Statistical Data Analysis
Problem sheet 4
Due Monday 30 October 2023

Please turn in a copy of the parts of the source code you wrote or modified and relevant output. Please do not turn in any large parts of source code that you use but did not modify.

Exercise 1(a) [2 marks] Let $x_N = \sum_{i=1}^{N} r_i$, where the r_i are independent and uniformly distributed between 0 and 1. Find the mean μ_N and standard deviation σ_N of x_N as a function of N.

1(b) [5 marks] Using the results from (a), construct the standardized variable

$$y_N = \frac{x_N - \mu_N}{\sigma_N} = \sqrt{\frac{12}{N}} \left(\sum_{i=1}^{N} r_i - \frac{N}{2} \right) .$$

Using the `simpleMC` program (either C++ or Python) from problem sheet 3 as a starting point, write a computer program to make histograms of 10000 values of y_N as defined above for $N = 1, 2, 4, 12$. Make sure to set the limits of the histogram such that the entire distribution is plotted. In the program, find the mean and standard deviation of y_N and verify that these are close to the values you expect. In python, you can put the values in a numpy array and use the functions `mean` and `std`. If you are using C++ with ROOT, you can use the `GetMean` and `GetStdDev` functions of the TH1D class (see `https://root.cern.ch/doc/master/classTH1.html`). Comment on the connection between your histograms and the central limit theorem.

Exercise 2 Consider the pdf $f(x) = 4x^3$, $0 \le x \le 1$.

2(a) [4 marks] Use the transformation method to find the function $x(r)$ to generate random numbers according $f(x)$. Implement the method in a short computer program and make a histogram with 10000 values.

2(b) [4 marks] Write a program to generate random numbers according to $f(x)$ using the acceptance-rejection technique. Plot a histogram of the results.

Exercise 3 [5 marks] Suppose $\vec{x} = (x_1, \ldots, x_n)$ follows an n-dimensional Gaussian distribution $f(\vec{x}; \vec{\mu}, V)$ with $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix $V_{ij} = \mathrm{cov}[x_i, x_j]$. (In the formulas below regard $\vec{x}$ and $\vec{\mu}$ to be column vectors.) Suppose we have two hypotheses for the vector of means, $\vec{\mu}_0$ and $\vec{\mu}_1$, where for both one uses the same covariance matrix V, and consider the test statistic

$$t(\vec{x}) = \ln \frac{f(\vec{x}|\vec{\mu}_1)}{f(\vec{x}|\vec{\mu}_0)} .$$

Show that this $t(\vec{x})$ can be written in the form

$$t(\vec{x}) = w_0 + \sum_{i=1}^{n} w_i x_i ,$$

or equivalently $t(\vec{x}) = w_0 + \vec{w}^T \vec{x}$, where $\vec{w}$ is a column vector of coefficients w_i, $i = 1, \ldots, n$.