Project 1
Computational Numerical Statistics
DM, FCT-UNL
Group: G2

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

Deadline: 28/10/19

Random variable generation

In the resolution of the following problems **do not** use any of the `R` random variable generation built-in functions, nor any `R` function referring to the densities of random variables.

Fix your `R` random seed to 123 in all simulations.

1. The **Box-Muller** method is a method for generating samples from two independent random variables $X,\ Y \sim N(0,1)$ from the $U(0,1)$ distribution. The algorithm proceeds as follows:

   1. `generate two observations` u_1 `and` u_2 `from` $U(0,1)$
   2. `set` $\theta = 2\pi u_1$ `and` $R = \sqrt{-2\log(u_2)}$
   3. `set` $x = R\cos(\theta)$ `and` $y = R\sin(\theta)$
   4. `repeat the previous steps until you reach the desired sample size`

   Often, when using this algorithm one is just interested in one of the variables X or Y. Also, if one is interested in generating a sample from $Z \sim N(\mu, \sigma^2)$, then the algorithm only requires the additional step $Z = \mu + \sigma X$.

   (a) Implement the Box-Muller method in `R`. Call your simulation routine `sim.norm()` and let it receive as input generic sample size m and parameters μ and σ from the Normal distribution. Provide both algorithm and **R** code.

   (b) Use routine `sim.norm()` to generate a sample of size $m = 10000$ from $X \sim N(0,4)$. Provide both algorithm and **R** code.

   (c) Plot the histogram referring to (b) with true p.d.f. superimposed.

2. Let X be a continuous random variable with probability density function (p.d.f.) proportional to $f(x) = xe^{-x}$, $x > 0$.

   (a) Describe and implement the **acceptance-rejection** method in `R` for generating a sample from f. Call your simulation routine `sim()` and let it receive as input generic sample size m. Provide both algorithm and **R** code.

   (b) Use routine `sim()` to generate a sample of size $m = 10000$ of X.. Compute the rejection rate.

   Suggestion: (i) search for the candidate function g in the χ^2_ν family; and (ii) use the following result

Let $Z_1, \ldots, Z_\nu$ be ν independent random variables such that $Z_i \sim N(0,1)$. Then

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu.$$

(c) Plot the histogram referring to (c) with true p.d.f. superimposed.

(d) Display the the hit-and-miss plot referring to `sim(10)`.

Monte Carlo integration

In the resolution of the problems in this section you may already use the R random variable generation built-in functions as well as any R function referring to the densities of random variables.

Fix your R random seed to 456 in all simulations.

3. Let $\mathcal{I} = \displaystyle\int_0^1 \frac{\sqrt{-\log(x)}}{2} \mathrm{d}x$.

   (a) Use the R function `integrate()` to compute the value of $\mathcal{I}$.

   (b) Describe and implement in R the Monte Carlo method of size $m = 10000$ for estimating $\mathcal{I}$. Report an estimate of the variance of the Monte Carlo estimator $\hat{\mathcal{I}}_{MC}$ of $\mathcal{I}$.

   (c) Describe and implement in R the Monte Carlo method of size $m = 10000$ based on **control variables** for estimating $\mathcal{I}$. Report an estimate of the variance of the Monte Carlo estimator $\hat{\mathcal{I}}_C$ of $\mathcal{I}$.

   (d) What's the percentage of variance reduction that is achieved when using $\hat{\mathcal{I}}_C$ instead of $\hat{\mathcal{I}}_{MC}$?

Hypothesis testing

In the resolution of the problems in this section you may already use the R random variable generation built-in functions as well as any R function referring to the densities of random variables.

Fix your R random seed to 789 in all simulations.

4. Let $X_1, \ldots, X_n$ be a random sample from the population $X \sim N(1,3)$.

   (a) Validate via a Monte Carlo simulation study that $\chi = \dfrac{(n-1)S^2}{2} \sim \chi^2_{n-1}$. Consider in your study the number of simulations $m = 1000$ and a sample size of $n = 25$. For validation purposes provide:

      - the histogram of the simulated $\chi_1, \ldots, \chi_{1000}$ with the theoretical density superimposed;
      - the plot of the empirical cumulative distribution (ecdf) of $\chi_1, \ldots, \chi_{1000}$ with the theoretical cumulative probability function (pdf) superimposed (use the R built-in `ecdf()` function to create the sample ecdf);

- perform a Kolmogorov-Smirnov two-sided test to test your hypothesis (use the `R` built-in `ks.test()` function);
- Compute the theoretical quantiles $0.90,\ 0.95, 0.975$ with the empirical quantiles (use the `R` built-in `qchisq()` and `quantile()` functions, respectively).

(b) Assume one wants to test the hypothesis

$$H_0 : \sigma^2 \leq 3 \qquad\qquad H_1 : \sigma^2 > 3$$

at the significance level $\alpha = 0.05$. Compute the empirical p-values for this test using the previous Monte Carlo simulation. Compute the empirical significance level $\hat{\alpha}$ and perform the binomial test to assess if $\hat{\alpha}$ departs significantly from α.

(c) For the hypothesis test above and the m simulations, construct a power plot for the alternative values $\sigma^2 = 4, \ldots, 50$. In face of these results, how far from H_0 does one need to be so that the power of the test gets higher than 90%?