

Bayesian modelling

Introduction

Léo Belzile, HEC Montréal
2025

Distribution and density function

Let $\mathbf{X} \in \mathbb{R}^d$ be a random vector with distribution function

$$F_{\mathbf{X}}(\mathbf{x}) = \Pr(\mathbf{X} \leq \mathbf{x}) = \Pr(X_1 \leq x_1, \dots, X_d \leq x_d).$$

If the distribution of \mathbf{X} is absolutely continuous,

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_d} \cdots \int_{-\infty}^{x_1} f_{\mathbf{X}}(z_1, \dots, z_d) dz_1 \cdots dz_d,$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the joint **density function**.

Mass function

By abuse of notation, we denote the mass function in the discrete case

$$0 \leq f_{\mathbf{X}}(\mathbf{x}) = \Pr (X_1 = x_1, \dots, X_d = x_d) \leq 1.$$

The support is the set of non-zero density/probability total probability over all points in the support,

$$\sum_{\mathbf{x} \in \text{supp}(\mathbf{X})} f_{\mathbf{X}}(\mathbf{x}) = 1.$$

Marginal distribution

The **marginal distribution** of a subvector $\mathbf{X}_{1:k} = (X_1, \dots, X_k)^\top$ is

$$\begin{aligned} F_{\mathbf{X}_{1:k}}(\mathbf{x}_{1:k}) &= \Pr(\mathbf{X}_{1:k} \leq \mathbf{x}_{1:k}) \\ &= F_{\mathbf{X}}(x_1, \dots, x_k, \infty, \dots, \infty). \end{aligned}$$

Marginal density

The **marginal density** $f_{X_{1:k}}(\mathbf{x}_{1:k})$ of an absolutely continuous subvector $X_{1:k} = (X_1, \dots, X_k)^\top$ is

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_k, z_{k+1}, \dots, z_d) dz_{k+1} \cdots dz_d.$$

through integration from the joint density.

Conditional distribution

The conditional distribution function of Y given $X = x$, is

$$f_{Y|X}(y; x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

for any value of x in the support of X .

Conditional and marginal for contingency table

Consider a bivariate distribution for (Y_1, Y_2) supported on $\{1, 2, 3\} \times \{1, 2\}$ whose joint probability mass function is given in **Table 1**

Table 1: Bivariate mass function with probability of each outcome for (Y_1, Y_2) .

	$Y_1 = 1$	$Y_1 = 2$	$Y_1 = 3$	row total
$Y_2 = 1$	0.20	0.3	0.10	0.6
$Y_2 = 2$	0.15	0.2	0.05	0.4
col. total	0.35	0.5	0.15	1.0

Calculations for the marginal distribution

The marginal distribution of Y_1 is obtained by looking at the total probability for each row/column, e.g.,

$$\Pr(Y_1 = i) = \Pr(Y_1 = i, Y_2 = 1) + \Pr(Y_1 = i, Y_2 = 2).$$

- $\Pr(Y_1 = 1) = 0.35$, $\Pr(Y_1 = 2) = 0.5$, $\Pr(Y_1 = 3) = 0.15$.
- $\Pr(Y_2 = 1) = 0.6$ and $\Pr(Y_2 = 2) = 0.4$

Conditional distribution

The conditional distribution

$$\Pr (Y_2 = i \mid Y_1 = 2) = \frac{\Pr (Y_1 = 2, Y_2 = i)}{\Pr (Y_1 = 2)},$$

so

$$\Pr (Y_2 = 1 \mid Y_1 = 2) = 0.3 / 0.5 = 0.6$$

$$\Pr (Y_2 = 2 \mid Y_1 = 2) = 0.4.$$

Independence

Vectors \mathbf{Y} and \mathbf{X} are independent if

$$F_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = F_{\mathbf{X}}(\mathbf{x})F_{\mathbf{Y}}(\mathbf{y})$$

for any value of \mathbf{x}, \mathbf{y} .

The joint density, if it exists, also factorizes

$$f_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y}).$$

If two subvectors \mathbf{X} and \mathbf{Y} are independent, then the conditional density $f_{\mathbf{Y} | \mathbf{X}}(\mathbf{y}; \mathbf{x})$ equals the marginal $f_{\mathbf{Y}}(\mathbf{y})$.

Expected value

If Y has density f_Y , then

$$E\{g(Y)\} = \int g(y)f_Y(y)dy$$

a weighted integral of g with weight f_Y .

The identity function gives the expected value $E(Y)$.

Covariance matrix

We define the covariance matrix of Y as

$$\text{Va}(Y) = \mathbf{E} \left[\{Y - \mathbf{E}(Y)\} \{Y - \mathbf{E}(Y)\}^\top \right],$$

which reduces in the unidimensional setting to

$$\text{Va}(Y) = \mathbf{E} \{Y - \mathbf{E}(Y)\}^2 = \mathbf{E}(Y^2) - \mathbf{E}(Y)^2.$$

Affine transformations

If Y is d -dimensional and A is $p \times d$ and b is a p vector, then

$$E(AY + b) = AE(Y) + b,$$

$$Va(AY + b) = AVa(Y)A^T.$$

Law of iterated expectation and variance

Let Z and Y be random vectors. The expected value of Y is

$$E_Y(Y) = E_Z \left\{ E_{Y|Z}(Y) \right\}.$$

The **tower** property gives a law of iterated variance

$$Va_Y(Y) = E_Z \left\{ Va_{Y|Z}(Y) \right\} + Va_Z \left\{ E_{Y|Z}(Y) \right\}.$$

Poisson distribution

The Poisson distribution has mass

$$f(x) = \Pr(Y = x) = \frac{\exp(-\lambda)\lambda^x}{\Gamma(x+1)}, \quad x = 0, 1, 2, \dots$$

where $\Gamma(\cdot)$ denotes the gamma function.

The parameter λ of the Poisson distribution is both the expectation and the variance of the distribution, meaning

$$E(Y) = \text{Va}(Y) = \lambda.$$

Gamma distribution

A gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$, denoted $Y \sim \text{gamma}(\alpha, \beta)$, has density

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \quad x \in (0, \infty),$$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$ is the gamma function.

Poisson with random scale

To handle overdispersion in count data, take

$$Y \mid \Lambda = \lambda \sim \text{Poisson}(\lambda)$$

$$\Lambda \sim \text{Gamma}(k\mu, k).$$

The joint density of Y and Λ on $\mathbf{N} = \{0, 1, \dots\} \times \mathbf{R}_+$ is

$$f(y, \lambda) = f(y \mid \lambda)f(\lambda)$$

$$= \frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu} \lambda^{k\mu-1} \exp(-k\lambda)}{\Gamma(k\mu)}$$

Conditional distribution

The conditional distribution of $\Lambda \mid Y = y$ can be found by considering only terms that are function of λ , whence

$$f(\lambda \mid Y = y) \propto \lambda^{y+k\mu-1} \exp\{- (k+1)\lambda\}$$

so $\Lambda \mid Y = y \sim \text{gamma}(k\mu + y, k + 1)$.

Marginal density of Poisson mean mixture

$$\begin{aligned}
 f(y) &= \frac{f(y, \lambda)}{f(\lambda \mid y)} = \frac{\frac{\lambda^y \exp(-\lambda)}{\Gamma(y+1)} \frac{k^{k\mu} \lambda^{k\mu-1} \exp(-k\lambda)}{\Gamma(k\mu)}}{\frac{(k+1)^{k\mu+y} \lambda^{k\mu+y-1} \exp\{- (k+1)\lambda\}}{\Gamma(k\mu+y)}} \\
 &= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} k^{k\mu} (k+1)^{-k\mu-y} \\
 &= \frac{\Gamma(k\mu+y)}{\Gamma(k\mu)\Gamma(y+1)} \left(1 - \frac{1}{k+1}\right)^{k\mu} \left(\frac{1}{k+1}\right)^y
 \end{aligned}$$

Marginally, $Y \sim \text{neg. binom}(p)$ where $p = (k+1)^{-1}$.

Moments of negative binomial

By the laws of iterated expectation and iterative variance,

$$\begin{aligned} \mathbf{E}(Y) &= \mathbf{E}_{\Lambda} \{ \mathbf{E}(Y \mid \Lambda) \} \\ &= \mathbf{E}(\Lambda) = \mu \end{aligned}$$

$$\begin{aligned} \mathbf{Va}(Y) &= \mathbf{E}_{\Lambda} \{ \mathbf{Va}(Y \mid \Lambda) \} + \mathbf{Va}_{\Lambda} \{ \mathbf{E}(Y \mid \Lambda) \} \\ &= \mathbf{E}(\Lambda) + \mathbf{Va}(\Lambda) \\ &= \mu + \mu/k. \end{aligned}$$

The marginal distribution of Y , unconditionally, has a variance which exceeds its mean.

Change of variable formula

Consider an injective (one-to-one) differentiable function $g: \mathbb{R}^d \rightarrow \mathbb{R}^d$, with inverse g^{-1} . Then, if $Y = g(X)$,

$$\Pr(Y \leq \mathbf{y}) = \Pr\{g(X) \leq \mathbf{y}\} = \Pr\{X \leq \mathbf{x} = g^{-1}(\mathbf{y})\}.$$

Using the chain rule, the density of Y is

$$f_Y(\mathbf{y}) = f_X\{g^{-1}(\mathbf{y})\} \left| \mathbf{J}_{g^{-1}}(\mathbf{y}) \right| = f_X(\mathbf{x}) \left| \mathbf{J}_g(\mathbf{x}) \right|^{-1}$$

where $\mathbf{J}_g(\mathbf{x})$ is the Jacobian matrix with i, j th element $\partial[g(\mathbf{x})]_i / \partial x_j$.

Gaussian location-scale

Consider d independent standard Gaussian variates $X_j \sim \text{Gauss}(0, 1)$ for $j = 1, \dots, d$, with joint density function

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{\mathbf{x}^\top \mathbf{x}}{2}\right).$$

Consider the transformation $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, with \mathbf{A} an invertible matrix.

Change of variable for Gaussian

- The inverse transformation is $g^{-1}(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$.
- The Jacobian $\mathbf{J}_g(\mathbf{x})$ is simply \mathbf{A} , so the joint density of \mathbf{Y} is

$$(2\pi)^{-d/2} |\mathbf{A}|^{-1} \exp \left\{ - \frac{(\mathbf{y} - \mathbf{b})^\top \mathbf{A}^{-\top} \mathbf{A}^{-1} (\mathbf{y} - \mathbf{b})}{2} \right\}.$$

Since $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$ and $\mathbf{A}^{-\top} \mathbf{A}^{-1} = (\mathbf{A} \mathbf{A}^\top)^{-1}$, we recover $\mathbf{Y} \sim \text{Gauss}_d(\mathbf{b}, \mathbf{A} \mathbf{A}^\top)$.

Conditional distribution of Gaussian subvectors

Let $\mathbf{Y} \sim \text{Gauss}_d(\boldsymbol{\mu}, \mathbf{Q}^{-1})$ and consider the partition

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix},$$

where \mathbf{Y}_1 is a $k \times 1$ and \mathbf{Y}_2 is a $(d - k) \times 1$ vector for some $1 \leq k < d$.

Then, we have the conditional distribution

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2 \sim \text{Gauss}_k(\boldsymbol{\mu}_1 - \mathbf{Q}_{11}^{-1} \mathbf{Q}_{12}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \mathbf{Q}_{11}^{-1})$$

Likelihood

The **likelihood** $L(\theta)$ is a function of the parameter vector θ that gives the ‘density’ of a sample under a postulated distribution, treating the observations as fixed,

$$L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta).$$

Likelihood for independent observations

If the joint density factorizes,

$$L(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^n f_i(y_i; \boldsymbol{\theta}) = f_1(y_1; \boldsymbol{\theta}) \times \cdots \times f_n(y_n; \boldsymbol{\theta}).$$

The corresponding log likelihood function for independent and identically distributions observations is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n \ln f(y_i; \boldsymbol{\theta})$$

Score

Let $\ell(\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$, be the log likelihood function. The gradient of the log likelihood, termed **score** is the p -vector

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}.$$

Information matrix

The **observed information matrix** is the hessian of the negative log likelihood,

$$j(\boldsymbol{\theta}; \mathbf{y}) = - \frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, so $j(\hat{\boldsymbol{\theta}})$.

Expected information

Under regularity conditions, the **expected information**, also called **Fisher information matrix**, is

$$i(\theta) = \mathbb{E} \left\{ U(\theta; Y) U(\theta; Y)^\top \right\} = \mathbb{E} \{ j(\theta; Y) \}$$

Note on information matrices

Information matrices are symmetric and provide information about the variability of $\hat{\theta}$.

The information of an iid sample of size n is n times that of a single observation

- information accumulates at a linear rate.

Information for the Gaussian distribution

Consider $Y \sim \text{Gauss}(\mu, \tau^{-1})$, parametrized in terms of precision τ . The likelihood contribution for an n sample is, up to proportionality,

$$\ell(\mu, \tau) \propto \frac{n}{2} \log(\tau) - \frac{\tau}{2} \sum_{i=1}^n (Y_i^2 - 2\mu Y_i + \mu^2)$$

Gaussian information matrices

The observed and Fisher information matrices are

$$j(\mu, \tau) = \begin{pmatrix} n\tau & -\sum_{i=1}^n (Y_i - \mu) \\ -\sum_{i=1}^n (Y_i - \mu) & \frac{n}{2\tau^2} \end{pmatrix},$$

$$i(\mu, \tau) = n \begin{pmatrix} \tau & 0 \\ 0 & \frac{1}{2\tau^2} \end{pmatrix}$$

Since $E(Y_i) = \mu$, the expected value of the off-diagonal entries of the Fisher information matrix are zero.

Example: random right-censoring

Consider a survival analysis problem for independent time-to-event data subject to (noninformative) random right-censoring. We observe

- failure times $Y_i (i = 1, \dots, n)$ drawn from $F(\cdot; \theta)$ supported on $(0, \infty)$
- independent binary censoring indicators $C_i \in \{0, 1\}$, with 0 indicating right-censoring and $C_i = 1$ observed failure time.

Likelihood contribution with censoring

If individual observation i has not experienced the event at the end of the collection period, then the likelihood contribution is

$\Pr(Y > y) = 1 - F(y; \theta)$, where y_i is the maximum time observed for Y_i . We write the log likelihood

$$\ell(\theta) = \sum_{i: c_i=0} \log\{1 - F(y_i; \theta)\} + \sum_{i: c_i=1} \log f(y_i; \theta)$$

Censoring and exponential data

Suppose for simplicity that $Y_i \sim \text{expo}(\lambda)$ and let $m = c_1 + \dots + c_n$ denote the number of observed failure times. Then, the log likelihood and the Fisher information are

$$\ell(\lambda) = \lambda \sum_{i=1}^n y_i + \log \lambda m$$

$$i(\lambda) = m / \lambda^2$$

and the right-censored observations for the exponential model do not contribute to the information.

Example: first-order autoregressive process

Consider an AR(1) model of the form

$$Y_t = \mu + \phi(Y_{t-1} - \mu) + \varepsilon_t,$$

where

- ϕ is the lag-one correlation,
- μ the global mean and
- ε_t is an iid innovation with mean zero and variance σ^2 .

If $|\phi| < 1$, the process is stationary, and the variance does not increase with t .

Markov property and likelihood decomposition

The Markov property states that the current realization depends on the past, $Y_t \mid Y_1, \dots, Y_{t-1}$, only through the most recent value Y_{t-1} .

The log likelihood thus becomes

$$\ell(\theta) = \ln f(y_1) + \sum_{i=2}^n f(y_i \mid y_{i-1}).$$

Marginal of AR(1)

The AR(1) stationarity process has unconditional moments

$$E(Y_t) = \mu, \quad \text{Var}(Y_t) = \sigma^2 / (1 - \phi^2).$$

The AR(1) process is first-order Markov since the conditional distribution $f(Y_t \mid Y_{t-1}, \dots, Y_{t-p})$ equals $f(Y_t \mid Y_{t-1})$.

Log likelihood of AR(1)

If innovations are Gaussian, we have

$$Y_t \mid Y_{t-1} = y_{t-1} \sim \text{Gauss}\{\mu(1 - \phi) + \phi y_{t-1}, \sigma^2\}, \quad t > 1.$$

so the log-likelihood is

$$\begin{aligned} \ell(\mu, \phi, \sigma^2) = & -\frac{n}{2}\log(2\pi) - n\log\sigma + \frac{1}{2}\log(1 - \phi^2) \\ & - \frac{(1 - \phi^2)(y_1 - \mu)^2}{2\sigma^2} - \sum_{i=2}^n \frac{(y_t - \mu(1 - \phi) - \phi y_{t-1})^2}{2\sigma^2} \end{aligned}$$

Estimation of integrals

Suppose we can simulate B i.i.d. variables with the same distribution, x_1, \dots, x_B with distribution F .

We want to compute $E\{g(X)\} = \int g(x)f(x)dx = \mu_g$ for some functional $g(\cdot)$

- $g(x) = x$ (mean)
- $g(x) = \mathbf{1}(x \in A)$ (probability of event)
- etc.

Vanilla Monte Carlo integration

We substitute expected value by sample average of

$$\hat{\mu}_g = \frac{1}{B} \sum_{b=1}^B g(x_b).$$

- law of large number guarantees convergence of $\hat{\mu}_g \rightarrow \mu_g$ if the latter is finite.
- Under finite second moments, central limit theorem gives

$$\sqrt{B}(\hat{\mu}_g - \mu_g) \sim \text{No}(0, \sigma_g^2).$$

Importance sampling

Consider density q instead with $\text{supp}(p) \subseteq \text{supp}(q)$. Then,

$$\mathbb{E}\{g(X)\} = \int_{\mathcal{X}} g(x) \frac{p(x)}{q(x)} q(x) dx$$

and we can proceed similarly by drawing samples from q .

Importance sampling estimator

An alternative Monte Carlo estimator uses the weighted average

$$\tilde{\mathbb{E}}\{g(X)\} = \frac{B^{-1} \sum_{b=1}^B w_b g(x_b)}{B^{-1} \sum_{b=1}^B w_b}.$$

with weights $w_b = p(x_b)/q(x_b)$. The latter equal 1 on average, so one could omit the denominator without harm.

Standard errors

If the variance of $g(X)$ is finite, we can approximate the latter by the sample variance of the simple random sample and obtain the Monte Carlo standard error of the estimator

$$\text{se}^2[\hat{\mathbb{E}}\{g(X)\}] = \frac{1}{B(B-1)} \sum_{b=1}^B \left[g(x_b) - \hat{\mathbb{E}}\{g(X)\} \right]^2.$$

Precision of Monte Carlo integration

We want to have an estimator as precise as possible.

- but we can't control the variance of $g(X)$, say σ_g^2
- the more simulations B , the lower the variance of the mean.
- sample average for i.i.d. data has variance σ_g^2 / B
- to reduce the standard deviation by a factor 10, we need 100 times more draws!

Remember: the answer is **random**.

Example: functionals of gamma distribution

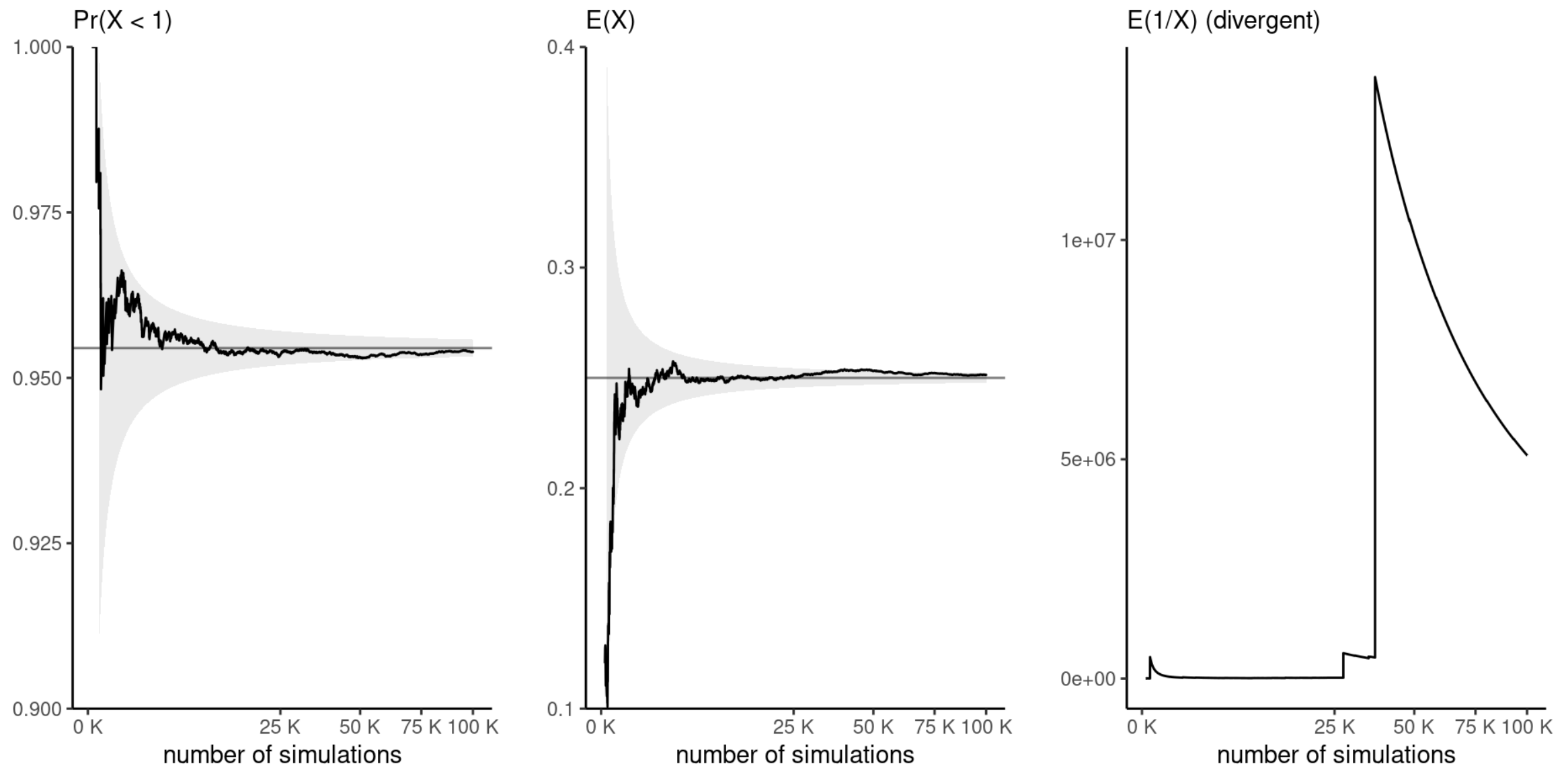


Figure 1: Running mean trace plots for $g(x) = I(x < 1)$ (left), $g(x) = x$ (middle) and $g(x) = 1/x$ (right) for a Gamma distribution with shape 0.5 and rate 2, as a function of the Monte Carlo sample size.

Recap

1. We can specify distribution using **hierarchies**, with marginal \times conditional.
2. Most density and mass functions for \mathbf{Y} can be identified from their support and their **kernel**, i.e., terms that depend on \mathbf{y} , ignoring normalizing constants. We then match expressions.
3. Expectations can be calculated analytically, or approximated via Monte Carlo simulations.