Project 1
Computational Numerical Statistics
DM, FCT-UNL
Group: G1

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. Let X be a discrete random variable with probability mass function (p.m.f.) f proportional to $g(x) = 1 - x + x^2$, $x = 0, 1, 2, 3, 4$.

   (a) Identify the p.m.f. f

   (b) Derive the cumulative distribution function (c.d.f.) F of X.

   (c) Describe and implement the **inverse-transform** method in `R` for generating a sample from f. Call your routine `sim.itm()` and let it receive as input a generic sample size m. Provide both algorithm and **R** code. Finally, use `sim.itm()` to generate a sample of size $m = 10000$ of f.

   (d) The same as in (c) but now using the **acceptance-rejection** method. Call you new simulation routine `sim.arm()`. Compute the rejection rate.

   (e) Plot the discrete histograms from (c) and (d) with the true p.m.f. superimposed.

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

2. Some random variables can be generated from the exponential distribution (**exponential-based method**), which we know how to obtain from the $U(0, 1)$ distribution. Such is the case of the random variable $X \sim Beta(a, b)$:

$$\text{If } Y_i \underset{iid}{\sim} Exp(1) \text{ then } X = \frac{\sum_{i=1}^{a} Y_i}{\sum_{i=1}^{a+b} Y_i} \sim Beta(a, b), \qquad a, b = 1, 2, \ldots$$

   (a) Implement the **exponential-based method** in `R` for generating a sample from $X \sim Beta(a, b)$, which has probability density function (p.d.f.)

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \qquad \alpha,\ \beta > 0, \qquad x \in [0, 1]$$

   where $B(a, b)$ is the *beta* function. Call your routine `sim.beta()` and let it receive as input a generic sample size m and the Beta distribution parameters a and b (note that here $a, b \in \mathbb{N}$). Provide both algorithm and **R** code.

(b) Use routine `sim.beta()` to generate a sample of size $m = 10000$ from $X \sim Beta(3, 2)$.

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

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 **antithetic variables** for estimating $\mathcal{I}$. Report an estimate of the variance of the Monte Carlo estimator $\hat{\mathcal{I}}_A$ of $\mathcal{I}$.

   (d) What's the percentage of variance reduction that is achieved when using $\hat{\mathcal{I}}_A$ 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 $X \sim N(0, 2)$.

   (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 $n = 20$. 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

$$H_0 : \sigma^2 = 2 \qquad\qquad H_1 : \sigma^2 \neq 2$$

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 $\sigma^2 = 3, 4, \ldots, 50$. 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 80%?