Math toolkit: probability

A collection of definitions, theorems, and formulae

March 2, 2026

Contents

1 Random variables

1.1 Discrete Distributions

If Ω is finite or countably infinite, say $\Omega = \{\omega_1, \omega_2, \dots\}$, then we have a *discrete probability space* and a *discrete distribution*. Let $X(\omega_j) = x_j$ and $p_j = \mathbb{P}(X = x_j), \quad j = 0, 1, \dots$ with $0 \le p_j \le 1, \quad \sum_j p_j = 1$.

Given a function f of X, its *expectation* is the following(provided the sum is well-defined):

$$\mathbb{E}f(X) = \sum_j f(x_j)\, p_j,$$

The p-th *moment* of the distribution is

$$m_p = \sum_j x_j^p\, p_j.$$

The *variance* of X is

$$\mathrm{Var}(X) = m_2 - m_1^2 = \sum_j (x_j - m_1)^2\, p_j.$$

1.2 Normal distribution

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then the PDF of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

and the CDF

$$\Phi(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(s-\mu)^2}{2\sigma^2}}\, ds$$

Properties of the CDF

- Limits: $\lim_{x\to\infty} \Phi(x) = 1$; $\lim_{x\to-\infty} \Phi(x) = 0$
- Symmetry: $\Phi(-x) = 1 - \Phi(x)$.
- Standardization: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{P}(X \le x) = \Phi\big((x-\mu)/\sigma\big)$.
- Not elementary in closed form; expressed via the error function `erf`.
- Φ is increasing so Φ^{-1} is decreasing

The MGF is

$$M_X(t) = \mathbb{E}\Big[e^{tX}\Big] = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$$

1.3 Bernoulli Distribution

The Bernoulli (binary) distribution is defined as

$$\mathbb{P}(X = j) = \begin{cases} p, & j = 1, \\ q, & j = 0, \end{cases}$$

where $p + q = 1$ and $p, q \geq 0$.

$$\mathbb{E}X = p, \quad \text{Var}(X) = pq.$$

The MGF is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = q + pe^t$$

1.4 Binomial Distribution

A discrete random variable X follows a *Binomial distribution* with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$, written

$$X \sim \text{Binomial}(n, p),$$

if X represents the number of successes in n independent Bernoulli trials, each with success probability p. The pmf is

$$\Pr(X = k) \;=\; \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

with

$$\mathbb{E}[X] = np, \quad \text{Var}(X) = npq.$$

and the moment generating function is

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \left(q + pe^t\right)^n.$$

Remarks:

- Models the count of successes in a fixed number of independent trials.
- As $n \to \infty$ and $p \to 0$ with $\lambda = np$ fixed, the Binomial distribution converges to Poisson(λ).
- For large n, by the Central Limit Theorem, X is approximately normal with mean np and variance npq.

1.5 Multinomial distribution

The multinomial distribution generalizes the binomial distribution over k categories. In particular, if we have a k sided die rolled n times, then a multinomial distribution models the probability of any particular combination of rolls across the k categories appears.

$$\vec{X} = (X_1, \ldots, X_k) \sim \text{Multinomial}(n, \mathbf{p})$$

where the parameter

$$\mathbf{p} = (p_1, \ldots, p_k), \qquad p_i \geq 0, \qquad \sum_{i=1}^{k} p_i = 1.$$

is interpreted as the probability of each of the k categories on each of the n trials. The pdf of multinomial$(n, \mathbf{p})$ is

$$\begin{aligned} \Pr(X = x) &= \binom{n}{x_1, \ldots, x_n} \prod_{i=1}^{k} p_i^{x_i} \\ &= \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \end{aligned}$$

where

$$x_i \in \mathbb{N}_0, \qquad \sum_{i=1}^{k} x_i = n.$$

The moments are

$$\mathbb{E}[X_i] = np_i, \qquad \mathrm{Var}(X_i) = np_i(1-p_i), \qquad \mathrm{Cov}(X_i, X_j) = -np_ip_j \ (i \neq j),$$

$$\mathrm{Cov}(X) = n\big(\,\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top\big).$$

The generating function and moment generating function is

$$G(\mathbf{t}) = \mathbb{E}\left[\prod_{i=1}^{k} t_i^{X_i}\right] = \Big(\sum_{i=1}^{k} p_i t_i\Big)^n, \qquad M(\mathbf{t}) = \mathbb{E}\Big[e^{\mathbf{t}^\top X}\Big] = \Big(\sum_{i=1}^{k} p_i e^{t_i}\Big)^n.$$

Common facts

- X counts outcomes from n i.i.d. categorical($\mathbf{p}$) trials; $\sum_i X_i = n$ a.s.
- Marginally, $X_i \sim \mathrm{Binomial}(n, p_i)$; conditionally on X_k, the remainder is $\mathrm{Multinomial}\Big(n - X_k,\ \frac{\mathbf{p}_{-k}}{1-p_k}\Big)$.
- As $n \to \infty$, $X \approx \mathcal{N}\big(n\mathbf{p},\ n(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)\big)$ (multivariate CLT).
- Conjugacy: if $\mathbf{p} \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$ and $X \mid \mathbf{p}$ is multinomial, then $\mathbf{p} \mid X \sim \mathrm{Dirichlet}(\boldsymbol{\alpha} + X)$.

1.6 Poisson distribution

The Poisson distribution $\mathcal{P}(\lambda)$ expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known constant average rate and independently of the time since the last event

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, \ldots$$

Here, λ is the average rate (floods per 100 years). Mean and variance:

$$\mathbb{E}X = \mathrm{Var}(X) = \lambda.$$

Poisson is the limit of binomial: If $n \to \infty$, $p \to 0$, and $np = \lambda$, then

$$\binom{n}{k} p^k q^{n-k} \ \longrightarrow\ \frac{\lambda^k}{k!} e^{-\lambda}.$$

Random Vectors. If $X = (X_1, X_2, \ldots, X_d)^T \in \mathbb{R}^d$ is a random vector, then the *covariance matrix* of X is

$$\mathrm{Cov}(X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T.$$

2 Conjugacy of random variabes

2.1 Gaussian ↔ chi-squared

Let $Z \in \mathbb{R}^k$ be a centered *and* normalized k-dimensional gaussian $Z \sim \mathcal{N}(0, I_k)$. Then

$$\|Z\|_2^2 = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2.$$

More generally, if $X \in \mathbb{R}^k$ is an arbitrary k-dimensional gaussian $X \sim \mathcal{N}(\mu, \Sigma)$ with covariance $\Sigma \succ 0$, then the squared Mahalanobis distance

$$D^2 := \|\Sigma^{-1/2}(X - \mu)\| = (X - \mu)^\top \Sigma^{-1}(X - \mu) \sim \chi_k^2.$$

3 Tools

3.1 Jensen's inequality for random variables

If X a random variable, then Jensen's inequality says that for any convex φ

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

"The function of the expectation is less than the expectation of the function"

3.2 Integration over Product Spaces

We collect here some basic manipulations of integrals on products of measure spaces.

(1) **Start with the integral over $X \times Y$:**

$$\int_{X \times Y} f(x, y)\, \mathrm{d}(x, y).$$

(2) **Express the integral as an iterated integral:**

$$\int_Y \Big(\int_X f(x, y)\, \mathrm{d}x \Big)\, \mathrm{d}y.$$

(3) **Since Y is finite, replace the integral over Y with a sum:**

$$\sum_{y \in Y} \int_X f(x, y)\, \mathrm{d}x.$$

(4) **Condition on each value of y:**

$$\int_X f(x, y_1)\, \mathrm{d}x \;+\; \int_X f(x, y_2)\, \mathrm{d}x \quad (\text{if } Y = \{y_1, y_2\}).$$

3.3 Fubini's Theorem (Simplified Statement):

If $f(x,y)$ is a function defined on the product of two measure spaces X and Y, and if f is integrable over $X \times Y$ (i.e. $\int_{X\times Y} |f(x,y)|\, d(x,y) < \infty$), then

$$\int_{X\times Y} f(x,y)\,\mathrm{d}(x,y) \;=\; \int_X \Big(\int_Y f(x,y)\,\mathrm{d}y\Big)\,\mathrm{d}x \;=\; \int_Y \Big(\int_X f(x,y)\,\mathrm{d}x\Big)\,\mathrm{d}y.$$

3.4 Probability as the Expectation of an Indicator Function

Identity. For any event A,

$$\Pr(A) \;=\; \mathbb{E}[\mathbf{1}_A],$$

where $\mathbf{1}_A$ is the indicator function of A:

$$\mathbf{1}_A(x) \;=\; \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

This identity states that the probability of A equals the expected value (mean) of its indicator function, bridging probability and expectation.

3.5 Expectation as an Integral over the Joint Distribution

Identity. For any measurable function $g(x,y) : X \times Y \to \mathbb{R}$,

$$\mathbb{E}[g(x,y)] \;=\; \int_X\!\int_Y g(x,y)\,p(x,y)\,\mathrm{d}y\,\mathrm{d}x,$$

where $p(x,y)$ is the joint probability density function of x and y.
This expresses the expectation of a function of two random variables as a double integral over their joint pdf.

3.6 Decomposition of the Joint Probability Density Function

Identity.

$$p(x,y) \;=\; p(y \mid x)\,p(x),$$

where

- $p(x,y)$ is the joint pdf of x and y,
- $p(y \mid x)$ is the conditional pdf of y given x,
- $p(x)$ is the marginal pdf of x.

This decomposes the joint pdf into the product of a conditional pdf and a marginal pdf, enabling separation of joint distributions into conditional and marginal components.

3.7 Integral of an Indicator Function Times a Conditional Density

Let X and Y be random variables over event space $\mathcal{X}$ and $\mathcal{Y}$, respectively. Define the conditional density $p(y \mid x)$. For any set-valued event $A(x) \subseteq \mathcal{Y}$ that may depend on x, define the indicator function

$$\mathbf{1}_{A(x)}(y) \;=\; \begin{cases} 1, & y \in A(x), \\ 0, & y \notin A(x). \end{cases}$$

Then

$$\int_{\mathcal{Y}} \mathbf{1}_{A(x)}(y)\,p(y \mid x)\,\mathrm{d}y \;=\; \int_{\{\,y : y\in A(x)\}} p(y \mid x)\,\mathrm{d}y \;=\; \Pr\big(Y \in A(x) \mid X = x\big).$$

Equivalently, you may write

$$\boxed{\int_{A(x)} p(y \mid x)\,\mathrm{d}y \;=\; \Pr(Y \in A(x) \mid X = x).}$$

3.8 Law of Total Probability for Expectations

FIX THIS **Identity.** For any measurable function $g(x, y)$,

$$\mathbb{E}[g(x,y)] \;=\; \int_X \mathbb{E}[g(x,y) \mid x]\, p(x)\,\mathrm{d}x,$$

where

$$\mathbb{E}[g(x,y) \mid x] \;=\; \int_Y g(x,y)\, p(y\,|\,x)\,\mathrm{d}y.$$

This expresses the overall expectation as an integral over x of the conditional expectation $\mathbb{E}[g(x, y) \mid x]$ weighted by the marginal pdf $p(x)$, and is a form of the law of total expectation.

3.9 Law of total expectation (tower law)

If X is a random variable and Y is any random variable on the same probability space, then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$$

3.10 Law of total variance

If X and Y are random variables on the same probability space, and X has finite variance, then

$$\mathrm{Var}(X) \;=\; \mathbb{E}\big[\mathrm{Var}(X \mid Y)\big] \;+\; \mathrm{Var}\big(\mathbb{E}[X \mid Y]\big).$$

Example: Let X be a coin flip taking values Heads with probability h and Tails with probability $1 - h$. Given Heads, $Y \sim \mathcal{N}(\mu_h, \sigma_h^2)$; given Tails, $Y \sim \mathcal{N}(\mu_t, \sigma_t^2)$. Then

$$\mathbb{E}[\mathrm{Var}(Y \,|\, X)] = h\,\sigma_h^2 + (1-h)\,\sigma_t^2, \qquad \mathrm{Var}(\mathbb{E}[Y \,|\, X]) = h(1-h)(\mu_h - \mu_t)^2,$$

so

$$\mathrm{Var}(Y) = h\,\sigma_h^2 + (1-h)\,\sigma_t^2 + h(1-h)(\mu_h - \mu_t)^2.$$

If X and Y are originally given on different spaces, you can place them on a common space by constructing a coupling that realizes their joint law; once a joint distribution is specified, the identity holds, which is what we do above.

3.11 Conditional Expectation

$$\mathbb{E}[Y \mid X = x] \;=\; \int_Y y\, p(y\,|\,x)\,\mathrm{d}y.$$

The conditional expectation of Y given $X = x$ is the expected value of Y under the conditional pdf $p(y \mid x)$, representing the best prediction of Y based on the knowledge that $X = x$. More generally, the conditional expecation is a function of X via $g(X) = \mathbb{E}[Y \mid X = x]$

3.12 Variance Decomposition (Bias–Variance Decomposition)

Identity:

$$\mathbb{E}\big[(f(x) - y)^2\big] \;=\; \mathbb{E}\big[(f(x) - f_p(x))^2\big] \;+\; \mathbb{E}\big[(f_p(x) - y)^2\big] \;+\; 2\,\mathbb{E}\big[(f(x) - f_p(x))\,(f_p(x) - y)\big].$$

Since by Identity 13 the cross-term $\mathbb{E}[(f(x) - f_p(x))(f_p(x) - y)] = 0$, we obtain the usual bias–variance split:

$$\mathbb{E}[(f(x) - y)^2] \;=\; \mathbb{E}[(f(x) - f_p(x))^2] \;+\; \mathbb{E}[(f_p(x) - y)^2].$$

Explanation. This shows that the total expected squared error decomposes into *variance* (the error due to variability of $f(x)$ around the predictor $f_p(x)$) and the *irreducible error* (the noise inherent in y).

3.13 Bounding an Integral by the Square Root of the Expected Value

Identity.

$$\int_X |g(x)|\, dp(x) \;\le\; \sqrt{\mathbb{E}[\,|g(x)|^2\,]}.$$

Explanation. When p is a probability measure (or more generally on a finite-measure space), the L^1 norm of g is bounded by its L^2 norm, since $\|g\|_1 \le \|g\|_2$ under normalization.

3.14 Chain rule of probability

The *chain rule of probability* states that for any random variables A, B and any conditioning information C,

$$p(A, B \mid C) \;=\; p(A \mid C)\, p(B \mid A, C) \;=\; p(B \mid C)\, p(A \mid B, C).$$

3.15 Boole's Inequality

Statement. For any countable collection of events $A_1, A_2, \cdots \in \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \;\le\; \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Interpretation. The probability that at least one of the events A_n occurs is no greater than the sum of their individual probabilities. The simplist case is

$$\mathbb{P}(A \cup B) \le \mathbb{P}(A) + \mathbb{P}(B)$$

3.16 Conditional Probability and Multiplication Rule

Let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) \neq 0$. The *conditional probability* of A given B is defined as

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

From the definition, we have the **multiplication rule**:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\, \mathbb{P}(B).$$

More generally, for three events $A, B, C \in \mathcal{F}$ with $\mathbb{P}(C) \neq 0$ and $\mathbb{P}(B \cap C) \neq 0$, we have

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \mid B \cap C)\; \mathbb{P}(B \mid C)\; \mathbb{P}(C).$$

3.17 Bayes' Rule

From the definition of conditional probability, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(B)},$$

which is called **Bayes' rule**.

Let $A_1, A_2, \ldots$ be disjoint events such that

$$\bigcup_{j=1}^{\infty} A_j = \Omega.$$

Then, for any $j \in \mathbb{N}$,

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)}{\sum_{n=1}^{\infty} \mathbb{P}(A_n)\,\mathbb{P}(B \mid A_n)}.$$

Here:

- $\mathbb{P}(A_j)$ is the *prior probability* of hypothesis A_j.
- $\mathbb{P}(A_j \mid B)$ is the *posterior probability* of A_j given event B.

3.18 Add-subtract expectation trick

If A, B, C are random variables, then

$$\mathbb{E}[(A-B)^2] = \mathbb{E}[(A-C)^2] + \mathbb{E}[(C-B)^2] + 2\,\mathbb{E}[(A-C)(C-B)].$$

When the cross term vanishes,

$$\mathbb{E}[(A-B)^2] = \mathbb{E}[(A-C)^2] + \mathbb{E}[(C-B)^2].$$

The cross term vanishes if

- Constant-times-zero-mean. If one factor is constant and the other has mean zero.
- Independence and zero mean. If $(A-C)$ and $(C-B)$ are independent and at least one has mean zero (e.g. $\mathbb{E}[A-C]=0$), then

3.19 Hoeffding's inequality (distribution-free bound on sums)

Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$ almost surely, and let

$$S_n = \sum_{i=1}^{n} (X_i - \mathbb{E}X_i).$$

Then for all $t \geq 0$, Hoeffding's inequality says that

$$\mathbb{P}(S_n \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

1. Can use the union bound to get the event $|S_n| \geq t$ then just multiply by 2
2. Hoeffding gives a distribution-free bound: only the range $(b_i - a_i)$ matter.
3. The probability behaves like a sub-Gaussian since the tails behave like $\exp(-c\,t^2)$
4. Hoeffding ignores the actual variance; if the X_i have much smaller variance than their ranges suggest, Hoeffding can be loose.

3.20 Bernstein's inequality (variance-sensitive bound on sums)

Let $X_1, \dots, X_n$ be independent, mean-zero random variables with $|X_i| \le M$ almost surely, and define

$$V = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Then for all $t \ge 0$, Bernstein's inequality says that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge t \right) \le \exp\left(-\frac{t^2}{2(V + \frac{Mt}{3})} \right).$$

1. Can use the union bound to get the event $|S_n| \ge t$ then just multiply by 2
2. It interpolates between sub-Gaussian and sub-exponential tails:

$$t \ll \frac{V}{M} \;\Rightarrow\; \exp\left(-\frac{t^2}{2V} \right), \qquad t \gg \frac{V}{M} \;\Rightarrow\; \exp\left(-\frac{3t}{2M} \right).$$

3. It is variance-adaptive: when V is small relative to $M^2 n$, it can be much sharper than Hoeffding.
4. A result, obtained by inverting the one sided bound to obtain t_* then using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ and the union bound is that, with probability $1 - \delta$,

$$\sum_{i=1}^{n} X_i \lesssim \sqrt{2V \log(1/\delta)} + \frac{2M}{3} \log(1/\delta).$$

4 Properties of random variables

4.1 Moment generating functions

$$M_X(t) := \mathbb{E}\Big[e^{tX}\Big] \quad \text{(when the expectation is w.r.t } X\text{)}$$

$$M_X(0) = 1, \qquad M_X(t) = \sum_{k=0}^{\infty} \frac{\mu'_k}{k!}\, t^k \ , \quad \mu'_k = \mathbb{E}[X^k], \quad M_X^{(k)}(0) = \mu'_k.$$

$$M_{aX+b}(t) = e^{bt}\, M_X(at) \ , \qquad M_{X+Y}(t) = M_X(t) M_Y(t) \ \text{ if } X \perp Y \ .$$

4.2 Quantile function

Let $F(x) = \Pr(X \le x)$ be a cumulative distribution function (cdf). Then the quantile function is

$$Q_F(p) \ := \ \inf\{x \in \mathbb{R} : \ F(x) \ge p\}, \qquad p \in (0,1),$$

If F is continuous and strictly increasing, then $Q_F(p) = F^{-1}(p)$ on $(0,1)$. Recall that the CDF is always increasing, but not always strictly increasing.

Essential identities (general cdf).

- $F(Q_F(p)) \ge p$ and $Q_F(F(x)) \le x$.
- Monotonicity & continuity: Q_F is nondecreasing and left–continuous on $(0,1)$.

When F has a density f with $f(Q_F(p)) > 0$.

$$Q'_F(p) \ = \ \frac{1}{f(Q_F(p))} \qquad \text{(a.e. on } (0,1)\text{)}.$$

Transformations.

- Inverse-transform sampling: if $U \sim \mathrm{Unif}(0,1)$, then $Q_F(U) \sim F$.
- Location–scale equivariance: for $Y = a + bX$,

$$Q_Y(p) = \begin{cases} a + b\, Q_X(p), & b > 0, \\ a + b\, Q_X(1-p), & b < 0. \end{cases}$$

Median Median(s): any m with $F(m^-) \le \frac{1}{2} \le F(m)$ since we might not have a m such that $F(m) = \frac{1}{2}$. Or alternatively, any m such that

$$\mathbb{P}(X \ge m) \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \le m) \ge \frac{1}{2}$$

note that this does not require the distribution to be continuous. If we do have continuity, then we can upgrade to equality: a median is any $m \in \mathbb{R}$ such that

$$\mathbb{P}(X \le m) = \int_{-\infty}^{m} f(x)dx = \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \ge m) = \int_{m}^{\infty} f(x)dx = \frac{1}{2}$$

Named quantiles. Lower/upper quartiles: $Q_F(\frac{1}{4})$, $Q_F(\frac{3}{4})$ is "the smallest value m such a

random draw from our distribution has a $\frac{1}{4}, \frac{3}{4}$, etc, chance of exceeding m.

Empirical quantile for data $x_1, \dots, x_n$

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \le x\}, \qquad \widehat{Q}_n(p) = \inf\{x : \widehat{F}_n(x) \ge p\}.$$

4.3 Probability generating function trick

Let X be a nonnegative integer–valued r.v. with probability generating function $G_X(t) = \mathbb{E}[t^X]$.

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{X+1}\Big] &= \sum_{k\ge 0} \frac{1}{k+1}\mathbb{P}(X=k) \\
&= \sum_{k\ge 0} \mathbb{P}(X=k)\int_0^1 t^k\,dt \\
&= \int_0^1 \mathbb{E}[t^X]\,dt \\
&= \int_0^1 G_X(t)\,dt
\end{aligned}$$

Intuition: replace the reciprocal via $\frac{1}{k+1} = \int_0^1 t^k\,dt$ to turn a sum into an integral of the probability generating function trick.

Quick examples.

$$\begin{aligned}
&X \sim \text{Pois}(\lambda) && \Rightarrow \quad \mathbb{E}\Big[\tfrac{1}{X+1}\Big] = \frac{1-e^{-\lambda}}{\lambda}, \\
&X \sim \text{Bin}(n,p) && \Rightarrow \quad \mathbb{E}\Big[\tfrac{1}{X+1}\Big] = \frac{1-(1-p)^{n+1}}{p(n+1)}, \\
&X \sim \text{Geom}(p) \text{ (failures before 1st success)} && \Rightarrow \quad \mathbb{E}\Big[\tfrac{1}{X+1}\Big] = \frac{p}{1-p}\,\ln\frac{1}{p}.
\end{aligned}$$

Useful bound (Jensen). Since $f(x) = \frac{1}{1+x}$ is convex on $(-1,\infty)$,

$$\mathbb{E}\Big[\frac{1}{1+X}\Big] \;\ge\; \frac{1}{1+\mathbb{E}[X]}.$$

5 Variance, covariance, correlation, and related topics

5.1 Independence of random variables

Let X, Y be random variables on the same probability space. We say X and Y are *independent*, written $X \perp\!\!\!\perp Y$, if for all Borel sets $A, B \subseteq \mathbb{R}$,

$$\Pr(X \in A,\ Y \in B) = \Pr(X \in A)\Pr(Y \in B).$$

Equivalently, when expectations exist, $X \perp\!\!\!\perp Y$ is equivalent to

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

5.2 Covariance

For two random variables X and Y, the *covariance* is

$$\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

We say X and Y are *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$. Note that in general we have that

$$\mathrm{Cov}(X, Y) \in (-\infty, \infty)$$

but given variances and using Cauchy–Schwarz we can obtain

$$|\mathrm{Cov}(X, Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

5.3 Correlation

Let X, Y be real-valued random variables. Define covariance between X and Y as

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

and define correlation as

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1]$$

which is well defined when $\mathrm{Var}(X) > 0$ and $\mathrm{Var}(Y) > 0$.

5.4 Independence $\implies$ Cov=0 $\iff$ Corr=0

The following are true of the relation between independence, covariance, and correlation:

1. Independence implies zero correlation:

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad \mathrm{Cov}(X, Y) = 0 \quad \iff \quad \mathrm{Corr}(X, Y) = 0$$

   assuming that $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$.

Does zero covariance imply independence? In genergal, zero covariance *does not* imply independence :

$$\mathrm{Cov}(X, Y) = 0 \ \not\Rightarrow \ X \perp\!\!\!\perp Y.$$

Zero covariance only rules out *linear* association; dependence can remain through nonlinear

relationships.

Standard counterexample. Let $X \sim \text{Unif}(-1, 1)$ and define $Y = X^2$. Then Y is a deterministic function of X, so X and Y are not independent. However,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0 - 0 \cdot \mathbb{E}[X^2] = 0,$$

since $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^3] = 0$ by symmetry of the uniform distribution on $(-1, 1)$.

When does zero covariance imply independence? It *does* hold under additional structure, e.g. if (X, Y) is jointly Gaussian:

$$(X, Y) \text{ jointly normal and } \text{Cov}(X, Y) = 0 \quad \Rightarrow \quad X \perp\!\!\!\perp Y.$$

5.5 Confounding variables

Let (X, Y, Z) be random variables defined on a common probability space, where

1. X is a proposed *cause*,
2. Y is an *outcome*,
3. Z is a *confounder* (typically unobserved or uncontrolled).

Intuitively, Z influences both X and Y, creating an association between X and Y that need not reflect a direct causal effect of X on Y. A precise way to formalize this is via the *law of total covariance*, which states that

$$\text{Cov}(X, Y) = \mathbb{E}[\text{Cov}(X, Y \mid Z)] + \text{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[Y \mid Z]) .$$

where $\text{Cov}(X, Y \mid Z)$ is a random variable (e.g. a function of Z), defined pointwise by

$$\text{Cov}(X, Y \mid Z) := \mathbb{E}[(X - \mathbb{E}[X \mid Z])(Y - \mathbb{E}[Y \mid Z]) \mid Z] = \mathbb{E}[XY \mid Z] - \mathbb{E}[X \mid Z]\mathbb{E}[Y \mid Z].$$

In particular, conditioning on Z makes the quantity afunction of Z, so the expectation in the first term $\mathbb{E}[\text{Cov}(X, Y \mid Z)]$ is taken over Z.

Interpretation of confounding variables

1. $\mathbb{E}[\text{Cov}(X, Y \mid Z)]$ measures the *within-Z association*: the average association between X and Y after conditioning on (i.e. holding fixed) the confounder Z.
2. $\text{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[Y \mid Z])$ measures the *between-Z association*: the component of $\text{Cov}(X, Y)$ induced solely by how variation in Z shifts the conditional means of X and Y.

A confounder corresponds precisely to the second term being nonzero. In particular, it is possible that

$$\text{Cov}(X, Y \mid Z) = 0 \quad \text{a.s.}$$

(i.e. no association between X and Y once Z is fixed), while

$$\text{Cov}(\mathbb{E}[X \mid Z], \mathbb{E}[Y \mid Z]) \neq 0,$$

so that $\text{Cov}(X, Y) \neq 0$ entirely due to the influence of Z on both variables.

5.6 Cvariance measures strength of linear relationship

A quick way to see why covariance captures the *strength of a linear relationship* is to view it through the lens of least squares. For random variables X, Y, consider approximating Y by an affine function of X,

$$\widehat{Y} = a + bX,$$

and choose (a, b) to minimize the mean squared error $\mathbb{E}[(Y - a - bX)^2]$. Differentiating in a gives $a = \mathbb{E}[Y] - b\,\mathbb{E}[X]$. Plugging this back in, the optimal slope is

$$b^\star = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \qquad \text{(assuming } \mathrm{Var}(X) > 0\text{)}.$$

Thus *covariance is exactly the numerator that determines the best linear slope* relating Y to X. Since covariance depends on scale, it is common to normalize it to obtain the *correlation*,

$$\mathrm{Corr}(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1],$$

which is dimensionless and equals ± 1 exactly when Y is an affine function of X almost surely (perfect linear relationship).

5.7 Empirical covariance formula

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be i.i.d. with mean μ and covariance

$$\Sigma = \mathbb{E}[(x - \bar{x})(x - \bar{x})^\top].$$

Define the sample mean and the (unbiased) sample covariance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad S_{\text{unb}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top.$$

We compute $\mathbb{E}[S_{\text{unb}}]$:

$$\begin{aligned}
\mathbb{E}[S_{\text{unb}}] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top\right] \\
&= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}\Big((x_i - \mu) - (\bar{x} - \mu)\Big)\Big((x_i - \mu) - (\bar{x} - \mu)\Big)^\top\right]
\end{aligned}$$

Using $(x_i - \bar{x}) = \Big((x_i - \mu) - (\bar{x} - \mu)\Big)$. Now expand the square and use $\sum_{i=1}^{n}(x_i - \mu) = n(\bar{x} - \mu)$

$$= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top - n(\bar{x} - \mu)(\bar{x} - \mu)^\top\right]$$

Apply linearity of expectation, $\mathbb{E}[(x_i - \mu)(x_i - \mu)^\top] = \Sigma$, and $\mathbb{E}[(\bar{x} - \mu)(\bar{x} - \mu)^\top] = \mathrm{Cov}(\bar{x}) = \Sigma/n$:

$$\begin{aligned}
&= \frac{1}{n-1}\left(n\Sigma - n \cdot \frac{\Sigma}{n}\right) \\
&= \frac{1}{n-1}\,(n-1)\Sigma \\
&= \Sigma.
\end{aligned}$$

Note that a similar calculation using the biased estimator

$$S_{\text{bias}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

gives that the estimators expectation is

$$\mathbb{E}[S_{\text{bias}}] = \frac{n-1}{n}\Sigma$$

so we underestimate the covariance by a constant factor in all feature dimensions.