Chapter 1
Introduction, Examples, and References

Qing Zhou*

Contents

*UCLA Department of Statistics (email: zhou@stat.ucla.edu).

This set of lecture notes, consisting of four chapters, is for an undergraduate course on Monte Carlo methods. Two main references are:

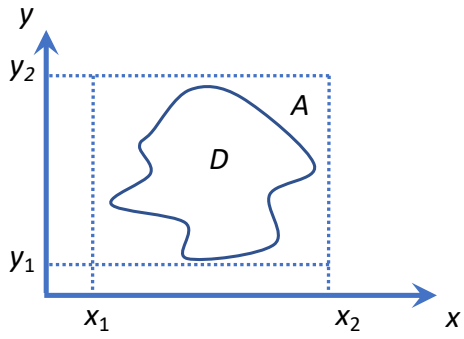
1. Jun S. Liu (2001) *Monte Carlo Strategies in Scientific Computing* (first edition), Springer.
2. Howard M. Taylor and Samuel Karlin (1998) *An Introduction to Stochastic Modeling* (third edition), Academic Press.

In particular, materials for sequential importance sampling and Markov chain Monte Carlo are mostly adapted from selected topics in chapters 2, 3, 5 and 6 of Liu (2001), supplemented with some simpler examples. A brief introduction to Markov chains is developed based on chapters 3 and 4 of Taylor and Karlin (1998).

1. Introduction

Goal of Monte Carlo: Use computer simulation to generate random variables from a given distribution $p(x)$.

1.1. Calculating Area



Want to compute the area of D in $\mathbb{R}^2$.

Find rectangle $A : [x_1, x_2] \times [y_1, y_2] \supset D$; Randomly generate n points in A and suppose M of them in D. Then we estimate the area of D as

$$\widehat{S}_n(D) = \frac{M}{n} \cdot S(A) = \frac{M}{n} \cdot (x_2 - x_1)(y_2 - y_1). \tag{1}$$

Why is $\widehat{S}_n(D)$ a reasonable estimate? Note that

$$\mathbb{P}(\text{a point in } D) = \frac{S(D)}{S(A)} := p. \tag{2}$$

If n is large, the fraction of points in D will be close to p, i.e.

$$\frac{S(D)}{S(A)} = p \approx M/n \;\Rightarrow\; S(D) \approx \frac{M}{n} \cdot S(A).$$

To implement this method, suppose the boundary of D is given by the curve $f(x, y) = 0$ and its interior is $\{(x, y) : f(x, y) < 0\}$. We first generate n uniform points $(x^{(i)}, y^{(i)}) \in A$, $i = 1, \ldots, n$. Let M be the number of points satisfying $f(x^{(i)}, y^{(i)}) \le 0$. Then we use (1) to estimate $S(D)$.

More rigorous justification: Let M (random variable) be the number of points in D if n points are uniformly generated in A. Then

$$M \sim \text{Bin}(n, p).$$

Apply the strong law of large numbers (SLLN), $M = \sum_{i=1}^n X_i$, $X_i \sim_{iid} \text{Bern}(p)$:

$$\frac{M}{n} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \mathbb{E}(X_1) = p \;\Rightarrow\; \widehat{S}_n(D) \xrightarrow{a.s.} S(D), \quad \text{as } n \to \infty.$$

Theorem 1 (The strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, each having a finite mean $\mu = \mathbb{E}(X_i)$. Then*

$$\mathbb{P}\Big[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i = \mu\Big] = 1.$$

For short, we write $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{a.s.} \mu$ (converge almost surely).

Or we can calculate the bias and variance of $\widehat{p} = M/n$ which estimates the probability p defined by (2):

$$\mathbb{E}(M) = n \cdot p \;\Rightarrow\; \mathbb{E}\left(\frac{M}{n}\right) = p: \text{ unbiased.}$$

$$\text{Var}(M) = np(1-p) \;\Rightarrow\; \text{Var}\left(\frac{M}{n}\right) = \frac{p(1-p)}{n} \;\to\; 0 \quad \text{as} \quad n \to \infty.$$

Recall that the mean squared error:

$$\mathbb{E}\big[\widehat{p} - p\big]^2 = \text{bias}(\widehat{p})^2 + \text{Var}(\widehat{p}) = \frac{p(1-p)}{n}.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{E}\big[\widehat{p}-p\big]^2 = 0 \;\Rightarrow\; \lim_{n\to\infty} \mathbb{E}\big[\widehat{S}_n(D) - S(D)\big]^2 = 0.$$

To give a concrete example, suppose we want to estimate the area of a circle:

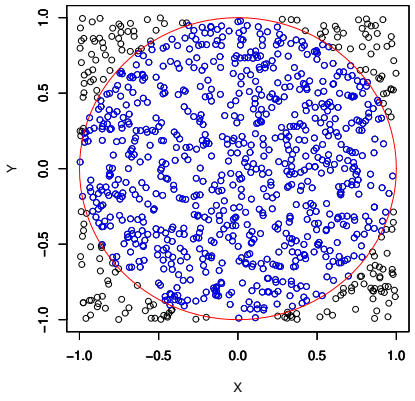
$$D = \{(x,y) : x^2 + y^2 \le 1\}.$$

Choose $A = [-1,1]\times[-1,1]$. Then $S(A) = 4$ and $\widehat{S}_n(D) = 4 \times M/n = 4\widehat{p}$. The standard error of $\widehat{S}_n$ is

$$\text{se} = \sqrt{\text{Var}(\widehat{S}_n)} = \sqrt{\text{Var}(4\widehat{p})} = 4\sqrt{\text{Var}(\widehat{p})} = 4\sqrt{p(1-p)/n},$$

which can be approximated as $4\sqrt{\widehat{p}(1-\widehat{p})/n}$.

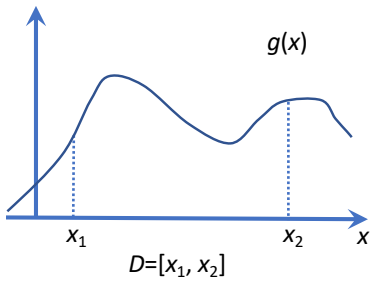
```
>
> n=1000;
> X=runif(n,-1,1);
> Y=runif(n,-1,1);
> f=X^2+Y^2;
> M=sum(f<=1);
> p_h=M/n;
> S_h=M/n*4;
> se=4*sqrt(p_h*(1-p_h)/n);
>
> p_h
[1] 0.788
> S_h
[1] 3.152
> se
[1] 0.05170006
```



1.2. Approximating Integrals

Want to estimate

$$I = \int_D g(x)dx.$$



Generate $x^{(1)}, x^{(2)}, \cdots, x^{(n)} \sim_{iid} \text{Unif}(D)$. Apply SLLN

$$\hat{g}_n := \frac{1}{n}\sum_{i=1}^{n} g(x^{(i)}) \xrightarrow{a.s.} \mathbb{E}[g(X)] = \int g(x) \cdot \frac{1}{|D|} dx,$$

$|D|$: volume of D and $X \sim \text{unif}(D)$. Then

$$|D| \cdot \hat{g}_n \xrightarrow{a.s.} \int g(x)dx.$$

Our estimate $\widehat{I}_n = |D| \cdot \hat{g}_n = \frac{|D|}{n}\sum_{i=1}^{n} g(x^{(i)}) \xrightarrow{a.s.} I$ as $n \to \infty$.

1.3. Estimating Expectations

$f(x)$: distribution of interest.

Want to estimate its mean and variance:

$\mathbb{E}(X) = \mu = \int x f(x)dx$, $\text{Var}(X) = \int (x-\mu)^2 f(x)dx = \mathbb{E}(X-\mu)^2$.

Generate samples $x^{(1)}, x^{(2)}, \cdots, x^{(n)} \sim_{iid} f$.

$$\hat{\mu}_n = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} \xrightarrow{a.s.} \mathbb{E}(X),$$

$$\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \hat{\mu}_n)^2 \xrightarrow{a.s.} \text{Var}(X) \quad \text{as} \quad n \to \infty.$$

In general, to estimate $\mathbb{E}(g(X))$ for some function g of X:

$$\frac{1}{n}\sum_{i=1}^{n} g(x^{(i)}) \xrightarrow{a.s.} \mathbb{E}(g(X)) = \int g(x)f(x)dx.$$

2. Bayesian Inference

Two major tasks of statistical inference is (i) to estimate unknown model parameters from data; (ii) to quantify the uncertainty in the estimates. Suppose we have collected data:

$$y_1, y_2, \cdots, y_n \overset{\text{iid}}{\sim} f(y \mid \theta),$$

where $f(y \mid \theta)$ is a pdf (or pmf) of a distribution parameterized by θ. Then we want to estimate θ and/or build a confidence interval for θ.

In general, denote the observed data by $\mathbf{y} = (y_1, y_2, \cdots, y_n)$. A common estimation method is the maximum likelihood estimate (MLE). Define the likelihood of $\mathbf{y}$ as

$$L(\theta \mid \mathbf{y}) := p(y_1, \cdots, y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta).$$

The MLE $\hat{\theta}_{\text{MLE}}$ is the maximizer of $L(\theta \mid \mathbf{y})$ over θ:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, L(\theta \mid \mathbf{y}).$$

Moreover, we often estimate the standard error of the MLE, denoted by $\widehat{\text{se}}$, and construct an approximate 95% confidence interval as

$$(\hat{\theta}_{\text{MLE}} - 2\widehat{\text{se}}, \hat{\theta}_{\text{MLE}} + 2\widehat{\text{se}})$$

as a way to quantify the uncertainty in our estimate. The interpretation of the interval is

$$\mathbb{P}[\theta \in (\hat{\theta}_{\text{MLE}} - 2\widehat{\text{se}}, \hat{\theta}_{\text{MLE}} + 2\widehat{\text{se}})] = 0.95.$$

Here, $\hat{\theta}_{\text{MLE}}$ is regarded as a random variable as a function of the random sample $\mathbf{y}$, while θ is an *unknown constant*.

2.1. Main steps

Bayesian inference relies on posterior distributions to provide solutions to the two inferential tasks (i) and (ii). The unknown parameter θ is regarded as a *random variable* and thus we need to specify a marginal distribution for θ, denoted by $p(\theta)$, which is called a prior distribution. Here, "prior" means before observing any data, as the prior distribution does not depend on the data $\mathbf{y}$. Therefore, a Bayesian model for the data $\mathbf{y}$ is set up by two distributions:

$$\text{Prior: } \theta \sim p(\theta), \tag{3}$$

$$\text{Data: } \mathbf{y} = (y_1, \ldots, y_n) \mid \theta \overset{\text{iid}}{\sim} f(y \mid \theta). \tag{4}$$

Together, they define a joint distribution for $(\theta, \mathbf{y})$:

$$p(\theta, \mathbf{y}) = p(\theta)p(\mathbf{y} \mid \theta) = p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta). \tag{5}$$

Based on (5), we find the conditional distribution $[\theta \mid \mathbf{y}]$ to perform inference on θ. This conditional distribution of θ given the data $\mathbf{y}$ is called the posterior distribution, where "posterior" means the distribution of θ is now updated after observing the data and thus depends on $\mathbf{y}$. Applying Bayes formula,

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\theta)p(\mathbf{y} \mid \theta)}{p(\mathbf{y})} = \frac{p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta)}{p(\mathbf{y})},$$

where the marginal density $p(\mathbf{y}) = \int p(\theta, \mathbf{y}) d\theta$ does not depend on θ and can be regarded as a normalizing constant. Consequently, it is more convenient to work with an unnormalized posterior density:

$$p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y} \mid \theta) = p(\theta) \cdot \prod_{i=1}^{n} f(y_i \mid \theta). \tag{6}$$

We may either recognize the posterior distribution via the unnormalized density on the right side or use Monte Carlo methods to draw samples given the unnormalized density.

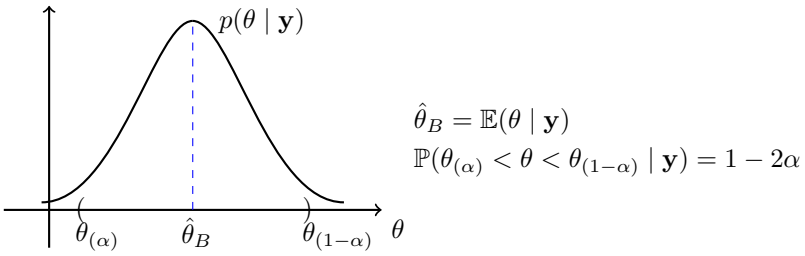
A Bayesian estimate of θ is usually constructed as the mean of the posterior distribution,

$$\hat{\theta}_B := \mathbb{E}(\theta|\mathbf{y}) = \int \theta \cdot p(\theta \mid \mathbf{y}) d\theta. \tag{7}$$

A $(1 - 2\alpha)$ Bayesian interval for θ can be constructed by the quantiles of the posterior distribution: $(\theta_{(\alpha)}, \theta_{(1-\alpha)})$, where $\theta_{(\alpha)}$ is the α-quantile for $\alpha \in (0, 1)$. The interpretation of a Bayesian interval is

$$\mathbb{P}(\theta \in (\theta_{(\alpha)}, \theta_{(1-\alpha)}) \mid \mathbf{y}) = 1 - 2\alpha, \tag{8}$$

where θ is a random variable following the posterior distribution $p(\theta \mid \mathbf{y})$.



For complicated problems, Monte Carlo simulation, such as MCMC, is applied to draw samples of θ from the posterior distribution $p(\theta \mid \mathbf{y})$, regarding (6) as the target density. From the Monte Carlo samples, one can easily calculate the sample mean and sample quantiles to approximate $\hat{\theta}_B$ and $(\theta_{(\alpha)}, \theta_{(1-\alpha)})$.

In summary, the main steps of Bayesian inference are:

1. Choose a prior distribution $p(\theta)$.
2. Find the posterior distribution $p(\theta \mid \mathbf{y})$ by (6).
3. Apply a Monte Carlo algorithm to draw samples from $p(\theta \mid \mathbf{y})$.
4. Construct Bayesian estimates and intervals from the Monte Carlo samples.

2.2. Some basic models

We will demonstrate the main steps of Bayesian inference with a few simple examples.

Example 1 (Binomial distribution)**.** Consider independent coin tossing with $\theta \in (0, 1)$ being the probability of heads. Suppose we toss n times and observe heads x times. How to estimate θ?

Let X (random variable) be the number of times we observe heads. The distribution of X given θ is

$$X \mid \theta \sim \text{Bin}(n, \theta).$$

Thus, the likelihood

$$\mathbb{P}(X = x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

MLE: $\hat{\theta}_{\text{MLE}} = \dfrac{x}{n}$.

Bayesian inference:

1. Choose a prior distribution for θ: Without any prior knowledge on θ, we usually choose a flat prior,

$$\theta \sim \text{Unif}(0, 1), \quad \text{i.e. } p(\theta) = 1,\ \theta \in (0, 1).$$

2. Then find the posterior distribution:

$$\begin{aligned} p(\theta \mid X = x) &\propto p(\theta) \cdot \mathbb{P}(X = x \mid \theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^x (1-\theta)^{n-x}, \end{aligned} \tag{9}$$

where θ is the random variable.

3. From (9), we recognize that it is an unnormalized Beta density. Therefore, the posterior distribution is

$$\theta \mid x \sim \text{Beta}(x+1, n-x+1). \tag{10}$$

   As a reference, the pdf of Beta(α, β) is

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

   and its mean is $\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$.

4. Given (10), we find Bayesian estimate

$$\hat{\theta}_B = \mathbb{E}(\theta|x) = \frac{x+1}{n+2}.$$

   To construct a 95% Bayesian interval, we use the 2.5% and 97.5% quantiles of Beta$(x+1, n-x+1)$. For example, if $n = 10, x = 3$, the posterior distribution is Beta$(4, 8)$, for which the two quantiles are

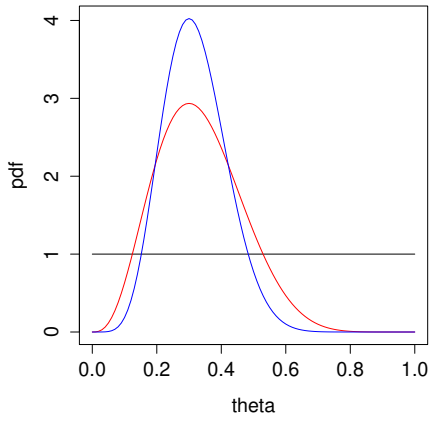
```
> qbeta(c(0.025,0.975),4,8)
[1] 0.1092634 0.6097426
```

   So the 95% Bayesian interval is $(0.109, 0.610)$. If $n = 20, x = 6$, the posterior distribution is Beta$(7, 15)$ with the quantiles given by

```
> qbeta(c(0.025,0.975),7,15)
[1] 0.1458769 0.5217511
```

   In this case, Bayesian interval is $(0.146, 0.522)$, which is shorter than the first case as the sample size n is larger.

The following figure shows the shape of the prior (black) and the posterior distributions: red for $n = 10, x = 3$ and blue for $n = 20, x = 6$.



A Bayesian interval can be used to do hypothesis test. Suppose we want to decide whether the coin is fair

$$H_0 : \theta = 0.5.$$

Based on the data $n = 20, x = 6$, the 95% Bayesian interval $(0.146, 0.522)$ covers 0.5, and therefore we will accept the null hypothesis H_0. If we collect more data and observe $n = 50, x = 15$, then $\theta \mid x \sim \text{Beta}(16, 36)$ and a 95% Bayesian interval will be $(0.191, 0.438)$. Because 0.5 falls outside this interval, we conclude with 95% probability that the coin is not fair (reject H_0).

The uniform distribution $\text{Unif}(0, 1)$ is equivalent to $\text{Beta}(1, 1)$. We may choose other Beta distribution as the prior for θ:

$$\theta \sim \text{Beta}(\alpha, \beta), \qquad p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Then the posterior distribution

$$\begin{aligned} p(\theta \mid X = x) &\propto p(\theta) \cdot \mathbb{P}(X = x \mid \theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}, \end{aligned}$$

and thus,

$$\theta \mid x \sim \text{Beta}(x + \alpha, n - x + \beta).$$

We see that the posterior is in the same family of the prior, both Beta distributions, in which case we say the prior is a *conjugate prior*. That is, Beta prior is conjugate to the Binomial distribution. The Bayesian estimate, i.e. the posterior mean, under this prior is

$$\hat{\theta}_B = \frac{x + \alpha}{n + \alpha + \beta}. \tag{11}$$

Compared to the MLE $\hat{\theta}_{\text{MLE}} = x/n$, the prior parameters (α, β) may be regarded as pseudo counts added to the two possible outcomes (heads or tails). If there is no prior knowledge about θ, we choose small pseudo counts, $\alpha, \beta \in (0, 1]$. If there is strong prior for θ, say from historical data, one may choose larger values of α, β to reflect such prior knowledge.

Example 2 (Multinomial distribution)**.** We generalize Example 1 to multinomial data. Let $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ be the probabilities of k possible outcomes in a random experiment, $\theta_j > 0, \sum_{j=1}^k \theta_j = 1$. Suppose we have done this experiment n times independently and observed the jth outcome x_j times. So the observations follow a multinomial distribution:

$$\mathbf{x} = (x_1, x_2, \cdots, x_k) \mid \theta \sim \text{M}(n, \theta), \quad \sum x_j = n.$$

The likelihood is

$$p(\mathbf{x} \mid \theta) \propto \theta_1^{x_1}\theta_2^{x_2}\cdots\theta_k^{x_k}, \tag{12}$$

and the MLE

$$(\hat{\theta}_j)_{\text{MLE}} = \frac{x_j}{n}, \qquad j = 1, \ldots, k.$$

To do Bayesian inference, let us first find a conjugate prior.

Definition 1 (Dirichlet distribution)**.** Let $\theta = (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k$ be a random vector such that $\theta_j \geq 0$ for all $j = 1, \ldots, k$ and $\sum_{j=1}^k \theta_j = 1$. Then θ follows the Dirichlet distribution $\text{Dir}(\alpha_1, \ldots, \alpha_k)$, $\alpha_j > 0$ for all j, if the pdf of θ is

$$p(\theta) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)}\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\cdots\theta_k^{\alpha_k - 1}.$$

The mean of θ is

$$\mathbb{E}(\theta_j) = \frac{\alpha_j}{\alpha_1 + \alpha_2 + \cdots + \alpha_k}, \quad j = 1, \ldots, k. \tag{13}$$

How to sample θ from $\text{Dir}(\alpha_1, \ldots, \alpha_k)$?

1. Draw $v_j \sim \text{Gamma}(\alpha_j, 1)$ independently for $j = 1, \ldots, k$.
2. Put $S = \sum_{j=1}^k v_j$ and define

$$\theta_j = \frac{v_j}{S} = \frac{v_j}{v_1 + \cdots + v_k}, \qquad j = 1, \ldots, k.$$

Then $\theta = (\theta_1, \theta_2, \cdots, \theta_k) \sim \text{Dir}(\alpha_1, \alpha_2, \cdots, \alpha_k)$.

It turns out the Dirichlet is a conjugate prior for multinomial distribution. To see that, let us assume the prior is $\theta \sim \text{Dir}(\alpha_1, \cdots, \alpha_k)$, i.e.

$$p(\theta) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\cdots\theta_k^{\alpha_k - 1}. \tag{14}$$

Then the posterior distribution, by multiplying (14) and (12),

$$\begin{aligned} p(\theta \mid \mathbf{x}) &\propto p(\theta)p(\mathbf{x} \mid \theta) \\ &\propto \theta_1^{x_1 + \alpha_1 - 1}\theta_2^{x_2 + \alpha_2 - 1}\cdots\theta_k^{x_k + \alpha_k - 1}, \end{aligned}$$

which is an unnormalized density of $\text{Dir}(x_1 + \alpha_1, \ldots, x_k + \alpha_k)$. Therefore,

$$\theta \mid \mathbf{x} \sim \text{ Dir}(x_1 + \alpha_1, \ldots, x_k + \alpha_k). \tag{15}$$

Put $\alpha_0 = \sum_{j=1}^k \alpha_j$. By (13), we find the Bayesian estimate of θ by the posterior mean:

$$(\hat{\theta}_j)_B = \frac{x_j + \alpha_j}{n + \alpha_0}, \quad j = 1, \ldots, k.$$

Similar to (11), here $\alpha_1, \cdots, \alpha_k$ are also interpreted as pseudo counts for the k possible outcomes. Without any prior knowledge, we choose $\alpha_j \in (0, 1]$. In particular, if $\alpha_j = 1$ for all j, the prior is a uniform distribution ($p(\theta) \propto 1$).

If we wish to build a Bayesian interval for θ_j, we can do so using the quantiles of the posterior distribution $[\theta_j \mid \mathbf{x}]$, which is simply a marginal distribution of the Dirichlet distribution (15). By properties of Dirichlet distributions, the marginal distribution is a Beta distribution:

$$\theta_j \mid \mathbf{x} \sim \text{Beta}(x_j + \alpha_j, n - x_j + \alpha_0 - \alpha_j).$$

Then we can use the same procedure in Example 1 to construct a Bayesian interval for each θ_j.

Example 3 (Normal data with known variance)**.** Suppose we have observed

$$y_1, \cdots y_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2),$$

where σ^2 is known. Our goal is to make inference on θ. The likelihood of the data is

$$\begin{aligned} p(y_1, \cdots, y_n \mid \theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}. \end{aligned}$$

The MLE $\hat{\theta}_{\text{MLE}} = \bar{y} = \frac{1}{n}\sum_i y_i$. The standard error (standard deviation) of $\bar{y}$ is $\text{se} = \sigma/\sqrt{n}$. Thus, we can construct a 95% confidence interval $(\bar{y} \pm 2\sigma/\sqrt{n})$.

Now consider Bayesian inference. A conjugate prior for θ is $\theta \sim \mathcal{N}(\mu_0, \tau_0^2)$. Let us consider a flat prior by choosing $\tau_0 \to \infty$:

$$p(\theta) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} \to 1, \text{ as } \tau_0 \to \infty.$$

Then, the posterior distribution $[\theta \mid \mathbf{y} = (y_1, \ldots, y_n)]$ is

$$p(\theta \mid \mathbf{y}) \propto p(\theta)p(\mathbf{y}|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\theta - y_i)^2\right\}.$$

Recall that θ is the random variable and $\mathbf{y}$ is constant. Using the equality

$$\begin{aligned} \sum_{i=1}^{n}(\theta - y_i)^2 &= \sum_i (\theta - \bar{y} + \bar{y} - y_i)^2 \\ &= n(\theta - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2, \end{aligned}$$

we get

$$p(\theta \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2} n(\theta - \bar{y})^2\right\} = \exp\left\{-\frac{(\theta - \bar{y})^2}{2\sigma^2/n}\right\}.$$

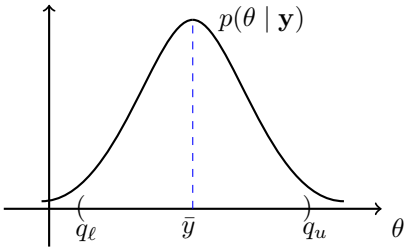
This shows that the posterior distribution

$$\theta \mid \mathbf{y} \sim \mathcal{N}(\bar{y}, \sigma^2/n).$$

Then, the Bayesian estimate is $\hat{\theta}_B = \mathbb{E}(\theta|\mathbf{y}) = \bar{y}$ and a 95% Bayesian interval, constructed by the quantiles (q_ℓ, q_u) of $\mathcal{N}(\bar{y}, \sigma^2/n)$, is

$$(\bar{y} - 2\sigma/\sqrt{n}, \bar{y} + 2\sigma/\sqrt{n}).$$

See below for illustration:



$\mathbb{E}(\theta \mid \mathbf{y}) = \bar{y}$
$\mathbb{P}(q_\ell < \theta < q_u \mid \mathbf{y}) = 0.95$

Again, the interval length $(4\sigma/\sqrt{n})$ shrinks when n increases. For this example, the Bayesian point and interval estimates both coincide with the MLE and the confidence interval.

3. Inverse-CDF Method

Consider a univariate random variable X. The cumulative distribution function (c.d.f.) F of X is defined as

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}. \tag{16}$$

Below are some important properties of the c.d.f. F:

1. F is nondecreasing: if $a < b$ then $F(a) \leq F(b)$.
2. $\lim_{a \to -\infty} F(a) = 0$.
3. $\lim_{b \to \infty} F(b) = 1$.
4. F is right continuous: For any b and any decreasing sequence $b_n, n \geq 1$ such that $b_n \to b$, we have $\lim_{n \to \infty} F(b_n) = F(b)$.

See cases 1 and 2 below for typical examples of F.

Now suppose we have calculated the c.d.f. F of X. Can we make use of F to simulate X? The following theorem shows how to do this by defining an inverse of the c.d.f.:

Theorem 2. *Let $F(x)$ denote a c.d.f. with inverse*

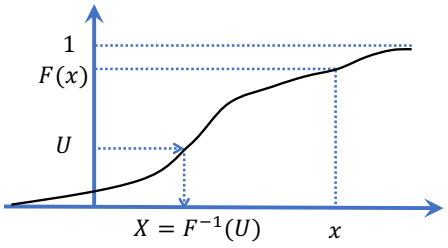
$$F^{-1}(u) := \min\{z : F(z) \geq u\} \quad \textit{for } u \in (0, 1]. \tag{17}$$

If $U \sim \text{Unif}(0, 1)$, then $X = F^{-1}(U) \sim F$, i.e., the c.d.f. of X is F.

Inverse-CDF method:

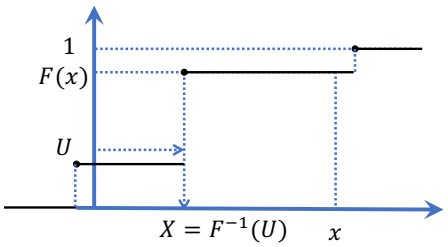
1. Find the c.d.f. $F(x) = \mathbb{P}(X \leq x)$ for any $x \in \mathbb{R}$.
2. Calculate its inverse c.d.f. $F^{-1}(u)$ by (17) for any $u \in (0, 1]$.
3. Simulate $U \sim \text{Unif}(0, 1)$ and let $X = F^{-1}(U)$.

Case 1: F is invertible, in which case F^{-1}, as defined in (17), is the inverse function of F.



$$\{X = F^{-1}(U) \leq x\} \;\Leftrightarrow\; \{U \leq F(x)\}$$

Case 2: F is discrete.



$$X = F^{-1}(U) := \min\{z : F(z) \geq U\}$$

Proof of Theorem 2. The overall idea of the proof is

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

To establish the second equality, we need to show that

$$\{F^{-1}(U) \leq x\} \;\Leftrightarrow\; \{U \leq F(x)\}.$$

This equivalence is easily seen in case 1 (the invertible case). For the general case, it is implied by (a) and (b) as follows:

(a) $F^{-1}(U) \leq x \;\Rightarrow\; U \leq F(x)$.

$X = F^{-1}(U) = \min\{z : F(z) \geq U\} \Rightarrow X \in \{z : F(z) \geq U\} \Rightarrow\; F(X) \geq U$. Since $X = F^{-1}(U) \leq x$ by assumption and F is non-decreasing, we have

$$F(x) \geq F(X) \geq U \;\Rightarrow\; U \leq F(x).$$

(b) $U \leq F(x) \;\Rightarrow\; F^{-1}(U) \leq x$.

$U \leq F(x) \Rightarrow\; x \in \{z : F(z) \geq U\} \Rightarrow\; x \geq \min\{z : F(z) \geq U\} = F^{-1}(U)$ $\Rightarrow F^{-1}(U) \leq x$.

□

We will use the inverse-CDF method to simulate random variables from a few distributions in the following examples:

Example 4. Unif (a, b):

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b]; \\ 0, & \text{otherwise.} \end{cases}$$

$$F(x) = \int_a^x \frac{1}{b-a} du = \frac{x-a}{b-a}$$

$$u = F(x) = \frac{x-a}{b-a} \implies \text{solve for } x\text{: } x = a + (b-a)u := F^{-1}(u).$$

1. Generate $U \sim \text{Unif}(0,1)$;
2. Let $X = a + (b-a)U$, then $X \sim \text{Unif}(a,b)$.

Check: $U \sim$ Unif $(0,1) \;\Rightarrow\; (b-a)U \sim$ Unif$(0,\,(b-a))$
$\Rightarrow\; a + (b-a)U \sim$ Unif (a,b).

Example 5. Exponential Distribution $\exp(\lambda)$.

$f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$,

$$F(x) = \int_0^x \lambda e^{-\lambda u} du = 1 - e^{-\lambda x} = u \;\Rightarrow\; x = -\frac{1}{\lambda}\log(1-u).$$

1. Generate $U \sim$ Unif$(0,\,1)$; [$(1-U \sim$Unif$(0,\,1))$]
2. Let $X = -\frac{1}{\lambda}\log U \sim \exp(\lambda)$.

Example 6. Geometric Distribution $\text{Ge}(p)$.

Let X be the number of trials until the first success in a sequence of independent $\text{Bern}(p)$ trials. Then

$$P(X = k) = (1-p)^{k-1}p = q^{k-1}p, \qquad k = 1, 2, \cdots,$$

where $q = 1 - p$. To find the c.d.f., note that

$$P(X \geq k) = q^{k-1}$$

is the probability that at least k trials are performed (first $k-1$ all failures). Then we have

$$P(X \leq k) = 1 - P(X \geq k+1) = 1 - q^k,$$

which shows that the c.d.f.

$$F(x) = P(X \leq x) = P(X \leq [x]) = 1 - q^{[x]}, \quad [x]: \text{ integer part of } x.$$

Now we work out the inverse c.d.f. defined by (17):

$$\begin{aligned} F^{-1}(U) &= \min\{z : F(z) \geq U\} \\ &= \min\{z : 1 - q^{[z]} \geq U\}. \end{aligned}$$

$$1 - q^{[z]} \geq U \;\Rightarrow\; [z] \geq \frac{\log(1-U)}{\log q}.$$

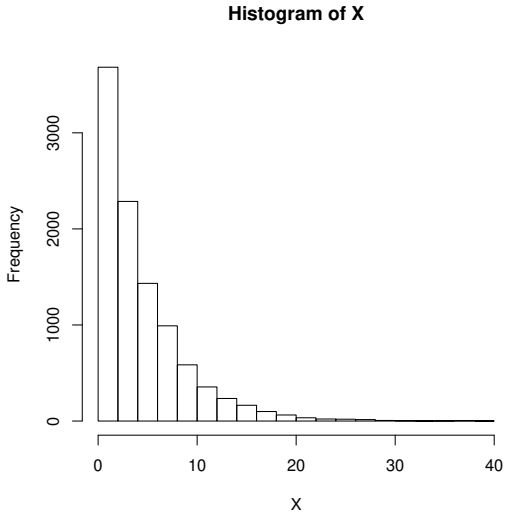
$F^{-1}(U) = \min\left\{z : [z] \geq \dfrac{\log(1-U)}{\log q}\right\} = \left[\dfrac{\log(1-U)}{\log q}\right] + 1$. (We can ignore the case that $\log(1-U)/\log q$ is an integer. Why?)

If $U \sim \text{Unif}(0,1)$, then $X = F^{-1}(U) = \left[\dfrac{\log(1-U)}{\log q}\right] + 1 \sim \text{Ge}(p)$:

1. $U \sim \text{Unif}(0,\, 1)$;
2. Let $X = [\log U / \log q] + 1 \sim \text{Ge}(p)$.

The following code simulates $n = 10,000$ samples from $\text{Ge}(0.2)$ and use the samples to estimate $\mathbb{E}(X) = 1/p = 5$ and $P(X = k)$ for $k = 1, \ldots, 10$:

```
> #input
> n=10000;p=0.2;q=1-p;
>
> #inverse cdf
> U=runif(n,0,1);
> X=floor(log(U)/log(q))+1;
>
> #estimates
> m_est=mean(X); #sample mean
> m_t=1/p; #true mean
> pr=numeric(10);
> pr_est=numeric(10);
> for(k in 1:10){
+     pr_est[k]=sum(X==k)/n; #estimated pr
+     pr[k]=q^(k-1)*p; # true pr
+ }
>
> m_t
[1] 5
> m_est
[1] 5.027
> pr
 [1] 0.20000000 0.16000000 0.12800000 0.10240000 0.08192000 0.06553600 0.05242880
 [8] 0.04194304 0.03355443 0.02684355
> pr_est
 [1] 0.1943 0.1586 0.1217 0.1068 0.0897 0.0697 0.0541 0.0415 0.0327 0.0250
>
```



4. Finite Discrete Distributions

$$\mathbb{P}(X = x_k) = p_k, \quad k = 1, 2, \cdots, m, \quad x_1 < x_2 < \cdots < x_m. \tag{18}$$

$$\sum_{k=1}^{m} p_k = 1, \quad p_k > 0. \tag{19}$$

4.1. Bernoulli Distribution

$X \sim \text{Bern}(p)$. $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$.

1. Generate $U \sim \text{Unif}(0, 1)$;
2. If $U \leq p$, $X = 1$; otherwise, $X = 0$.

Proof. $P(X = 1) = P(U \leq p) = p$, $P(X = 0) = P(U > p) = 1 - p$. □

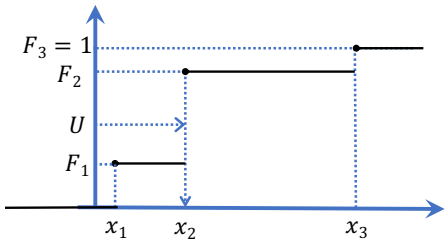
4.2. General Finite Discrete Distributions

$\mathbb{P}(X = x_k) = p_k, k \in [m] := \{1, \ldots, m\}$ as in (18).

Put $F_0 = 0$ and $F_k = \sum_{i=1}^{k} p_i$ for $k \in [m]$. Note $F_k = \mathbb{P}(X \leq x_k)$ and $F_m = 1$.

1. Generate $U \sim$Unif(0, 1);
2. If $F_{k-1} < U \leq F_k$, then $X = x_k$.

Proof. $P(X = x_k) = P(U \in (F_{k-1}, F_k]) = F_k - F_{k-1} = p_k$. □



This is in fact the inverse-cdf method: Let $I(x_i \leq z < x_{i+1})$ be the indicator function of $\{x_i \leq z < x_{i+1}\}$. Then the c.d.f. of X is $F(z) = F_i I(x_i \leq z < x_{i+1})$.

Thus, $F(z) = F_i I(x_i \leq z < x_{i+1}) \geq U \in (F_{k-1}, F_k]$ if and only if $z \geq x_k$. By Theorem 2, if $U \in (F_{k-1}, F_k]$,

$$X = F^{-1}(U) = \min\{z : F(z) \geq U\} = \min\{z : z \geq x_k\} = x_k.$$

Example 7. Suppose the joint distribution of X and Y is given by:

$X \backslash Y$	0	1
0	0.2	0.6
1	0.1	0.1

Then regard $x_1 = (0, 0)$, $x_2 = (0, 1)$, $x_3 = (1, 0)$ and $x_4 = (1, 1)$ and apply the same algorithm.

5. Composition Methods

5.1. Normal Distribution

5.1.1. Univariate Normal

Our goal is to simulate $X \sim \mathcal{N}(0,1)$. To do that, we consider two i.i.d. $X, Y \sim \mathcal{N}(0,1)$: Their joint pdf is

$$f_{XY}(x,y) = f_X(x) f_Y(y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right).$$

Consider polar coordinates: $\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases}$.

Jacobian of $(r,\theta) \mapsto (x,y)$ is

$$\det\left[\frac{\partial(x,y)}{\partial(r,\theta)}\right] = \det\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = r.$$

Apply change-of-variables: noting that $x^2+y^2 = r^2$ and $dxdy = rdrd\theta$,

$$\begin{aligned} f_{XY}(x,y)dxdy = \frac{1}{2\pi}\exp\Big(-\frac{r^2}{2}\Big) rdrd\theta &= \frac{1}{2\pi}d\theta \cdot \frac{1}{2} e^{-\frac{r^2}{2}} d(r^2) \\ &= f_{\Theta,R^2}(\theta, r^2) d(r^2) d\theta. \end{aligned}$$

Therefore, the density of (Θ, R^2) is

$$f_{\Theta,R^2}(\theta, r^2) = \left(\frac{1}{2\pi}\right) \cdot \left(\frac{1}{2} e^{-\frac{r^2}{2}}\right),$$

i.e. $\Theta \sim \text{Unif}(0, 2\pi)$ and $R^2 \sim \text{Exp}(1/2)$ are independent.

Now we can generate a pair of i.i.d. normal random variables by the following algorithm:

(1) Draw $\Theta \sim$ Unif $(0,\, 2\pi)$ and $R^2 \sim \text{Exp}(1/2)$ independently ($\Theta \perp\!\!\!\perp R^2$).

(2) Set $\begin{cases} X = \sqrt{R^2} \cdot \cos\Theta \\ Y = \sqrt{R^2} \cdot \sin\Theta \end{cases}$.

Then we have $X, Y \sim \mathcal{N}(0,1)$ and $X \perp\!\!\!\perp Y$.

Remark 1. We say a pdf $f(x)$ is spherically symmetric if

$$\|x\| = \|y\| \Rightarrow f(x) = f(y),$$

where $\|x\| = \sqrt{x_1^2 + x_2^2}$ is the Euclidean norm of $x = (x_1, x_2) \in \mathbb{R}^2$. In general, if a pdf $f(x)$ is spherically symmetric, usually it is convenient to use the polar coordinate system for simulation. This idea also applies to spherically symmetric tri-variate pdfs with $x \in \mathbb{R}^3$, for which we may consider using the spherical coordinate system for simulation.

5.1.2. Multivariate Normal

We want to simulate a random vector

$$X = (X_1, X_2, \cdots, X_p)^\top \sim \mathcal{N}(\mu, \Sigma),$$

where the mean vector and covariance matrix are

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{pmatrix}.$$

That is, $\mathbb{E}(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$ and

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \sigma_{ij}.$$

Lemma 3. *If $Z = (Z_1, \cdots, Z_p)^\top \sim \mathcal{N}(0, I_p)$ and A is an invertible $p \times p$ matrix, then $X = \mu + AZ \sim \mathcal{N}(\mu, AA^\top)$.*

Basic idea: Using the algorithm in Section 5.1.1, we can draw i.i.d. samples $Z_1, \ldots, Z_p$ from $\mathcal{N}(0, 1)$. Then put $Z = (Z_1, \ldots, Z_p)^\top$ and apply the transformation $X = \mu + AZ$, which gives us a random vector $X \sim \mathcal{N}(\mu, AA^\top)$ according to Lemma 3. If we choose A such that $AA^\top = \Sigma$, then we achieve the goal of simulating $X \sim \mathcal{N}(\mu, \Sigma)$. Cholesky decomposition is a way to find such an A given Σ:

Theorem 4 (Cholesky decomposition)**.** *If Σ is positive definite (and symmetric), there exists a unique lower triangular matrix $T = (t_{ij}), (t_{ij} = 0, i < j)$ with positive diagonal elements such that $\Sigma = TT^\top$.*

Algorithm to sample from $\mathcal{N}(\mu, \Sigma)$ (p-variate Normal):

1. Generate $Z_1, Z_2, \cdots, Z_p \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, and let $Z = (Z_1, \cdots, Z_p)^\top$;
2. Apply Cholesky decomposition of Σ to get a lower triangular matrix A such that $\Sigma = AA^\top$;
3. Let $X = \mu + AZ$. Then $X \sim \mathcal{N}(\mu, \Sigma)$.

Computation of Cholesky Decomposition

$B = (b_{ij})_{n\times n}$, B is symmetric and $B > 0$, $B = TT^\top$.

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} = \begin{bmatrix} t_{11} & & & 0 \\ t_{21} & t_{22} & & \\ \vdots & \vdots & \ddots & \\ t_{n1} & t_{n2} & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} t_{11} & t_{21} & \cdots & t_{n1} \\ & t_{22} & \cdots & t_{n2} \\ & & \ddots & \vdots \\ 0 & & & b_{nn} \end{bmatrix}$$

1. $t_{11} = \sqrt{b_{11}}$;
2. For $i = 2, \cdots, n$, $t_{i1} = b_{i1}/t_{11}$;
3. For $j = 2, \cdots, n$

$$t_{jj} = \sqrt{b_{jj} - \sum_{k=1}^{j-1} t_{jk}^2};$$

$$\text{for } i = j+1, \cdots, n, \quad t_{ij} = \left(b_{ij} - \sum_{k=1}^{j-1} t_{ik}t_{jk}\right)/t_{jj};$$

Example 8. Simulate from a tri-variate normal distribution:

$$\mu = (0,0,0)^\top \qquad \Sigma = \begin{bmatrix} 4 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 6 \end{bmatrix}.$$

First apply the Cholesky decomposition to find the lower-triangular matrix

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix}.$$

Then draw $Z_1, Z_2, Z_3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, put $Z = (Z_1, Z_2, Z_3)^\top$, and let $X = AZ$ (since $\mu = 0$). That is

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \begin{bmatrix} 2Z_1 \\ Z_1 + Z_2 \\ -Z_1 + 2Z_2 + Z_3 \end{bmatrix}, \tag{20}$$

which uses linear combinations of Z_1, Z_2, Z_3 to generate correlations among X_1, X_2, X_3.

The following code implements this method to generate $n = 1000$ samples. Note that the `chol` function in `R` returns an upper-triangular matrix (i.e. $A^\top$) so we need to use its transpose.

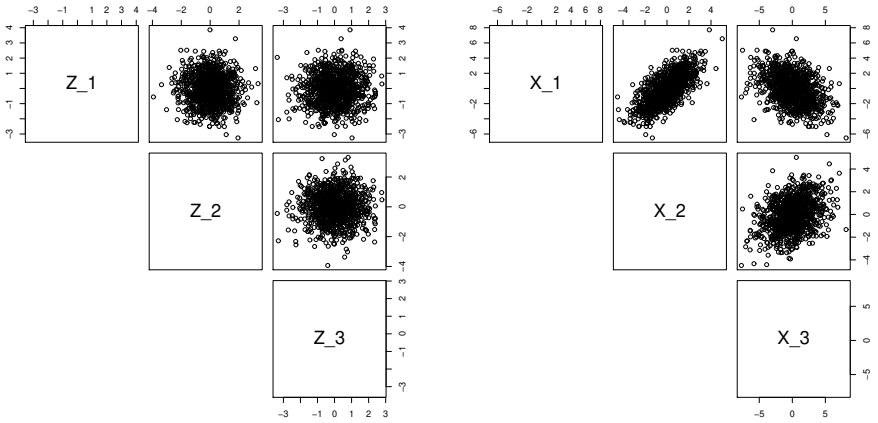
```
mu=c(0,0,0);
Sgm=matrix(c(4,2,-2,2,2,1,-2,1,6),nrow=3,ncol=3);

#Cholesky decomposition
A=t(chol(Sgm)); # A is lower triangular, A%*%t(A)=Sgm

n=1000; p=3;
Z=matrix(0,nrow=n,ncol=p);
X=matrix(0,nrow=n,ncol=p);

for(i in 1:n){
  Z[i,]=rnorm(p,mean=0,sd=1);
  X[i,]=mu+A%*%Z[i,];
}
```

The pairwise scatter plots for Z and X confirm that while Z_1, Z_2, Z_3 are independent, X_1, X_2, X_3 are indeed correlated due to the linear transformation (20).



In fact, the idea of using linear combinations of independent random variables to generate correlated random variables applies to any distributions, not just normal distributions. This is illustrated in the following example.

Example 9. Simulate two random variables such that their correlation coefficient is 0.8.

For simplicity, we assume both random variables X_1, X_2 have zero mean and unit variance. Then the covariance

$$\mathrm{Cov}(X_1, X_2) = \mathrm{cor}(X_1, X_2)\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)} = 0.8.$$

Thus the covariance matrix of (X_1, X_2) is

$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}.$$

Apply the Cholesky decomposition $\Sigma = AA^\top$ to get a lower-triangular matrix

$$A = \begin{bmatrix} 1 & 0 \\ 0.8 & 0.6 \end{bmatrix}.$$

First draw Z_1, Z_2 independently from some distributions with zero mean and unit variance, i.e. $\mathrm{Var}(Z_1) = \mathrm{Var}(Z_2) = 1$ and $Z_1 \perp Z_2$ (independence). Then put $Z = (Z_1, Z_2)^\top$ and let $X = AZ$:

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} Z_1 \\ 0.8Z_1 + 0.6Z_2 \end{bmatrix}.$$

One can easily verify that

$$\begin{aligned} \mathrm{Var}(X_1) &= \mathrm{Var}(Z_1) = 1, \\ \mathrm{Var}(X_2) &= \mathrm{Var}(0.8Z_1 + 0.6Z_2) = 0.8^2 + 0.6^2 = 1, \\ \mathrm{Cov}(X_1, X_2) &= \mathrm{Cov}(Z_1, 0.8Z_1 + 0.6Z_2) \\ &= \mathrm{Cov}(Z_1, 0.8Z_1) + \mathrm{Cov}(Z_1, 0.6Z_2) = 0.8\,\mathrm{Var}(Z_1) + 0 = 0.8, \end{aligned}$$

and thus

$$\mathrm{cor}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\,\mathrm{Var}(X_2)}} = 0.8.$$

5.2. Mixture Distributions

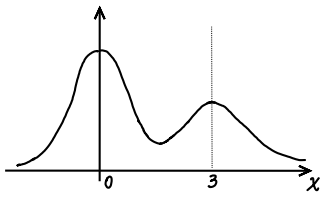
A mixture distribution has pdf

$$f(x) = \sum_{i=1}^{K} \theta_i f_i(x),$$

$f_i(x)$: pdf of a component distribution, $\int f_i dx = 1$.

θ_i : mixture proportion ($\theta_i > 0, \sum_{i=1}^K \theta_i = 1$).

Example 10. $f_1(x) : \mathcal{N}(0,1)$, $f_2(x) : \mathcal{N}(3, 2^2)$, $\theta_1 = \theta_2 = 1/2$.



$$X \sim f = \frac{1}{2} f_1 + \frac{1}{2} f_2$$

$$\Leftrightarrow X \sim \begin{cases} \mathcal{N}(0,1), & \text{with probability } 1/2 \\ \mathcal{N}(3, 2^2), & \text{with prpbability } 1/2 \end{cases}$$

Algorithm to draw from the mixture distribution f:

1. Generate $Z \sim$ Discrete $(\theta_1, \theta_2, \cdots, \theta_K)$; i.e., $P(Z = i) = \theta_i$ for $i = 1, \ldots, K$.
2. Generate $X \sim f_Z$ i.e. $X \sim f_i$ if $Z = i$.

To verify this algorithm, we need to confirm that the pdf of X, $p_X(x)$, is indeed $f(x)$. Note that the algorithm generates a pair of random variables (Z, X). Denote by $p_{X,Z}(x, i)$ the joint pdf of (X, Z). Then from the algorithm,

$$p_{X,Z}(x, i) = \mathbb{P}(Z = i) p_{X|Z}(x \mid i) = \theta_i f_i(x).$$

Accordingly, the distribution of X is given by marginalizing out the discrete random variable Z:

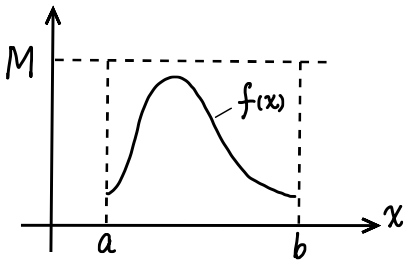
$$p_X(x) = \sum_{i=1}^{K} p_{X,Z}(x, i) = \sum_{i=1}^{K} \theta_i f_i(x) = f(x).$$

This shows that the pdf of X is $f(x)$.

6. Rejection Sampling

Suppose $f(x)$ is a pdf defined on $[a, b]$ and there exists $M \geq f(x)$ for all $x \in [a, b]$. Consider the following three-step algorithm:

(1) Draw $X \sim$ Unif (a, b), compute $r(X) = f(X)/M \in [0, 1]$.
(2) Draw $U \sim$ Unif (0, 1).
(3) If $U \leq r(X)$, accept X; otherwise, repeat (1) and (2).



Lemma 5. *If X is accepted in the above algorithm, then its pdf is $f(x)$.*

Proof. Want to show the conditional density $p_X(x \mid X \text{ is accepted}) = f(x)$. Let $p_X(x)$ denote the marginal probability density function of X. By Bayes rule:

$$p_X(x|\text{Accepted}) = \frac{P(X \text{ is accepted}|X = x)p_X(x)}{P(X \text{ is accepted})}. \tag{21}$$

By step (1), $X \sim$ Unif (a, b) and thus,

$$p_X(x) = 1/(b - a) \quad \text{for any } x \in [a, b]. \tag{22}$$

According to steps (2) and (3), X is accepted if and only if $U \leq r(X)$. Therefore,

$$P(X \text{ is accepted}|X = x) = P(U \leq r(x)) = r(x) = f(x)/M, \tag{23}$$

as $U \sim \text{Unif}(0, 1)$ and $r(x) \in [0, 1]$. Now note that

$$\begin{aligned} P(X \text{ is accepted}) &= \int_a^b P(X \text{ is accepted}|X = x)p_X(x)dx \\ &= \int_a^b \frac{f(x)}{M} \cdot \frac{1}{b-a} dx \\ &= \frac{1}{M(b-a)} \int_a^b f(x)dx = \frac{1}{M(b-a)}, \end{aligned} \tag{24}$$

since $\int_a^b f(x)dx = 1$ as a pdf. Now plugging (22), (23) and (24) into (21), we have

$$\begin{aligned} p_X(x|\text{Accepted}) &= \frac{P(X \text{ is accepted}|X = x)p_X(x)}{P(X \text{ is accepted})} \\ &= \frac{\frac{f(x)}{M} \cdot \frac{1}{b-a}}{\frac{1}{M(b-a)}} = f(x). \end{aligned}$$

□

Efficiency of this algorithm depends on the acceptance rate given in (24):

$$P(\text{Acceptance}) = \frac{1}{M(b-a)}.$$

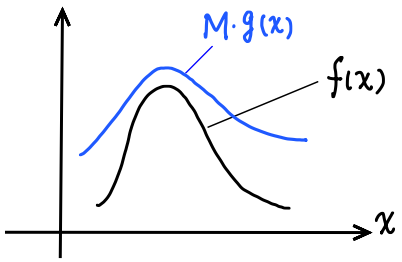
The higher the acceptance rate, the more efficient. Thus we shall choose the smallest M such that $f(x) \le M$ for all $x \in [a, b]$.

Rejection Sampling:

Let f be a pdf defined on D. Our goal is to simulate X from the distribution specified by the pdf $f(x)$. Let g be another pdf such that there is $M \ge f(x)/g(x)$ for all $x \in D$. The rejection sampling method is:

(1) Draw $X \sim g$, compute $r(X) = \dfrac{f(X)}{Mg(X)} \in [0, 1]$.
(2) Draw $U \sim$ Unif (0, 1).
(3) If $U \le r(X)$, accept X; otherwise, repeat (1) and (2).

Then the accepted X has pdf $f(x)$. [f is called the target distribution; g is called a trial distribution.]



Theorem 6. *If X is accepted in the rejection sampling method, then its pdf is $f(x)$.*

Proof. Let "Acceptance" be the event that X is accepted.

$$\begin{aligned}\mathbb{P}(\text{Acceptance}) &= \int \mathbb{P}(\text{Acceptance}|X=x)g(x)dx \\ &= \int \frac{f(x)}{Mg(x)}g(x)dx = \frac{1}{M}. \end{aligned} \tag{25}$$

Then

$$\begin{aligned}p_X(x|\text{Acceptance}) &= \frac{\mathbb{P}(\text{Acceptance}|X=x)g(x)}{\mathbb{P}(\text{Acceptance})} \\ &= \frac{(f(x)/Mg(x))\cdot g(x)}{1/M} = f(x).\end{aligned}$$

□

Efficiency of the rejection sampling algorithm depends on the acceptance rate given in (25):

$$\mathbb{P}(\text{Acceptance}) = \frac{1}{M}.$$

The higher the acceptance rate, the more efficient. Thus we shall choose the smallest M such that $f(x)/g(x) \le M$ for all $x \in D$. Therefore, a common choice is

$$M = \max_{x\in D} \frac{f(x)}{g(x)}. \tag{26}$$

Example 11. Use rejection sampling to simulation from the distribution with pdf $f(x) = \frac{1}{2}\sin x$, $x \in (0,\pi)$. [Verify this is indeed a pdf.]

Let $g(x)$ be the pdf of Unif $(0,\pi)$, so $g(x) = \frac{1}{\pi}$ for $x \in (0,\pi)$, which serves as our trial distribution.

Let

$$M = \max_{0<x<\pi} \frac{f(x)}{g(x)} = \max_{0<x<\pi} \frac{\pi}{2}\sin(x) = \frac{\pi}{2}.$$

Then $g(x)\cdot M = 1/2 \ge f(x) \quad \forall x \in (0,\pi)$.

Algorithm: Each iteration contains three steps; accepted X are the samples from $f(x)$.

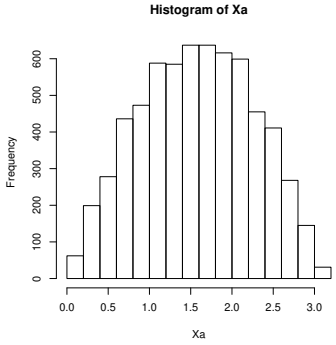
1. Draw $X \sim$ Unif$(0,\pi)$, and compute $r(X) = \frac{f(X)}{M\cdot g(X)} = \sin(X) \le 1$.
2. Draw $U \sim$ Unif$(0,1)$.
3. If $U \le \sin(X)$, then accept X; otherwise reject.

Acceptance probability:

$$\mathbb{P}(\text{Accetance}) = \frac{1}{M} = \frac{2}{\pi} \approx 0.64.$$

The following code implements this example with $n = 10,000$ trial samples, and $n_a = 6,420$ were accepted. The histogram of the accepted samples `Xa` shows that the pdf is $f(x) = \frac{1}{2}\sin x$, $x \in (0, \pi)$.

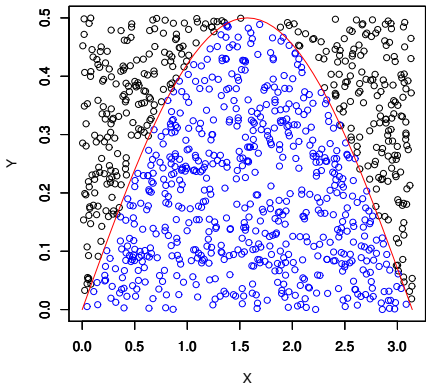
```
> n=10000; # sample size of trial (before acceptance)
> X=runif(n,0,pi); #X~ g
> r=sin(X); #r(X)
> U=runif(n,0,1);
> Xa=X[U<=r]; #acceptance
> na=sum(U<=r); # size of accepted samples
> Pa=na/n; #acceptance rate
>
> hist(Xa)
> na
[1] 6420
> Pa
[1] 0.642
>
```



Let us illustrate the intuition behind rejection sampling using this example: The acceptance criterion is $U \le r(X) = f(X)/[Mg(X)] \Leftrightarrow UMg(X) \le f(X)$. Putting $Y = UMg(X)$, then X will be accepted if $Y \le f(X)$, i.e., if the random point (X, Y) is under the curve $y = f(x)$. In this example,

$$X \sim \text{Unif}(0, \pi), \quad Y = U/2 \sim \text{Unif}(0, 1/2) \Rightarrow (X, Y) \sim \text{Unif}([0, \pi] \times [0, 1/2]).$$

In the following plot, the red curve is $y = f(x)$ (pdf). The black dots are rejected and the blue dots are accepted. The blue dots are uniformly distributed under the red curve $y = f(x)$, and their x-coordinates (accepted X) has pdf $f(x)$.



Code for the plot:

```
Y=U/2; #Y=U*[M*g(X)]
f=0.5*sin(X); #f(X)
plot(X[Y>f],Y[Y>f],col="black",xlim=c(0,pi),ylim=c(0,0.5),xlab="X",ylab="Y");
par(new=T)
plot(X[Y<=f],Y[Y<=f],col="blue",xlim=c(0,pi),ylim=c(0,0.5),xlab="",ylab="");
par(new=T)
x=seq(0,pi,by=0.02*pi);
y=0.5*sin(x); # pdf y=f(x)
plot(x,y,type="l",col="red",xlim=c(0,pi),ylim=c(0,0.5),xlab="",ylab="")
```

Example 12. Absolute normal Distribution.

$Z \sim \mathcal{N}(0,1)$. Let $X = |Z|$, then $f_X(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}, \quad (x \geq 0)$.

This is because

$$f_X(x) = \phi(x) + \phi(-x) = 2 \cdot \phi(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}},$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ is the pdf for $\mathcal{N}(0,1)$.

Let us choose $\text{Exp}(\lambda = 1)$ as a trial distribution, i.e. $g(x) = e^{-x}, \ x \geq 0$.

Then the minimum M is $M = \max_{x \geq 0} \frac{f(x)}{g(x)} = \max \sqrt{\frac{2}{\pi}} \ \frac{e^{-\frac{x^2}{2}}}{e^{-x}} = \sqrt{\frac{2}{\pi}} \ e^{-(\frac{x^2}{2} - x)}$

$\Leftrightarrow \min_{x>0}(\frac{x^2}{2} - x) \ \Rightarrow \ x^* = 1$

$\therefore M = \sqrt{\frac{2e}{\pi}} \approx 1.32 \ (1/M \approx 0.76)$

Example 13. Truncated Normal.

$\phi(x)$: density of $\mathcal{N}(0,1)$.

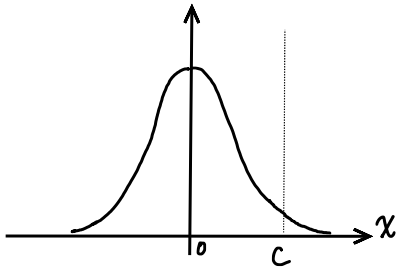
$$f(x) \propto \phi(x) I(x > c) = \begin{cases} \phi(x), & \text{if } x > c; \\ 0, & \text{otherwise.} \end{cases} \qquad \text{truncated normal}$$

$$\int \phi(x) I(x > c) dx = \int_c^\infty \phi(x) dx = 1 - \Phi(c), \quad \Phi(x) \text{: } c.d.f. \text{ of } N(0,1).$$

$\Rightarrow f(x) = \frac{1}{1-\Phi(c)} \phi(x) I(x > c)$

(A) If $c < 0$, generate $Z \sim \mathcal{N}(0,1)$. Accept Z if $Z > c$. $P(\text{Acceptance}) \geq 0.5$

(B) If $c \to \infty$, $P(\text{Acceptance}) = 1 - \Phi(c) \to 0$.



Rejection sampling: Use shifted Exp(λ) as the trial distribution, so

$$g(x) = \begin{cases} \lambda e^{-\lambda(x-c)}, & \text{for } x > c; \\ 0, & \text{otherwise.} \end{cases}$$

If $X' \sim \text{Exp}(\lambda)$, then $X = X' + c \sim g$. To find the optimal M:

$$M = \max_{x>c} \frac{f(x)}{g(x)} = \max_{x>c} \frac{\phi(x)}{1-\Phi(c)} \frac{e^{\lambda(x-c)}}{\lambda} = \frac{\max_{x>c} \exp(-\frac{x^2}{2} + \lambda x - \lambda c)}{\sqrt{2\pi}\lambda(1-\Phi(c))}.$$

This is equivalent to

$$\max_{x>c} \Big(-\frac{x^2}{2} + \lambda x - \lambda c \Big),$$

for which the maximizer $x^* = \lambda$ if $\lambda > c$. Plugging in $x = \lambda \Rightarrow$

$$M = \frac{\exp(\frac{\lambda^2}{2} - \lambda c)}{\sqrt{2\pi}\lambda(1-\Phi(c))}, \quad \lambda > c.$$

Since $M = M(\lambda)$ depends on λ, to maximize $P(\text{Acceptance}) = 1/M(\lambda)$, we choose $\lambda > c$ such that $M(\lambda)$ is minimized:

$$\Rightarrow \lambda^* = \frac{c + \sqrt{c^2+4}}{2} \; (> c).$$

Thus, we choose $g(x) = \lambda^* e^{-\lambda^*(x-c)}$ as the trial distribution.

Under this design, $P(\text{Acceptance}) = 1/M(\lambda^*) = 0.76, 0.88, 0.93$ for $c = 0, 1, 2$, respectively.

Remark 2. In fact, we do not need to know the normalizing constant to do rejection sampling. Suppose we want to draw from $p(x) = f(x)/Z$, where $Z = \int f dx < \infty$ is the normalizing constant, and $f(x)$ is given but Z is unknown or cannot be computed easily. We can apply the same rejection sampling method with $f(x)$ (unnormalized). *Then the distribution of an accepted sample X is $p(x)$.* This can be shown by modifying the proof of Theorem 6: The normalizing constant Z cancels when calculating $p_X(x \mid X \text{ is accepted})$.