Project 1
Computational Numerical Statistics
DM, FCT-UNL
Group: G5

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. 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 $X \sim \chi^2_{2\nu}$:

$$\text{If } Y_i \underset{iid}{\sim} Exp(1) \text{ then } X = 2\sum_{i=1}^{\nu} Y_i \sim \chi^2_{2\nu}, \qquad \nu = 1, 2, \ldots$$

   (a) Implement the **exponential-based** method in `R` for generating a sample from $\chi^2_{2\nu}$ (note that here $\nu \in \mathbb{N}$). Call your simulation routine `sim.chi()` and let it receive as input a generic sample size m and the χ^2 distribution ν parameter. Provide both algorithm and **R** code.

   (b) Use routine `sim.chi()` to generate a sample of size $m = 10000$ from $X \sim \chi^2_4$.

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

2. Let $X \sim Gamma(\alpha, \theta)$, which has probability density function

$$f(x) = \frac{\theta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x}, \qquad x \geq 0.$$

   (a) Describe the **acceptance-rejection** method for the continuous case. Proof that indeed the method generates a sample from the target density f.

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

   (c) Use routine `sim.gam()` to generate a sample of size $m = 10000$ of $X \sim Gamma(3/2, 1)$. Compute the rejection rate.

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

   (e) Display the the hit-and-miss plot referring to `sim.gam(10,3/2,1)`.

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 **importance sampling** for estimating $\mathcal{I}$. Take the importance function ϕ from the $Beta(a, b)$ distribution family and explain your criteria for choosing the a, b parameters.

   Report an estimate of the variance of the Monte Carlo estimator $\hat{\mathcal{I}}_{IS}$ of $\mathcal{I}$.

   (d) What's the percentage of variance reduction that is achieved when using $\hat{\mathcal{I}}_{IS}$ 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 $X \sim N(3, 1)$.

   (a) Validate via a Monte Carlo simulation study that $T = \dfrac{\overline{X} - 3}{S/\sqrt{n}} \sim t_{n-1}$. 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 $T_1, \ldots, T_{1000}$ with the theoretical density superimposed;
   - the plot of the empirical cumulative distribution (ecdf) of $T_1, \ldots, T_{1000}$ 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

$$H_0 : \mu \leq 3 \qquad\qquad H_1 : \mu > 3$$

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 $\mu = 4, 5, \ldots, 49, 50$. 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 85%?