One way to derive the saddlepoint approximation is to use an *Edgeworth expansion* (see Hall 1992 or Reid 1988 for details). As a result of a quite detailed derivation, we obtain the approximation to the density of $\bar{X}$ to be

$$f_{\bar{X}}(x) = \frac{\sqrt{n}}{\sigma}\varphi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right) \times \left[1 + \frac{\kappa}{6\sqrt{n}}\left\{\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)^3 - 3\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)\right\} + \mathcal{O}(1/n)\right]. \tag{3.28}$$

Ignoring the term within braces produces the usual normal approximation, which is accurate to $\mathcal{O}(1/\sqrt{n})$. If we are using (3.28) for values of x near μ, then the value of the expression in braces is close to zero, and the approximation will then be accurate to $\mathcal{O}(1/n)$. The trick of the saddlepoint approximation is to make this always be the case.

To do so, we use a family of densities such that, for each x, we can choose a density from the family to cancel the term in braces in (3.28). One method of creating such a family is through a technique known as *exponential tilting* (see Efron 1981, Stuart and Ord 1987, Section 11.13, Reid 1988, or Problem 3.37). The result of the exponential tilt is a family of Edgeworth expansions for $f_{\bar{X}}(x)$ indexed by a parameter τ, that is,

$$f_{\bar{X}}(x) = \exp\{-n[\tau x - K(\tau)]\}\frac{\sqrt{n}}{\sigma_\tau}\varphi\left(\frac{x-\mu_\tau}{\sigma_\tau/\sqrt{n}}\right) \times \left[1 + \frac{\kappa_\tau}{6\sqrt{n}}\left\{\left(\frac{x-\mu_\tau}{\sigma_\tau/\sqrt{n}}\right)^3 - 3\left(\frac{x-\mu_\tau}{\sigma_\tau/\sqrt{n}}\right)\right\} + \mathcal{O}(1/n)\right]. \tag{3.29}$$

As the parameter τ is free for us to choose in (3.29), for each x we choose $\tau = \tau(x)$ so that the mean satisfies $\mu_\tau = x$. This choice cancels the middle term in the square brackets in (3.29), thereby improving the order of the approximation. If $K(\tau) = \log(\mathbb{E}\exp(\tau X))$ is the cumulant generating function, we can choose τ so that $K'(\tau) = x$, which is the saddlepoint equation. Denoting this value by $\hat{\tau} = \hat{\tau}(x)$ and noting that $\sigma_\tau = K''(\tau)$, we get the saddlepoint approximation

$$\begin{aligned} f_{\bar{X}}(x) &= \frac{\sqrt{n}}{\sigma_{\hat{\tau}}}\varphi(0)\exp\{n[K(\hat{\tau}) - \hat{\tau}x]\}\,[1 + \mathcal{O}(1/n)] \\ &\approx \left(\frac{n}{2\pi\, K''(\hat{\tau}(x))}\right)^{1/2} \exp\left\{n\left[K\left(\hat{\tau}(x)\right) - \hat{\tau}(x)x\right]\right\}. \end{aligned} \tag{3.30}$$

Example 3.18. Saddlepoint tail area approximation. The noncentral chi squared density has the rather complex form

$$f(x|\lambda) = \sum_{k=0}^{\infty} \frac{x^{p/2+k-1}e^{-x/2}}{\Gamma(p/2+k)2^{p/2+k}} \frac{\lambda^k e^{-\lambda}}{k!}, \tag{3.31}$$

where p is the number of degrees of freedom and λ is the noncentrality parameter. It turns out that calculation of the moment generating function is simple, and it can be expressed in closed form as

$$\phi_X(t) = \frac{e^{2\lambda t/(1-2t)}}{(1-2t)^{p/2}}. \tag{3.32}$$

where the $Y_i^{(m)}$ are iid from t, as follows:

$$\hat{I} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{I}_{[0,\infty[}(J^{(m)}) e^{-n\theta_0 J^{(m)}} \lambda(\theta_0)^n, \tag{3.26}$$

with $\lambda(\theta) = \int f(x)e^{\theta_0 f(x)}dx$. The fact that (3.26) is unbounded follows from a regular importance sampling argument (Problem 3.36). Bucklew (1990) provides arguments about the fact that the variance of $\hat{I}$ goes to 0 exponentially twice as fast as the regular (direct sampling) estimate.

Example 3.17. Laplace distribution. Consider $h(x) = x$ and the sampling distribution $f(x) = \frac{1}{2a}\exp\{-|x-\mu|/a\}$, $\mu < 0$. We then have

$$\begin{aligned} t(x) &\propto \exp\{-|x-\mu|/a + \theta_0\}, \\ \theta_0 &= \sqrt{\mu^{-2} + a^{-2}} - \mu^{-1}, \\ \lambda(\theta_0) &= \frac{\mu^2}{2}\exp(-C)a^2 C\,, \end{aligned}$$

with $C = \sqrt{1 + \frac{\mu^2}{a^2}} - 1$. A large deviation computation then shows that (Bucklew 1990, p. 139)

$$\lim_n \frac{1}{n}\log(M\text{var}\hat{I}) = 2\log\lambda(\theta_0),$$

while the standard average $\bar{I}$ satisfies

$$\lim_n \frac{1}{n}\log(M\text{var}\bar{I}) = \log\lambda(\theta_0)\,. \quad \|$$

Obviously, this is not the entire story, Further improvements can be found in the theory, while the computation of θ_0 and $\lambda(\theta_0)$ and simulation from $t(x)$ may become quite intricate in realistic setups.

3.6.2 The Saddlepoint Approximation

The saddlepoint approximation, in contrast to the Laplace approximation, is mainly a technique for approximating a function rather than an integral, although it naturally leads to an integral approximation. (For introductions to the topic see Goutis and Casella 1999, the review papers of Reid 1988, 1991, or the books by Field and Ronchetti 1990, Kolassa 1994, or Jensen 1995.)

Suppose we would like to evaluate

$$g(\theta) = \int_A f(x|\theta)dx \tag{3.27}$$

for a range of values of θ. One interpretation of a *saddlepoint approximation* is that for each value of θ, we do a Laplace approximation centered at $\hat{x}_\theta$ (the *saddlepoint*).[5]

[5] The saddlepoint approximation got its name because its original derivation (Daniels 1954) used a complex analysis argument, and the point $\hat{x}_\theta$ is a saddlepoint in the complex plane.

(b) If $X_1, X_2, \ldots, X_n$ is an iid sample from $f(x)$, and $f_{\bar{X}}(x)$ is the density of the sample mean, show that $\int e^{n[\tau x - K(\tau)]} f_{\bar{X}}(x)dx = 1$ and hence $f_{\bar{X}}(x|\tau)$ of (3.24) is a density.

(c) Show that the mgf of $f_{\bar{X}}(x|\tau)$ is $e^{n[K(\tau+t/n)x - K(\tau)]}$

(d) In (3.29), for each x we choose τ so that μ_τ, the mean of $f_{\bar{X}}(x|\tau)$, satisfies $\mu_\tau = x$. Show that this value of τ is the solution to the equation $K'(\tau) = x$.

3.38 For the situation of Example 3.18:

(a) Verify the mgf in (3.32).

(b) Show that the solution to the saddlepoint equation is given by (3.33).

(c) Plot the saddlepoint density for $p = 7$ and $n = 1, 5, 20$. Compare your results to the exact density.

3.6 Notes

3.6.1 Large Deviations Techniques

When we introduced importance sampling methods in Section 3.3, we showed in Example 3.8 that alternatives to direct sampling were preferable when sampling from the tails of a distribution f. When the event A is particularly rare, say $P(A) \leq 10^{-6}$, methods like importance sampling are needed to get an acceptable approximation (see Problem 3.35). Since the optimal choice given in Theorem 3.12 is formal, in the sense that it involves the unknown constant I, more practical choices have been proposed in the literature. In particular, Bucklew (1990) indicates how the *theory of large deviations* may help in devising proposal distributions in this purpose.

Briefly, the theory of large deviations is concerned with the approximation of tail probabilities $P(|\bar{X}_n - \mu| > \varepsilon)$ when $\bar{X}_n = (X_1 + \cdots + X_n)/n$ is a mean of iid random variables, n goes to infinity, and ε is large. (When ε is small, the normal approximation based on the Central Limit Theorem works well enough.)

If $M(\theta) = \mathbb{E}[\exp(\theta X_1)]$ is the moment generating function of X_1 and we define $I(x) = \sup_\theta \{\theta x - \log M(\theta)\}$, the large deviation approximation is

$$\frac{1}{n} \log P(S_n \in F) \approx - \inf_F I(x).$$

This result is sometimes called *Cramér's Theorem* and a simulation device based on this result and called *twisted simulation* is as follows.

To evaluate

$$I = P\left(\frac{1}{n} \sum_{i=1}^{n} h(x_i) \geq 0 \right),$$

when $\mathbb{E}[h(X_1)] < 0$, we use the proposal density

$$t(x) \propto f(x) \exp\{\theta_0 h(x)\}, \tag{3.25}$$

where the parameter θ_0 is chosen such that $\int h(x) f(x) e^{\theta_0 h(x)} dx = 0$. (Note the similarity with exponential tilting in saddlepoint approximations in Section 3.6.2.) The corresponding estimate of I is then based on blocks ($m = 1, \ldots, M$)

$$J^{(m)} = \frac{1}{n} \sum_{i=1}^{n} h(Y_i^{(m)}),$$

(a) Show that conditionally on t, the joint density of $(Y_1, \ldots, Y_t)$ is indeed

$$\prod_{j=1}^{t-1} \left(\frac{Mg(y_j) - f(y_j)}{M-1} \right) f(y_t)$$

and the expectation of δ_2 of (3.15) is given by

$$\mathbb{E}\left[\frac{t-1}{t}\left\{\frac{M}{M-1}\,\mathbb{E}_f[h(X)] - \frac{1}{M-1}\,\mathbb{E}_f\left[h(X)\frac{f(X)}{g(X)}\right]\right\} + \frac{1}{t}\,\mathbb{E}_f\left[h(X)\frac{f(X)}{g(X)}\right]\right].$$

(b) If we denote the acceptance probability of the Accept–Reject algorithm by $\rho = 1/M$ and assume $\mathbb{E}_f[h(X)] = 0$, show that the *bias* of δ_2 is

$$\left(\frac{1}{1-\rho}\,\mathbb{E}[t^{-1}] - \frac{\rho}{1-\rho}\right)\,\mathbb{E}_f\left[h(X)\frac{f(X)}{g(X)}\right].$$

(c) Establish that for $t \sim \mathcal{G}eo(\rho)$, $\mathbb{E}[t^{-1}] = -\rho\log(\rho)/(1-\rho)$, and that the bias of δ_2 can be written as

$$-\frac{\rho}{1-\rho}\,(1+\log(\rho))\,\mathbb{E}_f\left[h(X)\frac{f(X)}{g(X)}\right].$$

(d) Assuming that $\mathbb{E}_f[h(X)] = 0$, show that the variance of δ_2 is

$$\mathbb{E}\left[\frac{t-1}{t^2}\right]\,\frac{1}{1-\rho}\mathbb{E}_f\left[h^2(X)\frac{f(X)}{g(X)}\right] + \mathbb{E}\left[\frac{1}{t^2}\left\{1 - \frac{\rho(t-1)}{1-\rho}\right\}\right]\,\mathrm{var}_f(h(X)f(X)/g(X)).$$

3.35 Using the information from Note 3.6.1, for a binomial experiment $X_n \sim \mathcal{B}(n,p)$ with $p = 10^{-6}$, determine the minimum sample size n so that

$$P\left(\left|\frac{X_n}{n} - p\right| \le \epsilon p\right) > .95$$

when $\epsilon = 10^{-1}, 10^{-2}$, and 10^{-3}.

3.36 When random variables Y_i are generated from (3.25), show that $J^{(m)}$ is distributed as $\lambda(\theta_0)^{-n}\exp(-n\theta J)$. Deduce that (3.26) is unbiased.

3.37 Starting with a density f of interest, we create the exponential family

$$\mathcal{F} = \{f(\cdot|\tau); f(x|\tau) = \exp[\tau x - K(\tau)]f(x)\}\ ,$$

where $K(\tau)$ is the cumulant generating function of f given in Section 3.6.2. It immediately follows that if $X_1, X_2, \ldots, X_n$ are iid from $f(x|\tau)$, the density of $\bar{X}$ is

$$f_{\bar{X}}(x|\tau) = \exp\{n[\tau x - K(\tau)]\}\, f_{\bar{X}}(x), \tag{3.24}$$

where $f_{\bar{X}}(x)$ is the density of the average of an iid sample from f

(a) Show that $f(x|\tau)$ is a density.

and that the joint distribution of (Z_i, Z_j) $(1 \le i \ne j < n+t)$ is

$$\frac{(n-1)(n-2)}{(n+t-1)(n+t-2)} f(z_i) f(z_j)$$
$$+\frac{(n-1)t}{(n+t-1)(n+t-2)} \left\{ f(z_i) \frac{g(z_j) - \rho f(z_j)}{1-\rho} \frac{g(z_i) - \rho f(z_i)}{1-\rho} f(z_j) \right\}$$
$$+\frac{n(n-1)}{(n+t-1)(n+t-2)} \frac{g(z_i) - \rho f(z_i)}{1-\rho} \frac{g(z_j) - \rho f(z_j)}{1-\rho}.$$

3.32 (Continuation of Problem 3.31) If $Z_1, \ldots, Z_{n+t}$ is the sample produced by an Accept–Reject algorithm to generate n values, based on (f, g, M), show that the Z_i's are negatively correlated in the sense that for every square integrable function h,

$$\begin{aligned} \operatorname{cov}(h(Y_i), h(Y_j)) &= -\mathbb{E}_g[h]^2 \mathbb{E}_N \left[\frac{(t-1)(n-t)}{(n-1)^2(n-2)} \right] \\ &= -\mathbb{E}_g[h]^2 \{ \rho^t {}_2F_1(t-1, t-1; t-1; 1-\rho) - \rho^2 \}, \end{aligned}$$

where ${}_2F_1(a, b; c; z)$ is the confluent hypergeometric function (see Abramowitz and Stegun 1964 or Problem 1.38).

3.33 Given an Accept–Reject algorithm based on (f, g, ρ), we denote by

$$b(y_j) = \frac{(1-\rho) f(y_j)}{g(y_j) - \rho f(y_j)}$$

the importance sampling weight of the rejected variables $(Y_1, \ldots, Y_t)$, and by $(X_1, \ldots, X_n)$ the accepted variables.

(a) Show that the estimator

$$\delta_1 = \frac{n}{n+t} \delta^{AR} + \frac{t}{n+t} \delta_0,$$

with

$$\delta_0 = \frac{1}{t} \sum_{j=1}^{t} b(Y_j) h(Y_j)$$

and

$$\delta^{AR} = \frac{1}{n} \sum_{i=1}^{n} h(X_i),$$

does not uniformly dominate δ^{AR}. (*Hint:* Consider the constant functions.)

(b) Show that

$$\delta_{2w} = \frac{n}{n+t} \delta^{AR} + \frac{t}{n+t} \sum_{j=1}^{t} b(Y_j) h(Y_j) \Big/ \sum_{j=1}^{t} b(Y_j)$$

is asymptotically equivalent to δ_1 in terms of bias and variance.

(c) Deduce that δ_{2w} asymptotically dominates δ^{AR} if (4.20) holds.

3.34 For the Accept–Reject algorithm of Section 3.3.3:

where $w_i = f(y_i)/Mg(y_i)$ and the sum is over all subsets of $\{1, \ldots, n-1\}$ of size $t-1$.

(b) There is also interest in the joint distribution of $(Y_i, U_i)|N = n$, for any $i = 1, \ldots, n-1$, as we will see in Problem 4.17. Since this distribution is the same for each i, we can just derive it for (Y_1, U_1). (Recall that $Y_n \sim f$.) Show that

$$P(N = n, Y_1 \le y, U_1 \le u_1)$$
$$= \binom{n-1}{t-1}\left(\frac{1}{M}\right)^{t-1}\left(1-\frac{1}{M}\right)^{n-t-1}$$
$$\times\left[\frac{t-1}{n-1}(w_1 \wedge u_1)\left(1-\frac{1}{M}\right)+\frac{n-t}{n-1}(u_1-w_1)^+\left(\frac{1}{M}\right)\right]\int_{-\infty}^{y} g(t_1)dt_1.$$

(c) Show that part (b) yields the negative binomial marginal distribution of N,

$$P(N = n) = \binom{n-1}{t-1}\left(\frac{1}{M}\right)^{t}\left(1-\frac{1}{M}\right)^{n-t},$$

the marginal distribution of Y_1, $m(y)$,

$$m(y) = \frac{t-1}{n-1}f(y) + \frac{n-t}{n-1}\frac{g(y)-\rho f(y)}{1-\rho},$$

and

$$P(U_1 \le w(y)|Y_1 = y, N = n) = \frac{g(y)w(y)M\frac{t-1}{n-1}}{m(y)}.$$

3.30 If $(Y_1, \ldots, Y_N)$ is the sample produced by an Accept–Reject method based on (f, g), where $M = \sup(f/g)$, $(X_1, \ldots, X_t)$ denotes the accepted subsample and $(Z_1, \ldots, Z_{N-t})$ the rejected subsample.

(a) Show that both

$$\delta_2 = \frac{1}{N-t}\sum_{i=1}^{N-t} h(Z_i)\frac{(M-1)f(Z_i)}{Mg(Z_i)-f(Z_i)}$$

and

$$\delta_1 = \frac{1}{t}\sum_{i=1}^{t} h(X_i)$$

are unbiased estimators of $I = \mathbb{E}_f[h(X)]$ (when $N > t$).

(b) Show that δ_1 and δ_2 are independent.

(c) Determine the optimal weight β^* in $\delta_3 = \beta\delta_1 + (1-\beta)\delta_2$ in terms of variance. (*Note*: β may depend on N but not on $(Y_1, \ldots, Y_N)$.)

3.31 Given a sample $Z_1, \ldots, Z_{n+t}$ produced by an Accept–Reject algorithm to accept n values, based on (f, g, M), show that the distribution of a rejected variable is

$$\left(1-\frac{f(z)}{Mg(z)}\right)g(z) = \frac{g(z)-\rho f(z)}{1-\rho},$$

where $\rho = 1/M$, that the marginal distribution of Z_i $(i < n+t)$ is

$$f_m(z) = \frac{n-1}{n+t-1}f(z) + \frac{t}{n+t-1}\frac{g(z)-\rho f(z)}{1-\rho},$$

3.27 (Gelfand and Dey 1994) Consider a density function $f(x|\theta)$ and a prior distribution $\pi(\theta)$ such that the marginal $m(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$ is finite a.e. The marginal density is of use in the comparison of models since it appears in the Bayes factor (see Section 1.3).

(a) Give a Laplace approximation of m and derive the corresponding approximation of the Bayes factor. (See Tierney et al. 1989 for details.)

(b) Give the general shape of an importance sampling approximation of m.

(c) Detail this approximation when the importance function is the posterior distribution and when the normalizing constant is unknown.

(d) Show that for a proper density τ,

$$m(x)^{-1} = \int_\Theta \frac{\tau(\theta)}{f(x|\theta)\pi(\theta)}\pi(\theta|x)d\theta\,,$$

and deduce that when the θ_i^*'s are generated from the posterior,

$$\hat{m}(x) = \left\{\frac{1}{T}\sum_{t=1}^{T}\frac{\tau(\theta_i^*)}{f(x|\theta_i^*)\pi(\theta_i^*)}\right\}^{-1}$$

is another importance sampling estimator of m.

3.28 (Berger et al. 1998) For Σ a $p \times p$ positive-definite symmetric matrix, consider the distribution

$$\pi(\theta) \propto \frac{\exp\left(-(\theta-\mu)^t\Sigma^{-1}(\theta-\mu)/2\right)}{||\theta||^{p-1}}\,.$$

(a) Show that the distribution is well defined; that is, that

$$\int_{\mathbb{R}^p}\frac{\exp\left(-(\theta-\mu)^t\Sigma^{-1}(\theta-\mu)/2\right)}{||\theta||^{p-1}}d\theta < \infty.$$

(b) Show that an importance sampling implementation based on the normal instrumental distribution $\mathcal{N}_p(\mu, \Sigma)$ is not satisfactory from both theoretical and practical points of view.

(c) Examine the alternative based on a Gamma distribution $\mathcal{G}a(\alpha, \beta)$ on $\eta = ||\theta||^2$ and a uniform distribution on the angles.

Note: Priors such as these have been used to derive *Bayes minimax estimators* of a multivariate normal mean. See Lehmann and Casella (1998).

3.29 From the Accept–Reject Algorithm we get a sequence $Y_1, Y_2, \ldots$ of independent random variables generated from g along with a corresponding sequence $U_1, U_2, \ldots$ of uniform random variables. For a fixed sample size t (i.e. for a fixed number of accepted random variables), the number of generated Y_i's is a random integer N.

(a) Show that the joint distribution of $(N, Y_1, \ldots, Y_N, U_1, \ldots, U_N)$ is given by

$$\begin{aligned}
&P(N = n, Y_1 \le y_1, \ldots, Y_n \le y_n, U_1 \le u_1, \ldots, U_n \le u_n)\\
&\quad = \int_{-\infty}^{y_n} g(t_n)(u_n \wedge w_n)dt_n \int_{-\infty}^{y_1}\cdots\int_{-\infty}^{y_{n-1}} g(t_1)\ldots g(t_{n-1})\\
&\qquad \times \sum_{(i_1,\cdots,i_{t-1})}\prod_{j=1}^{t-1}(w_{i_j}\wedge u_{i_j})\prod_{j=t}^{n-1}(u_{i_j}-w_{i_j})^+dt_1\cdots dt_{n-1},
\end{aligned}$$

3.21 Monte Carlo marginalization is a technique for calculating a marginal density when simulating from a joint density. Let $(X_i, Y_i) \sim f_{XY}(x, y)$, independent, and the corresponding marginal distribution $f_X(x) = \int f_{XY}(x, y)dy$.

(a) Let $w(x)$ be an arbitrary density. Show that

$$\lim_n \frac{1}{n}\sum_{i=1}^{n} \frac{f_{XY}(x^*, y_i)w(x_i)}{f_{XY}(x_i, y_i)} = \int\int \frac{f_{XY}(x^*, y)w(x)}{f_{XY}(x, y)} f_{XY}(x, y)dxdy = f_X(x^*$$

and so we have a Monte Carlo estimate of f_X, the marginal distribution of X, from only knowing the form of the joint distribution.

(b) Let $X|Y = y \sim \mathcal{G}a(y, 1)$ and $Y \sim \mathcal{E}xp(1)$. Use the technique of part (a) to plot the marginal density of X. Compare it to the exact marginal.

(c) Choosing $w(x) = f_{X|Y}(x|y)$ works to produce the marginal distribution, and it is optimal. In the spirit of Theorem 3.12, can you prove this?

3.22 Given a real importance sample $X_1, \ldots, X_n$ with importance function g and target density f,

(a) show that the sum of the weights $\omega_i = f(X_i)/g(X_i)$ is only equal to 1 in expectation and deduce that the weights need to be renormalized even when both densities have know normalizing constants.

(b) Assuming that the weights ω_i have been renormalized to sum to one, we sample, with replacement, n points $\tilde{X}_j$ from the X_i's using those weights. Show that the $\tilde{X}_j$'s satisfy

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} h(\tilde{X}_j)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \omega_i h(X_i)\right].$$

(c) Deduce that, if the above formula is satisfied for $\omega_i = f(X_i)/g(X_i)$ instead, the empirical distribution associated with the $\tilde{X}_j$'s is unbiased.

3.23 (Evans and Swartz 1995) Devise and implement a simulation experiment to approximate the probability $P(Z \in (0, \infty)^6)$ when $Z \sim \mathcal{N}_6(0, \Sigma)$ and

$$\Sigma^{-1/2} = \text{diag}(0, 1, 2, 3, 4, 5) + e \cdot e^t,$$

with $e^t = (1, 1, 1, 1, 1, 1)$:

(a) when using the $\Sigma^{-1/2}$ transform of a $\mathcal{N}_6(0, I_6)$ random variables;

(b) when using the Choleski decomposition of Σ;

(c) when using a distribution restricted to $(0, \infty)^6$ and importance sampling.

3.24 Using the facts

$$\int y^3 e^{-cy^2/2}dy = \frac{-1}{2c}\left[y^2 + \frac{1}{c}\right]e^{-cy^2/2},$$

$$\int y^6 e^{-cy^2/2}dy = \frac{-1}{2c}\left[y^5 + \frac{5y^3}{2c} + \frac{15y}{4c}\right]e^{-cy^2/2} + 30\sqrt{\frac{\pi}{c^7}}\Phi(\sqrt{2c}y),$$

derive expressions similar to (3.22) for the second- and third-order approximations (see also Problem 5.6).

3.25 By evaluating the normal integral for the first order approximation from (3.21), establish (3.22).

3.26 Referring to Example 3.16, derive the Laplace approximation for the Gamma density and reproduce Table 3.6.

(a) Show that to simulate $Y \sim \mathcal{TE}(a, 1)$, an exponential distribution left truncated at a, we can simulate $X \sim \mathcal{E}(1)$ and take $Y = a + X$.
(b) Use this method to calculate the probability that a χ^2_3 random variable is greater that 25, and that a t_5 random variable is greater than 50.
(c) Explore the gain in efficiency from this method. Take $a = 4.5$ in part (a) and run an experiment to determine how many random variables would be needed to calculate $P(Z > 4.5)$ to the same accuracy obtained from using 100 random variables in an importance sampler.

3.17 In this chapter, the importance sampling method is developed for an iid sample $(Y_1, \ldots, Y_n)$ from g.
(a) Show that the importance sampling estimator is still unbiased if the Y_i's are correlated while being marginally distributed from g.
(b) Show that the importance sampling estimator can be extended to the case when Y_i is generated from a conditional distribution $q(y_i|Y_{i-1})$.
(c) Implement a scheme based on an iid sample $(Y_1, Y_3, \ldots, Y_{2n-1})$ and a secondary sample $(Y_2, Y_4, \ldots, Y_{2n})$ such that $Y_{2i} \sim q(y_{2i}|Y_{2i-1})$. Show that the covariance

$$\text{cov}\left(h(Y_{2i-1})\frac{f(Y_{2i-1})}{g(Y_{2i-1})}, h(Y_{2i})\frac{f(Y_{2i})}{q(Y_{2i}|Y_{2i-1})}\right)$$

is null. Generalize.

3.18 For a sample $(Y_1, \ldots, Y_h)$ from g, the weights ω_i are defined as

$$\omega_i = \frac{f(Y_i)/g(Y_i)}{\sum_{j=1}^h f(Y_j)/g(Y_j)}\,.$$

Show that the following algorithm (Rubin 1987) produces a sample from f such that the empirical average

$$\frac{1}{M}\sum_{m=1}^{M} h(X_m)$$

is asymptotically equivalent to the importance sampling estimator based on $(Y_1, \ldots, Y_N)$:
For $m = 1, \ldots, M$,
take $X_m = Y_i$ with probability ω_i
(*Note:* This is the SIR algorithm.)

3.19 (Smith and Gelfand 1992) Show that, when evaluating an integral based on a posterior distribution

$$\pi(\theta|x) \propto \pi(\theta)\ell(\theta|x),$$

where π is the prior distribution and ℓ the likelihood function, the prior distribution can always be used as instrumental distribution (see Problem 2.29).
(a) Show that the variance is finite when the likelihood is bounded.
(b) Compare with choosing $\ell(\theta|x)$ as instrumental distribution when the likelihood is proportional to a density. (*Hint*: Consider the case of exponential families.)
(c) Discuss the drawbacks of this (these) choice(s) in specific settings.
(d) Show that a mixture between both instrumental distributions can ease some of the drawbacks.

3.20 In the setting of Example 3.13, show that the variance of the importance sampling estimator associated with an importance function g and the integrand $h(x) = \sqrt{x/(1-x)}$ is infinite for all g's such that $g(1) < \infty$.

depends only on (φ_1, φ_2), show that $\rho|\varphi_1, \varphi_2 \sim \mathcal{N}(x \cdot \xi, 1)$ and then integration of ρ then leads to

$$\pi(\varphi_1, \varphi_2|x) \propto \exp\{(x \cdot \xi)^2/2\} \sin(\varphi_1),$$

where $x \cdot \xi = x_1 \cos(\varphi_1) + x_2 \sin(\varphi_1) \cos(\varphi_2) + x_3 \sin(\varphi_1) \sin(\varphi_2)$.

(c) Show how to simulate from $\pi(\varphi_1, \varphi_2|x)$ using an Accept–Reject algorithm with instrumental function $\sin(\varphi_1) \exp\{||x||^2/2\}$.

(d) For $p = 3$ and $x = (0.1, 1.2, -0.7)$, demonstrate the convergence of the algorithm. Make plots of the iterations of the integral and its standard error.

3.9 For the situation of Example 3.10, recreate Figure 3.4 using the following simulation strategies with a sample size of 10,000 points:

(a) For each value of λ, simulate a sample from the $\mathcal{E}xp(1/\lambda)$ distribution and a separate sample from the log-normal $\mathcal{LN}(0, 2\log\lambda)$ distribution. Plot the resulting risk functions.

(b) For each value of λ, simulate a sample from the $\mathcal{E}xp(1/\lambda)$ distribution and then transform it into a sample from the $\mathcal{LN}(0, 2\log\lambda)$ distribution. Plot the resulting risk functions.

(c) Simulate a sample from the $\mathcal{E}xp(1)$ distribution. For each value of λ, transform it into a sample from $\mathcal{E}xp(1/\lambda)$, and then transform it into a sample from the $\mathcal{LN}(0, 2\log\lambda)$ distribution. Plot the resulting risk functions.

(d) Compare and comment on the accuracy of the plots.

3.10 Compare (in a simulation experiment) the performances of the regular Monte Carlo estimator of

$$\int_1^2 \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = \Phi(2) - \Phi(1)$$

with those of an estimator based on an optimal choice of instrumental distribution (see (3.11)).

3.11 In the setup of Example 3.10, give the two first moments of the log-normal distribution $\mathcal{LN}(\mu, \sigma^2)$.

3.12 In the setup of Example 3.13, examine whether or not the different estimators of the expectations $\mathbb{E}_f[h_i(X)]$ have finite variances.

3.13 Establish the equality (3.18) using the representation $b = \beta a/\alpha$.

3.14 (Ó Ruanaidh and Fitzgerald 1996) For simulating random variables from the density $f(x) = \exp\{-\sqrt{x}\}[\sin(x)]^2$, $0 < x < \infty$, compare the following choices of instrumental densities:

$$g_1(x) = \tfrac{1}{2}e^{-|x|}, \qquad g_2(x) = \tfrac{1}{2\sqrt{2}}\,\mathrm{sech}^2(x/\sqrt{2}),$$
$$g_3(x) = \tfrac{1}{2\pi}\tfrac{1}{1+x^2/4}, \qquad g_4(x) = \tfrac{1}{\sqrt{2\pi}}e^{-x/2}.$$

(a) For $M = 100, 1000$, and $10,000$, compare the standard deviations of the estimates based on simulating M random variables.

(b) For each of the instrumental densities, estimate the size of M needed to obtain three digits of accuracy in estimating $\mathbb{E}_f X$.

3.15 Use the techniques of Example 3.11 to redo Problem 3.3. Compare the number of variables needed to obtain three digits of accuracy with importance sampling to the answers obtained from Problem 3.3.

3.16 Referring to Example 3.11:

(b) Design a computer experiment to compare Monte Carlo error when using (*i*) the same random variables θ_i in numerator and denominator, or (*ii*) different random variables.

3.3 (a) For a standard normal random variable Z, calculate $P(Z > 2.5)$ using Monte Carlo sums based on indicator functions. How many simulated random variables are needed to obtain three digits of accuracy?

(b) Using Monte Carlo sums verify that if $X \sim \mathcal{G}(1,1)$, $P(X > 5.3) \approx .005$. Find the exact .995 cutoff to three digits of accuracy.

3.4 (a) If $X \sim \mathcal{N}(0, \sigma^2)$, show that

$$\mathbb{E}[e^{-X^2}] = \frac{1}{\sqrt{2\sigma^2 + 1}}.$$

(b) Generalize to the case $X \sim \mathcal{N}(\mu, \sigma^2)$.

3.5 Referring to Example 3.6:

(a) Verify the maximum of the likelihood ratio statistic.

(b) Generate 5000 random variables according to (3.7), recreating the left panel of Figure 3.2. Compare this distribution to a null distribution where we *fix* null values of p_1 and p_2, for example, $(p_1, p_2) = (.25, .75)$. For a range of values of (p_1, p_2), compare the histograms both with the one from (3.7) and the χ^2_1 density. What can you conclude?

3.6 An alternate analysis to that of Example 3.6 is to treat the contingency table as two binomial distributions, one for the patients receiving surgery and one for those receiving radiation. Then the test of hypothesis becomes a test of equality of the two binomial parameters. Repeat the analysis of the data in Table 3.2 under the assumption of two binomials. Compare the results to those of Example 3.6.

3.7 A famous medical experiment was conducted by Joseph Lister in the late 1800s to examine the relationship between the use of a disinfectant, carbolic acid, and surgical success rates. The data are

	Disinfectant	
	Yes	No
Success	34	19
Failure	6	16

Using the techniques of Example 3.6, analyze these data to examine the association between disinfectant and surgical success rates. Use both the multinomial model and the two-binomial model.

3.8 Referring to Example 3.3, we calculate the expected value of $\delta^\pi(x)$ from the posterior distribution $\pi(\theta|x) \propto \|\theta\|^{-2} \exp\{-\|x-\theta\|^2/2\}$, arising from a normal likelihood and noninformative prior $\|\theta\|^{-2}$ (see Example 1.12).

(a) Show that if the quadratic loss of Example 3.3 is normalized by $1/(2\|\theta\|^2 + p)$, the resulting Bayes estimator is

$$\delta^\pi(x) = \mathbb{E}^\pi\left[\frac{\|\theta\|^2}{2\|\theta\|^2 + p}\,\middle|\, x, \lambda\right] \Big/ \mathbb{E}^\pi\left[\frac{1}{2\|\theta\|^2 + p}\,\middle|\, x, \lambda\right].$$

(b) Simulation of the posterior can be done by representing θ in polar coordinates $(\rho, \varphi_1, \varphi_2)$ $(\rho > 0, \varphi_1 \in [-\pi/2, \pi/2], \varphi_2 \in [-\pi/2, \pi/2])$, with $\theta = (\rho\cos\varphi_1,\ \rho\sin\varphi_1\cos\varphi_2,\ \rho\sin\varphi_1\sin\varphi_2)$. If we denote $\xi = \theta/\rho$, which

$$h(x) \approx \frac{\hat{x}_\theta}{\beta} + (\alpha - 1)\log(\hat{x}_\theta) + \frac{\alpha - 1}{2\hat{x}_\theta^2}(x - \hat{x}_\theta)^2$$

Now substituting into (3.22) yields the Laplace approximation

$$\int_a^b \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta} dx = \hat{x}_\theta^{\alpha-1} e^{\hat{x}_\theta/\beta} \sqrt{\frac{2\pi\hat{x}_\theta^2}{\alpha - 1}} \times \left\{ \Phi\left(\sqrt{\frac{\alpha - 1}{\hat{x}_\theta^2}}(b - \hat{x}_\theta) \right) - \Phi\left(\sqrt{\frac{\alpha - 1}{\hat{x}_\theta^2}}(a - \hat{x}_\theta) \right) \right\}$$

For $\alpha = 5$ and $\beta = 2$, $\hat{x}_\theta = 8$, and the approximation will be best in that area. In Table 3.6 we see that although the approximation is reasonable in the central region of the density, it becomes quite unacceptable in the tails. ‖

Interval	Approximation	Exact
$(7, 9)$	0.193351	0.193341
$(6, 10)$	0.375046	0.37477
$(2, 14)$	0.848559	0.823349
$(15.987, \infty)$	0.0224544	0.100005

Table 3.6. Laplace approximation of a Gamma integral for $\alpha = 5$ and $\beta = 2$.

Thus, we see both the usefulness and the limits of the Laplace approximation. In problems where Monte Carlo calculations are prohibitive because of computing time, the Laplace approximation can be useful as a guide to the solution of the problem. Also, the corresponding Taylor series can be used as a proposal density, which is particularly useful in problems where no obvious proposal exists. (See Example 7.12 for a similar situation.)

3.5 Problems

3.1 For the normal-Cauchy Bayes estimator

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}{\int_{-\infty}^{\infty} \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}$$

(*a*) Plot the integrand and use Monte Carlo integration to calculate the integral.
(*b*) Monitor the convergence with the standard error of the estimate. Obtain three digits of accuracy with probability .95.

3.2 (Continuation of Problem 3.1)
(a) Use the Accept–Reject algorithm, with a Cauchy candidate, to generate a sample from the posterior distribution and calculate the estimator.

The cubic term in the exponent is now expanded in a series around $\hat{x}_\theta$. Recall that the second order Taylor expansion of e^y around 0 is $e^y \approx 1 + y + y^2/2!$, and hence expanding $\exp\{n(x-\hat{x}_\theta)^3 h'''(\hat{x}_\theta|\theta)/3!\}$ around $\hat{x}_\theta$, we obtain the approximation

$$1 + n\frac{(x-\hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) + n^2\frac{(x-\hat{x}_\theta)^6}{2!(3!)^2}[h'''(\hat{x}_\theta|\theta)]^2$$

and thus

$$\int_A e^{nh(x|\theta)}dx \simeq e^{nh(\hat{x}_\theta|\theta)} \int_A e^{n\frac{(x-\hat{x}_\theta)^2}{2}h''(\hat{x}_\theta|\theta)} \times \left[1 + n\frac{(x-\hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta) + n^2\frac{(x-\hat{x}_\theta)^6}{2!(3!)^2}[h'''(\hat{x}_\theta|\theta)]^2 + R_n\right] dx, \tag{3.21}$$

where R_n again denotes a remainder term.

Excluding R_n, we call the integral approximations in (3.21) a *first-order* approximation if it includes only the first term in the right-hand side, a *second-order* approximation if it includes the first two terms; and a *third-order* approximation if it includes all three terms.

Since the above integrand is the kernel of a normal density with mean $\hat{x}_\theta$ and variance $-1/n\,h''(\hat{x}_\theta|\theta)$, we can evaluate these expressions further. More precisely, letting $\Phi(\cdot)$ denote the standard normal cdf, and taking $A = [a, b]$, we can evaluate the integral in the first-order approximation to obtain (see Problem 3.25)

$$\int_a^b e^{nh(x|\theta)}dx \simeq e^{nh(\hat{x}_\theta|\theta)}\sqrt{\frac{2\pi}{-nh''(\hat{x}_\theta|\theta)}} \times \left\{\Phi[\sqrt{-nh''(\hat{x}_\theta|\theta)}(b-\hat{x}_\theta)] - \Phi[\sqrt{-nh''(\hat{x}_\theta|\theta)}(a-\hat{x}_\theta)]\right\}. \tag{3.22}$$

Example 3.16. Gamma approximation. As a simple illustration of the Laplace approximation, consider estimating a Gamma $\mathcal{G}a(\alpha, 1/\beta)$ integral, say

$$\int_a^b \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}\, e^{-x/\beta}dx. \tag{3.23}$$

Here we have $h(x) = -\frac{x}{\beta} + (\alpha-1)\log(x)$ with second order Taylor expansion (around a point x_0)

$$\begin{aligned} h(x) &\approx h(x_0) + h'(x_0)(x-x_0) + h''(x_0)\frac{(x-x_0)^2}{2!} \\ &= -\frac{x_0}{\beta} + (\alpha-1)\log(x_0) + \left(\frac{\alpha-1}{x_0} - \frac{1}{\beta}\right)(x-x_0) - \frac{\alpha-1}{2x_0^2}(x-x_0)^2. \end{aligned}$$

Choosing $x_0 = \hat{x}_\theta = (\alpha-1)\beta$ (the mode of the density and maximizer of h) yields

m	100			1000			5 000		
	δ_1	δ_4	δ_6	δ_1	δ_4	δ_6	δ_1	δ_4	δ_6
h_1	87.3	55.9	64.2	36.5	0.044	0.047	2.02	0.54	0.64
h_2	1.6	3.3	4.4	4.0	0.00	0.00	0.17	0.00	0.00
h_3	6.84	0.11	0.76	4.73	0.00	0.00	0.38	0.02	0.00

Table 3.5. Comparison of the performances of the Monte Carlo estimator (δ_1) with two importance sampling estimators (δ_4 and δ_6) under squared error loss after m iterations for $\alpha = 3.7$ and $\beta = 1$. The squared error loss is multiplied by 10^2 for the estimation of $\mathbb{E}[h_2(X)]$ and by 10^5 for the estimation of $\mathbb{E}[h_3(X)]$. The squared errors are actually the difference from the theoretical values (99.123, 5.3185, and 0.7497, respectively) and the three estimators are based on the same unique sample, which explains the lack of monotonicity (in m) of the errors. (*Source:* Casella and Robert 1998.)

Laplace approximation. It is based on the following argument: Suppose that we are interested in evaluating the integral

$$\int_A f(x|\theta)dx \tag{3.19}$$

for a fixed value of θ. (The function f needs to be non-negative and integrable; see Tierney and Kadane 1986 and Tierney et al. 1989 for extensions.). Write $f(x|\theta) = \exp\{nh(x|\theta)\}$, where n is the sample size or another parameter which can go to infinity, and use a Taylor series expansion of $h(x|\theta)$ about a point x_0 to obtain

$$h(x|\theta) \approx h(x_0|\theta) + (x - x_0)h'(x_0|\theta) + \frac{(x - x_0)^2}{2!}h''(x_0|\theta)$$
$$+\frac{(x - x_0)^3}{3!}h'''(x_0|\theta) + R_n(x)\,, \tag{3.20}$$

where we write

$$h'(x_0|\theta) = \left.\frac{\partial h(x|\theta)}{\partial x}\right|_{x=x_0},$$

and similarly for the other terms, while the remainder $R_n(x)$ satisfies

$$\lim_{x \to x_0} R_n(x)/(x - x_0)^3 = 0.$$

Now choose $x_0 = \hat{x}_\theta$, the value that satisfies $h'(\hat{x}_\theta|\theta) = 0$ and maximizes $h(x|\theta)$ for the given value of θ. Then, the linear term in (3.20) is zero and we have the approximation

$$\int_A e^{nh(x|\theta)}dx \simeq e^{nh(\hat{x}_\theta|\theta)} \int_A e^{n\frac{(x-\hat{x}_\theta)^2}{2}h''(\hat{x}_\theta|\theta)} e^{n\frac{(x-\hat{x}_\theta)^3}{3!}h'''(\hat{x}_\theta|\theta)}dx,$$

which is valid within a neighborhood of $\hat{x}_\theta$. (See Schervish 1995, Section 7.4.3, for detailed conditions.) Note the importance of choosing the point x_0 to be a maximum.

Figure 3.10 describes the convergence of the three estimators of h_3 in m for $\alpha = 3.7$ and $\beta = 1$ (which yields an Accept–Reject acceptance probability of $1/M = .10$). Both estimators δ_4 and δ_6 have more stable graphs than the empirical average δ_1 and they converge much faster to the theoretical expectation 0.7497, δ_6 then being equal to this value after 6,000 iterations. For $\alpha = 3.08$ and $\beta = 1$ (which yields an Accept–Reject acceptance probability of $1/M = .78$), Figure 3.11 illustrates the change of behavior of the three estimators of h_3 since they now converge at similar speeds. Note the proximity of δ_6 and δ_1, δ_6 again being the estimator closest to the theoretical expectation 0.7081 after 10,000 iterations.

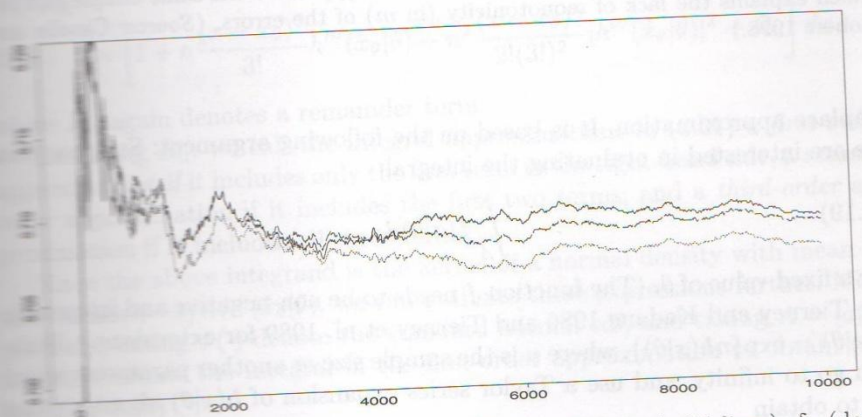

Fig. 3.11. Convergence of estimators of $\mathbb{E}[X/(1+X)]$, δ_1 (solid lines), δ_4 (dots) and δ_6 (dashes) for $\alpha = 3.08$ and $\beta = 1$. The final values are respectively 0.7087, 0.7069, and 0.7084, for a true value of the expectation equal to 0.7081.

Table 3.5 provides another evaluation of the three estimators in a case which is a priori very favorable to importance sampling, namely for $\alpha = 3.7$. The table exhibits, in most cases, a strong domination of δ_4 and δ_6 over δ_1 and a moderate domination of δ_4 over δ_6. ‖

In contrast to the general setup of Section 3.3, δ_4 (or its approximation δ_6) can always be used in an Accept–Reject sampling setup since this estimator does not require additional simulations. It provides a second evaluation of $\mathbb{E}_f[h]$, which can be compared with the Monte Carlo estimator for the purpose of convergence assessment.

3.4 Laplace Approximations

As an alternative to simulation of integrals, we can also attempt analytic approximations. One of the oldest and most useful approximations is the integral

Example 3.15. Gamma simulation. For illustrative purposes, consider the simulation of $\mathcal{G}a(\alpha,\beta)$ from the instrumental distribution $\mathcal{G}a(a,b)$, with $a=[\alpha]$ and $b=a\beta/\alpha$. (This choice of b is justified in Example 2.19 as maximizing the acceptance probability in an Accept–Reject scheme.) The ratio f/g is therefore

$$w(x) = \frac{\Gamma(a)}{\Gamma(\alpha)}\,\frac{\beta^\alpha}{b^a}\,x^{\alpha-a}e^{(b-\beta)x}\,,$$

which is bounded by

$$\begin{aligned} M &= \frac{\Gamma(a)}{\Gamma(\alpha)}\,\frac{\beta^\alpha}{b^a}\left(\frac{\alpha-a}{\beta-b}\right)^{\alpha-a} e^{-(\alpha-a)} \\ &= \frac{\Gamma(a)}{\Gamma(\alpha)}\,\exp\{\alpha(\log(\alpha)-1)-a(\log(a)-1)\}\,. \end{aligned} \tag{3.18}$$

Since the ratio $\Gamma(a)/\Gamma(\alpha)$ is bounded from above by 1, an approximate bound that can be used in the simulation is

$$M' = \exp\{a(\log(a)-1)-\alpha(\log(\alpha)-1)\}\,,$$

with $M'/M = 1+\varepsilon = \Gamma(\alpha)/\Gamma([\alpha])$. In this particular setup, the estimator δ_4 is available since f/g and M are explicitly known. In order to assess the effect of the approximation (3.17), we also compute the estimator δ_6 for the following functions of interest:

$$h_1(x)=x^3, \quad h_2(x)=x\log x\,, \quad \text{and} \quad h_3(x)=\frac{x}{1+x}\,.$$

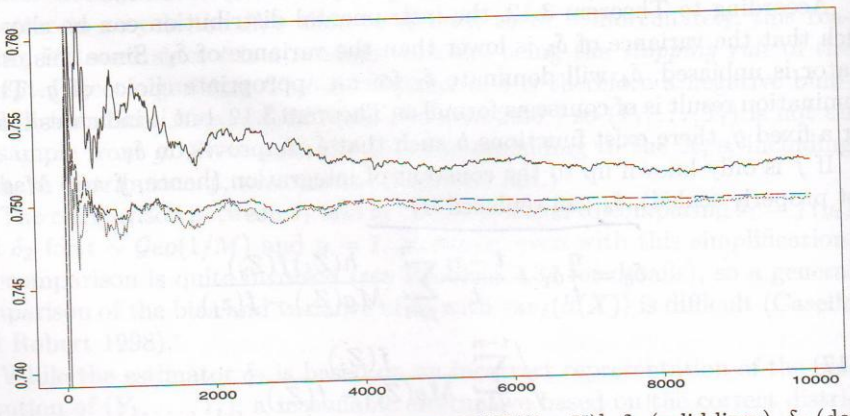

Fig. 3.10. Convergence of the estimators of $\mathbb{E}[X/(1+X)]$, δ_1 (solid lines), δ_4 (dots and δ_6 (dashes), for $\alpha=3.7$ and $\beta=1$. The final values are respectively 0.7518 0.7495, and 0.7497, for a true value of the expectation equal to 0.7497.

$$\delta_4 = \frac{n}{t}\,\delta_1 + \frac{1}{t}\sum_{j=1}^{t-n} h(Z_j)\,\frac{(M-1)f(Z_j)}{Mg(Z_j)-f(Z_j)}\,, \tag{3.16}$$

where the Z_j's are the elements of $(Y_1,\ldots,Y_t)$ that have been rejected. This estimator is also unbiased and the comparison with δ_1 can also be studied in the case $n = 1$; that is, through the comparison of the variances of $h(X_1)$ and of δ_4, which now can be written in the form

$$\delta_4 = \frac{1}{t}\,h(X_1) + (1-\rho)\,\frac{1}{t}\sum_{j=1}^{t-1} h(Z_j)\left(\frac{g(Z_j)}{f(Z_j)} - \rho\right)^{-1}.$$

Assuming again that $\mathbb{E}_f[h(X)] = 0$, the variance of δ_4 is

$$\text{var}(\delta_4) = \mathbb{E}\left[\frac{t-1}{t^2}\int h^2(x)\frac{f^2(x)(M-1)}{Mg(x)-f(x)}\,dx + \frac{1}{t^2}\,\mathbb{E}_f[h^2(X)]\right],$$

which is again too case-specific (that is, too dependent on f, g, and h) to allow for a general comparison.

The marginal distribution of the Z_i's from the Accept–Reject algorithm is $(Mg - f)/(M-1)$, and the importance sampling estimator δ_5 associated with this instrumental distribution is

$$\delta_5 = \frac{1}{t-n}\sum_{j=1}^{t-n}\frac{(M-1)f(Z_j)}{Mg(Z_j)-f(Z_j)}h(Z_j),$$

which allows us to write δ_4 as

$$\delta_4 = \frac{n}{t}\delta_1 + \frac{t-n}{t}\delta_5,$$

a weighted average of the usual Monte Carlo estimator and of δ_5.

According to Theorem 3.12, the instrumental distribution can be chosen such that the variance of δ_5 is lower than the variance of δ_1. Since this estimator is unbiased, δ_4 will dominate δ_1 for an appropriate choice of g. This domination result is of course as formal as Theorem 3.12, but it indicates that, for a fixed g, there exist functions h such that δ_4 improves on δ_1.

If f is only known up to the constant of integration (hence, f and M are not properly scaled), δ_4 can replaced by

$$\delta_6 = \frac{n}{t}\delta_1 + \frac{t-n}{t}\sum_{j=1}^{t-n}\frac{h(Z_j)f(Z_j)}{Mg(Z_j)-f(z_j)} \Bigg/ \sum_{j=1}^{t-n}\frac{f(Z_j)}{Mg(Z_j)-f(Z_j)}\,. \tag{3.17}$$

Although the above domination of δ_1 by δ_4 does not extend to δ_6, nonetheless, δ_6 correctly estimates constant functions while being asymptotically equivalent to δ_4. See Casella and Robert (1998) for additional domination results of δ_1 by weighted estimators.

To undertake a comparison of estimation using Accept–Reject and estimation using importance sampling, it is reasonable to start with the two traditional estimators

$$\delta_1 = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \qquad \text{and} \qquad \delta_2 = \frac{1}{t}\sum_{j=1}^{t} h(Y_j)\,\frac{f(Y_j)}{g(Y_j)}\,. \tag{3.15}$$

These estimators correspond to the straightforward utilization of the sample produced by Accept–Reject and to an importance sampling estimation derived from the overall sample, that is, to a recycling of the variables rejected by algorithm [A.4].[4] If the ratio f/g is only known up to a constant, δ_2 can be replaced by

$$\delta_3 = \sum_{j=1}^{t} h(Y_j)\,\frac{f(Y_j)}{g(Y_j)} \Bigg/ \sum_{j=1}^{t} \frac{f(Y_j)}{g(Y_j)}\,.$$

If we write δ_2 in the more explicit form

$$\delta_2 = \frac{n}{t}\left\{\frac{1}{n}\sum_{i=1}^{n} h(X_i)\,\frac{f(X_i)}{g(X_i)} + \frac{t-n}{n}\,\frac{1}{t-n}\sum_{i=1}^{t-n} h(Z_i)\,\frac{f(Z_i)}{g(Z_i)}\right\},$$

where $\{Y_1,\ldots,Y_t\} = \{X_1,\ldots,X_n\} \cup \{Z_1,\ldots,Z_{t-n}\}$ (the Z_i's being the variables rejected by the Accept–Reject algorithm [A.4]), one might argue that, based on sample size, the variance of δ_2 is smaller than that of the estimator

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i)\,\frac{f(X_i)}{g(X_i)}\,.$$

If we could apply Theorem 3.12, we could then conclude that this latter estimator dominates δ_1 (for an appropriate choice of g) and, hence, that it is better to recycle the Z_i's than to discard them. Unfortunately, this reasoning is flawed since t is a random variable, being the *stopping rule* of the Accept–Reject algorithm. The distribution of t is therefore a negative binomial distribution, $\mathcal{N}eg(n, 1/M)$ (see Problem 2.30) so $(Y_1,\ldots,Y_t)$ is not an iid sample from g. (Note that the Y_j's corresponding to the X_i's, including Y_t, have distribution f, whereas the others do not.)

The comparison between δ_1 and δ_2 can be reduced to comparing $\delta_1 = f(y_t)$ and δ_2 for $t \sim \mathcal{G}eo(1/M)$ and $n = 1$. However, even with this simplification, the comparison is quite involved (see Problem 3.34 for details), so a general comparison of the bias and variance of δ_2 with $\mathrm{var}_f(h(X))$ is difficult (Casella and Robert 1998).

While the estimator δ_2 is based on an incorrect representation of the distribution of $(Y_1,\ldots,Y_t)$, a reasonable alternative based on the correct distribution of the sample is

[4] This obviously assumes a relatively tight control on the simulation methods rather than the use of a (black box) pseudo-random generation software, which only delivers the accepted variables.

finiteness of the variance is ignored, and not detected, it may result in strong biases. For example, it can happen that the obvious divergence behavior of the previous examples does not occur. Thus, other measures, such as monitoring of the range of the weights $f(X_i)/g(X_i)$ (which are of mean 1 in all cases), can help to detect convergence problems. (See also Note 4.6.1.)

The finiteness of the ratio $\mathbb{E}_f[f(X)/g(X)]$ can be achieved by substituting a mixture distribution for the density g,

$$\rho g(x) + (1-\rho)\ell(x), \tag{3.14}$$

where ρ is close to 1 and ℓ is chosen for its heavy tails (for instance, a Cauchy or a Pareto distribution). From an operational point of view, this means that the observations are generated with probability ρ from g and with probability $1-\rho$ from ℓ. However, the mixture (g versus ℓ) does not play a role in the computation of the importance weights; that is, by construction, the estimator integrates out the uniform variable used to decide between g and ℓ. (We discuss in detail such a marginalization perspective in Section 4.2, where uniform variables involved in the simulation are integrated out in the estimator.) Obviously, (3.14) replaces $g(x)$ in the weights of (3.8) or (3.11), which can then ensure a finite variance for integrable functions h^2. Hesterberg (1998) studies the performances of this approach, called a *defensive mixture*.

3.3.3 Comparing Importance Sampling with Accept–Reject

Theorem[3] 3.12 formally solves the problem of comparing Accept–Reject and importance sampling methods, since with the exception of the constant functions $h(x) = h_0$, the optimal density g^* is always different from f. However, a more realistic comparison should also take account of the fact that Theorem 3.12 is of limited applicability in a practical setup, as it prescribes an instrumental density that depends on the function h of interest. This may not only result in a considerable increase of the computation time for every new function h (especially if the resulting instrumental density is not easy to generate from), but it also eliminates the possibility of reusing the generated sample to estimate a number of different quantities, as in Example 3.14. Now, when the Accept–Reject method is implemented with a density g satisfying $f(x) \leq Mg(x)$ for a constant $1 < M < \infty$, the density g can serve as the instrumental density for importance sampling. A positive feature is that f/g is bounded, thus ensuring finiteness of the variance for the corresponding importance sampling estimators. Bear in mind, though, that in the Accept–Reject method the resulting sample, $X_1, \ldots, X_n$, is a subsample of $Y_1, \ldots, Y_t$, where the Y_i's are simulated from g and where t is the (random) number of simulations from g required for produce the n variables from f.

[3] This section contains more specialized material and may be omitted on a first reading.

Distribution	h_1	h_2	h_3	h_4	h_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

Table 3.4. Comparison of the evaluations of $\mathbb{E}_f[h_j]$ for the estimators (3.10) corresponding to three instrumental distributions π_i and to the true distribution π (10,000 simulations).

Table 3.4, it shows the improvement brought by the distribution π_3 upon the alternative distributions, since the precision is of the same order as the true distribution, for a significantly lower simulation cost. The jumps in the graphs of the estimators associated with π_2 and, especially, with π_1 are characteristic of importance sampling estimators with infinite variance. ‖

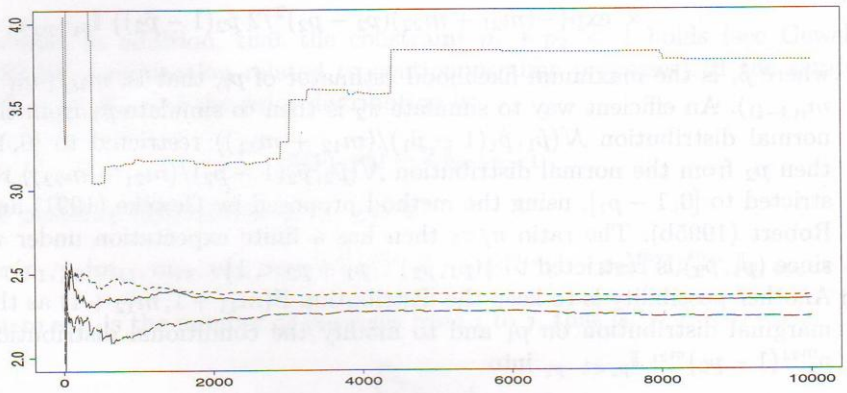

Fig. 3.9. Convergence of four estimators of $\mathbb{E}_f[h_5(X)]$ for the true distribution π (solid lines) and for the instrumental distributions π_1 (dots), π_2 (long dashes), and π_3 (short dashes). The final values after 10,000 iterations are 2.373, 3.184, 2.319, and 2.379, respectively.

We therefore see that importance sampling cannot be applied blindly. Rather, care must be taken in choosing an instrumental density as the almost sure convergence of (3.8) is only formal (in the sense that it may require an enormous number of simulations to produce an accurate approximation of the quantity of interest). These words of caution are meant to make the user aware of the problems that might be encountered if importance sampling is used when $\mathbb{E}_f[|f(X)/g(X)|]$ is infinite. (When $\mathbb{E}_f[f(X)/g(X)]$ is finite, the stakes are not so high, as convergence is more easily attained.) If the issue of

(i) The distribution $\pi(p_1, p_2|\mathcal{D})$ is the restriction of the product of two distributions $\mathcal{B}e(m_{11}+1, m_{12}+1)$ and $\mathcal{B}e(m_{22}+1, m_{21}+1)$ to the simplex $\{(p_1, p_2) : p_1+p_2 < 1\}$. So a reasonable first approach is to simulate these two distributions until the sum of two realizations is less than 1. Unfortunately, this naïve strategy is rather inefficient since, for the given data $(m_{11}, m_{12}, m_{21}, m_{22}) = (68, 28, 17, 4)$ we have $P^\pi(p_1 + p_2 < 1|\mathcal{D}) = 0.21$ (Geweke 1989). The importance sampling alternatives are to simulate distributions which are restricted to the simplex.

(ii) A solution inspired from the shape of $\pi(p_1, p_2|\mathcal{D})$ is a Dirichlet distribution $\mathcal{D}(m_{11}+1, m_{22}+1, m_{12}+m_{21}+1)$, with density

$$\pi_1(p_1, p_2|\mathcal{D}) \propto p_1^{m_{11}} p_2^{m_{22}} (1-p_1-p_2)^{m_{12}+m_{21}} .$$

However, the ratio $\pi(p_1, p_2|\mathcal{D})/\pi_1(p_1, p_2|\mathcal{D})$ is not bounded and the corresponding variance is infinite.

(iii) Geweke's (1989) proposal is to use the normal approximation to the binomial distribution, that is,

$$\begin{aligned}\pi_2(p_1, p_2|\mathcal{D}) \propto & \exp\{-(m_{11}+m_{12})(p_1-\hat{p}_1)^2/2\,\hat{p}_1(1-\hat{p}_1)\} \\ & \times \exp\{-(m_{21}+m_{22})(p_2-\hat{p}_2)^2/2\,\hat{p}_2(1-\hat{p}_2)\}\, \mathbb{I}_{p_1+p_2\le 1} ,\end{aligned}$$

where $\hat{p}_i$ is the maximum likelihood estimator of p_i, that is, $m_{ii}/(m_{ii} + m_{i(3-i)})$. An efficient way to simulate π_2 is then to simulate p_1 from the normal distribution $\mathcal{N}(\hat{p}_1, \hat{p}_1(1-\hat{p}_1)/(m_{12}+m_{11}))$ restricted to $[0,1]$, then p_2 from the normal distribution $\mathcal{N}(\hat{p}_2, \hat{p}_2(1-\hat{p}_2)/(m_{21}+m_{22}))$ restricted to $[0, 1-p_1]$, using the method proposed by Geweke (1991) and Robert (1995b). The ratio π/π_2 then has a finite expectation under π, since (p_1, p_2) is restricted to $\{(p_1, p_2) : p_1 + p_2 < 1\}$.

(iv) Another possibility is to keep the distribution $\mathcal{B}(m_{11}+1, m_{12}+1)$ as the marginal distribution on p_1 and to modify the conditional distribution $p_2^{m_{22}}(1-p_2)^{m_{21}}\, \mathbb{I}_{p_2<1-p_1}$ into

$$\pi_3(p_2|p_1, \mathcal{D}) = \frac{2}{(1-p_1)^2}\, p_2\, \mathbb{I}_{p_2<1-p_1} .$$

The ratio $w(p_1, p_2) \propto p_2^{m_{22}-1}(1-p_2)^{m_{21}}(1-p_1)^2$ is then bounded in (p_1, p_2).

Table 3.4 provides the estimators of the posterior expectations of the functions h_j evaluated for the true distribution π (simulated the naïve way, that is, until $p_1 + p_2 < 1$) and for the three instrumental distributions π_1, π_2 and π_3. The distribution π_3 is clearly preferable to the two other instrumental distributions since it provides the same estimation as the true distribution, at a lower computational cost. Note that π_1 does worse in all cases.

Figure 3.9 describes the evolution of the estimators (3.10) of $\mathbb{E}[h_5]$ as m increases for the three instrumental distributions considered. Similarly to

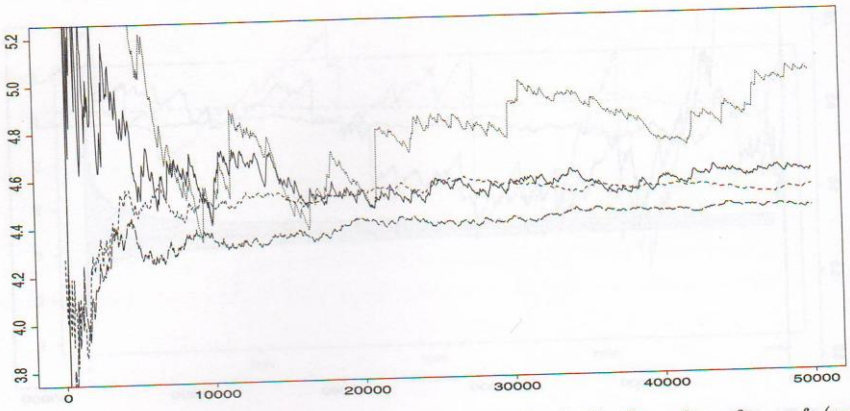

Fig. 3.8. Convergence of four estimators of $\mathbb{E}_f[h_3(X)]$: Sampling from f (solid lines), importance sampling with Cauchy instrumental distribution (short dashes), with normal instrumental distribution (dots), and with exponential instrumental distribution (long dashes). The final values after 50,000 iterations are respectively 4.58, 4.42, 4.99, and 4.52, for a true value of 4.64.

Assume, in addition, that the constraint $p_1 + p_2 < 1$ holds (see Geweke 1989 for a motivation related to continuous time processes). If the sample is $X_1, \ldots, X_m$ and the prior distribution is

$$\pi(p_1, p_2) = 2\, \mathbb{I}_{p_1+p_2<1}\,,$$

the posterior distribution of (p_1, p_2) is

$$\pi(p_1, p_2|m_{11}, m_{12}, m_{21}, m_{22}) \propto p_1^{m_{11}}(1-p_1)^{m_{12}}(1-p_2)^{m_{21}}p_2^{m_{22}}\ \mathbb{I}_{p_1+p_2<1}\,,$$

where m_{ij} is the number of passages from i to j, that is,

$$m_{ij} = \sum_{t=2}^{m} \mathbb{I}_{x_t=i}\mathbb{I}_{x_{t+1}=j},$$

and it follows that $\mathcal{D} = (m_{11}, \ldots, m_{22})$ is a sufficient statistic.

Suppose now that the quantities of interest are the posterior expectations of the probabilities and the associated odds:

$$h_1(p_1, p_2) = p_1,\ h_2(p_1, p_2) = p_2,\ h_3(p_1, p_2) = \frac{p_1}{1-p_1}$$

and

$$h_4(p_1, p_2) = \frac{p_2}{1-p_2}, \quad h_5(p_1, p_2) = \log\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right)\,,$$

respectively.

We now look at a number of ways in which to calculate these posterior expectations.

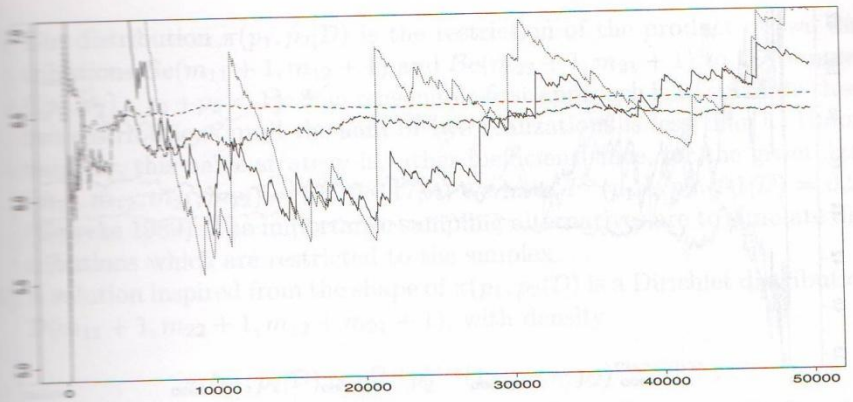

Fig. 3.7. Convergence of four estimators of $\mathbb{E}_f[X^5 \mathbb{I}_{X \geq 2.1}]$ for $\nu = 12$: Sampling from f (solid lines), importance sampling with Cauchy instrumental distribution (short dashes), importance sampling with uniform $\mathcal{U}([0, 1/2.1])$ instrumental distribution (long dashes) and importance sampling with normal instrumental distribution (dots). The final values are respectively 6.75, 6.48, 6.57, and 7.06, for an exact value of 6.54.

$$\frac{1}{m} \sum_{j=1}^{m} h_3(X_j)\, w(X_j)\,, \tag{3.13}$$

where the X_j's are iid $\mathcal{E}xp(1)$ and $w(x) = f(x)\exp(x)$. Figure 3.8 shows that, although this weight does not have a finite expectation under $\mathcal{T}(\nu, 0, 1)$, meaning that the variance is infinite, the estimator (3.13) provides a good approximation of $\mathbb{E}_f[h_3(X)]$, having the same order of precision as the estimation provided by the exact simulation, and greater stability. The estimator based on the Cauchy distribution is, as in the other case, stable, but its bias is, again, slow to vanish, and the estimator associated with the normal distribution once more displays large fluctuations which considerably hinder its convergence. ‖

Example 3.14. Transition matrix estimation. Consider a Markov chain with two states, 1 and 2, whose transition matrix is

$$T = \begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}\,,$$

that is,

$$P(X_{t+1} = 1|X_t = 1) = 1 - P(X_{t+1} = 2|X_t = 1) = p_1,$$
$$P(X_{t+1} = 2|X_t = 2) = 1 - P(X_{t+1} = 1|X_t = 2) = p_2\,.$$

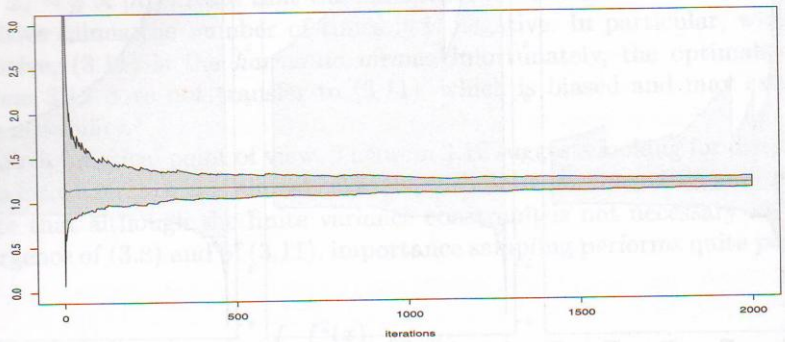

Fig. 3.6. Empirical range of the importance sampling estimator of $\mathbb{E}_f[|X/(1-X)|^{1/2}]$ for $\nu = 12$ and 500 replications based on the double Gamma $\mathcal{G}a(\alpha, 1)$ distribution folded at 1 when $\alpha = .5$. Average of the 500 series in overlay.

of h_2, a uniform distribution on $[0, 1/2.1]$ is reasonable, since the expectation $\mathbb{E}_f[h_2(X)]$ can be written as

$$\int_0^{1/2.1} u^{-7}\, f(1/u)\, du = \frac{1}{2.1} \int_0^{1/2.1} 2.1\, u^{-7}\, f(1/u)\, du\,,$$

as in Example 3.8. The corresponding importance sampling estimator is then

$$\delta_2 = \frac{1}{2.1m} \sum_{j=1}^{m} U_j^{-7} f(1/U_j)\,,$$

where the U_j's are iid $\mathcal{U}([0, 1/2.1])$. Figure 3.7 shows the improvement brought by this choice, with the estimator δ_2 converging to the true value after only a few hundred iterations. The importance sampling estimator associated with the Cauchy distribution is also quite stable, but it requires more iterations to achieve the same precision. Both of the other estimators (which are based on the true distribution and the normal distribution, respectively) fluctuate around the exact value with high-amplitude jumps, because their variance is infinite.

In the case of h_3, a reasonable candidate for the instrumental distribution is $g(x) = \exp(-x)\mathbb{I}_x \geq 0$, leading to the estimation of

$$\begin{aligned}\mathbb{E}_f[h_3(X)] &= \int_0^{\infty} \frac{x^5}{1+(x-3)^2}\, f(x)\, dx \\ &= \int_0^{\infty} \frac{x^5 e^x}{1+(x-3)^2}\, f(x)\, e^{-x}\, dx\end{aligned}$$

by

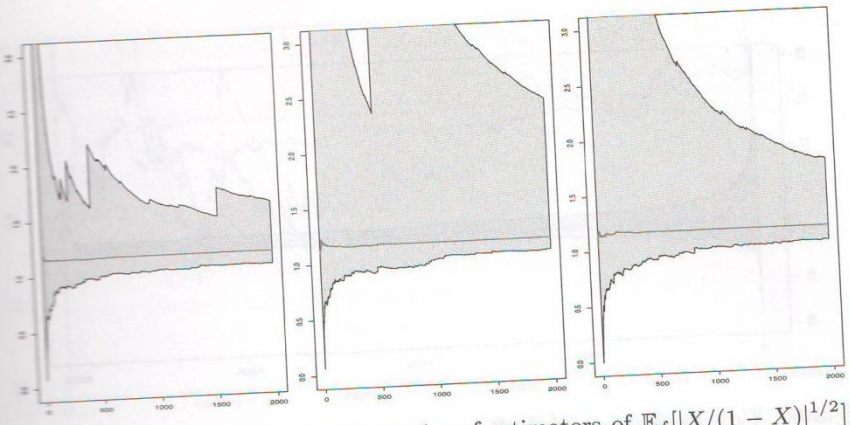

Fig. 3.5. Empirical range of three series of estimators of $\mathbb{E}_f[|X/(1-X)|^{1/2}]$ for $\nu = 12$ and 500 replications: sampling from f *(left)*, importance sampling with a Cauchy instrumental distribution *(center)* and importance sampling with normal importance distribution *(right)*. Average of the 500 series in overlay.

other hand, the $\mathcal{C}(0,1)$ distribution has larger tails than f and ensures that the variance of f/g is finite.

Figure 3.5 illustrates the performances of the three corresponding estimators for the function h_1 when $\nu = 12$ by representing the range of 500 series over 2000 iterations. The average of these series is quite stable over iterations and does not depend on the choice of the importance function, while the range exhibits wide jumps for all three. This phenomenon is due to the fact that the function h_1 has a singularity at $x = 1$ such that h_1^2 is not integrable under f but also such that none of the two other importance sampling estimators has a finite variance (Problem 3.20)! Were we to repeat this experiment with 5000 series rather than 500 series, we would then see larger ranges. There is thus no possible comparison between the three proposals in this case, since they all are inefficient. An alternative choice devised purposely for this function h_1 is to choose g such that $(1-x)g(x)$ is better behaved in $x = 1$. If we take for instance the double Gamma distribution folded at 1, that is, the distribution of X symmetric around 1 such that

$$|X - 1| \sim \mathcal{G}a(\alpha, 1)\,,$$

the ratio

$$h_1(x)\frac{f^2(x)}{g(x)} \propto \sqrt{x}\, f^2(x)\, |1-x|^{1-\alpha-1}\, exp|1-x|$$

is integrable around $x = 1$ when $\alpha < 1$. Obviously, the exponential part creates problems at ∞ and leads once more to an infinite variance, but it has much less influence on the stability of the estimator, as shown in Figure 3.6.

Since both h_2 and h_3 have restricted supports, we could benefit by having the instrumental distributions take this information into account. In the case

where $x_j \sim g \propto |h|f$. Note that the numerator is the number of times $h(x_j)$ is positive minus the number of times it is negative. In particular, when h is positive, (3.11) is the *harmonic mean*. Unfortunately, the optimality of Theorem 3.12 does not transfer to (3.11), which is biased and may exhibit severe instability.[2]

From a practical point of view, Theorem 3.12 suggests looking for distributions g for which $|h|f/g$ is almost constant with finite variance. It is important to note that although the finite variance constraint is not necessary for the convergence of (3.8) and of (3.11), importance sampling performs quite poorly when

$$\int \frac{f^2(x)}{g(x)}\, dx = +\infty\,, \tag{3.12}$$

whether in terms of behavior of the estimator (high-amplitude jumps, instability of the path of the average, slow convergence) or of comparison with direct Monte Carlo methods. Distributions g such that (3.12) occurs are therefore not recommended.

The next two examples show that importance sampling methods can bring considerable improvement over naïve Monte Carlo estimates when implemented with care. However, they can encounter disastrous performances and produce extremely poor estimates when the variance conditions are not met.

Example 3.13. Student's t distribution. Consider $X \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}.$$

Without loss of generality, we take $\theta = 0$ and $\sigma = 1$. We choose the quantities of interest to be $\mathbb{E}_f[h_i(X)]$ $(i = 1, 2, 3)$, with

$$h_1(x) = \sqrt{\left|\frac{x}{1-x}\right|}, \quad h_2(x) = x^5 \mathbb{I}_{[2.1,\infty[}(x), \quad h_3(x) = \frac{x^5}{1+(x-3)^2}\, \mathbb{I}_{x\geq 0}\,.$$

Obviously, it is possible to generate directly from f. Importance sampling alternatives are associated here with a Cauchy $\mathcal{C}(0,1)$ distribution and a normal $\mathcal{N}(0, \nu/(\nu-2))$ distribution (scaled so that the variance is the same as $\mathcal{T}(\nu, \theta, \sigma^2)$). The choice of the normal distribution is not expected to be efficient, as the ratio

$$\frac{f^2(x)}{g(x)} \propto \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}}$$

does not have a finite integral. However, this will give us an opportunity to study the performance of importance sampling in such a situation. On the

[2] In fact, the optimality only applies to the numerator, while another sequence should be used to better approximate the denominator.

$$\frac{\sum_{j=1}^{m} h(x_j)\, f(x_j)/g(x_j)}{\sum_{j=1}^{m} f(x_j)/g(x_j)}, \tag{3.10}$$

where we have replaced m with the sum of the weights. Since $(1/m)\sum_{j=1}^{m} f(x_j)/g(x_j)$ converges to 1 as $m \to \infty$, this estimator also converges to $\mathbb{E}_f h(X)$ by the Strong Law of Large Numbers.. Although this estimator is biased, the bias is small, and the improvement in variance makes it a preferred alternative to (3.8) (see also Lemma 4.3). In fact, Casella and Robert (1998) have shown that the weighted estimator (3.10) may perform better (when evaluated under squared error loss) in some settings. (See also Van Dijk and Kloeck 1984.) For instance, when h is nearly constant, (3.10) is close to this value, while (3.8) has a higher variation since the sum of the weights is different from one.

Among the distributions g leading to finite variances for the estimator (3.8), it is, in fact, possible to exhibit the optimal distribution corresponding to a given function h and a fixed distribution f, as stated by the following result of Rubinstein (1981); see also Geweke (1989).

Theorem 3.12. *The choice of g that minimizes the variance of the estimator (3.8) is*

$$g^*(x) = \frac{|h(x)|\, f(x)}{\int_{\mathcal{X}} |h(z)|\, f(z)\, dz}.$$

Proof. First note that

$$\operatorname{var}\left[\frac{h(X)f(X)}{g(X)}\right] = \mathbb{E}_g\left[\frac{h^2(X)f^2(X)}{g^2(X)}\right] - \left(\mathbb{E}_g\left[\frac{h(X)f(X)}{g(X)}\right]\right)^2,$$

and the second term does not depend on g. So, to minimize variance, we only need minimize the first term. From Jensen's inequality it follows that

$$\mathbb{E}_g\left[\frac{h^2(X)f^2(X)}{g^2(X)}\right] \geq \left(\mathbb{E}_g\left[\frac{|h(X)|f(X)}{g(X)}\right]\right)^2 = \left(\int |h(x)|f(x)dx\right)^2,$$

which provides a lower bound that is independent of the choice of g. It is straightforward to verify that this lower bound is attained by choosing $g = g^*$. □

This optimality result is rather formal since, when $h(x) > 0$, the optimal choice $g^*(x)$ requires the knowledge of $\int h(x)f(x)dx$, the integral of interest! A practical alternative taking advantage of Theorem 3.12 is to use the estimator (3.10) as

$$\frac{\sum_{j=1}^{m} h(x_j)\, f(x_j)/g(x_j)}{\sum_{j=1}^{m} f(x_j)/g(x_j)} = \frac{\sum_{j=1}^{m} h(x_j)|h(x_j)|^{-1}}{\sum_{j=1}^{m} |h(x_j)|^{-1}}, \tag{3.11}$$

Of course, the problem is that we are calculating the probability of a very rare event, and naïve simulation will need a lot of iterations to get a reasonable answer. However, with importance sampling we can greatly improve our accuracy.

Let $Y \sim \mathcal{TE}(4.5, 1)$, an exponential distribution (left) truncated at 4.5 with scale 1, with density

$$f_Y(y) = e^{-(x-4.5)} / \int_{4.5}^{\infty} e^{-x} dx \,.$$

If we now simulate from f_Y and use importance sampling, we obtain (see Problem 3.16)

$$P(Z > 4.5) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{\varphi(Y^{(i)})}{f_Y(Y^{(i)})} \mathbb{I}(Y^{(i)} > 4.5) = .000003377 \,. \quad \|$$

3.3.2 Finite Variance Estimators

Although the distribution g can be almost any density for the estimator (3.8) to converge, there are obviously some choices that are better than others, and it is natural to try to compare different distributions g for the evaluation of (3.4). First, note that, while (3.8) does converge almost surely to (3.4), its variance is finite only when the expectation

$$\mathbb{E}_g \left[h^2(X) \frac{f^2(X)}{g^2(X)} \right] = \mathbb{E}_f \left[h^2(X) \frac{f(X)}{g(X)} \right] = \int_{\mathcal{X}} h^2(x) \, \frac{f^2(x)}{g(x)} \, dx < \infty \,.$$

Thus, instrumental distributions with tails lighter than those of f (that is, those with unbounded ratios f/g) are not appropriate for importance sampling. In fact, in these cases, the variances of the corresponding estimators (3.8) will be infinite for many functions h. More generally, if the ratio f/g is unbounded, the weights $f(x_j)/g(x_j)$ will vary widely, giving too much importance to a few values x_j. This means that the estimator (3.8) may change abruptly from one iteration to the next one, even after many iterations. Conversely, distributions g with thicker tails than f ensure that the ratio f/g does not cause the divergence of $\mathbb{E}_f[h^2 f/g]$. In particular, Geweke (1989) mentions two types of sufficient conditions:

(a) $f(x)/g(x) < M \;\; \forall x \in \mathcal{X}$ and $\text{var}_f(h) < \infty$;
(b) $\mathcal{X}$ is compact, $f(x) < F$ and $g(x) > \varepsilon \;\; \forall x \in \mathcal{X}$.

These conditions are quite restrictive. In particular, $f/g < M$ implies that the Accept–Reject algorithm [A.4] also applies. (A comparison between the two approaches is given in Section 3.3.3.)

An alternative to (3.8) which addresses the finite variance issue, and generally yields a more stable estimator, is to use

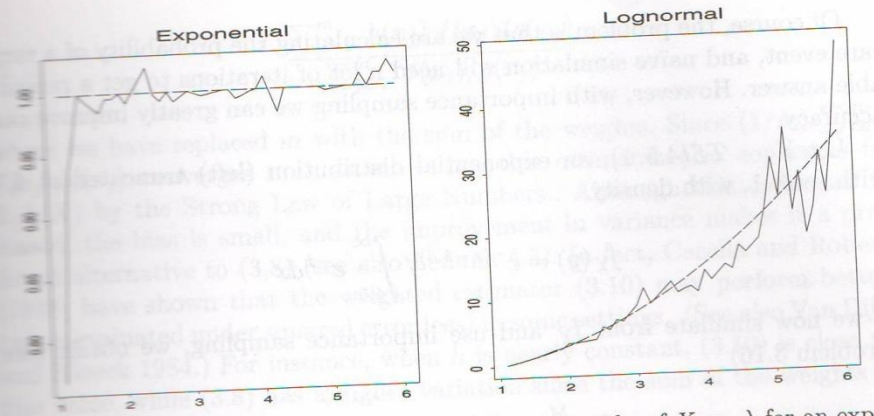

Fig. 3.4. Graph of approximate scaled squared error risks of X vs. λ for an exponential and a log-normal observation, compared with the theoretical values (dashes) for $\lambda \in [1,6]$ (10,000 simulations).

$$\hat{R}_2 = \frac{1}{T\lambda^2}\sum_{t=1}^{T}(X_t - \lambda)^2$$

in the log-normal case. In addition, the scale nature of the parameterization allows a single sample $(Y_1^0, \ldots, Y_T^0)$ from $\mathcal{N}(0,1)$ to be used for all σ's, with $X_t = \exp(\sigma Y_t^0)$.

The comparison of these evaluations is given in Figure 3.4 for $T = 10,000$, each point corresponding to a sample of size T simulated from $\mathcal{LN}(0,\sigma^2)$ by the above transformation. The exact values are given by 1 and $(\lambda+1)(\lambda-1)$, respectively. Note that implementing importance sampling in the opposite way offers little appeal since the weights $\exp\{-\log(X_t)^2/2\sigma^2\} \times \exp(\lambda X_t)/X_t$ have infinite variance (see below). The graph of the risk in the exponential case is then more stable than for the original sample from the log-normal distribution. ‖

We close this section by revisiting a previous example with a new twist.

Example 3.11. Small tail probabilities. In Example 3.5 we calculated normal tail probabilities with Monte Carlo sums, and found the method to work well. However, the method breaks down if we need to go too far into the tail. For example, if $Z \sim \mathcal{N}(0,1)$, and we are interested in the probability $P(Z > 4.5)$ (which we know is very small), we could simulate $Z^{(i)} \sim \mathcal{N}(0,1)$ for $i = 1, \ldots M$ and calculate

$$P(Z > 4.5) \approx \frac{1}{M}\sum_{i=1}^{M}\mathbb{I}(Z^{(i)} > 4.5)\,.$$

If we do this, a value of $M = 10,000$ usually produces all zeros of the indicator function.

occur in Bayesian analysis, only generic methods can be compared (that is to say, those which are independent of h).

The principal alternative to direct sampling from f for the evaluation of (3.4) is to use importance sampling, defined as follows:

Definition 3.9. The method of *importance sampling* is an evaluation of (3.4) based on generating a sample $X_1, \ldots, X_n$ from a given distribution g and approximating

$$\mathbb{E}_f[h(X)] \approx \frac{1}{m} \sum_{j=1}^{m} \frac{f(X_j)}{g(X_j)} h(X_j) \,. \tag{3.8}$$

This method is based on the alternative representation of (3.4):

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x)\, dx \,, \tag{3.9}$$

which is called the *importance sampling fundamental identity*, and the estimator (3.8) converges to (3.4) for the same reason the regular Monte Carlo estimator $\overline{h}_m$ converges, whatever the choice of the distribution g (as long as $\text{supp}(g) \supset \text{supp}(f)$).

Note that (3.9) is a very general representation that expresses the fact that a given integral is not intrinsically associated with a given distribution. Example 3.8 shows how much of an effect this choice of representation can have. Importance sampling is therefore of considerable interest since it puts very little restriction on the choice of the instrumental distribution g, which can be chosen from distributions that are easy to simulate. Moreover, the same sample (generated from g) can be used repeatedly, not only for different functions h but also for different densities f, a feature which is quite attractive for robustness and Bayesian sensitivity analyses.

Example 3.10. Exponential and log-normal comparison. Consider X as an estimator of λ, when $X \sim \mathcal{E}xp(1/\lambda)$ or when $X \sim \mathcal{LN}(0, \sigma^2)$ (with $e^{\sigma^2/2} = \lambda$, see Problem 3.11). If the goal is to compare the performances of this estimator under both distributions for the scaled squared error loss

$$L(\lambda, \delta) = (\delta - \lambda)^2/\lambda^2,$$

a *single* sample from $\mathcal{LN}(0, \sigma^2)$, $X_1, \ldots, X_T$, can be used for both purposes, the risks being evaluated by

$$\hat{R}_1 = \frac{1}{T\lambda^2} \sum_{t=1}^{T} X_t e^{-X_t/\lambda} \lambda^{-1} e^{\log(X_t)^2/2\sigma^2} \sqrt{2\pi}\sigma (X_t - \lambda)^2$$

in the exponential case and by

has variance $p(1-2p)/2m$ equal to $0.052/m$.

The (relative) inefficiency of these methods is due to the generation of values outside the domain of interest, $[2,+\infty)$, which are, in some sense, irrelevant for the approximation of p. If p is written as

$$p = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)}\, dx\,,$$

the integral above can be considered to be the expectation of $h(X) = 2/\pi(1+X^2)$, where $X \sim \mathcal{U}_{[0,2]}$. An alternative method of evaluation for p is therefore

$$\hat{p}_3 = \frac{1}{2} - \frac{1}{m}\sum_{j=1}^{m} h(U_j)$$

for $U_j \sim \mathcal{U}_{[0,2]}$. The variance of $\hat{p}_3$ is $(\mathbb{E}[h^2] - \mathbb{E}[h]^2)/m$ and an integration by parts shows that it is equal to $0.0285/m$. Moreover, since p can be written as

$$p = \int_0^{1/2} \frac{y^{-2}}{\pi(1+y^{-2})}\, dy\,,$$

this integral can also be seen as the expectation of $\frac{1}{4}\, h(Y) = 1/2\pi(1+Y^2)$ against the uniform distribution on $[0,1/2]$ and another evaluation of p is

$$\hat{p}_4 = \frac{1}{4m}\sum_{j=1}^{m} h(Y_j)$$

when $Y_j \sim \mathcal{U}_{[0,1/2]}$. The same integration by parts shows that the variance of $\hat{p}_4$ is then $0.95\ 10^{-4}/m$.

Compared with $\hat{p}_1$, the reduction in variance brought by $\hat{p}_4$ is of order 10^{-3}, which implies, in particular, that this evaluation requires $\sqrt{1000} \approx 32$ times fewer simulations than $\hat{p}_1$ to achieve the same precision. ‖

The evaluation of (3.4) based on simulation from f is therefore not necessarily optimal and Theorem 3.12 shows that this choice is, in fact, always suboptimal. Note also that the integral (3.4) can be represented in an infinite number of ways by triplets $(\mathcal{X}, h, f)$. Therefore, the search for an optimal estimator should encompass all these possible representations (as in Example 3.8). As a side remark, we should stress that the very notion of "optimality" of a representation is quite difficult to define precisely. Indeed, as already noted in Chapter 2, the comparison of simulation methods cannot be equated with the comparison of the variances of the resulting estimators. Conception and computation times should also be taken into account. At another level, note that the optimal method proposed in Theorem 3.12 depends on the function h involved in (3.4). Therefore, it cannot be considered as optimal when several integrals related to f are simultaneously evaluated. In such cases, which often

of the quantiles of the distribution of (3.5) under H_0 or to evaluate the power of a standard test. ‖

It may seem that the method proposed above is sufficient to approximate integrals like (3.4) in a controlled way. However, while the straightforward Monte Carlo method indeed provides good approximations of (3.4) in most regular cases, there exist more efficient alternatives which not only avoid a direct simulation from f but also can be used repeatedly for several integrals of the form (3.4). The repeated use can be for either a family of functions h or a family of densities f. In particular, the usefulness of this flexibility is quite evident in Bayesian analyses of *robustness*, of *sensitivity* (see Berger 1990, 1994), or for the computation of power functions of specific tests (see Lehmann 1986, or Gouriéroux and Monfort 1996).

3.3 Importance Sampling

3.3.1 Principles

The method we now study is called *importance sampling* because it is based on so-called *importance functions*, and although it would be more accurate to call it "weighted sampling," we will follow common usage. We start this section with a somewhat unusual example, borrowed from Ripley (1987), which shows that it may actually pay to generate from a distribution other than the distribution f of interest or, in other words, to modify the representation of an integral as an expectation against a given density. (See Note 3.6.1 for a global approach to the approximation of tail probabilities by *large deviation* techniques.)

Example 3.8. Cauchy tail probability. Suppose that the quantity of interest is the probability, p, that a Cauchy $\mathcal{C}(0,1)$ variable is larger than 2, that is,

$$p = \int_2^{+\infty} \frac{1}{\pi(1+x^2)}\, dx\,.$$

When p is evaluated through the empirical average

$$\hat{p}_1 = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}_{X_j>2}$$

of an iid sample $X_1,\ldots,X_m \sim \mathcal{C}(0,1)$, the variance of this estimator is $p(1-p)/m$ (equal to $0.127/m$ since $p = 0.15$). This variance can be reduced by taking into account the symmetric nature of $\mathcal{C}(0,1)$, since the average

$$\hat{p}_2 = \frac{1}{2m}\sum_{j=1}^{m} \mathbb{I}_{|x_j|>2}$$

$$p\,\mathcal{N}(\mu,1) + (1-p)\,\mathcal{N}(\mu+\theta,1)\,,$$

where the constraint $\theta > 0$ ensures *identifiability*. A test on the existence of a mixture cannot be easily represented in a hypothesis test since $H_0 : p = 0$ effectively eliminates the mixture and results in the identifiability problem related with $\mathcal{N}(\mu+\theta,1)$. (The inability to estimate the nuisance parameter p under H_0 results in the likelihood not satisfying the necessary regularity conditions; see Davies 1977. However, see Lehmann and Casella 1998, Section 6.6 for mixtures where it is possible to construct efficient estimators.)

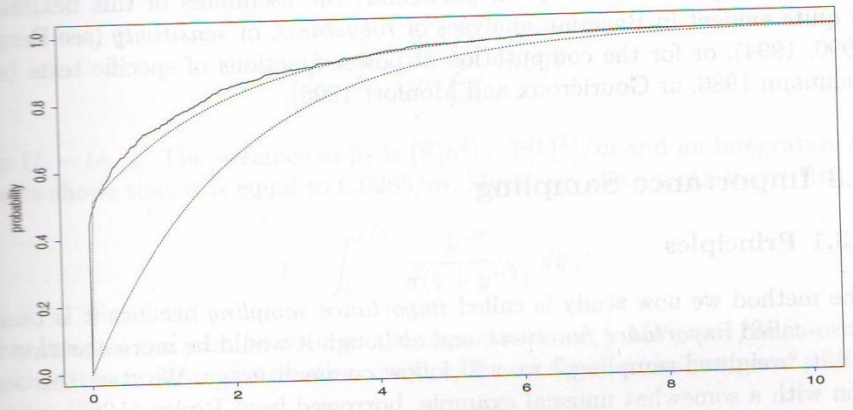

Fig. 3.3. Empirical cdf of a sample of log-likelihood ratios for the test of presence of a Gaussian mixture (solid lines) and comparison with the cdf of a χ^2_2 distribution (dotted lines, below) and with the cdf of a $.5 - .5$ mixture of a χ^2_2 distribution and of a Dirac mass at 0 (dotted lines, above) (based on 1000 simulations of a normal $\mathcal{N}(0,1)$ sample of size 100).

A slightly different formulation of the problem will allow a solution, however. If the identifiability constraint is taken to be $p \geq 1/2$ instead of $\theta > 0$, then H_0 can be represented as

$$H_0 : \quad p = 1 \quad \text{or} \quad \theta = 0\,.$$

We therefore want to determine the limiting distribution of (3.5) under this hypothesis and under a local alternative. Figure 3.3 represents the empirical cdf of $2\,\{\log\,\ell(\hat{p},\hat{\mu},\hat{\theta}|x) - \log\,\ell(\hat{\mu}^0|x)\}$ and compares it with the χ^2_2 cdf, where $\hat{p},\hat{\mu},\hat{\theta}$, and $\hat{\mu}^0$ are the respective MLEs for 1000 simulations of a normal $\mathcal{N}(0,1)$ sample of size 100. The poor agreement between the asymptotic approximation and the empirical cdf is quite obvious. Figure 3.3 also shows how the χ^2_2 approximation is improved if the limit (3.5) is replaced by an equally weighted mixture of a Dirac mass at 0 and a χ^2_2 distribution.

Note that the resulting sample of the log-likelihood ratios can also be used for inferential purposes, for instance to derive an exact test via the estimation

Percentile	Monte Carlo	χ^2_1
.10	2.84	3.87
.05	3.93	4.68
.01	6.72	6.36

Table 3.3. Cutoff points for the null distribution f_0 compared to χ^2_1.

To run the Monte Carlo experiment, we need to generate values from $f_0(\lambda)$. Since this distribution is not completely specified (the parameters p_1 and p_2 can be any value in $(0,1)$), to generate a value from $f_0(\lambda)$ we generate

$$p_i \sim \mathcal{U}(0,1), \quad i = 1,2, \tag{3.7}$$
$$\mathbf{X} \sim \mathcal{M}_4(p_1p_2, p_1(1-p_2), (1-p_1)p_2, (1-p_1)(1-p_2)),$$

and calculate $\lambda(\mathbf{x})$. The results, given in Table 3.3 and Figure 3.2, show that the Monte Carlo null distribution has a slightly different shape than the χ^2_1 distribution, being slightly more concentrated around 0 but with longer tails.

The analysis of the given data is somewhat anticlimactic, as the observed value of $\lambda(\mathbf{y})$ is .594, which according to any calibration gives overwhelming support to H_0. ‖

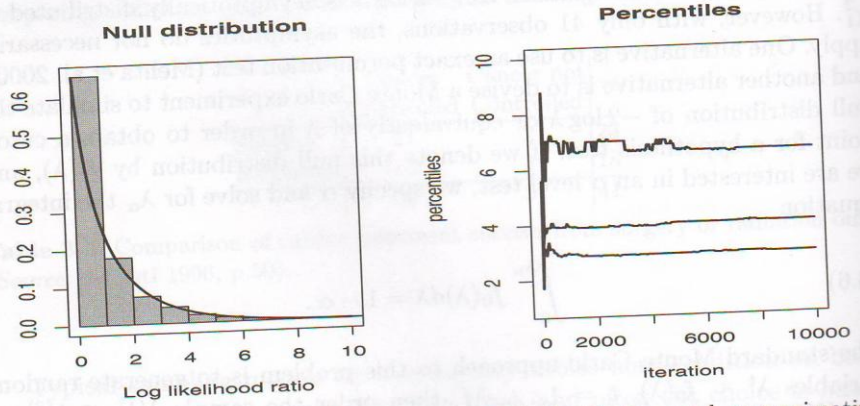

Fig. 3.2. For Example 3.6, histogram of null distribution and approximating χ^2_1 density(left panel). The right panel gives the running empirical percentiles (.90, .95, .99), from bottom to top. Notice the higher variability in the higher percentiles (10,000 simulations).

Example 3.7. Testing the number of components. A situation where the standard χ^2_r regularity conditions do not apply for the likelihood ratio test is that of the normal mixture (see Example 1.10)

$$X_i \sim \mathcal{M}_4(1, \mathbf{p}), \quad i = 1, \ldots, n\,.$$

If we denote by y_{ij} the number of x_i that are in cell ij, the likelihood function can be written

$$\ell(\mathbf{p}|\mathbf{y}) \propto \prod_{ij} p_{ij}^{y_{ij}}\,.$$

The null hypothesis to be tested is one of independence, which is to say that the treatment has no bearing on the control of cancer. To translate this into a parameter statement, we note that the full parameter space corresponding to Table 3.2 is

p_{11}	p_{12}	p_1
p_{21}	p_{22}	$1-p_1$
p_2	$1-p_2$	1

and the null hypothesis of independence is $H_0 : p_{11} = p_1 p_2$. The likelihood ratio statistic for testing this hypothesis is

$$\lambda(\mathbf{y}) = \frac{\max_{\mathbf{p}:p_{11}=p_1p_2} \ell(\mathbf{p}|\mathbf{y})}{\max_{\mathbf{p}} \ell(\mathbf{p}|\mathbf{y})}\,.$$

It is straightforward to show that the numerator maximum is attained at $\hat{p}_1 = (y_{11} + y_{12})/n$ and the denominator maximum at $\hat{p}_{ij} = y_{ij}/n$.

As mentioned above, under H_0, $-2\log\lambda$ is asymptotically distributed as χ^2_1. However, with only 41 observations, the asymptotics do not necessarily apply. One alternative is to use an exact permutation test (Mehta et al. 2000), and another alternative is to devise a Monte Carlo experiment to simulate the null distribution of $-2\log\lambda$ or equivalently of λ in order to obtain a cutoff point for a hypothesis test. If we denote this null distribution by $f_0(\lambda)$, and we are interested in an α level test, we specify α and solve for λ_α the integral equation

$$\int_0^{\lambda_\alpha} f_0(\lambda)d\lambda = 1 - \alpha\,. \tag{3.6}$$

The standard Monte Carlo approach to this problem is to generate random variables $\lambda^t \sim f_0(\lambda)$, $t = 1, \ldots, M$, then order the sample $\lambda^{(1)} \leq \lambda^{(2)} \leq \cdots \lambda^{(M)}$ and finally calculate the empirical $1-\alpha$ percentile $\lambda^{(\lfloor(1-\alpha)M\rfloor)}$. We then have

$$\lim_{M\to\infty} \lambda^{(\lfloor(1-\alpha)M\rfloor)} \to \lambda_\alpha\,.$$

(Note that this is a slightly unusual Monte Carlo experiment in that α is known and λ_α is not, but it is nonetheless based on the same convergence of empirical measures.)

We mentioned in Section 3.1 the potential of this approach in evaluating estimators based on a decision-theoretic formulation. The same applies for testing, when the level of significance of a test, and its power function, cannot be easily computed, and simulation thus can provide a useful improvement over asymptotic approximations when explicit computations are impossible. The following example illustrates this somewhat different application of Monte Carlo integration.

Many tests are based on an asymptotic normality assumption as, for instance, the *likelihood ratio test*. Given H_0, a null hypothesis corresponding to r independent constraints on the parameter $\theta \in \mathbb{R}^k$, denote by $\hat{\theta}$ and $\hat{\theta}^0$ the unconstrained and constrained (under H_0) maximum likelihood estimators of θ, respectively. The likelihood ratio $\ell(\hat{\theta}|x)/\ell(\hat{\theta}^0|x)$ then satisfies

$$\log[\ell(\hat{\theta}|x)/\ell(\hat{\theta}^0|x)] = 2\,\{\log\,\ell(\hat{\theta}|x) - \log\,\ell(\hat{\theta}^0|x)\} \xrightarrow{\mathcal{L}} \chi^2_r\,, \tag{3.5}$$

when the number of observations goes to infinity (see Lehmann 1986, Section 8.8, or Gouriéroux and Monfort 1996). However, the χ^2_r approximation only holds asymptotically and, further, this convergence only holds under regularity constraints on the likelihood function (see Lehmann and Casella 1998, Chapter 6, for a full development); hence, the asymptotics may even not apply.

Example 3.6. Contingency Tables. Table 3.2 gives the results of a study comparing radiation therapy with surgery in treating cancer of the larynx.

	Cancer Controlled	Cancer not Controlled	
Surgery	21	2	23
Radiation	15	3	18
	36	5	41

Table 3.2. Comparison of cancer treatment success from surgery or radiation only (*Source:* Agresti 1996, p.50).

Typical sampling models for contingency tables may condition on both margins, one margin, or only the table total, and often the choice is based on philosophical reasons (see, for example, Agresti 1992). In this case we may argue for conditioning on the number of patients in each group, or we may just condition on the table total (there is little argument for conditioning on both margins). Happily, in many cases the resulting statistical conclusion is not dependent on this choice but, for definiteness, we will choose to condition only on the table total, $n = 41$.

Under this model, each observation X_i comes from a multinomial distribution with four cells and cell probabilities $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$, with $\sum_{ij} p_{ij} = 1$, that is,

n	0.0	0.67	0.84	1.28	1.65	2.32	2.58	3.09	3.72
10^2	0.485	0.74	0.77	0.9	0.945	0.985	0.995	1	1
10^3	0.4925	0.7455	0.801	0.902	0.9425	0.9885	0.9955	0.9985	1
10^4	0.4962	0.7425	0.7941	0.9	0.9498	0.9896	0.995	0.999	0.9999
10^5	0.4995	0.7489	0.7993	0.9003	0.9498	0.9898	0.995	0.9989	0.9999
10^6	0.5001	0.7497	0.8	0.9002	0.9502	0.99	0.995	0.999	0.9999
10^7	0.5002	0.7499	0.8	0.9001	0.9501	0.99	0.995	0.999	0.9999
10^8	0.5	0.75	0.8	0.9	0.95	0.99	0.995	0.999	0.9999

Table 3.1. Evaluation of some normal quantiles by a regular Monte Carlo experiment based on n replications of a normal generation. The last line gives the exact values.

The approach followed in the above example can be successfully utilized in many cases, even though it is often possible to achieve greater efficiency through numerical methods (Riemann quadrature, Simpson method, etc.) in dimension 1 or 2. The scope of application of this Monte Carlo integration method is obviously not limited to the Bayesian paradigm since, similar to Example 3.3, the performances of complex procedures can be measured in any setting where the distributions involved in the model can be simulated. For instance, we can use Monte Carlo sums to calculate a normal cumulative distribution function (even though the normal cdf can now be found in all software and many pocket calculators).

Example 3.5. Normal cdf. Since the normal cdf cannot be written in an explicit form, a possible way to construct normal distribution tables is to use simulation. Consider the generation of a sample of size n, $(x_1, \ldots, x_n)$, based on the Box–Muller algorithm $[A_4]$ of Example 2.2.2.

The approximation of

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

by the Monte Carlo method is thus

$$\hat{\Phi}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{x_i \le t},$$

with (exact) variance $\Phi(t)(1-\Phi(t))/n$ (as the variables $\mathbb{I}_{x_i \le t}$ are independent Bernoulli with success probability $\Phi(t)$). For values of t around $t = 0$, the variance is thus approximately $1/4n$, and to achieve a precision of four decimals, the approximation requires on average $n = (\sqrt{2}\ 10^4)^2$ simulations, that is, 200 million iterations. Table 3.1 gives the evolution of this approximation for several values of t and shows an accurate evaluation for 100 million iterations. Note that greater (absolute) accuracy is achieved in the tails and that more efficient simulations methods could be used, as in Example 3.8 below. ‖

can also be estimated from the sample $(X_1, \ldots, X_m)$ through

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} [h(x_j) - \overline{h}_m]^2 .$$

For m large,

$$\frac{\overline{h}_m - \mathbb{E}_f[h(X)]}{\sqrt{v_m}}$$

is therefore approximately distributed as a $\mathcal{N}(0,1)$ variable, and this leads to the construction of a convergence test and of confidence bounds on the approximation of $\mathbb{E}_f[h(X)]$.

Example 3.4. A first Monte Carlo integration. Recall the function (1.26) that we saw in Example 1.17, $h(x) = [\cos(50x) + \sin(20x)]^2$. As a first example, we look at integrating this function, which is shown in Figure 3.1 *(left)*. Although it is possible to integrate this function analytically, it is a good first test case. To calculate the integral, we generate $U_1, U_2, \ldots, U_n$ iid $\mathcal{U}(0,1)$ random variables, and approximate $\int h(x)dx$ with $\sum h(U_i)/n$. The center panel in Figure 3.1 shows a histogram of the values of $h(U_i)$, and the last panel shows the running means and standard errors. It is clear that the Monte Carlo average is converging, with value of 0.963 after 10,000 iterations. This compares favorably with the exact value of 0.965. (See Example 4.1 for a more formal monitoring of convergence.) ||

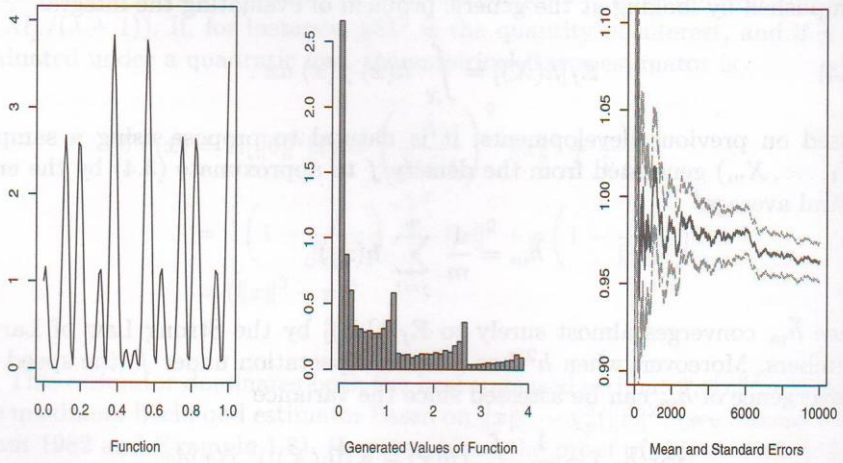

Fig. 3.1. Calculation of the integral of the function (1.26): *(left)* function (1.26), *(center)* histogram of 10,000 values $h(U_i)$, simulated using a uniform generation, and *(right)* mean $\pm$ one standard error.

allow for analytic expressions. This makes their evaluation under a given loss problematic.

Given a sampling distribution $f(x|\theta)$ and a conjugate prior distribution $\pi(\theta|\lambda,\mu)$, the empirical Bayes method estimates the *hyperparameters* λ and μ from the *marginal distribution*

$$m(x|\lambda,\mu) = \int f(x|\theta)\,\pi(\theta|\lambda,\mu)\,d\theta$$

by maximum likelihood. The estimated distribution $\pi(\theta|\hat{\lambda},\hat{\mu})$ is often used as in a standard Bayesian approach (that is, without taking into account the effect of the substitution) to derive a point estimator. See Searle et al. (1992, Chapter 9) or Carlin and Louis (1996) for a more detailed discussion on this approach. (We note that this approach is sometimes called *parametric* empirical Bayes, as opposed to the *nonparametric* empirical Bayes approach developed by Herbert Robbins. See Robbins 1964, 1983 or Maritz and Lwin 1989 for details.) The following example illustrates some difficulties encountered in evaluating empirical Bayes estimators (see also Example 4.12).

Example 3.3. Empirical Bayes estimator. Let X have the distribution $X \sim \mathcal{N}_p(\theta, I_p)$ and let $\theta \sim \mathcal{N}_p(\mu, \lambda I_p)$, the corresponding conjugate prior. The hyperparameter μ is often specified, and here we take $\mu = 0$. In the empirical Bayes approach, the scale hyperparameter λ is replaced by the maximum likelihood estimator, $\hat{\lambda}$, based on the marginal distribution $X \sim \mathcal{N}_p(0, (\lambda+1)I_p)$. This leads to the maximum likelihood estimator $\hat{\lambda} = (\|x\|^2/p - 1)^+$. Since the posterior distribution of θ given λ is $\mathcal{N}_p(\lambda x/(\lambda+1), \lambda I_p/(\lambda+1))$, empirical Bayes inference may be based on the pseudo-posterior $\mathcal{N}_p(\hat{\lambda}x/(\hat{\lambda}+1), \hat{\lambda}I_p/(\hat{\lambda}+1))$. If, for instance, $\|\theta\|^2$ is the quantity of interest, and if it is evaluated under a quadratic loss, the empirical Bayes estimator is

$$\begin{aligned}\delta^{eb}(x) = \mathbb{E}(\|\theta\|^2|x) &= \left(\frac{\hat{\lambda}}{\hat{\lambda}+1}\right)^2 \|x\|^2 + \frac{\hat{\lambda}p}{\hat{\lambda}+1} \\ &= \left[\left(1-\frac{p}{\|x\|^2}\right)^+\right]^2 \|x\|^2 + p\left(1-\frac{p}{\|x\|^2}\right)^+ \\ &= (\|x\|^2 - p)^+ .\end{aligned}$$

This estimator dominates both the best unbiased estimator, $\|x\|^2 - p$, and the maximum likelihood estimator based on $\|x\|^2 \sim \chi^2_p(\|\theta\|^2)$ (see Saxena and Alam 1982 and Example 1.8). However, since the proof of this second domination result is quite involved, one might first check for domination through a simulation experiment that evaluates the risk function,

$$R(\theta,\delta) = \mathbb{E}_\theta[(\|\theta\|^2 - \delta)^2] ,$$

Example 3.2. Piecewise linear and quadratic loss functions. Consider a loss function which is piecewise quadratic,

$$\text{(3.3)} \qquad \mathrm{L}(\theta,\delta) = w_i(\theta-\delta)^2 \quad \text{when} \quad \theta-\delta \in [a_i, a_{i+1}), \quad \omega_i > 0.$$

Differentiating the posterior expectation (3.3) shows that the associated Bayes estimator satisfies

$$\sum_i w_i \int_{a_i}^{a_{i+1}} (\theta - \delta^\pi(x))\, \pi(\theta|x)\, d\theta = 0\,,$$

that is,

$$\delta^\pi(x) = \frac{\sum_i w_i \int_{a_i}^{a_{i+1}} \theta\, \pi(\theta)\, f(x|\theta)\, d\theta}{\sum_i w_i \int_{a_i}^{a_{i+1}} \pi(\theta)\, f(x|\theta)\, d\theta}\,.$$

Although formally explicit, the computation of $\delta^\pi(x)$ requires the computation of the posterior means restricted to the intervals $[a_i, a_{i+1})$ and of the posterior probabilities of these intervals.

Similarly, consider a piecewise linear loss function,

$$\mathrm{L}(\theta,\delta) = w_i|\theta-\delta| \quad \text{if} \quad \theta-\delta \in [a_i, a_{i+1}),$$

or Huber's (1972) loss function,

$$\mathrm{L}(\theta,\delta) = \begin{cases} \rho(\theta-\delta)^2 & \text{if } |\theta-\delta| < c, \\ 2\rho c\{|\theta-\delta| - c/2\} & \text{otherwise,} \end{cases}$$

where ρ and c are specified constants. Although a specific type of prior distribution leads to explicit formulas, most priors result only in integral forms of δ^π. Some of these may be quite complex. ‖

Inference based on *classical decision theory* evaluates the performance of estimators (maximum likelihood estimator, best unbiased estimator, moment estimator, etc.) through the loss imposed by the decision-maker or by the setting. Estimators are then compared through their expected losses, also called risks. In most cases, it is impossible to obtain an analytical evaluation of the risk of a given estimator, or even to establish that a new estimator (uniformly) dominates a standard estimator.

It may seem that the topic of *James–Stein* estimation is an exception to this observation, given the abundant literature on the topic. In fact, for some families of distributions (such as exponential or spherically symmetric) and some types of loss functions (such as quadratic or concave), it is possible to analytically establish domination results over the maximum likelihood estimator or unbiased estimators (see Lehmann and Casella 1998, Chapter 5 or Robert 2001, Chapter 2). Nonetheless, in these situations, estimators such as *empirical Bayes estimators*, which are quite attractive in practice, will rarely

approach, may involve the integration of the empirical cdf. Similarly, alternatives to standard likelihood, such as *marginal* likelihood, may require the integration of the nuisance parameters (Barndorff-Nielsen and Cox 1994).

Although many calculations in Bayesian inference require integration, this is not always the case. Integration is clearly needed when the Bayes estimators are posterior expectations (see Section 1.3 and Problem 1.22), however Bayes estimators are not always posterior expectations. In general, the Bayes estimate under the loss function $\mathrm{L}(\theta,\delta)$ and the prior π is the solution of the minimization program

$$\min_{\delta} \int_{\Theta} \mathrm{L}(\theta,\delta)\,\pi(\theta)\,f(x|\theta)\,d\theta\,. \tag{3.1}$$

Only when the loss function is the quadratic function $\|\theta-\delta\|^2$ will the Bayes estimator be a posterior expectation. While some other loss functions lead to general solutions $\delta^\pi(x)$ of (3.1) in terms of $\pi(\theta|x)$ (see, for instance, Robert 1996b, 2001 for the case of *intrinsic losses*), a specific setup where the loss function is constructed by the decision-maker almost always precludes analytical integration of (3.1). This necessitates an approximate solution of (3.1) either by numerical methods or by simulation.

Thus, whatever the type of statistical inference, we are led to consider numerical solutions. The previous chapter has illustrated a number of methods for the generation of random variables with any given distribution and, hence, provides a basis for the construction of solutions to our statistical problems. Thus, just as the search for a stationary state in a dynamical system in physics or in economics can require one or several simulations of successive states of the system, statistical inference on complex models will often require the use of simulation techniques. (See, for instance, Bauwens 1984, Bauwens and Richard 1985 and Gouriéroux and Monfort 1996 for illustrations in econometrics.) We now look at a number of examples illustrating these situations before embarking on a description of simulation-based integration methods.

Example 3.1. L_1 loss. For $\theta\in\mathbb{R}$ and $\mathrm{L}(\theta,\delta)=|\theta-\delta|$, the Bayes estimator associated with π is the posterior median of $\pi(\theta|x)$, $\delta^\pi(x)$, which is the solution to the equation

$$\int_{\theta\le\delta^\pi(x)} \pi(\theta)\,f(x|\theta)\,d\theta = \int_{\theta\ge\delta^\pi(x)} \pi(\theta)\,f(x|\theta)\,d\theta\,. \tag{3.2}$$

In the setup of Example 1.7, that is, when $\lambda=\|\theta\|^2$ and $X\sim\mathcal{N}_p(\theta,I_p)$, this equation is quite complex, since, when using the reference prior of Example 1.12,

$$\pi(\lambda|x)\propto\lambda^{p-1/2}\int e^{-\|x-\theta\|^2/2}\prod_{i=1}^{p-2}\sin(\varphi_i)^{p-i-1}\,d\varphi_1\ldots d\varphi_{p-1}\,,$$

where $\lambda,\varphi_1,\ldots,\varphi_{p-1}$ are the polar coordinates of θ, that is, $\theta_1=\lambda\cos(\varphi_1)$, $\theta_2=\lambda\sin(\varphi_1)\cos(\varphi_2)$, ...

3

Monte Carlo Integration

> Cadfael had heard the words without hearing them and enlightenment fell on him so dazzlingly that he stumbled on the threshold.
> —Ellis Peter, *The Heretic's Apprentice*

While Chapter 2 focussed on developing techniques to produce random variables by computer, this chapter introduces the central concept of Monte Carlo methods, that is, taking advantage of the availability of computer generated random variables to approximate univariate and multidimensional integrals. In Section 3.2, we introduce the basic notion of Monte Carlo approximations as a byproduct of the Law of Large Numbers, while Section 3.3 highlights the universality of the approach by stressing the versatility of the representation of an integral as an expectation.

3.1 Introduction

Two major classes of numerical problems that arise in statistical inference are *optimization* problems and *integration* problems. (An associated problem, that of *implicit equations*, can often be reformulated as an optimization problem.) Although optimization is generally associated with the likelihood approach, and integration with the Bayesian approach, these are not strict classifications, as shown by Examples 1.5 and 1.15, and Examples 3.1, 3.2 and 3.3, respectively.

Examples 1.1–1.15 have also shown that it is not always possible to derive explicit probabilistic models and that it is even less possible to analytically compute the estimators associated with a given paradigm (maximum likelihood, Bayes, method of moments, etc.). Moreover, other statistical methods, such as *bootstrap* methods (see Note 1.6.2), although unrelated to the Bayesian

for the three estimators. This quadratic risk is often normalized by $1/(2\|\theta\|^2 + p)$ (which does not affect domination results but ensures the existence of a minimax estimator; see Robert 2001). Problem 3.8 contains a complete solution to the evaluation of risk. ‖

A general solution to the different computational problems contained in the previous examples and in those of Section 1.1 is to use simulation, of either the true or approximate distributions to calculate the quantities of interest. In the setup of Decision Theory, whether it is classical or Bayesian, this solution is natural, since risks and Bayes estimators involve integrals with respect to probability distributions. We will see in Chapter 5 why this solution also applies in the case of maximum likelihood estimation. Note that the possibility of producing an almost infinite number of random variables distributed according to a given distribution gives us access to the use of *frequentist* and *asymptotic* results much more easily than in usual inferential settings (see Serfling 1980 or Lehmann and Casella 1998, Chapter 6) where the sample size is most often fixed. One can, therefore, apply probabilistic results such as the Law of Large Numbers or the Central Limit Theorem, since they allow for an assessment of the convergence of simulation methods (which is equivalent to the deterministic bounds used by numerical approaches.)

3.2 Classical Monte Carlo Integration

Before applying our simulation techniques to more practical problems, we first need to develop their properties in some detail. This is more easily accomplished by looking at the generic problem of evaluating the integral

$$\mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx\,. \tag{3.4}$$

Based on previous developments, it is natural to propose using a sample $(X_1, \ldots, X_m)$ generated from the density f to approximate (3.4) by the empirical average[1]

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)\,,$$

since $\overline{h}_m$ converges almost surely to $\mathbb{E}_f[h(X)]$ by the Strong Law of Large Numbers. Moreover, when h^2 has a finite expectation under f, the speed of convergence of $\overline{h}_m$ can be assessed since the variance

$$\text{var}(\overline{h}_m) = \frac{1}{m} \int_{\mathcal{X}} (h(x) - \mathbb{E}_f[h(X)])^2\, f(x) dx$$

[1] This approach is often referred to as the *Monte Carlo method*, following Metropolis and Ulam (1949). We will meet Nicolas Metropolis (1915–1999) again in Chapters 5 and 7, with the simulated annealing and MCMC methods.

Solving the saddlepoint equation $\partial \log \phi x(t)/\partial$

$$\hat{t}(x) = \frac{-p + 2x - \sqrt{}}{4x} \tag{3.33}$$

and applying (3.30) yields the approximate de
3.38).

The saddlepoint can also be used to appro
From (3.30), we have the approximation

$$P(\bar{X} > a) = \int_a^\infty \left(\frac{n}{2\pi\, K_X''(\hat{\tau}(x))} \right)^{1/2}$$
$$= \int_{\hat{\tau}(a)}^{1/2} \left(\frac{n}{2\pi} \right)^{1/2} [K_X''(t)]^{1/}$$

where we make the transformation $K_X'(t) = x$
transformation was noted by Daniels (1983,
integral with only one saddlepoint evaluation

Interval	Approximation
$(36.225, \infty)$	0.1012
$(40.542, \infty)$	0.0505
$(49.333, \infty)$	0.0101

Table 3.7. Saddlepoint approximation of a
for $p = 6$ and $\lambda = 9$.

To examine the accuracy of the saddlepo
chi squared distribution of Example 3.18. Ta
by integrating the exact density and using th
As can be seen, the accuracy is quite impre

The discussion above shows only that th
not the $\mathcal{O}(n^{-3/2})$ that is often claimed. Th
malizing (3.30) so that it integrates to 1.

Saddlepoint approximations for tail ar
than given here. For example, the work of
very accurate approximation that only requ
no integration. There are other approaches
the work of Barndorff-Nielsen (1991) using
proximation of DiCiccio and Martin (1993)
of the Lugannani and Rice formula.