Data Analysis II - 2023

Exercise sheet no 1: Monte Carlo techniques — 21. February 2023

- `numpy.mean(data)`, `numpy.std(data)`, `numpy.var(data)`
- `numpy.random.uniform(low, high, size)`
- `numpy.random.power(a, size)`
- `numpy.random.norm(loc,scale,size)`

Exercise 1: Monte-Carlo integration (5 Points)

The equation of the circle with unit radius centered around $(0, 0)$ is given by

$$x^2 + y^2 = 1. \tag{1.1}$$

Exploiting the symmetry of the circle, one can define the area under the positive quadrant to be

$$\int_0^1 y dx = \int_0^1 \sqrt{1-x^2} dx = \pi/4. \tag{1.2}$$

Using the Monte Carlo technique evaluate the integral to find the value of π.

(a) Generate $N_{\text{smpl}} = 10000$ random values for x, distributed uniformly between 0 and 1. Use these values to calculate the average function value $\langle y \rangle$ and estimate π. (1.5 points)

(b) Repeat the exercise (a) $N_{\text{expt}} = 1000$ times to evaluate the uncertainty due to the limited sample size used in the estimation of π. (1.5 points)

(c) For different integration samples sizes (i.e. $N_{\text{smpl}} = 10^i;\ i = 1, ..., 6$) which are repeated $N_{\text{expt}} = 100$ times, evaluate the mean of the absolute difference between estimated π_{est} and exact result π (`np.pi`) i.e. $\Delta_\pi = |\pi_{est} - \pi|$. Make a log-log plot of Δ_π Vs N_{smpl} and show that the data can be fit to a straight line of slope -1/2. (2.5 points).

Exercise 2: Importance sampling (4 Points)

As done in Exercise 1(c), for different integration samples sizes (i.e. N_{smpl}[1]), repeated $N_{expt} = 100$ times, evaluate the mean of the absolute difference between the numerical estimate of the integral of the function $f(x) = x^3$ between $x = 0$ and $x = 1$ (I_{est}) and the theoretical estimate (I) i.e. $\Delta_I = |I_{est} - I|$. Plot Δ_I Vs N_{smpl} for two cases:

(a) One where the I_{est} is estimated by generating random values of x distributed uniformly between 0 and 1. (1.5 points)

(b) One where the I_{est} is estimated by generating random values of x distributed according to a power law (i.e. $q(x) = (k + 1)x^k$, with $k = 2.5$ and $x \in \{0, 1\}$). (2.5 points)

Which one of two cases converges to the true value faster?

---

[1] $N_{smpl} = \{2, 5, 10, 50, 100, 200, 400, 500, 700, 1000, 2000, 3000, 4000, 5000, 10000\}$

Exercise 3: Random sample generator (4 Points)

Write a Python script that generates random numbers distributed according to a Breit-Wigner distribution with the statistical median $m = 3$ and the full-width at half maximum $b = 7$. Use for the PDF and CDF the following expressions:

$$p(x) = \frac{1}{\pi} \frac{b}{(x-m)^2 + b^2} \,, \tag{3.1}$$

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-m}{b}\right) \,. \tag{3.2}$$

Generate $N_{\text{smpl}} = 10000$ accepted values of x, distributed according to $p(x)$ using the following methods discussed in the lecture:

(a) Accept-reject method. (2 points)

(b) Inverting the CDF. (2 points)

For each of the two methods, plot the histogram of the generated sample and superimpose the PDF after normalising it to match the histogram.

Exercise 4: Uncertainty estimation comparison (7 Points)

The file `gauss_data.txt` contains 100 random numbers generated according to a Gaussian distribution with $\mu = 10$ and $\sigma = 3$. Your task is to estimate the uncertainty of these parameters using the data in three different ways.

(a) First, calculate the mean and standard deviation (S_x) of the data. (1 point)

(b) Calculate the uncertainty of these estimates using the standard formulae[2]. (1 points)

(c) Use Monte Carlo to generate repetitions of the dataset $N_{\text{expt}} = 1000$ times (i.e. 1000×100 data points) according to the underlying Gaussian distribution. Calculate the mean and standard deviation for each generated dataset and use the RMS of those values to estimate the uncertainties of the mean and standard deviation of the original dataset in the text file. (2 points)

(d) Bootstrap the dataset 1000 times by resampling the dataset with replacement. Calculate the mean and standard deviation for each bootstrapped dataset and use the variance of those values to estimate the uncertainties of the mean and standard deviation of the original dataset in the text file. (2 points)

(e) Which uncertainty technique do you think can be trusted in most situations? (1 point)

Deadline for submission: Tuesday, 7 March 2023 18:00
Form: Please submit your solutions to OLAT. The solutions should be submitted as a single python script which creates the answers to the questions.

[2] $\sigma(\bar{x}) = S_x/\sqrt{N}$, $\sigma(S_x) = \frac{S_x}{\sqrt{2N-2}}$