Expectation

Last modified: 2026-09-28 23:45:35 (PDT)

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

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

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

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

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.

Remark. See also https://en.wikipedia.org/wiki/Expected_value

Example 0.1 (The standard Cauchy distribution has no expectation)**.** A random variable whose expectation is not defined is not exotic. Let X have the standard Cauchy distribution, with density $\mathrm{p}(X = x) = \frac{1}{\pi(1+x^2)}$ 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:

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

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

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,* $\text{P}(X=1) = \pi$ *and* $\text{P}(X=0) = 1-\pi$*, then:*

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

Proof

Proof.

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

□

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

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

Proof

Proof. We prove the continuous case, in which T has a density f. The integrand $\mathrm{f}(u) \cdot \mathbb{1}(0 \le t \le u)$ is non-negative on $[0,\infty) \times [0,\infty)$, so Tonelli's theorem (the non-negative case of the Fubini–Tonelli theorem; Billingsley (1995), Theorem 18.3) lets us exchange the order of integration:

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

Every step also holds when $\int_{u=0}^{\infty} u\ \mathrm{f}(u)\, du$ is infinite, because Tonelli's theorem needs only a non-negative integrand; so $\int_{t=0}^{\infty} \mathrm{S}(t)\, dt$ is finite exactly when $\mathrm{E}[T]$ is defined.
See (Soch 2020) for an alternative presentation of this proof. □

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

$$
\begin{aligned}
\mathrm{E}[T] &= \int_0^\infty \mathrm{S}(t)\,dt && \text{(expectation via the survival function)}\\
&= \int_0^\infty \mathrm{e}^{-\lambda t}\,dt && \text{(substitute the survival function)}\\
&= \left[-\frac{1}{\lambda}\mathrm{e}^{-\lambda t}\right]_0^\infty && \text{(antiderivative)}\\
&= 0 - \left(-\frac{1}{\lambda}\right) && \text{(evaluate at the bounds)}\\
&= \frac{1}{\lambda} && \text{(simplify)}
\end{aligned}
$$

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** $\mathrm{E}\,\mathbf{A}$ is the $m \times n$ matrix whose (i, j)-th element is $\mathrm{E}\big[A_{ij}\big]$:

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

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

Example 0.3 (Expectation of a random vector of coin flips)**.** For the two fair coin flips in the independence page's example, the random vector $\tilde{X} = (X_1, X_2)^\top$ is a 2×1 random matrix, and:

$$
\mathrm{E}\,\tilde{X} = \begin{pmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}
$$

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:*

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

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

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

Proof

Proof. We prove the discrete case. 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.

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

The last step regroups a possibly infinite sum, which is valid because $E[g(X)]$ is defined only when the sum converges absolutely. □

Remark. LOTUS says that to compute $E[g(X)]$, we do not need to first find the distribution of $g(X)$; we can compute the expectation directly using the distribution of X. The continuous case is the density-weighted analogue of this argument, but a fully rigorous proof needs the general 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 (Billingsley 1995; Gut 2013; Casella and Berger 2002; Wikipedia contributors 2026).

Example 0.4 (Expected value of X^2 for a Bernoulli variable)**.** Let $X \sim \mathrm{Ber}(\pi)$. By LOTUS (Theorem 0.3, discrete case):

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

Example 0.5 (Expected value of X^2 for a Uniform(0,1) variable)**.** Let $X \sim \mathrm{Uniform}(0, 1)$ (uniform distribution), so $\mathrm{p}(X = x) = 1$ for $x \in [0, 1]$ (uniform density). By LOTUS (Theorem 0.3, continuous case):

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

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

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

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 $g(x, y) = ax + by + c$:

$$
\begin{aligned}
\mathrm{E}[aX+bY+c] &= \sum_x \sum_y (ax+by+c)\cdot \mathrm{P}(X=x,Y=y) \\
&= a\sum_x\sum_y x\cdot \mathrm{P}(X=x,Y=y) + b\sum_x\sum_y y\cdot \mathrm{P}(X=x,Y=y) + c\sum_x\sum_y \mathrm{P}(X=x,Y=y) \\
&= a\sum_x\sum_y x\cdot \mathrm{P}(X=x,Y=y) + b\sum_y\sum_x y\cdot \mathrm{P}(X=x,Y=y) + c\sum_x\sum_y \mathrm{P}(X=x,Y=y) \\
&= a\sum_x x\sum_y \mathrm{P}(X=x,Y=y) + b\sum_y y\sum_x \mathrm{P}(X=x,Y=y) + c\cdot 1 \\
&= a\sum_x x\cdot \mathrm{P}(X=x) + b\sum_y y\cdot \mathrm{P}(Y=y) + c \\
&= a\,\mathrm{E}[X] + b\,\mathrm{E}[Y] + c
\end{aligned}
$$

Splitting and exchanging the sums is valid because the sums converge absolutely, since $|ax+by+c| \le |a||x|+|b||y|+|c|$ and $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ 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 (Billingsley 1995), which is beyond these notes' scope. □

Example 0.6 (Expected total of two dice)**.** Let X and Y be the results of two six-sided die rolls. Each has expectation $\sum_{x=1}^{6} x\cdot\frac{1}{6} = \frac{21}{6} = 3.5$, so:

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

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:*

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

Proof

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

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

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

Example 0.7 (Expectation of a rescaled uniform variable)**.** Let $X \sim \mathrm{Uniform}(0,1)$, so $\mathrm{E}[X] = \int_0^1 x\,dx = \frac{1}{2}$. By Corollary 0.1 with $a = 2$ and $c = 1$:

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

Conditional distributions and expectations

Definition 0.3 (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 $\mathrm{P}(X = x) > 0$) is:

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

Example 0.8 (Conditional PMF from a vaccine trial)**.** The `vaccine` dataset in the `dobson` package (Dobson and Barnett (2018), 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).

```r
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

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

The joint PMF of (X, Y) is the table of frequencies divided by the total $n = 73$ trial participants. By the marginal PMF theorem, the marginal probability $\text{P}(X = \text{placebo})$ is:

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

By Definition 0.3, the conditional PMF of Y given $X = \text{placebo}$ is:

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

Definition 0.4 (Conditional probability density function)**.** Let X and Y be jointly distributed continuous random variables with joint density $\text{p}(X = x, Y = y)$ and marginal density $\text{p}(X = x)$. The **conditional probability density function** of Y given $X = x$ (for values of x with $\text{p}(X = x) > 0$) is:

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

Example 0.9 (Conditional PDF from a bivariate normal model of birthweight data)**.** The `birthweight` dataset in the `dobson` package (Dobson and Barnett (2018), 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.

```r
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)
pander::pander(data.frame(
  parameter = c("mu_x", "sigma_x", "mu_y", "sigma_y", "rho"),
  estimate = round(c(mu_x, sigma_x, mu_y, sigma_y, rho), 4)
))
```

parameter	estimate
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 sample moments, the joint density is (a standard result; e.g. Casella and Berger (2002)):

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

A further standard fact about the bivariate normal (Casella and Berger (2002)) 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:

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

By Definition 0.4, dividing the joint density by this marginal density and simplifying the exponent (completing the square in y; Casella and Berger (2002)) 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 $1 - \rho^2$:

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

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

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

Continuous case. Let X and Y be jointly distributed continuous random variables. The **conditional expectation** of Y given $X = x$, using Definition 0.4, is:

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

Example 0.10 (Conditional expectation from real trial and birthweight data)**. Discrete case.** Continuing Example 0.8, score the vaccine trial's response levels small = 1, moderate = 2, large = 3. The conditional PMF of Y given X = placebo (from Example 0.8) is:

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

so:

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

Continuous case. Continuing Example 0.9, $Y \mid X = 40 \sim \mathrm{N}(3136.15, 188.37^2)$. The mean of a normal distribution is its location parameter (Casella and Berger (2002)), so integrating y against this conditional density gives:

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

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 0.6 (Conditional expectation: mixed case)**.** Suppose exactly one of X, Y is discrete and the other is continuous, with joint density-mass function $\mathrm{p}(X = x,\, Y = y)$.
X **discrete,** Y **continuous.** Here $\mathrm{p}(X = x,\, Y = y)$ is, for each fixed x, a probability density in y, with $\int_y \mathrm{p}(X = x,\, Y = y)\, dy = \mathrm{P}(X = x)$. The conditional PDF of Y given $X = x$ (for values of x with $\mathrm{P}(X = x) > 0$) is:

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

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

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

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

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

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

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

Example 0.11 (Conditional expectation, one discrete variable and one continuous variable)**.** X **discrete,** Y **continuous.** The `plasma` dataset in the `dobson` package (Dobson and Barnett (2018), 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.

```
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 $\mathrm{P}(X = \text{H-O}) = 11/23$:

$$Y \mid X = \text{H-O} \sim \mathrm{N}(3.95,\, 0.78^2)$$
$$Y \mid X = \text{C} \sim \mathrm{N}(2.78,\, 0.41^2)$$

By Definition 0.6, since the mean of a normal distribution is its location parameter:

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

***X* continuous, *Y* discrete.** The `senility` dataset in the `dobson` package (Dobson and Barnett (2018), 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 $\mathrm{P}(Y=1 \mid X=x)$ with the fitted logistic regression:

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

By Definition 0.6, since $Y \mid X=x$ is Bernoulli:

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

```
senility_p10 <- predict(
  senility_fit,
  newdata = data.frame(x = 10),
  type = "response"
)
senility_p15 <- predict(
  senility_fit,
  newdata = data.frame(x = 15),
  type = "response"
)
```

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

Definition 0.7 (Conditional expectation function)**.** The **conditional expectation function** $\mathrm{E}[Y \mid X]$ is the function (and hence random variable) of X obtained by evaluating $\mathrm{E}[Y \mid X=x]$ at X; specifically, $\mathrm{E}[Y \mid X] = g(X)$ where $g(x) \stackrel{\text{def}}{=} \mathrm{E}[Y \mid X=x]$.

Example 0.12 (The conditional expectation function of the birthweight model)**.** Continuing Example 0.9 and Example 0.10, (X, Y) (gestational age, birthweight) is modeled as bivariate normal. The general form of the conditional mean derived in Example 0.9, 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 0.10's result:

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

Definition 0.8 (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 0.3, Definition 0.4, or Definition 0.6). The **conditional expectation** of $h(X, Y)$ given $X = x$ is, for discrete Y:

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

and for continuous Y:

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

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

Example 0.13 (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 $h(x, y) = (y - x)^2$:

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

Corollary 0.2 (With $h(x, y) = y$, the general definition gives $\mathrm{E}[Y \mid X = x]$)**.** *For jointly distributed X and Y and $h(x, y) = y$, Definition 0.8 gives the conditional expectation of Y given $X = x$ from Definition 0.5 (when X and Y are both discrete or both continuous) or from Definition 0.6 (when one is discrete and the other continuous):*

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

Proof

Proof. For discrete Y:

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

The last step is Definition 0.5 when X is discrete and Definition 0.6 when X is continuous. For continuous Y, the same steps hold with integrals over y in place of the sums and $\mathrm{p}(Y=y \mid X=x)$ in place of $\mathrm{P}(Y=y \mid X=x)$. $\square$

Theorem 0.5 (Conditional LOTUS, discrete case)**.** *Let X and Y be jointly distributed discrete random variables, h a function of two arguments, $W \overset{def}{=} h(X,Y)$, and x a value with $\mathrm{P}(X=x)>0$. Then W is discrete, and $\mathrm{E}[h(X,Y) \mid X=x]$ from Definition 0.8 equals $\mathrm{E}[W \mid X=x]$ from Definition 0.5, 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:

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

Every $y \in \mathcal{R}(Y)$ with $\mathrm{P}(Y=y \mid X=x)>0$ 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:

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

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. □

Remark. Definition 0.8 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 0.5, results about $\mathrm{E}[W \mid X]$ apply to it. Y can also be a pair (Y, Z), with the sum or integral taken over both.

Theorem 0.6 (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:*

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

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

Proof

Proof. For discrete Y:

$$
\begin{aligned}
&\mathrm{E}[ah_1(X,Y) + bh_2(X,Y) \mid X = x] \\
&= \sum_y (ah_1(x,y) + bh_2(x,y)) \cdot \mathrm{P}(Y = y \mid X = x) && \text{(definition of conditional expectatio} \\
&= a\sum_y h_1(x,y) \cdot \mathrm{P}(Y = y \mid X = x) + b\sum_y h_2(x,y) \cdot \mathrm{P}(Y = y \mid X = x) && \text{(split the sum; factor out the consta} \\
&= a\,\mathrm{E}[h_1(X,Y) \mid X = x] + b\,\mathrm{E}[h_2(X,Y) \mid X = x] && \text{(definition of conditional expectatio}
\end{aligned}
$$

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 0.14 (Linearity, given the first flip)**.** Continuing Example 0.13, given $X = 1$, Y is 1 or 2 with probability $1/2$ each, so $\mathrm{E}[Y \mid X = 1] = 3/2$ and $\mathrm{E}[Y^2 \mid X = 1] = (1+4)/2 = 5/2$. By Theorem 0.6 with $a = b = 1$:

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

As a check, directly: $(1+1) \cdot \frac{1}{2} + (2+4) \cdot \frac{1}{2} = 1 + 3 = 4$.

Theorem 0.7 (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:*

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

Evaluating both sides at X gives $\mathrm{E}[g(X)\,h(X,Y) \mid X] = g(X) \cdot \mathrm{E}[h(X,Y) \mid X]$. In particular, taking $h = 1$ gives $\mathrm{E}[g(X) \mid X] = g(X)$.

ℹ Proof

Proof. For discrete Y:

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

For $h = 1$, the remaining sum is $\sum_y \mathrm{P}(Y = y \mid X = x) = 1$:

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

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. This result is often summarized as "taking out what is known": given $X = x$, any function of X is the known constant $g(x)$.

Example 0.15 (Taking out the first flip)**.** Continuing Example 0.14, with $g(x) = x$ and $h(x, y) = y$:

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

and $\mathrm{E}[XY \mid X = 0] = 0 \cdot \mathrm{E}[Y \mid X = 0] = 0$. 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 $\mathrm{E}[X + Y]$.

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:

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

As a check: $\mathrm{E}[X] = 0(0.5) + 1(0.5) = 0.5$ and $\mathrm{E}[Y] = 0(0.3) + 1(0.7) = 0.7$, so $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 1.2$ 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 $\mathrm{P}(X = x, Y = y)$. Marginal totals: $\mathrm{P}(X = 0) = 0.5$, $\mathrm{P}(X = 1) = 0.5$, $\mathrm{P}(Y = 0) = 0.3$, $\mathrm{P}(Y = 1) = 0.7$.

Remark. Note that X and Y are **not** independent here: $\mathrm{P}(X = 0, Y = 0) = 0.2 \neq 0.15 = \mathrm{P}(X = 0)\ \mathrm{P}(Y = 0)$. 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 0.8 (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:

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

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

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 (Billingsley 1995, 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 0.3 (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) $\mathrm{E}[|h(X, Y)|] < \infty$*,*

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

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

The choice of reference measures covers three cases:

- ***Both continuous:*** $\mu_X = \mu_Y =$ *Lebesgue measure;* $f_{X,Y}$ *is the joint probability density function (PDF), and* $\int g(x) \, d\mu_X(x) = \int g(x) \, dx$*.*
- ***Both discrete:*** $\mu_X = \mu_Y =$ *counting measure;* $f_{X,Y}(x, y) = \mathrm{P}(X = x, \, Y = y)$ *is the joint probability mass function (PMF), and* $\int g(x) \, d\mu_X(x) = \sum_{x \in \mathcal{R}(X)} g(x)$*.*
- ***Mixed*** *(one continuous, one discrete): one reference measure is Lebesgue and the other is counting;* $f_{X,Y}$ *is the joint density-mass function,* $f_{X,Y}(x, y) = f_{X|Y}(x \mid$

$y)$ $\mathrm{P}(Y = y)$ (or $\mathrm{P}(X = x \mid Y = y)\, f_Y(y)$ 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 0.6, just conditioning on Y instead of X.

Proof

Proof. Apply Theorem 0.8 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 $\mathrm{E}[|h(X, Y)|] < \infty$. Independence is not required. When X and Y are independent, $f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$ (or $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\, \mathrm{P}(Y = y)$ in the discrete case), and the iterated integrals factor into separate integrals over the marginals. $\square$

Example 0.16 (Expectation of a product of independent variables)**.** Let $X \sim \mathrm{Uniform}(0, 1)$ and $Y \sim \mathrm{Uniform}(0, 2)$, independently distributed. Compute $\mathrm{E}[XY]$.
We apply Corollary 0.3 (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 $f_{X,Y}(x, y) = f_X(x)\, f_Y(y) = \frac{1}{2}$, and $\mu_X = \mu_Y =$ Lebesgue measure:

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

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

Example 0.17 (When independence fails: a counterexample)**.** Correctly applying Corollary 0.3 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:

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

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 0.3 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 $f_X = f_Y = 1$, 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 $\mathrm{E}[XY]$ for *independent* uniforms ($\frac{1}{4}$), not $\mathrm{E}[XX]$ for the perfectly correlated pair ($\frac{1}{3}$). The lesson is that Corollary 0.3 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 0.3 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 (grey, 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 grey distribution).

Example 0.18 (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 $\mathrm{E}[X + Y]$.
By Corollary 0.3:

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

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

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

Proof

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

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

Continuous case. When X and Y are continuous:

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

Mixed case, X discrete and Y continuous. With the joint density-mass function $\mathrm{p}(X = x,\, Y = y)$, Corollary 0.3 applies with $h(x, y) = y$, counting measure for X, and Lebesgue measure for Y; its condition (b) holds because $\mathrm{E}[|Y|] < \infty$. The sum runs over the x with $\mathrm{P}(X = x) > 0$, where Definition 0.6 defines $\mathrm{E}[Y \mid X = x]$:

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X]] &= \sum_x \mathrm{E}[Y \mid X = x] \cdot \mathrm{P}(X = x) && \text{(LOTUS, discrete case)} \\
&= \sum_x \left(\int_y y \cdot \mathrm{p}(Y = y \mid X = x)\, dy \right) \cdot \mathrm{P}(X = x) && \text{(definition of conditional expectation, mixed cas} \\
&= \sum_x \int_y y \cdot \mathrm{p}(Y = y \mid X = x) \cdot \mathrm{P}(X = x)\, dy && \text{(move the constant } \mathrm{P}(X = x) \text{ inside the integra} \\
&= \sum_x \int_y y \cdot \mathrm{p}(X = x,\, Y = y)\, dy && \text{(definition of the conditional density, mixed cas} \\
&= \mathrm{E}[Y] && \text{(joint-distribution form of Fubini–Tonelli, with}
\end{aligned}
$$

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

$$
\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X]] &= \int_x \mathrm{E}[Y \mid X = x] \cdot \mathrm{p}(X = x)\, dx && \text{(LOTUS, continuous case)} \\
&= \int_x \left(\sum_y y \cdot \mathrm{P}(Y = y \mid X = x) \right) \cdot \mathrm{p}(X = x)\, dx && \text{(definition of conditional expectation, mixed cas} \\
&= \int_x \sum_y y \cdot \mathrm{P}(Y = y \mid X = x) \cdot \mathrm{p}(X = x)\, dx && \text{(move the constant } \mathrm{p}(X = x) \text{ inside the sum)} \\
&= \int_x \sum_y y \cdot \mathrm{p}(X = x,\, Y = y)\, dx && \text{(definition of the conditional PMF, mixed case)} \\
&= \mathrm{E}[Y] && \text{(joint-distribution form of Fubini–Tonelli, with}
\end{aligned}
$$

□

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

Theorem 0.10 (Conditional law of iterated expectations)**.** *For random variables X, Y, and Z with* $\mathrm{E}[|Y|] < \infty$*:*

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

Proof

Proof. Fix 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 0.9 goes through under it, line by line.

Discrete case.

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

Continuous case.

$$\begin{aligned}
\mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z = z] &= \int_x \mathrm{E}[Y \mid X = x, Z = z] \cdot \mathrm{p}(X = x \mid Z = z)\, dx && \text{(expectation given } Z \\
&= \int_x \left(\int_y y \cdot \mathrm{p}(Y = y \mid X = x, Z = z)\, dy \right) \cdot \mathrm{p}(X = x \mid Z = z)\, dx && \text{(definition of E[}Y \mid X \\
&= \int_x \int_y y \cdot \mathrm{p}(X = x, Y = y \mid Z = z)\, dy\, dx && \text{(product of condition} \\
&= \int_y y \cdot \left(\int_x \mathrm{p}(X = x, Y = y \mid Z = z)\, dx \right) dy && \text{(Fubini's theorem)} \\
&= \int_y y \cdot \mathrm{p}(Y = y \mid Z = z)\, dy && \text{(marginalize over } x\text{)} \\
&= \mathrm{E}[Y \mid Z = z] && \text{(definition of conditio}
\end{aligned}$$

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

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

Example 0.19 (Marginal expectation from conditional expectations)**.** Suppose X is a binary random variable indicating treatment assignment ($X = 1$ treated, $X = 0$ control), with $\mathrm{P}(X = 1) = 0.5$, and suppose the outcome Y has conditional expectations:

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

By the law of iterated expectations (Theorem 0.9):

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

Exercise 0.2 (Both-discrete case, infinite support: joint PMF)**.** Let X and Y be independent, each Geometric on $\mathcal{R}(X) = \mathcal{R}(Y) = \{0, 1, 2, \dots\}$ with $\mathrm{P}(X = x) = (1 - p)\, p^x$ 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 $\mathrm{P}(X = x,\, Y = y) = (1 - p)^2\, p^{x+y}$. Compute $\mathrm{E}[X + Y]$.

Solution

Solution. Compute $\mathrm{E}[X + Y]$ using Corollary 0.3 with $\mu_X = \mu_Y =$ counting measure 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 0.3 (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 (2002)). By Corollary 0.3 (both-discrete case), summing over y first for each fixed x:

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

As a check, $\mathrm{E}[X]=\mathrm{E}[Y]=\frac{p}{1-p}$ (the mean of this Geometric distribution; Casella and Berger (2002)), so $\mathrm{E}[X+Y]=\mathrm{E}[X]+\mathrm{E}[Y]=\frac{2p}{1-p}$ 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))

pander::pander(data.frame(
  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)
))
```

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 0.2 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 0.3's condition (b), $\mathrm{E}[|h(X, Y)|] < \infty$, 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. (Rudin 1976, 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 0.3 (Mixed case: one continuous variable, one discrete variable)**.** Let $Y \sim \text{Bernoulli}(0.6)$ and, given $Y = y$, let $X \mid Y = y \sim \text{Uniform}(0,\, y+1)$.
Compute $\mathrm{E}[X]$.

Solution

Solution. Compute $\mathrm{E}[X]$ using Corollary 0.3 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x, y) = x$.
The joint density w.r.t. Lebesgue $\times$ counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\ \mathrm{P}(Y = y)$:

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

By Corollary 0.3 (mixed case):

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

As a check using the law of iterated expectations (Theorem 0.9): $\mathrm{E}[X \mid Y = 0] = \frac{1}{2}$ and $\mathrm{E}[X \mid Y = 1] = 1$, so $\mathrm{E}[X] = \frac{1}{2}(0.4) + 1(0.6) = 0.2 + 0.6 = 0.8$.

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

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

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 0.3's guarantee would be load-bearing.

Exercise 0.4 (Mixed case, infinite discrete support)**.** Let Y be Geometric on $\{0, 1, 2, \dots\}$ with $\mathrm{P}(Y = y) = (1-q)\, q^y$ for a fixed $q \in (0, 1)$ and, given $Y = y$, let $X \mid Y = y \sim \mathrm{Uniform}(0,\, y+1)$. Unlike Exercise 0.3, Y's range here is countably infinite.
Compute $\mathrm{E}[X]$.

Solution

Solution. Compute $\mathrm{E}[X]$ using Corollary 0.3 with μ_X = Lebesgue measure, μ_Y = counting measure, and $h(x, y) = x$.
The joint density w.r.t. Lebesgue $\times$ counting measure is $f_{X,Y}(x, y) = f_{X|Y}(x \mid y)\, \mathrm{P}(Y = y) = \frac{(1-q)\, q^y}{y+1}$ for $x \in [0, y+1]$.
Since $h(x, y) = x \geq 0$ on this support, condition (a) holds, so Corollary 0.3 (via Tonelli's theorem) guarantees the now-infinite sum-of-integrals expression is valid — unlike in Exercise 0.3, this Fubini–Tonelli justification is required because the sum is infinite rather than finite.
By Corollary 0.3 (mixed case):

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

using the same geometric-series facts as Exercise 0.2 (e.g. Casella and Berger (2002)). As a check using the law of iterated expectations (Theorem 0.9) and the expectation of a linear function (Corollary 0.1): $\mathrm{E}[X \mid Y = y] = \frac{y+1}{2}$ (the mean of Uniform$(0, y+1)$) and $\mathrm{E}[Y] = \frac{q}{1-q}$ (the mean of this Geometric distribution; Casella and Berger (2002)), so:

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

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)

pander::pander(data.frame(
  quantity = c("exact (closed form)", "truncated sum-of-integrals"),
  value = round(c(exact_ex, trunc_ex), 6)
))
```

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

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

References

Billingsley, Patrick. 1995. *Probability and Measure*. 3rd ed. Wiley Series in Probability and Mathematical Statistics. Wiley.

Casella, George, and Roger Berger. 2002. *Statistical Inference*. 2nd ed. Cengage Learning. https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Gut, Allan. 2013. *Probability: A Graduate Course*. 2nd ed. Springer Texts in Statistics. Springer. https://doi.org/10.1007/978-1-4614-4708-5.

Rudin, Walter. 1976. *Principles of Mathematical Analysis*. 3rd ed. International Series in Pure and Applied Mathematics. McGraw-Hill.

Soch, Joram. 2020. *Proof: Expected Value of a Non-Negative Random Variable*. The Book of Statistical Proofs. https://doi.org/10.5281/zenodo.4305949.

Wikipedia contributors. 2026. *Law of the Unconscious Statistician — Wikipedia, the Free Encyclopedia.* https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician.