 1$:

$$\begin{aligned}
\sum_{y \in \mathcal{R}(Y)} \mathbb{P}(Y = y \mid X = x) &= \sum_{y \in \mathcal{R}(Y)} \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} && \text{(definition of the conditional PMF)} \\
&= \frac{\mathbb{P}(X = x)}{\mathbb{P}(X = x)} && \text{(marginal PMF from a joint PMF)} \\
&= 1 && \text{(divide)}
\end{aligned}$$

For continuous Y, the same steps hold with integrals over y in place of the sums; the conditional density integrates to 1 by the marginal density theorem (or, in the mixed case, by the joint density-mass function's marginal identity). □

Remark

Remark. This result is often summarized as "taking out what is known" (e.g., Ross [10], Section 5.6.4): given $X = x$, any function of X is the known constant $g(x)$.

Example 0.8 (Taking out the first flip)**.** Continuing Example 0.7, with $g(x) = x$ and $h(x, y) = y$:

$$\begin{aligned}
\mathbb{E}[XY \mid X = 1] &= 1 \cdot \mathbb{E}[Y \mid X = 1] && \text{(a function of } X \text{ factors out)} \\
&= \tfrac{3}{2} && \text{(substitute } \mathbb{E}[Y \mid X = 1] = \tfrac{3}{2})
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[XY \mid X = 0] &= 0 \cdot \mathbb{E}[Y \mid X = 0] \\
&= 0.
\end{aligned}$$

As a check, given $X = 1$ the product XY equals Y, which is 1 or 2 with probability $1/2$ each.

Exercise 0.1 (Expectation of a sum, given a joint PMF)**.** Let (X, Y) be discrete with joint probability mass function:

	$Y = 0$	$Y = 1$
$X = 0$	0.2	0.3
$X = 1$	0.1	0.4

Compute $\mathbb{E}[X + Y]$.

Solution

Solution. Treating (X, Y) as a single discrete random object taking one of the four values $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, LOTUS (Theorem 0.3) applied to $g(x, y) = x + y$ gives directly:

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{x\in\{0,1\}} \sum_{y\in\{0,1\}} (x + y)\, \mathbb{P}(X = x,\, Y = y) && \text{(LOTUS)} \\
&= (0{+}0)(0.2) + (0{+}1)(0.3) + (1{+}0)(0.1) + (1{+}1)(0.4) && \text{(expand the four terms)} \\
&= 0 + 0.3 + 0.1 + 0.8 && \text{(multiply)} \\
&= 1.2 && \text{(add)}
\end{aligned}$$

As a check:

$$\begin{aligned}
\mathbb{E}[X] &= 0(0.5) + 1(0.5) \\
&= 0.5
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[Y] &= 0(0.3) + 1(0.7) \\
&= 0.7,
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}[X + Y] &= \mathbb{E}[X] + \mathbb{E}[Y] \\
&= 1.2
\end{aligned}$$

by linearity of expectation.

(This bar chart is interactive; see the HTML or slide version of this page to view it.)

Figure 1: Joint probability mass function $\mathbb{P}(X = x, Y = y)$. Marginal totals: $\mathbb{P}(X = 0) = 0.5$, $\mathbb{P}(X = 1) = 0.5$, $\mathbb{P}(Y = 0) = 0.3$, $\mathbb{P}(Y = 1) = 0.7$.

Remark

Remark. Note that X and Y are **not** independent here:

$$\begin{aligned}\mathbb{P}(X = 0, Y = 0) &= 0.2 \\ &\neq 0.15 \\ &= \mathbb{P}(X = 0)\,\mathbb{P}(Y = 0).\end{aligned}$$

LOTUS applies regardless, since it requires only the *actual* joint mass function, not independence.

There are only four (x, y) pairs here, so summing them in any order — row by row, column by column, or any other listing — gives the same total by ordinary commutativity and associativity of addition; no result about interchanging summation order is needed. Once the support is countably infinite, exchanging the order of summation needs a justification, which the Fubini–Tonelli theorem provides.

Fubini–Tonelli for expectations

Theorem 0.1 (Fubini–Tonelli theorem, measure-theoretic form)**.** *Let μ_1 and μ_2 be σ-finite measures on σ-algebras of subsets of sets S_1 and S_2, and let f be a measurable function on $S_1 \times S_2$. If either*

(a) $f \geq 0$ (Tonelli), or

(b) $\int_{S_1 \times S_2} |f| \, d(\mu_1 \otimes \mu_2) < \infty$ (Fubini),

then the integral of f against the product measure $\mu_1 \otimes \mu_2$ equals both iterated integrals:

$$\begin{aligned} \int_{S_1 \times S_2} f \, d(\mu_1 \otimes \mu_2) &= \int_{S_1} \left(\int_{S_2} f(s_1, s_2) \, d\mu_2(s_2) \right) d\mu_1(s_1) \\ &= \int_{S_2} \left(\int_{S_1} f(s_1, s_2) \, d\mu_1(s_1) \right) d\mu_2(s_2). \end{aligned}$$

In case (b), each inner integral is finite except on a set of measure 0, which does not affect the outer integral.

Remark

Remark. For expectations, we use this measure-theoretic form of the Fubini–Tonelli theorem, which lets us exchange the order of integration (or summation) over a product of σ-finite measure spaces, provided the integrand is non-negative (Tonelli) or absolutely integrable (Fubini). Its proof is beyond these notes' scope [2], Theorem 18.3. Lebesgue measure on the real line and counting measure on a countable set are both σ-finite, which gives the theorem a form stated in terms of a joint distribution.

Corollary 0.1 (Joint-distribution form (without independence; corollary of Fubini–Tonelli))**.** *Let (X, Y) be jointly distributed random variables whose joint distribution has a density $f_{X,Y}$ with respect to a product of σ-finite reference measures $\mu_X \otimes \mu_Y$ on $\mathcal{R}(X) \times \mathcal{R}(Y)$, and let $h : \mathcal{R}(X) \times \mathcal{R}(Y) \to \mathbb{R}$ be measurable. If either*

(a) $h(X, Y) \geq 0$ almost surely, or

(b) $\mathbb{E}[|h(X, Y)|] < \infty$,

then the expectation of $h(X, Y)$ can be written as an iterated integral against $f_{X,Y}$, with the order of integration exchangeable:

$$\mathbb{E}[h(X,Y)] = \int_{\mathcal{R}(X)} \left(\int_{\mathcal{R}(Y)} h(x,y)\, f_{X,Y}(x,y)\, d\mu_Y(y) \right) d\mu_X(x)$$
$$= \int_{\mathcal{R}(Y)} \left(\int_{\mathcal{R}(X)} h(x,y)\, f_{X,Y}(x,y)\, d\mu_X(x) \right) d\mu_Y(y).$$

The choice of reference measures covers three cases:

- ***Both continuous:***

$$\mu_X = \mu_Y$$
$$= \textit{Lebesgue measure};$$

  $f_{X,Y}$ *is the joint probability density function (PDF), and* $\int_{\mathcal{R}(X)} g(x)\, d\mu_X(x) = \int_{\mathcal{R}(X)} g(x)\, dx$.

- ***Both discrete:***

$$\mu_X = \mu_Y$$
$$= \textit{counting measure};$$

  $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ *is the joint probability mass function (PMF), and* $\int_{\mathcal{R}(X)} g(x)\, d\mu_X(x) = \sum_{x \in \mathcal{R}(X)} g(x)$.

- ***Mixed*** *(one continuous, one discrete): one reference measure is Lebesgue and the other is counting;* $f_{X,Y}$ *is the joint density-mass function,* $f_{X,Y}(x,y) = f_{X|Y}(x \mid y)\, \mathbb{P}(Y = y)$ *(or* $\mathbb{P}(X = x \mid Y = y)\, f_Y(y)$ *if* X *is discrete and* Y *continuous), and the iterated integrals combine an ordinary integral with a sum. The conditional densities/PMFs here are defined the same way as in Definition 0.4, just conditioning on* Y *instead of* X.

Proof

Proof. Apply Theorem 0.1 with $\mu_1 = \mu_X$ and $\mu_2 = \mu_Y$ to the integrand $h(x,y)\, f_{X,Y}(x,y)$ on $\mathcal{R}(X) \times \mathcal{R}(Y)$. Lebesgue measure and counting measure on a countable set are each σ-finite, so $\mu_X \otimes \mu_Y$ is σ-finite in all three cases. The relevant condition is (a) when $h \geq 0$ and (b) when $\mathbb{E}[|h(X,Y)|] < \infty$. Independence is not required. When X and Y are independent, $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ (or $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\, \mathbb{P}(Y = y)$ in the discrete case), and the iterated integrals factor into separate integrals over the marginals. □

Example 0.1 (Expectation of a product of independent variables)**.** Let $X \sim \text{Uniform}(0,1)$ and $Y \sim \text{Uniform}(0,2)$, independently distributed. Compute $\mathbb{E}[XY]$.
We apply Corollary 0.1 (both-continuous case) with $h(x,y) = xy$. Since X and Y are independent with densities $f_X(x) = 1$ on $[0,1]$ and $f_Y(y) = \frac{1}{2}$ on $[0,2]$, the joint density factors as

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)\, f_Y(y) \\ &= \tfrac{1}{2}, \end{aligned}$$

and

$$\begin{aligned} \mu_X &= \mu_Y \\ &= \text{Lebesgue measure} : \end{aligned}$$

$$\begin{aligned} \mathbb{E}[XY] &= \int_0^1 \left(\int_0^2 xy \cdot \tfrac{1}{2}\, dy \right) dx && \text{(joint-distribution form of Fubini–Tonelli)} \\ &= \int_0^1 x \left(\frac{1}{2} \int_0^2 y\, dy \right) dx && \text{(factor constants out of the inner integral)} \\ &= \int_0^1 x \cdot \frac{1}{2} \cdot \left[\frac{y^2}{2} \right]_0^2 dx && \text{(antiderivative of } y\text{)} \\ &= \int_0^1 x \cdot \frac{1}{2} \cdot 2\, dx && \text{(evaluate at the bounds)} \\ &= \int_0^1 x\, dx && \text{(simplify)} \\ &= \frac{1}{2} && \text{(integrate)} \end{aligned}$$

As a check: $\mathbb{E}[X] = \frac{1}{2}$, $\mathbb{E}[Y] = 1$, and $\mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{2}$.

Example 0.2 (When independence fails: a counterexample)**.** Correctly applying Corollary 0.1 requires the *actual* joint density $f_{X,Y}$ — not the product of marginals $f_X(x)\, f_Y(y)$, which is valid only when X and Y are independent. Using the wrong joint density gives the wrong answer.
Let $X \sim \text{Uniform}(0,1)$ and set $Y = X$ (so X and Y are perfectly correlated and **not** independent).
True expectation:

$$\begin{aligned}
\mathbb{E}[XY] &= \mathbb{E}[X \cdot X] && (Y = X) \\
&= \mathbb{E}[X^2] && (X \cdot X = X^2) \\
&= \int_0^1 x^2 \, dx && \text{(LOTUS, with the uniform density 1 on } [0,1]) \\
&= \frac{1}{3} && \text{(integrate)}
\end{aligned}$$

Erroneously applying the product-measure formula:
Note that Fubini–Tonelli's own conditions still hold here ($h(x,y) = xy$ is nonnegative and integrable), so the error is not a failure of Fubini–Tonelli itself. Rather, the error is using the *wrong measure*: the joint distribution of (X, X) is concentrated on the diagonal $\{(x,x) : x \in [0,1]\} \subset [0,1]^2$, which has Lebesgue measure zero in $\mathbb{R}^2$. The joint distribution is therefore **not** absolutely continuous with respect to two-dimensional Lebesgue measure, so **no joint density $f_{X,Y}$ on $[0,1]^2$ exists**, which is the reference density Corollary 0.1 requires.
The following calculation is what someone would *erroneously* write if they assumed independence and used $f_X(x)\, f_Y(y)$ as a "joint density" — a function that does not in fact correspond to the joint distribution of (X, X). The marginals $X \sim \text{Uniform}(0,1)$ and $Y \sim \text{Uniform}(0,1)$ do have densities

$$\begin{aligned}
f_X &= f_Y \\
&= 1,
\end{aligned}$$

but the *product* $f_X(x)\, f_Y(y) = 1$ on $[0,1]^2$ is the density of an *independent* pair, not of (X, X):

$$\begin{aligned}
\int_0^1 \int_0^1 xy \cdot f_X(x) \cdot f_Y(y) \, dy \, dx &= \int_0^1 \int_0^1 xy \, dy \, dx && (f_X(x)\, f_Y(y) = 1 \text{ on } [0,1]^2) \\
&= \int_0^1 x \left(\int_0^1 y \, dy \right) dx && \text{(factor } x \text{ out of the inner integral)} \\
&= \int_0^1 x \cdot \frac{1}{2} \, dx && \left(\textstyle\int_0^1 y \, dy = \frac{1}{2} \right) \\
&= \frac{1}{4} && \left(\textstyle\int_0^1 \frac{x}{2} \, dx = \frac{1}{4} \right)
\end{aligned}$$

This calculation recovers $\mathbb{E}[XY]$ for *independent* uniforms ($\frac{1}{4}$), not $\mathbb{E}[XX]$ for the perfectly correlated pair ($\frac{1}{3}$). The lesson is that Corollary 0.1 requires the *actual* joint density $f_{X,Y}$. For independent (X, Y), this density factors as $f_X(x)\, f_Y(y)$; for dependent (X, Y), $f_{X,Y}$ need not factor — and for (X, X), no joint density on $\mathbb{R}^2$ exists at all, so Corollary 0.1 simply does not apply.

(This scatter plot is interactive; see the HTML or slide version of this page to view it.)

Figure 2: Samples from the joint distribution of (X, X) (red, on the diagonal) versus an independent pair (X_1, X_2) with the same marginals (gray, scattered over $[0,1]^2$). The actual joint mass for (X, X) is concentrated on a 1-dimensional diagonal — a set of Lebesgue measure zero in $\mathbb{R}^2$ — so no joint density on $[0,1]^2$ exists, and the "$f_X(x)\, f_Y(y) = 1$" calculation integrates against the wrong measure (the gray distribution).

Example 0.3 (Both-continuous case: joint PDF on a non-rectangular support)**.** Let (X, Y) have joint density $f_{X,Y}(x, y) = 2$ for $0 \le x \le y \le 1$ (and 0 otherwise). Compute $\mathbb{E}[X + Y]$.
By Corollary 0.1:

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \int_0^1 \int_0^y (x+y) \cdot 2 \, dx \, dy && \text{(joint-distribution form of Fubini–Tonelli, over the support } 0 \le x \le y \le \\
&= 2 \int_0^1 \left[\frac{x^2}{2} + xy \right]_{x=0}^{x=y} dy && \text{(antiderivative in } x\text{)} \\
&= 2 \int_0^1 \left(\frac{y^2}{2} + y^2 \right) dy && \text{(evaluate at the bounds)} \\
&= 2 \int_0^1 \frac{3y^2}{2} \, dy && \text{(add)} \\
&= 3 \int_0^1 y^2 \, dy && \text{(simplify the constant)} \\
&= 3 \cdot \frac{1}{3} && \text{(integrate)} \\
&= 1 && \text{(multiply)}
\end{aligned}
$$

(This 3D surface is interactive; see the HTML or slide version of this page to view it.)

Figure 3: Joint density $f_{X,Y}(x, y) = 2$ on the triangular support $\{(x, y) : 0 \le x \le y \le 1\}$, and zero elsewhere. The total "volume" under the density is $2 \cdot \frac{1}{2} = 1$, as required.

Theorem 0.2 (Law of iterated expectations)**.** *For any two random variables X and Y with $\mathbb{E}[|Y|] < \infty$:*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$$

Proof

Proof. **Discrete case.** When X and Y are discrete, applying Definition 0.1 to $\mathbb{E}[\mathbb{E}[Y \mid X]]$ and then the law of total probability applied to the countable partition $\{X = x : x \in \mathcal{R}(X)\}$:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_{x \in \mathcal{R}(X)} \mathbb{E}[Y \mid X = x] \cdot \mathbb{P}(X = x) && \text{(LOTUS for the function } x \mapsto \mathbb{E}[Y \mid X = x]\text{)} \\
&= \sum_{x \in \mathcal{R}(X)} \left(\sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x) \right) \cdot \mathbb{P}(X = x) && \text{(definition of conditional expectation } \mathbb{E}[Y \mid \text{)} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{x \in \mathcal{R}(X)} \mathbb{P}(Y = y \mid X = x) \cdot \mathbb{P}(X = x) && \text{(exchange order of summation: Fubini, since} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y) && \text{(law of total probability over } X\text{)} \\
&= \mathbb{E}[Y] && \text{(definition of expectation of } Y\text{)}
\end{aligned}$$

Continuous case. When X and Y are continuous:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \int_{x \in \mathcal{R}(X)} \mathbb{E}[Y \mid X = x] \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS)} \\
&= \int_{x \in \mathcal{R}(X)} \left(\int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x)\, dy \right) \cdot \mathbb{p}(X = x)\, dx && \text{(definition of conditional expectation} \\
&= \int_{x \in \mathcal{R}(X)} \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(X = x, Y = y)\, dy\, dx && \text{(definition of conditional density: } \mathbb{p}\text{(} \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \left(\int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x, Y = y)\, dx \right) dy && \text{(Fubini's theorem, since } \mathbb{E}[|Y|] < \infty\text{)} \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y)\, dy && \text{(marginal density from a joint densi} \\
&= \mathbb{E}[Y] && \text{(definition of expectation)}
\end{aligned}$$

Mixed case, X discrete and Y continuous. With the joint density-mass function $\mathbb{p}(X = x, Y = y)$, Corollary 0.1 applies with $h(x, y) = y$, counting measure for X, and Lebesgue measure for Y; its condition (b) holds because $\mathbb{E}[|Y|] < \infty$. The sum runs over the x with $\mathbb{P}(X = x) > 0$, where Definition 0.4 defines $\mathbb{E}[Y \mid X = x]$:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \sum_{x \in \mathcal{R}(X)} \mathbb{E}[Y \mid X = x] \cdot \mathbb{P}(X = x) && \text{(LOTUS, discrete case)} \\
&= \sum_{x \in \mathcal{R}(X)} \left(\int_{\mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x)\, dy \right) \cdot \mathbb{P}(X = x) && \text{(definition of conditional expectation, mix} \\
&= \sum_{x \in \mathcal{R}(X)} \int_{\mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x) \cdot \mathbb{P}(X = x)\, dy && \text{(move the constant } \mathbb{P}(X = x) \text{ inside the i} \\
&= \sum_{x \in \mathcal{R}(X)} \int_{\mathcal{R}(Y)} y \cdot \mathbb{p}(X = x, Y = y)\, dy && \text{(definition of the conditional density, mixe} \\
&= \mathbb{E}[Y] && \text{(joint-distribution form of Fubini–Tonelli,}
\end{aligned}
$$

Mixed case, X continuous and Y discrete. The same steps apply with the integral and the sum swapped, over the x with $\mathbb{p}(X = x) > 0$; at any other x, $\mathbb{p}(X = x, Y = y) = 0$ for every y, because these non-negative terms sum to $\mathbb{p}(X = x)$:

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X]] &= \int_{\mathcal{R}(X)} \mathbb{E}[Y \mid X = x] \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS, continuous case)} \\
&= \int_{\mathcal{R}(X)} \left(\sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x) \right) \cdot \mathbb{p}(X = x)\, dx && \text{(definition of conditional expectation, mix} \\
&= \int_{\mathcal{R}(X)} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x) \cdot \mathbb{p}(X = x)\, dx && \text{(move the constant } \mathbb{p}(X = x) \text{ inside the s} \\
&= \int_{\mathcal{R}(X)} \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(X = x, Y = y)\, dx && \text{(definition of the conditional PMF, mixed} \\
&= \mathbb{E}[Y] && \text{(joint-distribution form of Fubini–Tonelli,}
\end{aligned}
$$

□

Remark

Remark. Alternate names for this identity include: the **tower rule**, the **tower property**, the **law of total expectation**, and the **smoothing theorem**.

Example 0.4 (Fuel economy by number of cylinders)**.** The Auto data record the number of engine cylinders and the fuel economy (miles per gallon) of 392 car models sold between 1970 and 1982. They come from the StatLib library at Carnegie Mellon University and ship with the book *An Introduction to Statistical Learning*, in the Python package `ISLP` (`ISLP.load_data("Auto")`) and the R package `ISLR2` (`ISLR2::Auto`). This site has a

copy at data/auto.csv, and both code tabs below read that file.
Choose one car at random, each with the same probability, and let X be its number of cylinders and Y its miles per gallon. Then $\mathbb{P}(X = x)$ is the fraction of cars with x cylinders, and $\mathbb{E}[Y \mid X = x]$ from Definition 0.3 is the average miles per gallon among those cars. The code below computes both for every x, then checks Theorem 0.2 by averaging the conditional means with weights $\mathbb{P}(X = x)$.

Python

```python
import pandas as pd

auto = pd.read_csv("data/auto.csv")

by_cyl = auto.groupby("cylinders")["mpg"].agg(["count", "mean"])
by_cyl["P(X = x)"] = by_cyl["count"] / by_cyl["count"].sum()
by_cyl = by_cyl.rename(columns={"mean": "E(Y | X = x)"})
by_cyl = by_cyl[["count", "P(X = x)", "E(Y | X = x)"]]

lie_rhs = (by_cyl["P(X = x)"] * by_cyl["E(Y | X = x)"]).sum()
lie_lhs = auto["mpg"].mean()

by_cyl
#>            count  P(X = x)  E(Y | X = x)
#> cylinders
#> 3              4  0.010204     20.550000
#> 4            199  0.507653     29.283920
#> 5              3  0.007653     27.366667
#> 6             83  0.211735     19.973494
#> 8            103  0.262755     14.963107
```

R

```r
auto <- read.csv("data/auto.csv")

counts <- table(auto$cylinders)
cond_mean <- tapply(auto$mpg, auto$cylinders, mean)
by_cyl <- data.frame(
  count = as.vector(counts),
  "P(X = x)" = as.vector(counts) / sum(counts),
  "E(Y | X = x)" = as.vector(cond_mean),
  row.names = names(counts),
  check.names = FALSE
)

lie_rhs <- sum(by_cyl[["P(X = x)"]] * by_cyl[["E(Y | X = x)"]])
lie_lhs <- mean(auto$mpg)

by_cyl
#>   count    P(X = x) E(Y | X = x)
#> 3     4 0.010204082     20.55000
#> 4   199 0.507653061     29.28392
#> 5     3 0.007653061     27.36667
#> 6    83 0.211734694     19.97349
#> 8   103 0.262755102     14.96311
```

The conditional mean $\mathbb{E}[Y \mid X = x]$ is the function $g(x)$ of Definition 0.5. Four-cylinder cars average 29.3 miles per gallon, and eight-cylinder cars average 15.0. Weighting the 5 conditional means by $\mathbb{P}(X = x)$ gives 23.446, which equals the overall average miles per gallon, 23.446, as Theorem 0.2 states: $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$.

Theorem 0.3 (Conditional law of iterated expectations)**.** *For random variables X, Y, and Z with $\mathbb{E}[|Y|] < \infty$:*

$$\mathbb{E}[Y \mid Z] = \mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid Z]$$

Proof

Proof. Choose a value z with positive probability (discrete case) or positive density (continuous case). Conditioning every probability on $Z = z$ gives a probability measure, and the proof of Theorem 0.2 goes through under it, line by line.
Discrete case.

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid Z = z] &= \sum_{x \in \mathcal{R}(X)} \mathbb{E}[Y \mid X = x, Z = z] \cdot \mathbb{P}(X = x \mid Z = z) && \text{(expectation given} \\
&= \sum_{x \in \mathcal{R}(X)} \left(\sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid X = x, Z = z) \right) \cdot \mathbb{P}(X = x \mid Z = z) && \text{(definition of } \mathbb{E}[Y \mid \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \sum_{x \in \mathcal{R}(X)} \mathbb{P}(Y = y \mid X = x, Z = z) \cdot \mathbb{P}(X = x \mid Z = z) && \text{(exchange order of} \\
&= \sum_{y \in \mathcal{R}(Y)} y \cdot \mathbb{P}(Y = y \mid Z = z) && \text{(law of total proba} \\
&= \mathbb{E}[Y \mid Z = z] && \text{(definition of condi}
\end{aligned}
$$

Continuous case.

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid Z = z] &= \int_{x \in \mathcal{R}(X)} \mathbb{E}[Y \mid X = x, Z = z] \cdot \mathbb{p}(X = x \mid Z = z)\, dx && \text{(expectatio} \\
&= \int_{x \in \mathcal{R}(X)} \left(\int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid X = x, Z = z)\, dy \right) \cdot \mathbb{p}(X = x \mid Z = z)\, dx && \text{(definition} \\
&= \int_{x \in \mathcal{R}(X)} \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(X = x, Y = y \mid Z = z)\, dy\, dx && \text{(product o} \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \left(\int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x, Y = y \mid Z = z)\, dx \right) dy && \text{(Fubini's t} \\
&= \int_{y \in \mathcal{R}(Y)} y \cdot \mathbb{p}(Y = y \mid Z = z)\, dy && \text{(marginaliz} \\
&= \mathbb{E}[Y \mid Z = z] && \text{(definition}
\end{aligned}
$$

Since the two sides agree at every such z, they agree as random variables (functions of Z): $\mathbb{E}[Y \mid Z] = \mathbb{E}[\mathbb{E}[Y \mid X, Z] \mid Z]$. □

i Remark

Remark. This identity is the tower rule applied conditionally on Z.

Example 0.5 (Marginal expectation from conditional expectations)**.** Suppose X is a binary random variable indicating treatment assignment ($X = 1$ treated, $X = 0$ control), with $\mathbb{P}(X = 1) = 0.5$, and suppose the outcome Y has conditional expectations:

$$\mathbb{E}[Y \mid X = 1] = 10, \quad \mathbb{E}[Y \mid X = 0] = 6$$

By the law of iterated expectations (Theorem 0.2):

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[Y \mid X = 1] \cdot \mathbb{P}(X = 1) + \mathbb{E}[Y \mid X = 0] \cdot \mathbb{P}(X = 0) && \text{(LOTUS over the two values of } X\text{)} \\
&= 10 \cdot 0.5 + 6 \cdot 0.5 && \text{(substitute)} \\
&= 5 + 3 && \text{(multiply)} \\
&= 8 && \text{(add)}
\end{aligned}$$

Exercise 0.1 (Both-discrete case, infinite support: joint PMF)**.** Let X and Y be independent, each Geometric on

$$\begin{aligned}
\mathcal{R}(X) &= \mathcal{R}(Y) \\
&= \{0, 1, 2, \dots\}
\end{aligned}$$

with $\mathbb{P}(X = x) = (1-p)\,p^x$ for a fixed $p \in (0, 1)$ (X counts the number of failures before the first success in a sequence of independent trials with success probability $1-p$; likewise for Y). Unlike Exercise 0.1, the support here is countably infinite. The joint PMF is $\mathbb{P}(X = x,\, Y = y) = (1-p)^2\,p^{x+y}$.
Compute $\mathbb{E}[X + Y]$.

Solution

Solution. Compute $\mathbb{E}[X + Y]$ using Corollary 0.1 with

$$\begin{aligned}
\mu_X &= \mu_Y \\
&= \text{counting measure}
\end{aligned}$$

and $h(x, y) = x + y$. Since $h(x, y) = x + y \geq 0$ for every (x, y) in this support, condition (a) holds, so Corollary 0.1 (via Tonelli's theorem) guarantees the order of this now-infinite double sum is exchangeable — unlike the finite case, elementary algebra alone could not establish this.
The derivation uses the standard geometric-series facts $\sum_{y=0}^{\infty} p^y = \frac{1}{1-p}$ and $\sum_{y=0}^{\infty} y\,p^y = \frac{p}{(1-p)^2}$ (e.g. Casella and Berger [6]). By Corollary 0.1 (both-discrete case), summing over y first for each fixed x:

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(x+y)\,\mathbb{P}(X=x,\,Y=y) && \text{(joint-distribution form of Fubini–Tonelli)}\\
&= \sum_{x=0}^{\infty}\sum_{y=0}^{\infty}(x+y)(1-p)^2p^{x+y} && \text{(substitute the joint PMF)}\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\sum_{y=0}^{\infty}(x+y)\,p^y && \text{(factor } (1-p)^2p^x \text{ out of the inner sum)}\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\left(x\sum_{y=0}^{\infty}p^y+\sum_{y=0}^{\infty}y\,p^y\right) && \text{(split the inner sum; factor out } x)\\
&= \sum_{x=0}^{\infty}(1-p)^2p^x\left(\frac{x}{1-p}+\frac{p}{(1-p)^2}\right) && \text{(geometric-series facts)}\\
&= \sum_{x=0}^{\infty}p^x[x(1-p)+p] && \text{(multiply } (1-p)^2 \text{ into the parentheses)}\\
&= (1-p)\sum_{x=0}^{\infty}x\,p^x+p\sum_{x=0}^{\infty}p^x && \text{(distribute the sum over the terms; factor out the constant}\\
&= (1-p)\cdot\frac{p}{(1-p)^2}+p\cdot\frac{1}{1-p} && \text{(geometric-series facts)}\\
&= \frac{p}{1-p}+\frac{p}{1-p} && \text{(cancel } 1-p)\\
&= \frac{2p}{1-p} && \text{(add)}
\end{aligned}
$$

As a check,

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[Y]\\
&= \frac{p}{1-p}
\end{aligned}
$$

(the mean of this Geometric distribution; Casella and Berger [6]), so

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \mathbb{E}[X]+\mathbb{E}[Y]\\
&= \frac{2p}{1-p}
\end{aligned}
$$

by linearity of expectation, matching.

```r
p <- 0.4
exact_sum <- 2 * p / (1 - p)

n_trunc <- 500
support <- 0:n_trunc
joint_pmf_fn <- function(x, y) (1 - p)^2 * p^(x + y)
joint_probs <- outer(support, support, joint_pmf_fn)
h_vals <- outer(support, support, function(x, y) x + y)

sum_y_first <- sum(rowSums(h_vals * joint_probs))
sum_x_first <- sum(colSums(h_vals * joint_probs))

tibble::tibble(
  quantity = c("exact (closed form)", "sum over y first, then x",
               "sum over x first, then y"),
  value = round(c(exact_sum, sum_y_first, sum_x_first), 6)
) |>
  pander::pander()
```

quantity	value
exact (closed form)	1.333
sum over y first, then x	1.333
sum over x first, then y	1.333

With $p = 0.4$, the truncated sums agree with the closed form $\frac{2p}{1-p} = 1.3333$ up to truncation error, which confirms the closed form. They cannot illustrate Tonelli's guarantee, though: each truncation is a finite sum, and a finite sum gives the same total in either order, whether or not the infinite sums would agree. That the infinite sums agree is what Tonelli's theorem supplies.

i Tonelli's nonnegativity condition, and what happens without it

The calculation in Exercise 0.1 only needed condition (a), $h(X, Y) \geq 0$, because $h(x, y) = x + y$ is nonnegative on this support. For a **signed** h, interchanging an infinite double sum is not automatically valid — Corollary 0.1's condition (b), $\mathbb{E}[|h(X, Y)|] < \infty$, is what licenses it in that case. Without either condition, the two orders can genuinely disagree. A standard example (a signed array, not a probability distribution; see e.g. [11], Theorem 3.54, p. 76 for the general theory of rearranging series): let $a_{m,n} = 1$ if $m = n$, $a_{m,n} = -1$ if $m = n + 1$, and $a_{m,n} = 0$ otherwise, for $m, n = 0, 1, 2, \ldots$. Summing each row m first:

row 0 has only the term $a_{0,0} = 1$ (there is no valid $n = -1$), so its row sum is 1; every row $m \geq 1$ has $a_{m,m} = 1$ and $a_{m,m-1} = -1$, so its row sum is 0. Summing the rows then gives $1 + 0 + 0 + \cdots = 1$. Summing each column n first: every column $n \geq 0$ has $a_{n,n} = 1$ and $a_{n+1,n} = -1$, so its column sum is always 0, and summing the columns then gives $0 + 0 + \cdots = 0$. The two orders give 1 and 0: genuinely different answers, confirming that a condition like (a) or (b) really is needed once the terms are no longer all nonnegative.

Exercise 0.2 (Mixed case: one continuous variable, one discrete variable)**.** Let $Y \sim$ Bernoulli(0.6) and, given $Y = y$, let $X \mid Y = y \sim \text{Uniform}(0,\, y+1)$.
Compute $\mathbb{E}[X]$.

Solution

Solution. Compute $\mathbb{E}[X]$ using Corollary 0.1 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x, y) = x$.
The joint density w.r.t. Lebesgue × counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\, \mathbb{P}(Y = y)$:

$$\begin{aligned} f_{X,Y}(x,\,0) &= 1 \cdot 0.4 \\ &= 0.4 \qquad \text{for } x \in [0,1]; \\ f_{X,Y}(x,\,1) &= \tfrac{1}{2} \cdot 0.6 \\ &= 0.3 \qquad \text{for } x \in [0,2]. \end{aligned}$$

By Corollary 0.1 (mixed case):

$$\begin{aligned} \mathbb{E}[X] &= \sum_{y \in \{0,1\}} \int_0^{y+1} x\, f_{X,Y}(x,\,y)\, dx && \text{(joint-distribution form of Fubini–Tonelli, mixed case)} \\ &= \int_0^1 x \cdot 0.4\, dx + \int_0^2 x \cdot 0.3\, dx && \text{(expand the sum over } y \in \{0,1\}\text{; substitute } f_{X,Y}) \\ &= 0.4 \cdot \frac{1}{2} + 0.3 \cdot 2 && (\textstyle\int_0^b x\, dx = b^2/2) \\ &= 0.2 + 0.6 && \text{(multiply)} \\ &= 0.8 && \text{(add)} \end{aligned}$$

As a check using the law of iterated expectations (Theorem 0.2): $\mathbb{E}[X \mid Y = 0] = \frac{1}{2}$ and $\mathbb{E}[X \mid Y = 1] = 1$, so

$$\begin{aligned} \mathbb{E}[X] &= \tfrac{1}{2}(0.4) + 1(0.6) \\ &= 0.2 + 0.6 \\ &= 0.8. \end{aligned}$$

(This line plot is interactive; see the HTML or slide version of this page to view it.)

Figure 4: Joint density $f_{X,Y}(x,y) = f_{X|Y}(x \mid y)\,\mathbb{P}(Y=y)$ for each value of the discrete variable Y. The area under each component integrates to $\mathbb{P}(Y=y)$: $0.4 \cdot 1 = 0.4$ (blue) and $0.3 \cdot 2 = 0.6$ (red), summing to 1.

i Remark

Remark. Y takes only finitely many values here (two), so $\sum_{y\in\{0,1\}} \int_0^{y+1} x\, f_{X,Y}(x,\, y)\, dx$ is just linearity of the integral applied to a two-term sum — $\int g + \int k = \int (g+k)$ — not a genuine interchange of summation and integration order. If Y had a countably infinite range instead, the sum of integrals would be an infinite series, and Corollary 0.1's guarantee would be load-bearing.

Exercise 0.3 (Mixed case, infinite discrete support)**.** Let Y be Geometric on $\{0,1,2,\dots\}$ with $\mathbb{P}(Y=y) = (1-q)\,q^y$ for a fixed $q \in (0,1)$ and, given $Y=y$, let $X \mid Y=y \sim \text{Uniform}(0,\, y+1)$. Unlike Exercise 0.2, Y's range here is countably infinite.
Compute $\mathbb{E}[X]$.

Solution

Solution. Compute $\mathbb{E}[X]$ using Corollary 0.1 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x,y) = x$.
The joint density w.r.t. Lebesgue × counting measure is

$$\begin{aligned} f_{X,Y}(x,y) &= f_{X|Y}(x \mid y)\,\mathbb{P}(Y=y) \\ &= \frac{(1-q)\,q^y}{y+1} \end{aligned}$$

for $x \in [0, y+1]$.
Since $h(x,y) = x \geq 0$ on this support, condition (a) holds, so Corollary 0.1 (via Tonelli's theorem) guarantees the now-infinite sum-of-integrals expression is valid — unlike in Exercise 0.2, this Fubini–Tonelli justification is required because the sum is infinite rather than finite.
By Corollary 0.1 (mixed case):

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{y=0}^{\infty} \int_0^{y+1} x\, f_{X,Y}(x,\, y)\, dx && \text{(joint-distribution form of Fubini–Tonelli, mixed case)} \\
&= \sum_{y=0}^{\infty} \frac{(1-q)\, q^y}{y+1} \int_0^{y+1} x\, dx && (f_{X,Y}(x,y) \text{ is constant in } x \text{ on } [0, y+1]) \\
&= \sum_{y=0}^{\infty} \frac{(1-q)\, q^y}{y+1} \cdot \frac{(y+1)^2}{2} && \text{(integrate)} \\
&= \frac{1-q}{2} \sum_{y=0}^{\infty} (y+1)\, q^y && \text{(cancel } y+1\text{; factor out } \tfrac{1-q}{2}) \\
&= \frac{1-q}{2} \left(\sum_{y=0}^{\infty} y\, q^y + \sum_{y=0}^{\infty} q^y \right) && \text{(distribute the sum over the terms)} \\
&= \frac{1-q}{2} \left(\frac{q}{(1-q)^2} + \frac{1}{1-q} \right) && \text{(geometric-series facts)} \\
&= \frac{1-q}{2} \cdot \frac{q + (1-q)}{(1-q)^2} && \text{(common denominator)} \\
&= \frac{1-q}{2} \cdot \frac{1}{(1-q)^2} && \text{(simplify the numerator)} \\
&= \frac{1}{2(1-q)} && \text{(cancel } 1-q\text{)}
\end{aligned}
$$

using the same geometric-series facts as Exercise 0.1 (e.g. Casella and Berger [6]).
As a check using the law of iterated expectations (Theorem 0.2) and the expectation of a linear function (Corollary 0.1): $\mathbb{E}[X \mid Y = y] = \frac{y+1}{2}$ (the mean of Uniform$(0, y+1)$) and $\mathbb{E}[Y] = \frac{q}{1-q}$ (the mean of this Geometric distribution; Casella and Berger [6]), so:

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid Y]] && \text{(law of iterated expectations)} \\
&= \mathbb{E}\left[\frac{Y+1}{2} \right] && (\mathbb{E}[X \mid Y = y] = \tfrac{y+1}{2}) \\
&= \frac{\mathbb{E}[Y] + 1}{2} && \text{(expectation of a linear function of one random variable)} \\
&= \frac{1}{2} \left(\frac{q}{1-q} + 1 \right) && \text{(substitute } \mathbb{E}[Y]\text{)} \\
&= \frac{1}{2} \cdot \frac{q + (1-q)}{1-q} && \text{(common denominator)} \\
&= \frac{1}{2(1-q)} && \text{(simplify the numerator)}
\end{aligned}
$$

matching.

```r
q <- 0.5
exact_ex <- 1 / (2 * (1 - q))

n_trunc <- 2000
y_vals <- 0:n_trunc
py_vals <- (1 - q) * q^y_vals
cond_mean_x <- (y_vals + 1) / 2
trunc_ex <- sum(cond_mean_x * py_vals)

tibble::tibble(
  quantity = c("exact (closed form)", "truncated sum-of-integrals"),
  value = round(c(exact_ex, trunc_ex), 6)
) |>
  pander::pander()
```

quantity	value
exact (closed form)	1
truncated sum-of-integrals	1

With $q = 0.5$, the truncated sum-of-integrals matches the closed form $\frac{1}{2(1-q)} = 1$.

Loss and risk

Definition 0.1 (Prediction)**.** A **prediction** of a random variable Y is a constant $\hat{y}$, or a function $\hat{Y} \stackrel{\text{def}}{=} g(X)$ of another observable random variable X, used as a guess for the value Y takes:

$$\hat{y} \quad \text{or} \quad \hat{Y} \stackrel{\text{def}}{=} g(X)$$

Example 0.1 (Predictions of a die roll)**.** Let Y be the number shown by a fair six-sided die. The constant $\hat{y} = 3$ is a prediction of Y.
Let X be the indicator that Y is even. The function $g(X) = 2 + 2X$ is also a prediction of Y: it guesses 2 when Y is odd and 4 when Y is even.

Remark (A prediction from a random input is random)

Remark 0.1 (A prediction from a random input is random). When no covariate or feature is available, a prediction is a single number $\hat{y}$ (such as a constant c). When an informative random variable X is observed, a prediction is a function $g(X)$, which is itself a random variable because X is random.

Definition 0.2 (Loss function)**.** A **loss function** $L(y, \hat{y})$ is a rule that gives a number that is 0 or larger: the cost of predicting $\hat{y}$ when the true value is y [12].

Example 0.2 (A loss function that charges more for overpredicting)**.** Let $L(y, \hat{y})$ cost 2 per unit of overprediction and 1 per unit of underprediction:

$$L(y, \hat{y}) = \begin{cases} 2(\hat{y} - y) & \text{if } \hat{y} > y \\ y - \hat{y} & \text{if } \hat{y} \leq y. \end{cases}$$

This is a loss function because it is never negative. For the true value $y = 2$ and the prediction $\hat{y} = 7$, the loss is $L(2, 7) = 2 \cdot 5 = 10$. For $y = 7$ and $\hat{y} = 2$, the loss is $L(7, 2) = 5$.

Definition 0.3 (Squared error loss)**.** The **squared error loss** is the loss function that charges the square of the difference between the true value y and the prediction $\hat{y}$:

$$L(y, \hat{y}) \stackrel{\text{def}}{=} (y - \hat{y})^2$$

Example 0.3 (Squared error loss for one prediction)**.** For the true value $y = 2$ and the prediction $\hat{y} = 7$, the squared error loss is

$$(2 - 7)^2 = 25.$$

Definition 0.4 (Absolute error loss)**.** The **absolute error loss** is the loss function that charges the absolute value of the difference between the true value y and the prediction $\hat{y}$:

$$L(y, \hat{y}) \stackrel{\text{def}}{=} |y - \hat{y}|$$

Example 0.4 (Absolute error loss for one prediction)**.** For the true value $y = 2$ and the prediction $\hat{y} = 7$, the absolute error loss is

$$|2-7| = 5.$$

Remark (Squared error loss is the most common choice)

Remark 0.2 (Squared error loss is the most common choice). Hastie et al. [13, p. 18] call squared error loss "by far the most common and convenient" choice.

Definition 0.5 (Risk (expected loss))**.** For a random variable Y and a prediction $\hat{y}$ (a constant, or a function $g(X)$ of a random input X), the **risk** is the expectation of the loss [14]:

$$R \stackrel{\text{def}}{=} \mathbb{E}[L(Y, \hat{y})]$$

Example 0.5 (Risk of a constant prediction)**.** Let Y be a random variable with $\mathbb{P}(Y = 0) = 0.5$ and $\mathbb{P}(Y = 2) = 0.5$, and use the constant prediction $\hat{y} = 0$. The risk under squared error loss is

$$\begin{aligned}\mathbb{E}\left[(Y-0)^2\right] &= 0.5 \cdot (0-0)^2 + 0.5 \cdot (2-0)^2 \\ &= 2.\end{aligned}$$

The risk under absolute error loss is

$$\begin{aligned}\mathbb{E}[|Y-0|] &= 0.5 \cdot |0-0| + 0.5 \cdot |2-0| \\ &= 1.\end{aligned}$$

Remark (Risk measures the quality of a prediction)

Remark 0.3 (Risk measures the quality of a prediction). The quality of a prediction is evaluated using a loss function and its expected value, the risk.

Exercise 0.1 (Squared and absolute loss)**.**

1. For a true value $y = 3$ and a prediction $\hat{y} = 5$, compute the squared error loss $(y - \hat{y})^2$ and the absolute error loss $|y - \hat{y}|$.
2. A random variable Y has $\mathbb{P}(Y = 0) = 0.5$ and $\mathbb{P}(Y = 4) = 0.5$. For the constant prediction $c = 1$, compute the risk under squared error loss $\mathbb{E}\left[(Y-c)^2\right]$ and the risk under absolute error loss $\mathbb{E}[|Y - c|]$.
3. Repeat part 2 for $c = 2$.

4. Which of $c = 1$ and $c = 2$ gives the smaller risk under each loss?

Solution

Solution 0.1.

1. For squared error loss:
$$(3-5)^2 = 4$$
For absolute error loss:
$$|3-5| = 2$$

2. For $c = 1$: The risk under squared error loss is:
$$\begin{aligned}\mathbb{E}\left[(Y-1)^2\right] &= 0.5 \cdot (0-1)^2 + 0.5 \cdot (4-1)^2 \\ &= 0.5 \cdot 1 + 0.5 \cdot 9 \\ &= 5\end{aligned}$$
The risk under absolute error loss is:
$$\begin{aligned}\mathbb{E}[|Y-1|] &= 0.5 \cdot |0-1| + 0.5 \cdot |4-1| \\ &= 0.5 \cdot 1 + 0.5 \cdot 3 \\ &= 2\end{aligned}$$

3. For $c = 2$: The risk under squared error loss is:
$$\begin{aligned}\mathbb{E}\left[(Y-2)^2\right] &= 0.5 \cdot (0-2)^2 + 0.5 \cdot (4-2)^2 \\ &= 0.5 \cdot 4 + 0.5 \cdot 4 \\ &= 4\end{aligned}$$
The risk under absolute error loss is:
$$\begin{aligned}\mathbb{E}[|Y-2|] &= 0.5 \cdot |0-2| + 0.5 \cdot |4-2| \\ &= 0.5 \cdot 2 + 0.5 \cdot 2 \\ &= 2\end{aligned}$$

4. The prediction $c = 2$ has the smaller risk under squared error loss (4 versus 5). The two predictions have the same risk under absolute error loss (2), so under absolute error loss this exercise does not choose one of them. The prediction $c = 2$ equals the expectation:
$$\begin{aligned}\mathbb{E}[Y] &= 0.5 \cdot 0 + 0.5 \cdot 4 \\ &= 2\end{aligned}$$

Remark (The best prediction depends on the loss)

Remark 0.4 (The best prediction depends on the loss). In Exercise 0.1 (parts 2 and 3), the risk of the prediction $c = 1$ is 5 for squared error loss and 2 for absolute error loss. The risk of the prediction $c = 2$ is 4 for squared error loss and 2 for absolute error loss. The prediction with the smallest risk depends on the loss. In Exercise 0.1, every constant c from 0 to 4 has absolute error risk 2, because

$$\begin{aligned}\mathbb{E}[|Y - c|] &= 0.5\,c + 0.5\,(4 - c) \\ &= 2.\end{aligned}$$

Only $c = 2$ has the smallest squared error risk. Under squared error loss, the prediction function that minimizes risk is the conditional mean $\mathbb{E}[Y \mid X]$, proved below (Theorem 0.2). Hastie et al. [13, p. 20] state that, for absolute error loss, the best prediction function is the conditional median instead of the conditional mean. This statement is given here without proof.

Exercise 0.2 (Comparing prediction functions under squared-error loss)**.** Let (X, Y) be jointly distributed discrete random variables. X takes values in $\{0, 1\}$ with $\mathbb{P}(X = 0) = 0.5$ and $\mathbb{P}(X = 1) = 0.5$. Given $X = 0$, Y takes values in $\{0, 2\}$ with conditional probabilities $\mathbb{P}(Y = 0 \mid X = 0) = 0.5$ and $\mathbb{P}(Y = 2 \mid X = 0) = 0.5$. Given $X = 1$, Y takes values in $\{4, 6\}$ with conditional probabilities $\mathbb{P}(Y = 4 \mid X = 1) = 0.5$ and $\mathbb{P}(Y = 6 \mid X = 1) = 0.5$.

1. Compute the marginal mean $\mathbb{E}[Y]$ and the conditional expectation function $f^*(X) \overset{\text{def}}{=} \mathbb{E}[Y \mid X]$.

2. Consider the constant prediction $f_1(X) \overset{\text{def}}{=} \mathbb{E}[Y] = 3$, which ignores X. Compute its squared-error risk $\mathbb{E}\left[(Y - f_1(X))^2\right]$.

3. Consider the alternate prediction $f_2(X) \overset{\text{def}}{=} 4X$. Compute its squared-error risk $\mathbb{E}\left[(Y - f_2(X))^2\right]$.

4. Compute the squared-error risk $\mathbb{E}\left[(Y - f^*(X))^2\right]$ of the conditional mean predictor $f^*(X)$. Which of f_1, f_2, and f^* achieves the lowest risk?

Solution

Solution 0.2.

1. For each value of X, compute the conditional expectation:

$$\begin{aligned}\mathbb{E}[Y \mid X = 0] &= 0 \cdot 0.5 + 2 \cdot 0.5 \\ &= 1\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Y \mid X = 1] &= 4 \cdot 0.5 + 6 \cdot 0.5 \\ &= 5\end{aligned}$$

   By the law of iterated expectations (Theorem 0.2), the marginal mean is:

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid X]] \\ &= 1 \cdot 0.5 + 5 \cdot 0.5 \\ &= 3\end{aligned}$$

   The conditional expectation function is $f^*(0) = 1$ and $f^*(1) = 5$, which can also be written $f^*(X) = 1 + 4X$.

2. For the constant prediction $f_1(X) = 3$: Given $X = 0$:

$$\begin{aligned}\mathbb{E}\left[(Y-3)^2 \mid X = 0\right] &= (0-3)^2 \cdot 0.5 + (2-3)^2 \cdot 0.5 \\ &= 9 \cdot 0.5 + 1 \cdot 0.5 \\ &= 5\end{aligned}$$

   Given $X = 1$:

$$\begin{aligned}\mathbb{E}\left[(Y-3)^2 \mid X = 1\right] &= (4-3)^2 \cdot 0.5 + (6-3)^2 \cdot 0.5 \\ &= 1 \cdot 0.5 + 9 \cdot 0.5 \\ &= 5\end{aligned}$$

   By the law of iterated expectations (Theorem 0.2):

$$\begin{aligned}\mathbb{E}\left[(Y - f_1(X))^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y-3)^2 \mid X\right]\right] \\ &= 5 \cdot 0.5 + 5 \cdot 0.5 \\ &= 5\end{aligned}$$

3. For the prediction $f_2(X) = 4X$: Here $f_2(0) = 0$ and $f_2(1) = 4$. Given $X = 0$:

$$\begin{aligned}\mathbb{E}\left[(Y-0)^2 \mid X = 0\right] &= (0-0)^2 \cdot 0.5 + (2-0)^2 \cdot 0.5 \\ &= 0 \cdot 0.5 + 4 \cdot 0.5 \\ &= 2\end{aligned}$$

Given $X = 1$:

$$\begin{aligned}\mathbb{E}\left[(Y-4)^2 \mid X=1\right] &= (4-4)^2 \cdot 0.5 + (6-4)^2 \cdot 0.5 \\ &= 0 \cdot 0.5 + 4 \cdot 0.5 \\ &= 2\end{aligned}$$

By Theorem 0.2:

$$\begin{aligned}\mathbb{E}\left[(Y-f_2(X))^2\right] &= 2 \cdot 0.5 + 2 \cdot 0.5 \\ &= 2\end{aligned}$$

4. For the conditional mean predictor $f^*(X) = \mathbb{E}[Y \mid X]$: Given $X = 0$, $f^*(0) = 1$:

$$\begin{aligned}\mathbb{E}\left[(Y-1)^2 \mid X=0\right] &= (0-1)^2 \cdot 0.5 + (2-1)^2 \cdot 0.5 \\ &= 1 \cdot 0.5 + 1 \cdot 0.5 \\ &= 1\end{aligned}$$

Given $X = 1$, $f^*(1) = 5$:

$$\begin{aligned}\mathbb{E}\left[(Y-5)^2 \mid X=1\right] &= (4-5)^2 \cdot 0.5 + (6-5)^2 \cdot 0.5 \\ &= 1 \cdot 0.5 + 1 \cdot 0.5 \\ &= 1\end{aligned}$$

By Theorem 0.2:

$$\begin{aligned}\mathbb{E}\left[(Y-f^*(X))^2\right] &= 1 \cdot 0.5 + 1 \cdot 0.5 \\ &= 1\end{aligned}$$

The conditional mean predictor f^* achieves the lowest risk (1), strictly outperforming f_2 (2) and the constant prediction f_1 (5).

Definition 0.6 (Outcome)**.** The **outcome** Y of a data point is the random variable whose value a prediction tries to guess.

Example 0.6 (The outcome in the best-predictor exercise)**.** In Exercise 0.2, the outcome Y takes the values 0 and 2 when the input is 0, and the values 4 and 6 when the input is 1.

Definition 0.7 (Input)**.** The **input** X of a data point (also called its **predictor** or **covariate**) is the observed random variable from which a prediction of its outcome is

computed.

Example 0.7 (The input in the best-predictor exercise)**.** In Exercise 0.2, the input X takes the values 0 and 1, each with probability 0.5.

Definition 0.8 (Prediction function)**.** A **prediction function** f is a function that takes a value x of the input and returns a prediction $f(x)$ of the outcome.

Example 0.8 (A constant prediction function)**.** In Exercise 0.2, the function $f_1(x) = 3$ is a prediction function: it returns the prediction 3 for every value of the input.

Definition 0.9 (New observation)**.** A **new observation** is an input and outcome pair (X, Y) with the same joint distribution as the observations used to choose the prediction, statistically independent of them, whose input is observed and whose outcome is not yet observed.

Example 0.9 (A new observation in the best-predictor exercise)**.** In Exercise 0.2, suppose the new observation (X, Y) satisfies the suppositions of its definition, with the joint distribution given in the exercise as the population it is drawn from. Suppose also that its input is observed to be $X = 1$. Then its unobserved outcome is 4 or 6 with equal probability, and a prediction of 5 is a guess for that outcome.

Definition 0.10 (Prediction error)**.** The **prediction error** of a prediction $\hat{Y}$ of an outcome Y is the prediction minus the outcome:

$$\varepsilon_{\text{pred}}\left(\hat{Y}\right) \stackrel{\text{def}}{=} \hat{Y} - Y$$

Example 0.10 (Prediction error in the best-predictor exercise)**.** In Exercise 0.2, consider the constant prediction $f_1(X) = 3$. Its prediction error is $\varepsilon_{\text{pred}}(f_1) = 3 - Y$. This is 3 when $Y = 0$. It is -3 when $Y = 6$.

Remark (The prediction error of a prediction function)

Remark 0.5 (The prediction error of a prediction function). A prediction function f makes the prediction $f(X)$ for the outcome Y of a new observation with input X. We write $\varepsilon_{\text{pred}}(f)$ for the prediction error of that prediction, meaning $\varepsilon_{\text{pred}}(f(X))$. It is

$$\varepsilon_{\text{pred}}(f) = f(X) - Y$$

The reference value in a prediction error is the observed outcome Y, not the conditional mean of that outcome.

Definition 0.11 (Squared-error risk (expected prediction error))**.** The **squared-error risk** of a prediction function f, also called its **expected prediction error** [13, p. 18], is the expected square of its prediction error:

$$R(f) \overset{\text{def}}{=} \mathbb{E}\left[\left(\varepsilon_{\text{pred}}(f)\right)^2\right]$$

Example 0.11 (Squared-error risk in the best-predictor exercise)**.** In Exercise 0.2, consider the constant prediction $f_1(X) = 3$. Its squared-error risk is $R(f_1) = 5$.

Remark (The squared-error risk is a risk)

Remark 0.6 (The squared-error risk is a risk). The squared-error risk is the risk under squared error loss.

Lemma 0.1 (Prediction error splits at the conditional mean)**.** *Let $\mu(X) \overset{def}{=} \mathbb{E}[Y \mid X]$ be the conditional expectation function of Y given X. For any prediction function f, the prediction error of f is the prediction error of μ plus the difference $(f(X) - \mu(X))$ between the prediction $f(X)$ and $\mu(X)$:*

$$\varepsilon_{pred}(f) = \varepsilon_{pred}(\mu) + (f(X) - \mu(X))$$

Proof

Proof. The notation $\varepsilon_{\text{pred}}(f)$ is introduced in Remark 0.5.

$$\begin{aligned}
\varepsilon_{\text{pred}}(f) &= f(X) - Y && \text{(prediction error notation)} \\
&= (\mu(X) - Y) + (f(X) - \mu(X)) && \text{(add and subtract } \mu(X)) \\
&= \varepsilon_{\text{pred}}(\mu) + (f(X) - \mu(X)) && \text{(prediction error notation)}
\end{aligned}$$

□

Example 0.12 (The split in the best-predictor exercise)**.** In Exercise 0.2, when $X = 0$ and $Y = 0$, the constant prediction function f_1, with $f_1(X) = 3$, has prediction error

$\varepsilon_{\text{pred}}(f_1) = 3$. The conditional mean is $\mu(0) = 1$, so $\varepsilon_{\text{pred}}(\mu) = 1$ and $(f_1(0) - \mu(0)) = 2$, and $3 = 1 + 2$.

Lemma 0.2 (The conditional mean's error is uncorrelated with any function of the input)**.** *Let $\mu(X) \overset{def}{=} \mathbb{E}[Y \mid X]$ be the conditional expectation function of Y given X. For any prediction function f with $\mathbb{E}\left[f(X)^2\right] < \infty$:*

$$\mathbb{E}\left[\varepsilon_{pred}(\mu) \cdot (f(X) - \mu(X))\right] = 0$$

Proof

Proof. Because $f(X) - \mu(X)$ is a function of X, by Theorem 0.3 it factors out of the conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_{\text{pred}}(\mu)(f(X) - \mu(X)) \mid X\right] &= (f(X) - \mu(X)) \cdot \mathbb{E}\left[\varepsilon_{\text{pred}}(\mu) \mid X\right] && \text{(a function of } X \text{ factors out)} \\
&= (f(X) - \mu(X)) \cdot \mathbb{E}[\mu(X) - Y \mid X] && \text{(definition of } \varepsilon_{\text{pred}}(\mu)) \\
&= (f(X) - \mu(X)) \cdot (\mathbb{E}[\mu(X) \mid X] - \mathbb{E}[Y \mid X]) && \text{(linearity of conditional expectat} \\
&= (f(X) - \mu(X)) \cdot (\mu(X) - \mu(X)) && \text{(a function of } X \text{ is its own condi} \\
&= 0 && \text{(evaluate)}
\end{aligned}$$

By the law of iterated expectations (Theorem 0.2):

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_{\text{pred}}(\mu)(f(X) - \mu(X))\right] &= \mathbb{E}\left[\mathbb{E}\left[\varepsilon_{\text{pred}}(\mu)(f(X) - \mu(X)) \mid X\right]\right] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[0] && \text{(the conditional expectation is 0)} \\
&= 0 && \text{(expectation of a constant)}
\end{aligned}$$

□

Example 0.13 (The vanishing cross term in the best-predictor exercise)**.** In Exercise 0.2, take the constant prediction $f_1(X) = 3$. When $X = 0$, $\mu(0) = 1$, so $f_1(0) - \mu(0) = 2$, and $\varepsilon_{\text{pred}}(\mu) = 1 - Y$ is 1 or -1 with equal probability, so the product has conditional mean 0. The same cancellation happens when $X = 1$, so the expectation of the product is 0.

Theorem 0.1 (Squared-error risk decomposition)**.** *Let $\mu(X) \overset{def}{=} \mathbb{E}[Y \mid X]$ be the conditional expectation function of Y given X. For any prediction function f with $\mathbb{E}\left[f(X)^2\right] < \infty$:*

$$R(f) = R(\mu) + \mathbb{E}\Big[\big(f(X) - \mu(X)\big)^2\Big]$$

Proof

Proof. The steps use the squared-error risk definition, Lemma 0.1, Lemma 0.2, and Theorem 0.4:

$$\begin{aligned}
R(f) &= \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(f)\big)^2\Big] && \text{(definition of the squared-e} \\
&= \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(\mu) + (f(X) - \mu(X))\big)^2\Big] && \text{(the prediction error splits} \\
&= \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(\mu)\big)^2\Big] + 2 \cdot \mathbb{E}\big[\varepsilon_{\text{pred}}(\mu) \cdot (f(X) - \mu(X))\big] + \mathbb{E}\Big[\big((f(X) - \mu(X))\big)^2\Big] && \text{(expand the square; linearit} \\
&= \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(\mu)\big)^2\Big] + 2 \cdot 0 + \mathbb{E}\Big[\big((f(X) - \mu(X))\big)^2\Big] && \text{(the cross term is 0)} \\
&= R(\mu) + \mathbb{E}\Big[\big((f(X) - \mu(X))\big)^2\Big] && \text{(definition of the squared-e}
\end{aligned}$$

□

Example 0.14 (The risk decomposition in the best-predictor exercise)**.** In Exercise 0.2, the constant prediction $f_1(X) = 3$ has squared-error risk $R(f_1) = 5$, which Theorem 0.1 splits as $5 = 1 + 4$: the risk 1 of the conditional expectation function plus the excess 4 from using f_1.

Theorem 0.2 (Conditional mean minimizes squared-error risk)**.** *Let X and Y be jointly distributed random variables with $\mathbb{E}[Y^2] < \infty$. Let $\mu(X) \overset{def}{=} \mathbb{E}[Y \mid X]$ be the conditional expectation function of Y given X. Write $f^* \overset{def}{=} \mu$ for the prediction function that uses it. Then for any prediction function f with $\mathbb{E}\Big[f(X)^2\Big] < \infty$:*

$$R(f) \geq R(f^*)$$

with equality if and only if $\mathbb{P}(f(X) = f^*(X)) = 1$. That is, f^* minimizes the squared-error risk among all prediction functions of X.

Proof

Proof. By Theorem 0.1:

$$
\begin{aligned}
R(f) - R(f^*) &= \mathbb{E}\Big[\big(f(X) - f^*(X)\big)^2\Big] && \text{(squared-error risk decomposition)} \\
&\geq 0 && \text{(the expectation of a non-negative variable)}
\end{aligned}
$$

The expectation of the non-negative variable $\big(f(X) - f^*(X)\big)^2$ is zero if and only if the variable is 0 with probability 1, meaning $\mathbb{P}(f(X) = f^*(X)) = 1$. □

Example 0.15 (The best predictor in the best-predictor exercise)**.** In Exercise 0.2, consider the constant prediction $f_1(X) = 3$. Its squared-error risk is $R(f_1) = 5$. The risk of the conditional mean is $R(f^*) = 1$. So f_1 has the larger risk. The inequality is strict because $f_1(X) \neq f^*(X)$.

Definition 0.12 (Residual risk (irreducible risk))**.** The **residual risk** (also called the **irreducible risk**) is the squared-error risk of the conditional expectation function μ of Theorem 0.1:

$$
R(\mu) = \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(\mu)\big)^2\Big]
$$

Theorem 0.3 (The residual risk is the expected conditional variance)**.** *Let X and Y be as in Theorem 0.2. Then the residual risk is the expectation of the conditional variance of Y given X:*

$$
R(\mu) = \mathbb{E}[\mathrm{Var}(Y \mid X)]. \tag{1}
$$

Proof

Proof.

$$
\begin{aligned}
R(\mu) &= \mathbb{E}\Big[\big(\varepsilon_{\text{pred}}(\mu)\big)^2\Big] && \text{(definition of the residual risk)} \\
&= \mathbb{E}\Big[\big(\mu(X) - Y\big)^2\Big] && \text{(definition of the prediction error)} \\
&= \mathbb{E}\Big[\big(Y - \mu(X)\big)^2\Big] && \text{(a square is unchanged by negation)} \\
&= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2\Big] && \text{(definition of } \mu\text{)} \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X]\big)^2 \mid X\Big]\Big] && \text{(law of total expectation)} \\
&= \mathbb{E}[\mathrm{Var}(Y \mid X)] && \text{(definition of conditional variance)}
\end{aligned}
$$

□

Remark (The residual risk does not depend on the prediction function)

Remark 0.7 (The residual risk does not depend on the prediction function). The residual risk depends only on the joint distribution of (X, Y) and not on the prediction function f, so no choice of f can achieve a smaller risk. It is the minimum only among prediction functions of the given X: the outcome may also depend on variables that X leaves out, and those variables are part of the unexplained variation $Y - \mathbb{E}[Y \mid X]$. Measuring one of them and adding it to the variables in X can lower the residual risk, as in Example 0.17. [15, Sec. 2.1.1], p. 18 makes the same point: the error term "may contain unmeasured variables that are useful in predicting $[Y]$".

Example 0.16 (Residual risk in the best-predictor exercise)**.** In Exercise 0.2, $\mathrm{Var}(Y \mid X = 0) = 1$ and $\mathrm{Var}(Y \mid X = 1) = 1$, so the residual risk is $\mathbb{E}[\mathrm{Var}(Y \mid X)] = 1$.

Example 0.17 (Adding a predictor lowers the residual risk)**.** Let X_1, X_2 and ε be independent, with $\mathrm{Var}(X_2) = 1$ and $\mathrm{Var}(\varepsilon) = 1$, and let $Y = X_1 + X_2 + \varepsilon$. A prediction from X_1 alone has conditional variance $\mathrm{Var}(Y \mid X_1) = \mathrm{Var}(X_2) + \mathrm{Var}(\varepsilon) = 2$, so the residual risk is 2. A prediction from (X_1, X_2) has conditional variance $\mathrm{Var}(Y \mid X_1, X_2) = \mathrm{Var}(\varepsilon) = 1$, so the residual risk is 1.

Definition 0.13 (Reducible risk)**.** The **reducible risk** of a prediction function f is the expectation of the squared difference between $f(X)$ and the conditional mean $\mu(X)$ of the outcome:

$$\mathbb{E}\Big[(f(X) - \mu(X))^2\Big]$$

Proposition 0.1 (The reducible risk is zero only for the conditional mean)**.** *The reducible risk is zero if and only if the prediction function f equals the conditional expectation function with probability 1:*

$$\mathbb{E}\Big[(\mathbb{E}[Y \mid X] - f(X))^2\Big] = 0 \iff \mathbb{P}(f(X) = \mathbb{E}[Y \mid X]) = 1$$

Proof

Proof. Write $Z = (\mathbb{E}[Y \mid X] - f(X))^2$. Since $x^2 \geq 0$ for every x, with $x^2 = 0$ only at $x = 0$, $Z \geq 0$, and $Z = 0$ exactly when $f(X) = \mathbb{E}[Y \mid X]$.
($\Leftarrow$) If $\mathbb{P}(Z = 0) = 1$, then Z takes the value 0 with probability 1, so $\mathbb{E}[Z] = 0$ by the definition of expectation.

($\Rightarrow$) Suppose $\mathbb{E}[Z] = 0$, and fix $a > 0$. Since $Z \geq 0$, and $Z \geq a$ on the event $\{Z \geq a\}$:

$$\begin{aligned} Z &\geq a \cdot 1_{Z \geq a} && \text{(on } \{Z \geq a\} \text{ the right side is } a \leq Z\text{; elsewhere it is } 0 \leq Z) \\ \mathbb{E}[Z] &\geq \mathbb{E}\big[a \cdot 1_{Z \geq a}\big] && \text{(expectation preserves inequalities)} \\ &= a \cdot \mathbb{P}(Z \geq a) && \text{(the expectation of an indicator is the probability of its event)} \end{aligned}$$

With $\mathbb{E}[Z] = 0$, the displayed chain gives $a \cdot \mathbb{P}(Z \geq a) \leq 0$, so $\mathbb{P}(Z \geq a) = 0$. The events $\{Z \geq 1/n\}$ for $n = 1, 2, \ldots$ increase to the event $\{Z > 0\}$, so by the continuity of probability:

$$\begin{aligned} \mathbb{P}(Z > 0) &= \lim_{n \to \infty} \mathbb{P}(Z \geq 1/n) && \text{(continuity of probability)} \\ &= 0 && \text{(each term is 0, taking } a = 1/n\text{)} \end{aligned}$$

Hence $\mathbb{P}(Z = 0) = 1 - \mathbb{P}(Z > 0) = 1$. □

Example 0.18 (Reducible risk in the best-predictor exercise)**.** In Exercise 0.2, the constant prediction function f_1, with $f_1(X) = 3$, has reducible risk $5 - 1 = 4$, and the prediction function f_2, with $f_2(X) = 4X$, has reducible risk $2 - 1 = 1$.

Remark

Remark. This characterization of the best predictor under squared error loss follows Hastie et al. [13, Sec. 2.4], eq. 2.11. The risk identity in Theorem 0.1

$$\mathbb{E}\Big[(Y - f(X))^2\Big] = \mathbb{E}\Big[(Y - \mathbb{E}[Y \mid X])^2\Big] + \mathbb{E}\Big[(\mathbb{E}[Y \mid X] - f(X))^2\Big]$$

shows that the total squared-error risk is the sum of the residual risk (Definition 0.12) and the reducible risk (Definition 0.13). James et al. [15, Sec. 2.1.1], pp. 18–19 calls the squared difference between a prediction function and the conditional mean the "reducible error".

Corollary 0.1 (Conditional mean as the minimizer of squared-error risk)**.** *Let X and Y be jointly distributed random variables with $\mathbb{E}[Y^2] < \infty$. Among prediction functions f with $\mathbb{E}\big[f(X)^2\big] < \infty$, the conditional expectation function is a minimizer of the squared-error risk:*

$$\mathbb{E}[Y \mid X] \in \arg\min_f \mathbb{E}\Big[(Y - f(X))^2\Big],$$

where arg min *is the set of minimizers.*

Proof

Proof. By Theorem 0.2, $f^*(X) = \mathbb{E}[Y \mid X]$ has risk no larger than that of any prediction function f, so it is a minimizer. □

Example 0.19 (The minimizer in the best-predictor exercise)**.** In Exercise 0.2, the conditional expectation function $f^*(X)$ has risk 1. The constant prediction function f_1, with $f_1(X) = 3$, has risk 5 and $f_2(X) = 4X$ has risk 2, both larger than 1. So f^* is the minimizer among the three.

Corollary 0.2 (The minimizer of squared-error risk is unique up to probability-1 equality)**.** *Let X and Y be jointly distributed random variables with $\mathbb{E}[Y^2] < \infty$. Every prediction function f with $\mathbb{E}\left[f(X)^2\right] < \infty$ that is a minimizer of the squared-error risk (see Corollary 0.1) satisfies:*

$$\mathbb{P}(f(X) = \mathbb{E}[Y \mid X]) = 1$$

Proof

Proof. If f is a minimizer, its risk equals that of f^*, so the equality case of Theorem 0.2 gives $\mathbb{P}(f(X) = f^*(X)) = 1$. □

Example 0.20 (Uniqueness in the best-predictor exercise)**.** In Exercise 0.2, $f_2(X) = 4X$ differs from $f^*(X)$ when $X = 0$, which has probability 0.5. So f_2 is not a minimizer, which matches its risk 2 being larger than 1.

Entropy and cross-entropy

Definition 0.1 (Surprisal (**self-information**, **information content**))**.** The **surprisal** of a possible value x of a discrete random variable X with $\mathbb{P}(X = x) > 0$ is

$$\mathrm{s}(x) \stackrel{\text{def}}{=} -\log \mathbb{P}(X = x).$$

The surprisal is also called the **self-information** [16, p. 73] or the **information content** [17, p. 49] of x. With the natural logarithm, the surprisal is measured in nats, and with base 2 in bits. Evaluating $\mathrm{s}(\cdot)$ at the random variable X itself gives the random variable $\mathrm{s}(X)$, which equals $\mathrm{s}(x)$ when $X = x$. The notation leaves the probability mass function of X implicit.

Example 0.1 (Surprisal of the values of a biased coin flip)**.** Let X be 1 if one flip of a coin lands heads and 0 if it lands tails, with $\mathbb{P}(X = 1) = \frac{1}{4}$ and $\mathbb{P}(X = 0) = \frac{3}{4}$. Using logarithms to base 2, so that surprisal is measured in bits, the surprisal of the value 1 (heads) is $\mathbb{s}(1) = -\log_2 \frac{1}{4} = 2$ bits, and the surprisal of the value 0 (tails) is $\mathbb{s}(0) = -\log_2 \frac{3}{4} = \log_2 \frac{4}{3} \approx 0.415$ bits. For a fair coin, each value of X has surprisal $-\log_2 \frac{1}{2} = 1$ bit.

Remark (Rarer values are more surprising)

Remark 0.1 (Rarer values are more surprising). A possible value x with probability $\mathbb{P}(X = x) = 1$ has surprisal 0, and the surprisal grows without bound as $\mathbb{P}(X = x) \to 0$.

Definition 0.2 (Surprisal of a value of a continuous random variable)**.** The **surprisal** of a possible value x of a continuous random variable X with probability density function $\mathbb{p}(X = x)$ and $\mathbb{p}(X = x) > 0$ is

$$\mathbb{s}(x) \stackrel{\text{def}}{=} -\log \mathbb{p}(X = x).$$

This extends the surprisal of a value of a discrete random variable, with the density in place of the probability.

Example 0.2 (Surprisal of a value of a uniform random variable)**.** Let X be uniformly distributed on the interval from 0 to $a > 0$, so that its density is $\mathbb{p}(X = x) = 1/a$ for $x \in (0, a)$. Every value in the interval has surprisal

$$\mathbb{s}(x) = -\log \tfrac{1}{a} = \log a.$$

With $a = \frac{1}{2}$ and logarithms to base 2, the surprisal of every value is $\log_2 \frac{1}{2} = -1$ bit. A density value can exceed 1, so unlike the surprisal of a discrete value, it can be negative.

Definition 0.3 (Entropy (differential entropy, for continuous random variables))**.** The **entropy** of a random variable X is the expectation of its surprisal (discrete or continuous) [18], [19, Sec. 8.1], when that expectation exists:

$$\mathbb{H}(X) \stackrel{\text{def}}{=} \mathbb{E}[\mathbb{s}(X)].$$

For a continuous random variable, the entropy is also called the **differential entropy**.

Theorem 0.1 (Entropy of a discrete random variable)**.** *Let X be a discrete random variable with probability mass function $\mathbb{P}(X = x)$. Then*

$$\mathbb{H}(X) = -\sum_{x \in \mathcal{R}(X)} \mathbb{P}(X = x) \log \mathbb{P}(X = x).$$

A term with $\mathbb{P}(X = x) = 0$ is taken to be 0, because $p \log p \to 0$ as $p \to 0$.

Proof

Proof. Let $g(x) \stackrel{\text{def}}{=} \mathbb{s}(x) = -\log \mathbb{P}(X = x)$, which is finite for each value x with $\mathbb{P}(X = x) > 0$. Then LOTUS gives

$$\begin{aligned}
\mathbb{H}(X) &= \mathbb{E}[\mathbb{s}(X)] && \text{(definition of entropy)} \\
&= \mathbb{E}[g(X)] && \text{(definition of } g\text{)} \\
&= \sum_{x \in \mathcal{R}(X)} g(x) \cdot \mathbb{P}(X = x) && \text{(LOTUS)} \\
&= -\sum_{x \in \mathcal{R}(X)} \mathbb{P}(X = x) \log \mathbb{P}(X = x). && \text{(definition of } g\text{)}
\end{aligned}$$

□

Theorem 0.2 (Entropy of a continuous random variable)**.** *Let X be a continuous random variable with probability density function $\mathbb{p}(X = x)$. Then*

$$\mathbb{H}(X) = -\int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x) \log \mathbb{p}(X = x)\, dx.$$

The integrand is taken to be 0 where $\mathbb{p}(X = x) = 0$, because $p \log p \to 0$ as $p \to 0$.

Proof

Proof. Let $g(x) \stackrel{\text{def}}{=} \mathbb{s}(x) = -\log \mathbb{p}(X = x)$. Then LOTUS gives

$$\begin{aligned}
\mathbb{H}(X) &= \mathbb{E}[\mathbb{s}(X)] && \text{(definition of entropy)} \\
&= \mathbb{E}[g(X)] && \text{(definition of } g\text{)} \\
&= \int_{x \in \mathcal{R}(X)} g(x) \cdot \mathbb{p}(X = x)\, dx && \text{(LOTUS)} \\
&= -\int_{x \in \mathcal{R}(X)} \mathbb{p}(X = x) \log \mathbb{p}(X = x)\, dx. && \text{(definition of } g\text{)}
\end{aligned}$$

□

Example 0.3 (Entropy of a uniform random variable)**.** Let X be uniformly distributed on the interval from 0 to $a > 0$, as in Example 0.2. By Theorem 0.2,

$$\begin{aligned}\mathbb{H}(X) &= -\int_0^a \tfrac{1}{a} \log \tfrac{1}{a}\, dx \\ &= \log a.\end{aligned}$$

Every value has the same surprisal $\log a$, so its expectation is also $\log a$. With $a = \frac{1}{2}$ and logarithms to base 2, the entropy is -1 bit. The entropy of a continuous random variable can be negative [19], Example 8.1.1.

Remark (The base of the logarithm is a unit)

Remark 0.2 (The base of the logarithm is a unit). With the natural logarithm, entropy is measured in *nats*. With base 2, it is measured in *bits*. Changing the base multiplies every entropy by the same positive constant. It does not change which distribution has the larger entropy.

Example 0.4 (Entropy of a coin flip)**.** Let X be the result of one flip of a coin with $\mathbb{P}(X = 1) = \pi$ and $\mathbb{P}(X = 0) = 1 - \pi$ (Bernoulli distribution). By Theorem 0.1, using logarithms to base 2,

$$H(X) = -\pi \log_2 \pi - (1 - \pi) \log_2 (1 - \pi).$$

For a fair coin, $\pi = 1/2$, and

$$\begin{aligned}H(X) &= -\tfrac{1}{2} \log_2 \tfrac{1}{2} - \tfrac{1}{2} \log_2 \tfrac{1}{2} \\ &= \tfrac{1}{2} + \tfrac{1}{2} \\ &= 1 \text{ bit}.\end{aligned}$$

For a coin that always lands heads, $\pi = 1$, and

$$\begin{aligned}H(X) &= -1 \cdot \log_2 1 - 0 \\ &= 0.\end{aligned}$$

The outcome is known in advance, so there is no uncertainty to measure. An entropy of 0 means that all of the probability is on one outcome.

Theorem 0.3 (Entropy of a discrete random variable is nonnegative)**.** *Let X be a discrete random variable. Then*

$$\mathbb{H}(X) \geq 0.$$

Proof

Proof. For each value x with $\mathbb{P}(X = x) > 0$, the probability satisfies $\mathbb{P}(X = x) \leq 1$, so $\log \mathbb{P}(X = x) \leq 0$. Hence the surprisal $\mathfrak{s}(x) = -\log \mathbb{P}(X = x)$ is nonnegative. The expectation of a nonnegative random variable is nonnegative, so $\mathbb{H}(X) = \mathbb{E}[\mathfrak{s}(X)] \geq 0$. □

Example 0.5 (Entropy of a random variable that is not random)**.** Let X be a discrete random variable that takes one value x_0 with probability 1. By Theorem 0.1, the entropy of X is

$$\mathbb{H}(X) = -1 \cdot \log 1 = 0.$$

The value 0 is the smallest that Theorem 0.3 allows, and the fair coin in Example 0.4 has a larger entropy.

Theorem 0.4 (Rescaling changes entropy by the log of the scale)**.** *Let X be a continuous random variable with entropy $\mathbb{H}(X)$, and let $a \neq 0$ be a real number. Then*

$$\mathbb{H}(aX) = \mathbb{H}(X) + \log|a|.$$

[19], Theorem 8.6.4

Proof

Proof. Let $Y \stackrel{\text{def}}{=} aX$. The density of Y is

$$\mathbb{p}(Y = y) = \frac{1}{|a|}\mathbb{p}\Big(X = \frac{y}{a}\Big).$$

At the value $y = ax$, the density of Y is therefore $1/|a|$ times the density of X at x. That relation holds at every value x, so it holds between the random variables. Taking $-\log$ of both sides turns this relation into one between surprisals:

$$\mathfrak{s}(aX) = \mathfrak{s}(X) + \log|a|.$$

Hence

$$\begin{aligned}
\mathbb{H}(aX) &= \mathbb{E}[\mathbb{s}(aX)] && \text{(definition of entropy)} \\
&= \mathbb{E}[\mathbb{s}(X) + \log|a|] && \text{(surprisal of } aX\text{)} \\
&= \mathbb{E}[\mathbb{s}(X)] + \log|a| && \text{(linearity of expectation)} \\
&= \mathbb{H}(X) + \log|a|. && \text{(definition of entropy)}
\end{aligned}$$

□

Theorem 0.5 (Shifting does not change entropy)**.** *Let X be a continuous random variable with entropy $\mathbb{H}(X)$, and let c be a real number. Then*

$$\mathbb{H}(X + c) = \mathbb{H}(X).$$

[19], Theorem 8.6.3

Proof

Proof. Let $Y \overset{\text{def}}{=} X + c$. The density of Y is

$$\mathbb{p}(Y = y) = \mathbb{p}(X = y - c).$$

At the value $y = x + c$, the density of Y therefore matches the density of X at x. That relation holds at every value x, so it holds between the random variables. So the surprisal of $X + c$ matches the surprisal of X:

$$\mathbb{s}(X + c) = \mathbb{s}(X).$$

Hence

$$\begin{aligned}
\mathbb{H}(X + c) &= \mathbb{E}[\mathbb{s}(X + c)] && \text{(definition of entropy)} \\
&= \mathbb{E}[\mathbb{s}(X)] && \text{(surprisal of } X + c\text{)} \\
&= \mathbb{H}(X). && \text{(definition of entropy)}
\end{aligned}$$

□

Example 0.6 (Rescaling and shifting a uniform random variable)**.** Let X be uniformly distributed between 0 and 1. By Example 0.3 with $a = 1$, the entropy of X is $\mathbb{H}(X) = \log 1 = 0$.
For a number $a > 0$, the random variable aX is uniformly distributed between 0 and a, so Example 0.3 gives $\mathbb{H}(aX) = \log a$. That value matches Theorem 0.4, which gives $\mathbb{H}(aX) = \mathbb{H}(X) + \log a = 0 + \log a$.
For a number c, the random variable $X + c$ is uniformly distributed between c and $c + 1$.

Its density is 1 between c and $c+1$, so by Theorem 0.2, its entropy is $-\int_c^{c+1} 1 \cdot \log 1 \, dx = 0 = \mathbb{H}(X)$, as Theorem 0.5 states.

Theorem 0.6 (Entropy of a quantized random variable)**.** *Let X be a continuous random variable whose density is continuous and is 0 outside a bounded interval. For a bin width $\Delta > 0$, let X^Δ be the discrete random variable that equals i when $i\Delta \le X < (i+1)\Delta$. Then*

$$\mathbb{H}\big(X^\Delta\big) + \log \Delta \to \mathbb{H}(X) \quad \text{as } \Delta \to 0.$$

[19], Theorem 8.3.1

Proof

Proof. Only finitely many bins meet the bounded interval, so $\mathbb{H}\big(X^\Delta\big)$ is a finite sum over those bins. Because the density is continuous, in every bin it takes the value $\mathbb{P}(i\Delta \le X < (i+1)\Delta)/\Delta$ at some point x_i of that bin. If the density is constant on the bin, any point of the bin takes that value. Otherwise $\mathbb{P}(i\Delta \le X < (i+1)\Delta)/\Delta$ equals $\frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} \mathbb{p}(X = x)\, dx$, the density's integral over the bin divided by the bin's width, so the value lies strictly between the minimum and the maximum of the density on the closed bin, so by continuity the density takes it at a point strictly between the two, which is inside the bin. By the choice of x_i, the probability mass of X^Δ at i is $\mathbb{P}(X^\Delta = i) = \mathbb{p}(X = x_i)\,\Delta$. By Theorem 0.1, with each sum over the bins of positive probability (a bin of probability 0 adds nothing to the entropy),

$$\begin{aligned}
\mathbb{H}\big(X^\Delta\big) &= -\sum_i \mathbb{p}(X = x_i)\,\Delta \log(\mathbb{p}(X = x_i)\,\Delta) \\
&= -\sum_i \Delta\, \mathbb{p}(X = x_i) \log \mathbb{p}(X = x_i) - \sum_i \mathbb{p}(X = x_i)\,\Delta \log \Delta \quad \text{(log of a product)} \\
&= -\sum_i \Delta\, \mathbb{p}(X = x_i) \log \mathbb{p}(X = x_i) - \log \Delta. \quad \text{(the masses sum to 1)}
\end{aligned}$$

Adding $\log \Delta$ to both sides leaves the first sum. A bin of probability 0 has density 0 throughout, because a continuous density is never negative, so that bin adds 0 to this sum too, by the convention $0 \log 0 = 0$ of Theorem 0.2. The sum is therefore a Riemann sum for $-\int \mathbb{p}(X = x) \log \mathbb{p}(X = x)\, dx$ over the bounded interval, with mesh Δ (the bin grid need not line up with where the interval starts and ends, because the density is 0 outside it), whose integrand is continuous. By Theorem 0.2, the value of that integral is $\mathbb{H}(X)$, and a Riemann sum converges to its Riemann integral as $\Delta \to 0$. □

Corollary 0.1 (Entropy of a quantized random variable grows without bound)**.** *In the setting of Theorem 0.6, the entropy of the quantized variable tends to ∞ as the bin width shrinks:*

$$\mathbb{H}\big(X^{\Delta}\big) \to \infty \quad \text{as } \Delta \to 0.$$

Proof

Proof. By Theorem 0.6, $\mathbb{H}\big(X^{\Delta}\big) + \log \Delta \to \mathbb{H}(X)$ as $\Delta \to 0$. Moving $\log \Delta$ to the other side,

$$\mathbb{H}\big(X^{\Delta}\big) = \big(\mathbb{H}\big(X^{\Delta}\big) + \log \Delta\big) - \log \Delta.$$

The first quantity on the right tends to $\mathbb{H}(X)$, not $\pm\infty$: the density is continuous and 0 outside a bounded interval, so $-\mathbb{p}(X = x) \log \mathbb{p}(X = x)$ is continuous there, with the convention $0 \log 0 = 0$, and a continuous function has an integral over a bounded interval. Also, $-\log \Delta \to \infty$ as $\Delta \to 0$, so $\mathbb{H}\big(X^{\Delta}\big) \to \infty$. □

Example 0.7 (Quantizing a uniform random variable)**.** Let X be uniformly distributed between 0 and 1, and let the bin width be $\Delta = 1/n$ for a positive whole number n. Then X^{Δ} takes each of n values with probability $1/n$, so by Theorem 0.1, its entropy is

$$\mathbb{H}\big(X^{\Delta}\big) = -n \cdot \tfrac{1}{n} \log \tfrac{1}{n} = \log n = -\log \Delta.$$

Hence $\mathbb{H}\big(X^{\Delta}\big) + \log \Delta = 0$, which equals $\mathbb{H}(X) = 0$. The density of X jumps at 0 and 1, so it is not continuous, and Theorem 0.6 does not apply as stated. Its conclusion still holds here, because the density takes one value on every bin. Measured in bits, with $n = 2^k$, the entropy is k. It tends to ∞ as k grows, which agrees with Corollary 0.1.

Remark (Where the discrete formula comes from)

Remark 0.3 (Where the discrete formula comes from)*.* Murphy [20, Sec. 2.8.1], p. 56 gives the discrete formula of Theorem 0.1 as the definition, and notes that among distributions on K outcomes it is largest, $\log K$, for the uniform distribution.

Definition 0.4 (Cross-entropy)**.** Let X be a random variable with probability mass function or density p, and let q be a mass function or density on the same values. The **cross-entropy** of q relative to p [21] is the expectation, when X has distribution p, of the surprisal (or its continuous form) $\mathbb{s}_q$ measured by q, where $\mathbb{s}_q(x) \stackrel{\text{def}}{=} -\log q(x)$:

$$\mathbb{H}(p,q) \overset{\text{def}}{=} \mathbb{E}\big[\mathbb{s}_q(X)\big].$$

We take $-\log 0 = \infty$, so when $\mathbb{P}(q(X) = 0) > 0$, the cross-entropy is infinite.

Exercise 0.1 (Mass function of the negative log probability for two mass functions)**.** Let X have probability mass function $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ on $\{1, 2, 3\}$, and let $q = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ be another probability mass function on $\{1, 2, 3\}$. Write q_X for the random variable that equals q_k when $X = k$.
Find the probability mass function of $Y = -\log_2 q_X$.

Solution

Solution 0.1. The variable Y equals $-\log_2 \frac{1}{4} = 2$ when $X \in \{1, 3\}$, and $-\log_2 \frac{1}{2} = 1$ when $X = 2$. So $\mathbb{P}(Y = 2) = p_1 + p_3 = \frac{3}{4}$ and $\mathbb{P}(Y = 1) = p_2 = \frac{1}{4}$.

Exercise 0.2 (Expected negative log probability for two mass functions)**.** Using the mass function of $Y = -\log_2 q_X$ from Exercise 0.1, compute $\mathbb{E}[Y]$.

Solution

Solution 0.2. From the mass function in Solution 0.1,

$$\begin{aligned}\mathbb{E}[Y] &= 2 \cdot \tfrac{3}{4} + 1 \cdot \tfrac{1}{4} \\ &= \tfrac{7}{4}.\end{aligned}$$

Exercise 0.3 (Cross-entropy of two mass functions from the definition)**.** Let $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ and $q = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ be probability mass functions on $\{1, 2, 3\}$.
Compute the cross-entropy $\mathbb{H}(p, q)$ in bits from its definition.

Solution

Solution 0.3. From the definition,

$$\begin{aligned}\mathbb{H}(p,q) &= -\tfrac{1}{2}\log_2 \tfrac{1}{4} - \tfrac{1}{4}\log_2 \tfrac{1}{2} - \tfrac{1}{4}\log_2 \tfrac{1}{4} \\ &= \tfrac{1}{2} \cdot 2 + \tfrac{1}{4} \cdot 1 + \tfrac{1}{4} \cdot 2 \\ &= \tfrac{7}{4}.\end{aligned}$$

Exercise 0.4 (Cross-entropy and the expected negative log probability)**.** Compare your value of $\mathbb{H}(p, q)$ from Exercise 0.3 with your value of $\mathbb{E}\left[-\log_2 q_X\right]$ from Exercise 0.2, where X has probability mass function p.

Solution

Solution 0.4. The cross-entropy equals $\mathbb{E}\left[-\log_2 q_X\right]$. The two calculations group the same three terms differently: the expectation first combines outcomes 1 and 3, which have the same value of $-\log_2 q_X$.

Exercise 0.5 (Cross-entropy as a sum when q is positive where p is)**.** Let p and q be probability mass functions on $\{1, \dots, K\}$ with $q_k > 0$ whenever $p_k > 0$. Let X have mass function p, and write q_X for the random variable that equals q_k when $X = k$. Show that the cross-entropy, the expectation $\mathbb{E}[-\log q_X]$, equals $-\sum_{k=1}^{K} p_k \log q_k$.

Solution

Solution 0.5. Let $g(k) \stackrel{\text{def}}{=} -\log q_k$, so that $-\log q_X = g(X)$. Since $q_k > 0$ whenever $p_k > 0$, $g(k)$ is finite for every value k that X takes with positive probability. By LOTUS,

$$\begin{aligned}
\mathbb{H}(p, q) &= \mathbb{E}[-\log q_X] \\
&= \mathbb{E}[g(X)] \\
&= \sum_{k:p_k>0} g(k) \cdot \mathbb{P}(X = k) \\
&= -\sum_{k:p_k>0} p_k \log q_k \\
&= -\sum_{k=1}^{K} p_k \log q_k.
\end{aligned}$$

The last line adds the terms with $p_k = 0$, which are 0 by the convention in Theorem 0.7.

Exercise 0.6 (Cross-entropy when q gives zero probability to a possible outcome)**.** Let p and q be probability mass functions on $\{1, \dots, K\}$ with $q_k = 0$ for some k with $p_k > 0$. Let X have mass function p, and write q_X for the random variable that equals q_k when $X = k$.
Find the cross-entropy $\mathbb{H}(p, q)$.

Solution

Solution 0.6. By Definition 0.4, $\mathbb{H}(p, q) = \infty$.

Exercise 0.7 (Expected negative log probability when q gives zero probability to a possible outcome)**.** In the setting of Exercise 0.6, find the expectation $\mathbb{E}[-\log q_X]$.

Solution

Solution 0.7. Here $-\log q_X \geq 0$ always, and $-\log q_X = \infty$ with probability at least $p_k > 0$, so $\mathbb{E}[-\log q_X] = \infty$.

Theorem 0.7 (Cross-entropy of two mass functions)**.** *Let p and q be probability mass functions on $\{1, \dots, K\}$, with p_k and q_k the probabilities of outcome k. Then the cross-entropy is*

$$\mathbb{H}(p, q) = -\sum_{k=1}^{K} p_k \log q_k.$$

A term with $p_k = 0$ is taken to be 0 whatever q_k is, as for entropy.

Proof

Proof. Let X have mass function p, and let $g(k) \stackrel{\text{def}}{=} -\log q_k$, so that $-\log q(X) = g(X)$. First suppose $q_k > 0$ whenever $p_k > 0$. Then $g(k)$ is finite for every value k that X takes with positive probability, and LOTUS gives

$$\begin{aligned}
\mathbb{H}(p, q) &= \mathbb{E}[g(X)] && \text{(definition)} \\
&= \sum_{k:p_k>0} g(k) \cdot \mathbb{P}(X = k) && \text{(LOTUS)} \\
&= -\sum_{k:p_k>0} p_k \log q_k && \text{(definition of } g \text{ and } \mathbb{P}(X = k) = p_k) \\
&= -\sum_{k=1}^{K} p_k \log q_k. && \text{(terms with } p_k = 0 \text{ are 0)}
\end{aligned}$$

Now suppose $q_k = 0$ for some k with $p_k > 0$. Then $-\log q(X) \geq 0$, and $-\log q(X) = \infty$ with probability at least $p_k > 0$, so $\mathbb{H}(p, q) = \infty$. The right side is also ∞, because its term for k is $-p_k \log 0 = \infty$ and its other terms are finite or ∞ but never $-\infty$. □

Theorem 0.8 (Cross-entropy of two densities)**.** *Let X be a continuous random variable with density p on its range $\mathcal{R}(X)$, and let q be a density on $\mathcal{R}(X)$. Then the cross-entropy is*

$$\mathbb{H}(p, q) = -\int_{x \in \mathcal{R}(X)} p(x) \log q(x)\, dx.$$

The integrand is taken to be 0 where $p(x) = 0$. If $\mathbb{P}(q(X) = 0) > 0$, both sides are ∞.

Proof

Proof. Let $g(x) \stackrel{\text{def}}{=} -\log q(x)$, so that $-\log q(X) = g(X)$. Then LOTUS gives

$$\begin{aligned} \mathbb{H}(p, q) &= \mathbb{E}[g(X)] && \text{(definition)} \\ &= \int_{x \in \mathcal{R}(X)} g(x) \cdot p(x)\, dx && \text{(LOTUS)} \\ &= -\int_{x \in \mathcal{R}(X)} p(x) \log q(x)\, dx. && \text{(definition of } g\text{)} \end{aligned}$$

Now suppose $\mathbb{P}(q(X) = 0) > 0$. Then $-\log q(X) \geq 0$, and $-\log q(X) = \infty$ with positive probability, so $\mathbb{H}(p, q) = \infty$. The integrand $-p(x) \log q(x)$ is ∞ wherever $q(x) = 0$ and $p(x) > 0$, which is a region of positive probability under p, and it is never $-\infty$, so the right side is also ∞. □

Example 0.8 (Cross-entropy of two densities on the unit interval)**.** Let $p(x) = 1$ and $q(x) = 2x$ for $x \in (0, 1)$ be densities. By Theorem 0.8, using the natural logarithm,

$$\begin{aligned} \mathbb{H}(p, q) &= -\int_0^1 \log(2x)\, dx \\ &= -\log 2 - \int_0^1 \log x\, dx \\ &= -\log 2 + 1 \\ &\approx 0.307 \text{ nats.} \end{aligned}$$

Exercise 0.8 (Cross-entropy of a mass function with itself)**.** Let $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ be a probability mass function on $\{1, 2, 3\}$. Compute the cross-entropy $\mathbb{H}(p, p)$ in bits.

Solution

Solution 0.8.

$$\begin{aligned}\mathbb{H}(p,p) &= -\tfrac{1}{2}\log_2\tfrac{1}{2} - \tfrac{1}{4}\log_2\tfrac{1}{4} - \tfrac{1}{4}\log_2\tfrac{1}{4}\\ &= \tfrac{1}{2}\cdot 1 + \tfrac{1}{4}\cdot 2 + \tfrac{1}{4}\cdot 2\\ &= \tfrac{3}{2} \text{ bits.}\end{aligned}$$

Exercise 0.9 (Entropy of a random variable with a given mass function)**.** Let X be a discrete random variable with probability mass function $p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ on $\{1, 2, 3\}$. Compute the entropy $\mathbb{H}(X)$ in bits from its definition.

Solution

Solution 0.9. The range of X is $\{1, 2, 3\}$, and $\mathbb{P}(X = k) = p_k$, so

$$\begin{aligned}\mathbb{H}(X) &= -\sum_{k\in\mathcal{R}(X)} \mathbb{P}(X=k)\log_2\mathbb{P}(X=k)\\ &= -\tfrac{1}{2}\log_2\tfrac{1}{2} - \tfrac{1}{4}\log_2\tfrac{1}{4} - \tfrac{1}{4}\log_2\tfrac{1}{4}\\ &= \tfrac{3}{2} \text{ bits.}\end{aligned}$$

Exercise 0.10 (Entropy and the cross-entropy of a mass function with itself)**.** Compare your value of $\mathbb{H}(X)$ from Exercise 0.9 with your value of $\mathbb{H}(p, p)$ from Exercise 0.8.

Solution

Solution 0.10. The entropy equals $\mathbb{H}(p, p)$, because the two sums have the same terms.

Exercise 0.11 (Entropy as a cross-entropy, for any mass function)**.** Let X be a discrete random variable with probability mass function p on $\{1, \dots, K\}$. Show that $\mathbb{H}(p, p) = \mathbb{H}(X)$.

Solution

Solution 0.11. By definition, $\mathbb{H}(p, p) = \mathbb{E}[-\log p(X)]$. Since $\mathbb{P}(X = k) = p_k$, we have $-\log p(X) = \mathfrak{s}(X)$, the surprisal of X. So $\mathbb{H}(p, p) = \mathbb{E}[\mathfrak{s}(X)] = \mathbb{H}(X)$ by the definition of entropy.

Corollary 0.2 (Entropy is the cross-entropy of a distribution with itself)**.** *Let X be a random variable with probability mass function or density p. Then*

$$\mathbb{H}(p, p) = \mathbb{H}(X).$$

Proof

Proof. By the definition of cross-entropy with $q = p$, $\mathbb{H}(p, p) = \mathbb{E}[-\log p(X)]$. The random variable $-\log p(X)$ is the surprisal (or its continuous form) $\mathrm{s}(X)$. So $\mathbb{H}(p, p) = \mathbb{E}[\mathrm{s}(X)] = \mathbb{H}(X)$ by the definition of entropy. □

Remark (Cross-entropy is an expected surprisal)

Remark 0.4 (Cross-entropy is an expected surprisal). By definition, the cross-entropy $\mathbb{H}(p, q)$ is the average surprisal of the values drawn from p, with each surprisal measured by q as $-\log q(x)$. By Corollary 0.2, measuring with p itself gives the entropy.

Definition 0.5 (Kullback–Leibler divergence (KL divergence, relative entropy))**.** For X, p and q as in Definition 0.4, the **Kullback–Leibler divergence** (**KL divergence**, or **relative entropy**) of q from p is the expectation, when X has distribution p, of the amount by which the surprisal measured by q exceeds the surprisal measured by p:

$$D_{\mathrm{KL}}(p \,\|\, q) \stackrel{\text{def}}{=} \mathbb{E}\big[\mathrm{s}_q(X) - \mathrm{s}_p(X)\big].$$

As for cross-entropy, we take $-\log 0 = \infty$, so when $\mathbb{P}(q(X) = 0) > 0$, the divergence is infinite.

Theorem 0.9 (KL divergence is cross-entropy minus entropy)**.** *For X, p and q as in Definition 0.4, suppose the cross-entropy $\mathbb{H}(p, q)$ and the entropy $\mathbb{H}(X)$ are real numbers. Then the KL divergence is*

$$D_{\mathrm{KL}}(p \,\|\, q) = \mathbb{H}(p, q) - \mathbb{H}(X).$$

This result does not cover the case $\mathbb{P}(q(X) = 0) > 0$, in which the cross-entropy is ∞.

Proof

Proof. By the definition of cross-entropy, $\mathbb{H}(p, q) = \mathbb{E}\big[\mathrm{s}_q(X)\big]$. By Corollary 0.2, $\mathbb{H}(p, p) = \mathbb{H}(X) = \mathbb{E}\big[\mathrm{s}_p(X)\big]$. By assumption, both of these expectations exist, so linearity of expectation applies:

$$
\begin{aligned}
D_{\mathrm{KL}}(p \,\|\, q) &= \mathbb{E}\big[\mathtt{s}_q(X) - \mathtt{s}_p(X)\big] && \text{(definition)} \\
&= \mathbb{E}\big[\mathtt{s}_q(X)\big] - \mathbb{E}\big[\mathtt{s}_p(X)\big] && \text{(linearity of expectation)} \\
&= \mathbb{H}(p, q) - \mathbb{H}(X). && \text{(the two expectations above)}
\end{aligned}
$$

□

Theorem 0.10 (KL divergence of two mass functions)**.** *Let p and q be probability mass functions on $\{1, \dots, K\}$, with p_k and q_k the probabilities of outcome k, and suppose $q_k > 0$ whenever $p_k > 0$. Then the KL divergence is*

$$
D_{\mathrm{KL}}(p \,\|\, q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}.
$$

A term with $p_k = 0$ is taken to be 0.

Proof

Proof. Let X have mass function p, and let $g(k) \overset{\text{def}}{=} -\log q_k + \log p_k$ for each k with $p_k > 0$, so that $\mathtt{s}_q(X) - \mathtt{s}_p(X) = g(X)$. Since $q_k > 0$ whenever $p_k > 0$, $g(k)$ is a real number for every value k that X takes with positive probability. By LOTUS,

$$
\begin{aligned}
D_{\mathrm{KL}}(p \,\|\, q) &= \mathbb{E}[g(X)] && \text{(definition)} \\
&= \sum_{k : p_k > 0} g(k) \cdot \mathbb{P}(X = k) && \text{(LOTUS)} \\
&= \sum_{k : p_k > 0} p_k (\log p_k - \log q_k) && \text{(definition of } g \text{ and } \mathbb{P}(X = k) = p_k) \\
&= \sum_{k : p_k > 0} p_k \log \frac{p_k}{q_k} && \text{(difference of logarithms)} \\
&= \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}. && \text{(terms with } p_k = 0 \text{ are 0)}
\end{aligned}
$$

□

Theorem 0.11 (Gibbs' inequality: the divergence is nonnegative)**.** *For p and q as in Theorem 0.10, the KL divergence satisfies*

$$
D_{\mathrm{KL}}(p \,\|\, q) \geq 0,
$$

with $D_{\mathrm{KL}}(p \,\|\, q) = 0$ exactly when $p_k = q_k$ for every k. The logarithm is natural here. Any other logarithm multiplies the divergence by a positive number.

Proof

Proof. Every sum below runs over the k with $p_k > 0$. By Theorem 0.10, $D_{\mathrm{KL}}(p \,\|\, q) = \sum_k p_k \log \frac{p_k}{q_k} = -\sum_k p_k \log \frac{q_k}{p_k}$.
By the bound on the logarithm, $\log t \le t - 1$ for every $t > 0$, with equality exactly when $t = 1$. Take $t = q_k/p_k$, and multiply by $p_k > 0$:

$$p_k \log \frac{q_k}{p_k} \le q_k - p_k,$$

with equality exactly when $q_k = p_k$. Summing over k gives

$$\begin{aligned} \sum_k p_k \log \frac{q_k}{p_k} &\le \sum_k q_k - \sum_k p_k && \text{(sum the bound over } k\text{)} \\ &= \sum_k q_k - 1 && (p \text{ sums to one}) \\ &\le 0. && (q \text{ sums to at most one}) \end{aligned}$$

Hence $D_{\mathrm{KL}}(p \,\|\, q) = -\sum_k p_k \log \frac{q_k}{p_k} \ge 0$.
Suppose $D_{\mathrm{KL}}(p \,\|\, q) = 0$. Then both relations above hold with equality. Equality in the first forces $q_k = p_k$ whenever $p_k > 0$. Equality in the second then gives $\sum_k q_k = 1$, so $q_k = 0$ whenever $p_k = 0$. Hence $q_k = p_k$ for every k. Conversely, if $q_k = p_k$ for every k, every term $p_k \log \frac{p_k}{q_k}$ is 0. □

Remark (The divergence is not symmetric)

Remark 0.5 (The divergence is not symmetric). In general, $D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$, so KL divergence is not a distance between p and q.

Example 0.9 (Cross-entropy and divergence for two coins)**.**
Let $p = (\frac{1}{2}, \frac{1}{2})$ be a fair coin and $q = (\frac{1}{4}, \frac{3}{4})$ a coin biased toward the second outcome. Using logarithms to base 2, $\log_2 \frac{1}{4} = -2$ and $\log_2 \frac{3}{4} = \log_2 3 - 2$, so

$$\begin{aligned} \mathbb{H}(p, q) &= -\tfrac{1}{2} \log_2 \tfrac{1}{4} - \tfrac{1}{2} \log_2 \tfrac{3}{4} \\ &= 1 + \tfrac{1}{2}\big(2 - \log_2 3\big) \\ &= 2 - \tfrac{1}{2} \log_2 3 \\ &\approx 1.208, \\ \mathbb{H}(p, p) &= -\tfrac{1}{2} \log_2 \tfrac{1}{2} - \tfrac{1}{2} \log_2 \tfrac{1}{2} \\ &= 1, \\ D_{\mathrm{KL}}(p \,\|\, q) &= \mathbb{H}(p, q) - \mathbb{H}(p, p) = 1 - \tfrac{1}{2} \log_2 3 \approx 0.208. \end{aligned}$$

Here the cross-entropy is larger than the entropy of p, and by Theorem 0.9 the divergence is the difference. It is positive, as Theorem 0.11 requires for $p \neq q$.

Remark (Where the definitions come from)

Remark 0.6 (Where the definitions come from). Murphy [20, Sec. 2.8.2], pp. 57–58 defines the KL divergence and the cross-entropy as above, and gives the coding interpretation: the cross-entropy is the average number of bits needed to encode data from a source with distribution p using a code designed for q.

Further reading

Another explanation

Lange's textbook [22] covers conditioning as a tool for computing expectations in Section 2.3 and tail probability methods in Section 2.5. Evans and Rosenthal's free textbook [23] covers expectation in Chapter 3.

References

[1] Wikipedia contributors, "Expected value — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Expected_value

[2] P. Billingsley, *Probability and measure*, 3rd ed. in Wiley series in probability and mathematical statistics. New York: Wiley, 1995.

[3] J. Soch, "Proof: Expected value of a non-negative random variable." The Book of Statistical Proofs, 2020. doi: 10.5281/zenodo.4305949.

[4] Wikipedia contributors, "Law of the unconscious statistician — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician

[5] A. Gut, *Probability: A graduate course*, 2nd ed. in Springer texts in statistics. New York: Springer, 2013. doi: 10.1007/978-1-4614-4708-5.

[6] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[7] Wikipedia contributors, "Law of the unconscious statistician — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician

[8] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.

[9] Wikipedia contributors, "Conditional expectation — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Conditional_expectation

[10] K. Ross, "An introduction to probability and simulation." Bookdown, 2022. Available: https://bookdown.org/kevin_davisross/probsim-book/

[11] W. Rudin, *Principles of mathematical analysis*, 3rd ed. in International series in pure and applied mathematics. New York: McGraw-Hill, 1976.

[12] Wikipedia contributors, "Loss function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Loss_function

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[14] Wikipedia contributors, "Risk function — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Risk_function

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: With applications in R*, 2nd ed. Springer, 2021. doi: 10.1007/978-1-0716-1418-1.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, MA: MIT Press, 2016. Available: https://www.deeplearningbook.org/

[17] C. M. Bishop, *Pattern recognition and machine learning*. in Information science and statistics. New York: Springer, 2006.

[18] Wikipedia contributors, "Entropy (information theory) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory)

[19] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Wiley-Interscience, 2006. doi: 10.1002/047174882X.

[20] K. P. Murphy, *Machine learning: A probabilistic perspective*. in Adaptive computation and machine learning. MIT Press, 2012.

[21] Wikipedia contributors, "Cross-entropy — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Cross-entropy

[22] K. Lange, *Applied probability*, 3rd ed. in Springer texts in statistics. Springer, 2024. doi: 10.1007/978-1-0716-4172-9.

[23] M. J. Evans and J. S. Rosenthal, *Probability and statistics: The science of uncertainty*, 2nd ed. Freeman, 2010. Available: https://utstat.utoronto.ca/mikevans/jeffrosenthal/