Random variables and expectations

Robert J. Carroll

Last revised: 5 May 2026

---

1 Motivation

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the setting in which uncertain outcomes live, but the quantities one actually wants to compute are usually *numbers attached to those outcomes* — a vote count, a candidate's quality score, a signal realization, a turnout share, a policy distance. A *random variable* is the formal version of "a real-valued quantity that depends on the random outcome." The *distribution* of a random variable summarizes what values it takes and with what probabilities — abstracted away from the underlying Ω, since one rarely needs to refer back to it. The *expectation* is the average value, weighted by probability; for a continuous random variable it is the Lebesgue integral against the probability measure.

These three objects — random variable, distribution, expectation — plus their conditional analogues are what the rest of probability theory is about, and they are what every applied political-science model uses. A polling estimator is a random variable; its expectation is what's being estimated; its variance is what determines the confidence interval. Expected utility under a lottery is the expectation of the utility random variable. A posterior belief is a conditional distribution, and an updated estimate is a conditional expectation. The framework is general and the same vocabulary serves the discrete and continuous cases in one breath.

This handout develops the basic machinery. Section 2 defines random variables as measurable functions; §3 introduces distributions and the cumulative distribution function (the standard concrete handle on a distribution); §4 develops the Lebesgue integral and expectation; §5 works through the standard inequalities (Markov, Chebyshev, Jensen, Cauchy–Schwarz) and conditional expectation. The next handout uses these objects for the law of large numbers, the central limit theorem, and the various modes of convergence.

2 Random variables

When a political scientist writes "let X be the number of voters supporting the incumbent" or "let θ be the politician's true type" or "let ϵ be the demand shock," a random variable is being introduced: a real-valued quantity that depends on which random outcome obtains. The formal apparatus that makes "depends on" precise is the notion of a *measurable function*: pre-images of events on the $\mathbb{R}$ side should be events on the Ω side, so that the question "what is the probability that X falls in this range?" has a definite answer. Without measurability, the joint use of the probability space and the real line would not connect, and the question "what is $\mathbb{P}(X \le x)$?" would be incoherent rather than easy.

Definition 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \in \mathcal{B}(\mathbb{R})$,

$$X^{-1}(B) := \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

We say such a function is *$\mathcal{F}$-measurable* (or just *measurable*, when the σ-algebra is understood).

The condition is equivalent — and in practice usually checked — in the following form.

Proposition 2. *$X : \Omega \to \mathbb{R}$ is a random variable iff for every $x \in \mathbb{R}$, the set $\{X \le x\} := \{\omega : X(\omega) \le x\}$ is in $\mathcal{F}$.*

Proof sketch. ($\Rightarrow$) The half-line $(-\infty, x]$ is Borel. ($\Leftarrow$) The collection of Borel sets B such that $X^{-1}(B) \in \mathcal{F}$ is a σ-algebra (verify the three closure properties). It contains the half-lines, which generate $\mathcal{B}(\mathbb{R})$, so it equals $\mathcal{B}(\mathbb{R})$. □

Example 3 (Indicator and simple random variables)**.** For $A \in \mathcal{F}$, the *indicator* $\mathbf{1}_A : \Omega \to \{0, 1\}$ defined by $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise is a random variable. A *simple* random variable is a finite linear combination $\sum_{i=1}^{k} c_i \mathbf{1}_{A_i}$ for $c_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$ — equivalently, a measurable function taking only finitely many values.

Example 4 (A vote count)**.** Let Ω be the sample space of an electorate of n voters with each voter independently voting for L with probability p, as in the previous handout. The function $X(\omega) =$ "number of voters in ω who vote for L" is a random variable: $X = \sum_{i=1}^{n} \mathbf{1}_{\{\text{voter } i \text{ votes for } L\}}$, a finite sum of indicators.

Example 5 (A continuous voter type)**.** Let $\Omega = [0, 1]$ with the Borel σ-algebra and Lebesgue measure (uniform on $[0, 1]$). Then the identity function $X(\omega) = \omega$ is a random variable, called the *uniform random variable on* $[0, 1]$. So is any continuous function of ω — e.g., $Y(\omega) = -\log(1 - \omega)$, which has the standard exponential distribution.

The class of random variables is closed under the operations one would hope.

Proposition 6. *If X, Y are random variables and $c \in \mathbb{R}$, then cX, $X + Y$, XY, $\max(X, Y)$, $\min(X, Y)$, and $|X|$ are random variables. If $g : \mathbb{R} \to \mathbb{R}$ is Borel-measurable (in particular, if g is continuous), then $g \circ X$ is a random variable. The pointwise limit $\lim_{n\to\infty} X_n$ of a sequence of random variables, when it exists, is a random variable.*

The closure under pointwise limits is the substantive property and is what makes the limit theorems of the next handout possible. (The Riemann-integrable functions are not closed under pointwise limits, which is one reason measure theory was developed.)

3 Distributions

When we report on a random quantity in a model, we usually want to summarize how it behaves without dragging the underlying probability space into the discussion. "The number of voters supporting the incumbent is binomially distributed with parameters n and p." "The candidate's true ability is normally distributed with mean μ and variance σ^2." "Voter turnout in this district has a beta distribution." Each of these is a statement about the *distribution* of the random variable — the probability measure that X induces on $\mathbb{R}$ — and the distribution carries the substantive content even when the probability space Ω is left unspecified. This section formalizes the distribution and the standard concrete handle on it, the cumulative distribution function.

Definition 7. The *distribution* (or *law*) of a random variable X is the probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

This is also called the *pushforward* of $\mathbb{P}$ under X, written $X_*\mathbb{P}$ or $\mathbb{P} \circ X^{-1}$.

Verifying $\mathbb{P}_X$ is a probability measure: $\mathbb{P}_X(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = \mathbb{P}(\Omega) = 1$; non-negativity is inherited; countable additivity follows because X^{-1} commutes with disjoint unions.

The standard concrete handle on a distribution is the *cumulative distribution function*:

Definition 8. The *cumulative distribution function* (CDF) of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by $F_X(x) = \mathbb{P}(X \leq x)$.

The CDF determines the distribution: every probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ corresponds to exactly one CDF, and vice versa. (One direction: given the measure, the CDF reads off as $F_X(x) = \mathbb{P}_X((-\infty, x])$. The other direction uses Carathéodory extension: a function on $\mathbb{R}$ that is monotone non-decreasing, right-continuous, with limits 0 at $-\infty$ and 1 at $+\infty$ extends uniquely to a probability measure on the Borel σ-algebra, by the same machinery used for Lebesgue measure.)

Two important special cases of distributions.

Discrete random variables. If X takes countably many values $x_1, x_2, \ldots$, then $\mathbb{P}_X = \sum_n p_X(x_n)\delta_{x_n}$, where $p_X(x) := \mathbb{P}(X = x)$ is the *probability mass function* (PMF). The PMF is non-negative, sums to 1 over the support, and determines $\mathbb{P}_X$. Standard examples: Bernoulli(p) (the indicator of an event of probability p); Binomial(n, p) (sum of n independent Bernoulli(p); see Example 4); Poisson(λ).

Absolutely continuous random variables. If there exists a non-negative integrable function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{for every } x \in \mathbb{R},$$

we say X is *absolutely continuous* and call f_X its *probability density function* (PDF). For absolutely continuous X, $\mathbb{P}(X \in B) = \int_B f_X(t)\, dt$ for every Borel B. Standard examples: Uniform(a, b) (density $1/(b - a)$ on $[a, b]$); Normal(μ, σ^2) (the Gaussian); Exponential(λ).[1]

[1] The existence of a PDF for $\mathbb{P}_X$ is the condition that $\mathbb{P}_X$ is *absolutely continuous* with respect to Lebesgue measure ($\lambda(B) = 0 \Rightarrow \mathbb{P}_X(B) = 0$). The corresponding PDF is the *Radon–Nikodym derivative* $d\mathbb{P}_X/d\lambda$, whose existence is the content of the Radon–Nikodym theorem (a substantive measure-theoretic result; see Folland (1999) or Billingsley (1995)). Distributions on $\mathbb{R}$ partition into three pieces by Lebesgue's decomposition theorem: a discrete part (atoms with positive mass), an absolutely continuous part (with a PDF), and a singular continuous part (atomless, but supported on a Lebesgue-null set — the Cantor distribution is the canonical example). Most political-economy models use distributions of the first two types only; the singular case is theoretically real but rarely encountered in practice.

4 Expectation

The expected vote share of an incumbent. The expected utility of a lottery. The expected number of qualifying respondents in a survey. The expected payoff to a candidate at a given platform. Whenever a political-economy model summarizes a random quantity by a single number, that number is almost always an *expectation* of some random variable: the average value of the variable, weighted by probability. For a discrete random variable the formula is the familiar weighted sum $\sum_n x_n p_X(x_n)$. For a continuous random variable it is an integral against the density, $\int x f_X(x)\,dx$. The unified definition that handles both cases — and supplies a rigorous foundation under each — is the Lebesgue integral against the underlying probability measure, built in three stages from simple functions upward.

We build the integral in three stages, as in the standard development of Lebesgue integration.

Definition 9. For a simple random variable $X = \sum_{i=1}^{k} c_i \mathbf{1}_{A_i}$ (with the A_i pairwise disjoint), the *expectation* is

$$\mathbb{E}[X] = \sum_{i=1}^{k} c_i \mathbb{P}(A_i).$$

For a non-negative random variable $X \geq 0$, $\mathbb{E}[X] := \sup\{\mathbb{E}[Y] : 0 \leq Y \leq X, Y \text{ simple}\}$, with values in $[0, \infty]$. For a general random variable X, write $X = X^+ - X^-$ with $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$ (both non-negative), and set $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$, provided at least one of $\mathbb{E}[X^+], \mathbb{E}[X^-]$ is finite. When $\mathbb{E}[|X|] < \infty$ we say X is *integrable.*

The notation $\int_\Omega X\,d\mathbb{P}$ is also used for $\mathbb{E}[X]$. The integral is constructed once and then specialized: for X discrete, $\mathbb{E}[X] = \sum_x x p_X(x)$; for X absolutely continuous, $\mathbb{E}[X] = \int_\mathbb{R} x f_X(x) dx$ (where now the right-hand side is a Riemann integral coinciding with the Lebesgue integral for nice f_X). The change-of-variable identity is the connection:

$$\mathbb{E}[g(X)] = \int_\Omega g(X(\omega))\,d\mathbb{P}(\omega) = \int_\mathbb{R} g(x)\,d\mathbb{P}_X(x),$$

which lets us forget about Ω and work with the distribution alone.

Theorem 10 (Properties of expectation)**.** *For integrable random variables X, Y and constants $a, b \in \mathbb{R}$:*

1. *(Linearity)* $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.
2. *(Monotonicity) If $X \leq Y$ almost surely, then* $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
3. *(Triangle)* $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$.
4. *($\mathbb{E}$ of a constant)* $\mathbb{E}[c] = c$.

The dominated convergence theorem and monotone convergence theorem (Lebesgue) say when $\mathbb{E}[\lim X_n] = \lim \mathbb{E}[X_n]$ for sequences X_n converging pointwise — the canonical limit-versus-integral results of measure-theoretic probability. We will use them in the next handout for limit theorems.

Definition 11. The *variance* of an integrable random variable with $\mathbb{E}[X^2] < \infty$ is

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

The *standard deviation* is $\sigma_X := \sqrt{\mathrm{Var}(X)}$. For two such random variables, the *covariance* is $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Example 12 (A polling estimator)**.** Suppose we sample n voters independently from a large electorate, and let X_i be 1 if voter i supports L, and 0 otherwise, with $\mathbb{E}[X_i] = p$ (the population vote share). The polling estimator $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ is a random variable with $\mathbb{E}[\bar{X}_n] = p$ and (by independence) $\mathrm{Var}(\bar{X}_n) = p(1-p)/n$. The decay of $\mathrm{Var}(\bar{X}_n)$ as $n \to \infty$ is the substantive content of polling: more voters in the sample produce a tighter estimator. The next handout's law of large numbers makes this precise as a convergence theorem.

5 Inequalities and conditional expectation

Probabilistic arguments in political economy frequently want to bound a tail probability — "how likely is it that the polling estimator misses the true vote share by more than five percentage points?" — or to compare an expected outcome to its certain counterpart — "is the expected utility of a fair lottery higher or lower than the utility of its expected value?" The standard inequalities below (Markov, Chebyshev, Jensen, Cauchy–Schwarz) handle these and many more. They are short to state, easy to prove, and assume nothing about the distribution beyond mild integrability. We close the section with conditional expectation, the formal version of "best forecast given the available information," which is the working object of every Bayesian-updating argument and every dynamic-programming argument that involves uncertainty.

Theorem 13 (Markov's inequality)**.** *For any non-negative random variable X and $c > 0$,*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[X]}{c}.$$

Proof. $\mathbb{E}[X] \geq \mathbb{E}[X\mathbf{1}_{\{X \geq c\}}] \geq \mathbb{E}[c\mathbf{1}_{\{X \geq c\}}] = c\mathbb{P}(X \geq c)$. □

Theorem 14 (Chebyshev's inequality)**.** *For an integrable random variable X with $\mathbb{E}[X^2] < \infty$ and $c > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

Proof. Apply Markov to $(X - \mathbb{E}[X])^2 \geq 0$ with threshold c^2: $\mathbb{P}((X - \mathbb{E}[X])^2 \geq c^2) \leq \mathbb{E}[(X - \mathbb{E}[X])^2]/c^2$. □

Theorem 15 (Jensen's inequality)**.** *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex and X an integrable random variable. Then*

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)],$$

provided $\mathbb{E}[\varphi(X)]$ exists.

Proof sketch. A convex φ admits a supporting line at every point: there is $a, b \in \mathbb{R}$ with $\varphi(\mathbb{E}[X]) = a\mathbb{E}[X] + b$ and $\varphi(x) \geq ax + b$ for every x. Take expectations of the inequality. □

The Jensen inequality is the formal version of "a risk-averse agent prefers the expected outcome to the gamble." For $\varphi(x) = x^2$, $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$, which is just the non-negativity of variance; for φ concave (e.g., utility functions of risk-averse agents in expected-utility theory), the inequality reverses: $\mathbb{E}[u(X)] \leq u(\mathbb{E}[X])$.

Theorem 16 (Cauchy–Schwarz)**.** *For random variables X, Y with $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$,*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]}.$$

Equivalently, $|\operatorname{Cov}(X,Y)| \leq \sigma_X \sigma_Y$*, so the* correlation $\rho_{X,Y} := \operatorname{Cov}(X,Y)/(\sigma_X \sigma_Y) \in [-1, 1]$.

We turn finally to conditional expectation. The intuition is the natural one: given a σ-sub-algebra $\mathcal{G} \subseteq \mathcal{F}$ representing "the information available," the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ is the best forecast of X using only that information.

Definition 17. Let X be an integrable random variable and $\mathcal{G} \subseteq \mathcal{F}$ a σ-sub-algebra. The *conditional expectation of X given $\mathcal{G}$*, written $\mathbb{E}[X \mid \mathcal{G}]$, is any random variable Y such that:

1. Y is $\mathcal{G}$-measurable;
2. for every $A \in \mathcal{G}$, $\mathbb{E}[Y \mathbf{1}_A] = \mathbb{E}[X \mathbf{1}_A]$.

Such a Y exists and is unique up to almost-sure equality, by the Radon–Nikodym theorem.

The two requirements together pin down $\mathbb{E}[X \mid \mathcal{G}]$ uniquely. The first says the forecast must be measurable with respect to the available information (no peeking at things you don't see). The second says the forecast must agree with X on average, on every $\mathcal{G}$-event — a kind of unbiasedness that is calibrated to $\mathcal{G}$.

Example 18 (Conditional expectation as a function)**.** If $\mathcal{G} = \sigma(Y)$ is generated by another random variable Y, then $\mathbb{E}[X \mid \mathcal{G}]$ is a function of Y — usually written $\mathbb{E}[X \mid Y]$. For discrete Y, $\mathbb{E}[X \mid Y = y] = \sum_x x\, \mathbb{P}(X = x \mid Y = y)$, the elementary conditional expectation; for jointly continuous (X, Y) with joint density $f_{X,Y}$, $\mathbb{E}[X \mid Y = y] = \int x\, f_{X|Y}(x \mid y)\, dx$ where $f_{X|Y}(x \mid y) = f_{X,Y}(x, y)/f_Y(y)$.

The basic properties: $\mathbb{E}[X \mid \mathcal{G}]$ is linear in X; takes constants out ($\mathbb{E}[c \mid \mathcal{G}] = c$); and satisfies the *tower property* $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = \mathbb{E}[X \mid \mathcal{G}_2]$ when $\mathcal{G}_2 \subseteq \mathcal{G}_1$ (iterated expectations agree with the smaller information set). The tower property is the workhorse for arguments involving sequential information revelation, including dynamic programming, martingale theory, and Bayesian updating in signaling games.

6 What's next

The next handout brings convergence into the picture. With random variables and expectations in hand, we can ask: when does a sequence of random variables "settle down," and in what sense? The four standard answers — almost-sure convergence, convergence in probability, convergence in L^p, and convergence in distribution — are different sufficient conditions for "the sample average behaves like the population average," and the relationships among them are the structure of the convergence story. The headline theorems — the law of large numbers, the central limit theorem, and Borel–Cantelli — live there.

A piece we mention but do not develop here is *random vectors*: a random variable taking values in $\mathbb{R}^n$, which is just a tuple $(X_1, \dots, X_n)$ of real-valued random variables on the same probability space. Joint distributions, marginal distributions, and joint expectations work as one would expect, and the theorems above carry over to the multidimensional setting with no surprises. Independence of random variables — the natural extension of independence of events — means joint distribution factors as a product of marginals, and is what makes the LLN and CLT in the next handout work for sums of independent random variables.

For broader treatments, Billingsley (1995), Durrett (2019), and Williams (1991) all develop random variables, expectations, and inequalities at this level; Folland (1999) for the underlying integration theory.

7 Exercises

Exercise 19. Let $X : \Omega \to \mathbb{R}$ be a function. Show that the collection $\{B \in \mathcal{B}(\mathbb{R}) : X^{-1}(B) \in \mathcal{F}\}$ is a σ-algebra. Use this to prove Proposition 2.

Exercise 20. Show that if X and Y are random variables, then $X + Y$ is a random variable. (*Hint*: $\{X + Y \le z\} = \bigcup_{q \in \mathbb{Q}} (\{X \le q\} \cap \{Y \le z - q\})$.)

Exercise 21. Prove that $\mathbb{E}[X^2] \ge (\mathbb{E}[X])^2$ from the definition of variance and the fact that $\operatorname{Var}(X) \ge 0$. Conclude that $\sigma_X \ge |\mathbb{E}[X]|$ when $\mathbb{E}[X]$ has the same sign as X.

Exercise 22 (Vote count)**.** Continuing Example 4: with n independent voters each voting for L with probability p, let $X = \sum_{i=1}^n \mathbf{1}_{\{\text{voter } i \text{ for } L\}}$ be the total vote count for L. Compute $\mathbb{E}[X]$ and $\operatorname{Var}(X)$ from the formulas $X = \sum X_i$, the linearity of expectation, and the fact that $\operatorname{Var}(\sum X_i) = \sum \operatorname{Var}(X_i)$ for independent X_i.

Exercise 23 (CDF of a uniform random variable)**.** Let X be uniform on $[0, 1]$ (Example 5). Compute the CDF $F_X(x)$. Now let $Y = X^2$. Compute the CDF $F_Y(y)$ and the PDF $f_Y(y)$ for $y \in [0, 1]$. Identify the distribution of Y as a member of a standard family.

Exercise 24 (Polling and Chebyshev)**.** Continuing Example 12: the polling estimator $\bar{X}_n$ has $\mathbb{E}[\bar{X}_n] = p$ and $\operatorname{Var}(\bar{X}_n) = p(1-p)/n$. Apply Chebyshev's inequality to bound $\mathbb{P}(|\bar{X}_n - p| \ge 0.02)$ in terms of n and p. With $p = 0.5$ (the worst case for polling variance), how large must n be for the bound to be at most 0.05? Compare with the more familiar normal-approximation bound (which we will derive from the CLT in the next handout).

Exercise 25 (Jensen and risk aversion)**.** Let $u : \mathbb{R} \to \mathbb{R}$ be a strictly concave utility function ($u'' < 0$ on the interior of its domain), and let W be a non-degenerate random variable representing wealth. Show that $\mathbb{E}[u(W)] < u(\mathbb{E}[W])$ — i.e., the agent strictly prefers the expected wealth $\mathbb{E}[W]$ for sure to the random W. This is the foundational inequality of expected-utility theory, identifying "risk averse" with "concave utility."

Exercise 26. Show that for any random variables X, Y on the same probability space with finite second moments, $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X, Y) + \operatorname{Var}(Y)$. Deduce that X and Y are uncorrelated ($\operatorname{Cov} = 0$) iff $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$. Argue that independence implies uncorrelatedness, and find a counter-example to the converse.

Exercise 27 (Bayesian updating revisited)**.** Returning to the signaling setting of the previous handout: the politician's type T is G (good) or B (bad), with prior $\mathbb{P}(T = G) = 1/2$. The signal S is H or L, with $\mathbb{P}(S = H \mid T = G) = 0.8$ and $\mathbb{P}(S = H \mid T = B) = 0.3$. Let $X = \mathbf{1}_{\{T=G\}}$ be the indicator of type G. Compute $\mathbb{E}[X]$ (the prior probability of G). Compute $\mathbb{E}[X \mid S = H]$ and $\mathbb{E}[X \mid S = L]$ (the posterior probabilities). Verify the tower property: $\mathbb{E}[\mathbb{E}[X \mid S]] = \mathbb{E}[X]$ — the unconditional expectation is the average of the conditional expectations weighted by the signal's distribution.

Exercise 28. Prove the tower property of conditional expectation: if $\mathcal{G}_2 \subseteq \mathcal{G}_1 \subseteq \mathcal{F}$, then $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = \mathbb{E}[X \mid \mathcal{G}_2]$ almost surely. (*Hint*: verify the two defining conditions for $\mathbb{E}[X \mid \mathcal{G}_2]$. Measurability with respect to $\mathcal{G}_2$ follows because $\mathbb{E}[Y \mid \mathcal{G}_2]$ is $\mathcal{G}_2$-measurable for any Y. The integration condition follows because $\mathcal{G}_2 \subseteq \mathcal{G}_1$, so any $A \in \mathcal{G}_2$ is also in $\mathcal{G}_1$.)

References

Billingsley, Patrick (1995). *Probability and Measure*. 3rd ed. New York: Wiley.

Durrett, Rick (2019). *Probability: Theory and Examples*. 5th ed. Cambridge: Cambridge University Press.

Folland, Gerald B. (1999). *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. New York: Wiley.

Williams, David (1991). *Probability with Martingales*. Cambridge: Cambridge University Press.