2.4 Simulation of Random Samples from Parametric Distributions

This project requires an understanding of the Part IB Statistics course.

1 Introduction

Let X be a random variable and let $F(x)$ be the distribution function of X, so that $\mathbb{P}(X \leqslant x) = F(x)$. We will assume that $F(x)$ is a continuous strictly increasing function, and define $U = F(X)$, which is therefore a random variable with values in $[0, 1]$. It is easy to find the distribution function of U, for

$$\begin{aligned} \mathbb{P}(U \leqslant u) &= \mathbb{P}\big(F(X) \leqslant u\big) \\ &= \mathbb{P}\big(X \leqslant F^{-1}(u)\big) \\ &= F\big(F^{-1}(u)\big) = u, \quad \text{for } 0 \leqslant u \leqslant 1. \end{aligned}$$

Hence U has the *uniform* or *rectangular* distribution on $[0, 1]$. We write this as

$$U \sim \text{Unif}[0, 1].$$

Clearly, given $U \sim \text{Unif}[0, 1]$, if we define X by $X = F^{-1}(U)$, then X will have distribution function $F(x)$. We can use this fact to generate $X_1, \ldots, X_n$, (pseudo-)random variables from a given distribution function $F(x)$.

Note: The MATLAB function `rand` *generates* Unif$[0, 1]$ *(pseudo-)random variables which may be assumed independent. To generate any other kind of random variable you will need to write your own routines. The use of existing routines, for example from the statistics toolbox or R will not earn you any credit. However, you may wish to compare your random number generators to pre-existing ones.*

2 The Exponential Distribution

Take $F(x) = 1 - e^{-\theta x}$, $x \geqslant 0$, corresponding to an exponential density with rate θ, mean θ^{-1}, and probability density function

$$f(x \mid \theta) = \theta e^{-\theta x}, \quad x \geqslant 0.$$

Question 1 Suppose that instead of indexing the probability distribution function by its rate θ, we decide to index it by its median m given by

$$\int_0^m f(x \mid \theta)\, dx = \frac{1}{2}.$$

Find θ as a function of m and hence find $g(x \mid m) = f\big(x \mid \theta(m)\big)$.

Question 2 Take $(u_1, \ldots, u_n)$, sampled from Unif$[0, 1]$, and hence compute the x_i defined by $u_i = 1 - e^{-\theta_0 x_i}$, giving $(x_1, \ldots, x_n)$, sampled from $f(x \mid \theta_0)$. Try this for $n = 6$, $\theta_0 = 1.2$. Plot the resulting *log likelihood* function $\ell_n(m)$ against m where

$$\ell_n(m) = \log \prod_{i=1}^{n} g(x_i \mid m).$$

Derive analytically $\widehat{m}_n$, the value of m which maximises $\ell_n(m)$, and compare this with m_0, the true value of the median.

Question 3 Repeat all of question 2 above for $n = 25, 50, 100$, and comment on the qualitative changes you observe (if any) in the shape of $\ell_n(m)$.

Question 4 Suppose that X, Y are independent random variables, each with a probability distribution function corresponding to an exponential with mean $1/\theta$. Calculate the moment generating function $M_X(\lambda) = E(e^{\lambda X})$ of X. Show that $X + Y \sim \Gamma(2, \theta)$.

Note that if X is distributed as a $\Gamma(n, \theta)$ random variable, then it has density function $f(x) = \theta^n x^{n-1} e^{-\theta x}/(n-1)!$ and moment generating function $E(e^{tX}) = (1 - t/\theta)^{-n}$, $(t < \theta)$.

Question 5 Take $f(x \mid \theta) = \theta^2 x e^{-\theta x}$, $x \geqslant 0$, and integrate it to find $F(x)$. Can you compute F^{-1} in closed form?

Question 6 The *log-likelihood* function is now

$$\ell_n(\theta) = \log \prod_{i=1}^{n} f(x_i \mid \theta)$$

Calculate the maximum likelihood estimator for θ.

Question 7 Take $\theta_0 = 2.2$, generate a random sample of $x_1, \ldots, x_n$ from $f(x \mid \theta_0)$ and plot $\ell_n(\theta)$ against θ, for $n = 10, 30, 50$. For each sample, calculate the maximum likelihood estimator for θ, and compare it with θ_0, describing any similarities or differences between this case and that in question 3.

Question 8 We investigate the distribution of $\widehat{\theta}_n$ as follows. Take $\theta_0 = 2.2$ and $N = 200$. Take $\mathbf{x}(1), \ldots, \mathbf{x}(N)$ as N independent random samples each of size $n = 10$ from $f(x \mid \theta_0)$. Let $\widehat{\theta}_n(1), \ldots, \widehat{\theta}_n(N)$ be the corresponding maximum likelihood estimators of θ. Generate the histogram* of $\widehat{\theta}_n(1), \ldots, \widehat{\theta}_n(N)$. How does this histogram change if we increase n from 10 to 50?

3 The Normal Distribution

Since the normal or Gaussian distribution plays such an important role in probability and statistics, it is clearly of interest to know how to generate, say, $X_1, \ldots, X_n$, a random sample from $N(\mu, \sigma^2)$, the normal distribution with mean μ and variance σ^2. An *unsubtle* method would be to use $X = F^{-1}(U)$, where $U \sim \text{Unif}[0, 1]$, and $F(x) = \int_{-\infty}^{x} e^{-\frac{1}{2}v^2} dv/\sqrt{2\pi}$. This method is *not* recommended, because $F(x)$ here is not of closed form. So what do we do instead?

Recall the following from Part IA Probability: if (Φ, V) have joint density $f(\phi, v)$, and we define

$$X = X(\Phi, V)$$
$$Y = Y(\Phi, V)$$

*For a definition of the histogram see, for example, *Statistical Theory* (1976) by B.W. Lindgren, pp. 206–7. You can use `hist` in MATLAB to draw one.

so that (X, Y) is a 1–1 function of (Φ, V), then (X, Y) has joint density $g(x, y)$ where

$$g(x, y) = f\big(\phi(x, y), v(x, y)\big) \left|\frac{\partial(\phi, v)}{\partial(x, y)}\right| .$$

Question 9 Show that if $f(\phi, v) = \frac{1}{4\pi}\, e^{-v/2}$, $0 \leqslant \phi \leqslant 2\pi, v \geqslant 0$, and if we define

$$X = \mu_1 + \sigma\sqrt{V}\cos\Phi,$$
$$Y = \mu_2 + \sigma\sqrt{V}\sin\Phi,$$

then X, Y are independent $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ random variables, i.e.,

$$g(x, y) = \frac{1}{2\pi\sigma^2}\, e^{-\{(x-\mu_1)^2+(y-\mu_2)^2\}/2\sigma^2}, \quad -\infty < x, y < \infty.$$

We apply this by

- generating A, B, independent Unif$[0, 1]$ variables;
- defining $\Phi = 2\pi A$ and $V = -2\log(1 - B)$;
- defining $X = \mu + \sigma\sqrt{V}\cos\Phi$ and $Y = \mu + \sigma\sqrt{V}\sin\Phi$.

Question 10 Write a program to generate a random sample $(x_1, \ldots, x_n)$ of size n from distribution $N(\mu, 1)$. Explain how to construct an 80% confidence interval for μ.

Question 11 For $\mu = 0$, generate a sample of size $n = 100$ from distribution $N(\mu, 1)$ and check whether the confidence interval does indeed contain μ. Repeat this procedure 25 times and display the results in a table with four columns, containing the sample mean, the lower and upper bound of the confidence interval, and 1 or 0 to indicate whether or not the interval contained the true mean. How many times did the interval *not* contain μ?

Question 12 If questions 10 and 11 were to be repeated with $n = 50$ and $\mu = 4$, how many times would you expect the confidence interval *not* to contain μ?

4 The χ^2 Distribution

Question 13 Write a program to generate a random sample of size n from each of the following distributions:

(a) chi-square with 1 degree of freedom (χ^2_1);
(b) chi-square with 5 degrees of freedom (χ^2_5);
(c) chi-square with 40 degrees of freedom (χ^2_{40}).

Run your program for $n = 100, 300, 500$ and include a histogram in each case. How do these histograms change in shape as you change the degrees of freedom?

