Project 1
Computational Numerical Statistics
DM, FCT-UNL
Group: G4

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.

   Consider the following results:

   Let $Z_1, \ldots, Z_\nu$ be ν independent random variables such that $Z_i \sim N(0,1)$. Then
   $$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2_\nu.$$

   Let $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ such that Z and V are independent random variables. Then,
   $$\frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t_\nu.$$

   (b) Implement a routine `sim.t()` in `R` for generating a sample of the t-student distribution t_ν. Let `sim.t()` receive as input the sample size m and the number of degrees of freedom ν of the t-student distribution. Use the results above and routine `sim.norm()` developed in (a). Provide both algorithm and **R** code.

(c) Use routine `sim.t()` to generate a sample of size $m = 10000$ of $X \sim t_3$.

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

2. Let $X \sim Gumbel(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma > 0$, which has cumulative distribution function (c.d.f.)

$$F(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}, \qquad x \in \mathbb{R}.$$

(a) Describe and implement the **inverse-transform** method in R for generating a sample from F. Call your routine `sim.gum()` and let it receive as input a generic sample size m and the Gumbel μ and σ parameters. Provide both algorithm and **R** code.

(b) Use routine `sim.gum()` to generate a sample of size $m = 10000$ of $X \sim Gumbel(0, 1)$.

(c) Plot the histogram referring to (c) 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 $\mathcal{I} = \displaystyle\int_0^1 \frac{e^{-x}}{1+x^2} \mathrm{d}x$.

(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 $\mathcal{I}$.

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 $X \sim N(-2, 1)$.

(a) Validate via a Monte Carlo simulation study that $T = \dfrac{\overline{X}+2}{S/\sqrt{n}} \sim t_{n-1}$. Consider in your study the number of simulations $m = 1000$ and a sample size of $n = 15$. 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

$$H_0 : \mu \geq -2 \qquad\qquad H_1 : \mu < -2$$

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 $\mu = -50, -49 \ldots, -4, -3$. 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 75%?