Advanced Methods II: Statistics, Data Analysis, Numerical Methods (PHYS 632)

| | |
|---|---|
| Department | Physics & Astronomy |
| Credit hours | 3 |
| Term | |
| Venue | |
| Instructor | |
| Office Hours | |
| Recommended Literature | See information under “Literature”. |
| Final Exam | |

## Summary and Target Audience

The course introduces basic techniques in statistics, data analysis, and numerical methods at the graduate level. It intends to provide students with the concepts behind common techniques encountered in their research. The course material does not attempt to treat any of the topics exhaustively, rather, the goal is to (a) enable students to follow up on, modify, or develop certain techniques of interest, and use them in an informed way in academia or in the private sector, and (b) make them aware of possible limitations such that they can interpret research results in an informed way.

PHYS 632 is targeted at graduate students in their first or second year, and at advanced (senior) undergraduate students.

The course satisfies the experimental requirement for theory graduate students.

## Technical Aspects

Students’ mastery of the subject will be assessed via homeworks, an in-class semester research project presentation, and a final exam. Homework assignments usually include coding. Homework solutions and code examples will be provided in Python or R, depending on the topic. Tutorials (see below) will be provided for students to get familiar with Linux, Python, or R, in the first week of class.

The course will include TA-led lab sessions to help students getting familiar with the programming languages.

| Assessment tool | Percentage of total grade |
|---|---|
| homeworks | 60% |
| Project & presentation | 20% |
| Final | 20% |

## Prerequisites

The course does not have formal prerequisites, but you should be familiar with the basic programming concepts in Python, listed below. Familiarity with R will also help.

- variable definitions and assignments
- data types
- if-statements, for-loops
- plotting
- functions
- array/vector operations
- reading and writing data

There are many tutorials out on the web. Appendix A of (3.) offers an introduction to Python. More resources can be found here, here, and here. For R, this is a decent starting point, covering the basics. Sample programs will be provided in class. An introduction to the departmental Linux environment can be found here.

## Literature

The texts below are examples on what literature the course is based. None of the texts below cover all the material presented in class. The instructor will make specific recommendations, and additional material will be provided in class.

Statistics & Data Analysis:

1. Statistical Data Analysis; Cowan, G., Clarendon Press, Oxford (1998)
2. Bayesian Data Analysis; Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin, CRC Press (2013). Online materials.
3. Statistics, Data Mining, and Machine Learning in Astronomy; Ivezić, Connolly, VanderPlas, Gray, Princeton U Press (2014). Online materials.
4. Information Theory, Inference, and Learning Algorithms; MacKay, D, Cambridge U Press (2004). Online materials.

Numerical Methods:

5. Applied Partial Differential Equations; Haberman, R. Pearson (2013).

6. Numerical Recipes in C; Teucholsky et al. (2nd edition, Princeton, 1992). Online materials.

## Content Details

Not all topics will be covered in class, but may be offered as research projects. In the following, content is listed according to field (statistics, data analysis, numerical methods).

**1. STAT: Probability Basics.** Probability distribution functions, cumulative pdfs, transformation of probabilities. Laws of probability, Bayes theorem.

**2. NUM: Computational and Numerical Basics:** Introduction to Python and R. Number representation on computers, types of errors.

**3. STAT: Gaussian and non-Gaussian statistics.** Binomial, Poisson and Gaussian statistics. Law of Large Numbers, Central Limit Theorem.

**4. STAT: Error analysis.** Monte Carlo simulations. One-sided and two-sided significance. Signal detection, replicability ("p-value crisis"). Signal-to-noise ratio and determination of dominant noises.

**5. DAT: Correlations and Statistical Tests.** Hypothesis testing and significance. Correlation tests. Realms of validity, Anscombe's quartet. Binning (standard, adaptive, detection of systematics). Effects on signal-to-noise ratio. 1D and 2D binning, Kernel density estimation. Distribution tests and applicability.

**6. DAT: Model Fitting and Parameter Estimation Techniques.** Maximum likelihood/least squares, median absolute deviations, outlier rejection vs. robust fitting methods, simplex methods. Sensitivity to x,y errors and selection. Bootstrapping, Polynomial fitting.

**7. NUM: Singular Value Decomposition.** Eigenvalues, -vectors, -spaces. Null spaces and rank ranges. Application to least squares fitting.

**8. DAT: Model Fitting Tests and Model Comparison**. Goodness of fit, chi-squared (statistics vs distribution, repeatability/significance, degrees of freedom). Solution stability testing. Overfitting and underfitting. Cross validation. Parameter covariance.

**9. STAT: Bayesian Inference: Parameter Estimation and Model Comparison.** Likelihoods, priors, posteriors, normalization. Marginalization, joint distributions. Choice of priors, stability testing. Confidence intervals, unified approach, unphysical regions. Model comparison and confidence. Model fitting including noise and selection biases.

**10. NUM: Numerical Approximation of Integrals.** Gaussian quadrature, Simpson, Monte Carlo. Multi-dimensional integrals. Markov Chains, MCMC. Initial conditions, burn-in, convergence. Marginalization, covariance matrices. Application to Bayesian modeling.

**11. DAT: Noise, Noise Mitigation, Experiment Design.** Identifying noise sources and levels of noise. Combining noise estimates to assess experiment sensitivity. Noise scaling as experimental parameters change.

**12. DAT: Statistical Evidence and Experiment Design.** Designing experiments to achieve required levels of statistical evidence (numbers of measurements, etc.). Look elsewhere effect. False positive rate estimation and effects on required detection significance. Upper limits, missing data, selection limits, survival analysis.

**13. NUM: Fourier Techniques.** Fourier series, discrete Fourier transform, fast Fourier transform. Nyquist frequency, multi-dimensional spectra. Filters. Window functions.

**14. DAT: Signal Processing**. Matched filters, Lomb-Scargle, NUFFT, wavelet decomposition. Time series analysis, image analysis.

**15. DAT: Detecting Data Features.** Principal component analysis, clustering, classification and machine learning.

**16. NUM: Ordinary Differential Equations**. Runge-Kutta, error control, adaptive stepsize, implicit, Hamiltonian (symplectic) integrators.

**17. NUM: Partial Differential Equations.** Finite differences, elliptic, hyperbolic, parabolic PDEs, strategies for mixed systems, finite volume, finite element methods.

## Sample Schedule

Given ~30 sessions, 2 sessions at the end of the semester are reserved for project presentations. The remaining sessions will be divided roughly equally between the 17 topics listed above, with emphases according to the preferences of the instructor and students. Some topics won't require even a whole session, while others may take a week or more.