Modern Statistical Data Analysis For Practitioners

Dr Barłomiej Wacław, IChF PAN

Outline:

The advent of fast and accessible computing has revolutionized statistical data analysis, making it possible to go (with little effort) beyond descriptive statistics, the square root law of error propagation, and non-linear curve fitting taught at university undergraduate courses not so long ago. This course will familiarize the students with a few modern data analysis tools that are quite powerful yet relatively easy to use.

Range of material covered:

- A brief statistics recap (probability distributions, descriptive statistics, standard error of mean versus standard deviation, statistical hypotheses),
- Going beyond uncorrelated random variables – how to estimate errors if your data points are correlated
- Stochastic processes frequently encountered in statistical data analysis: random walk, Poisson process, birth-and-death process. Their properties and how to recognize patterns their presence in your data.
- How to build a statistical model of your experiment (physics/chemistry/biology model + modelling errors)
- Model fitting – what to do if we cannot assume that errors are normally distributed?
- Monte-Carlo methods: error propagation and statistical hypotheses testing
- Bayesian inference: how it is related to the maximum likelihood method and least squares, Bayesian model fitting and model comparison.

The course will be focused on solving practical problems, with examples from physics, chemistry, and biology. Theoretical background will be reduced to a minimum.

This course is for experimentalists and computer modellers alike as similar problems to the ones discussed here arise in data analysis of real-world and computer-model generated data.

Assessment:

There will be three assignments that must be completed before a set date (last lecture + two weeks). The assignments will be marked and the total mark calculated as the average of the three marks.

Prerequisites:

You must be numerically and computationally literate, i.e., able to use computers to solve basic numerical problems. To complete most of the assignments you must be able to code – Python would be ideal but C/C++/Java/Matlab/Wolfram Mathematica will do. If you do not program at all, you may be able to get a pass mark as some assignments won't require coding, but you won't enjoy this course. Basic knowledge of calculus and linear algebra will help to understand various concepts presented during the course, though this is not required to complete the assignments.

How to tell if your computer skills are sufficient for this course? Try to solve the following problem: generate 100,000 pairs of random numbers (x,y), with each x,y uniformly distributed on the interval [-1,1). What is the fraction of (x,y) obeying the condition $x^2 + y^2 < 1$? You should get a number close to 0.785 +/- 0.005 (bonus question: why this number?).