# Data Analysis Techniques for Physics and Astronomy

## **PHYS 515** *Course Syllabus — Fall 2026 Term*

**Instructor:** Pavel Denisenkov (pavelden@uvic.ca, office TBD)

**Format:** 12 weekly sessions, 3 hours each (36 contact hours)

**Meeting Times:** Mondays and Thursdays, 11:30am – 12:50pm

**Level:** Graduate (Physics & Astronomy)

**Prerequisites:** Undergraduate calculus and linear algebra; undergraduate mechanics/E&M-level comfort with mathematical modeling. Basic Python is assumed (the standard language of the field); anyone who needs to catch up go to https://github.com/UVic-CompPhys/PHYS248.

---

## Course Description

This course provides a rigorous introduction to the theory and practice of data analysis, built specifically around the problems physics and astronomy researchers encounter, such as propagating measurement uncertainty, extracting parameters from noisy data, fitting models to spectra and light curves, distinguishing signal from background, and deciding between competing physical models. The course begins from first principles in probability theory and builds toward the statistical and computational methods used throughout experimental, observational, and computational physics — with heavy emphasis on maximum likelihood and chi-square fitting, and a substantial Bayesian/MCMC component, since both frequentist and Bayesian reasoning are in everyday use across subfields (e.g., particle physics leans frequentist/likelihood-ratio; astronomy and cosmology lean heavily Bayesian). Each session pairs theory with hands-on analysis of real or realistic physics/astronomy datasets.

## Learning Objectives

By the end of the course, students will be able to:

1. Reason rigorously about randomness using the axioms and tools of probability theory, including the counting-statistics (Poisson) framework central to physical measurement.
2. Propagate measurement uncertainty through arbitrary functions and derive properties of estimators (bias, variance, consistency, efficiency).
3. Construct and interpret confidence/credible intervals and hypothesis tests (including significance and discovery thresholds as used in physics, e.g. "5-sigma"), and understand their assumptions and limitations.
4. Perform least-squares and maximum-likelihood/chi-square fitting of physical models to data, with full error and covariance analysis.
5. Apply resampling, Monte Carlo, and MCMC methods for problems without closed-form solutions.
6. Analyze time-series and spectral data (periodicity, power spectra) as encountered in astronomical and experimental signals.
7. Select, justify, and communicate an appropriate analytical approach for a given physical dataset, including model comparison and selection.
8. Complete an independent data analysis project, ideally using real or simulated physics/astronomy data, from question formulation to written report.

## Textbooks & References

No single required text — readings drawn from:

- John R. Taylor, *An Introduction to Error Analysis* (online pdf)

- Luca Lista, *Statistical Methods For Data Analysis* (online pdf)
- Philip R. Bevington & D. Keith Robinson, *Data Reduction and Error Analysis for the Physical Sciences*
- Željko Ivezić, Andrew J. Connolly, Jacob T. VanderPlas & Alexander Gray, *Statistics, Data Mining, and Machine Learning in Astronomy*
- Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*
- Larry Wasserman, *All of Statistics* (general reference, online pdf)
- Andrew Gelman et al., *Bayesian Data Analysis* (selected chapters, online pdf)
- Instructor-provided lecture notes and datasets (e.g., photon-counting data, spectra, light curves, particle-detector data)

## Software

Python, with `numpy`, `scipy`, `matplotlib`, `astropy`, `emcee` or `PyMC` (MCMC), and `scikit-learn` (Week 12). Students with a strong existing R workflow may substitute where reasonable, but in-class examples will be given in Python given its standard role in the field.

## Assessment

| Component | Weight |
|---|---|
| Problem sets (6, biweekly) | 30% |
| Midterm exam (Week 6) | 20% |
| Data analysis labs (in-class, cumulative) | 15% |
| Final project (proposal + written report + presentation) | 35% |

## Course Policies

- Collaboration on problem sets is encouraged; work submitted must be written independently.
- Late final project components are not accepted except for documented emergencies.
- Academic integrity policy follows the university graduate school guidelines.

## Generative AI

In this course I welcome the use of generative AI tools such as ChatGPT to help learn course material and complete course assignments. Please note that you can opt for not using generative AI at all. In the case you opt to use generative AI for assignments, you must provide proper citation of the tools you used and describe how you used it. The written exams are held in class during the term, without computer access, and only the course lecture notes or a textbook are allowed as a reference.

---

## Weekly Schedule

### Part I — Foundations of Probability Theory (Weeks 1–3)

### Week 1: Sample Spaces, Axioms, and Conditional Probability

- Sample spaces, events, set-theoretic operations
- Axioms of probability (Kolmogorov)
- Conditional probability, independence
- Bayes' theorem, with early framing of the frequentist/Bayesian distinction as it plays out in physics (e.g., particle physics vs. cosmology conventions)
- Example: detector false-positive/false-negative rates as a Bayes'-theorem case study
- Lab: Python environment setup (`numpy`/`scipy`); simulate a detector with known false-positive/false-negative rates and use Bayes' theorem to compute the probability a "trigger" is real. **Dataset:** synthetic, calibrated against published false-alarm rates from the LIGO/Virgo *Gravity Spy* glitch-classification project (real-world motivation for the exercise). **Access:** gravityspy.org

### Week 2: Random Variables and Distributions

- Discrete and continuous random variables; PMFs/PDFs, CDFs
- Distributions central to physics: Poisson (counting/shot noise), Binomial, Exponential (decay times), Gaussian, and the Poisson to Gaussian limit at high counts
- Expectation, variance, moment generating functions
- Lab: simulating photon-counting / radioactive-decay data and comparing to theoretical distributions. **Dataset:** the *Fermi* Gamma-ray Burst Monitor (GBM) Burst Catalog background/trigger count rates (via NASA HEASARC), compared against simulated Poisson counts at matched mean rates. **Access:** heasarc.gsfc.nasa.gov/W3Browse/fermi/fermigbrst.html

### Week 3: Error Propagation, Joint Distributions, and Limit Theorems

- Joint, marginal, and conditional distributions; covariance and correlation
- Propagation of errors (linear and general functions of random variables) — the physicist's everyday tool
- Law of Large Numbers; Central Limit Theorem and why Gaussian errors are ubiquitous in physics
- Lab: propagating uncertainty through a multi-step calculation. **Dataset:** the CODATA compilation of independent historical measurements of a fundamental constant (e.g., Newton's constant *G*, or early speed-of-light determinations) — students combine measurements with different quoted uncertainties and propagate to a derived quantity. **Access:** physics.nist.gov/cuu/Constants

## Part II — Statistical Inference (Weeks 4–7)

### Week 4: Descriptive Statistics and Exploratory Data Analysis

- Measures of location, spread, shape; robust statistics for data with outliers/cosmic-ray hits or glitches
- Visualization principles for scientific data (error bars, log scales, residual plots)
- Data cleaning: instrumental artifacts, missing data, outlier rejection
- Lab: EDA on a real dataset. **Dataset:** a photometric/spectroscopic subsample from the Sloan Digital Sky Survey (SDSS DR17, retrieved via `astroquery`) — magnitude distributions, color-color diagrams, and identification of artifacts/outliers. **Access:** skyserver.sdss.org/dr17

### Week 5: Estimation Theory and Maximum Likelihood

- Point estimation: method of moments, maximum likelihood estimation (MLE)
- Properties of estimators: bias, variance, consistency, efficiency; the Cramér–Rao bound
- MLE for physics contexts: fitting decay constants, spectral line parameters
- Lab: deriving and numerically computing MLEs for a physical model. **Dataset:** cosmic-ray muon decay-time measurements from a CosmicWatch/QuarkNet desktop muon detector (publicly shared classroom datasets) — MLE fit of the exponential decay constant (muon lifetime). **Access:** cosmicwatch.lns.mit.edu

### Week 6: Confidence Intervals and Hypothesis Testing I

- Logic of hypothesis testing; Type I/II error, power
- Confidence intervals vs. Bayesian credible intervals (previewed)
- Significance conventions in physics: p-values, sigma levels, and the "5-sigma" discovery threshold
- **Midterm exam**

### Week 7: Chi-Square Fitting and Hypothesis Testing II

- The chi-square statistic and goodness-of-fit testing
- Least-squares fitting as a special case of maximum likelihood under Gaussian errors
- Nonparametric alternatives (Kolmogorov–Smirnov, Anderson–Darling) for distribution comparison
- Multiple testing / look-elsewhere effect (as encountered in searches for new signals)
- Lab: chi-square fit of a physical model to noisy data, with goodness-of-fit assessment. **Dataset:** the COBE/FIRAS cosmic microwave background spectrum (Fixsen et al. 1996 data) — fit a Planck blackbody function and evaluate goodness-of-fit via chi-square, the classic precision-cosmology example. **Access:** lambda.gsfc.nasa.gov/product/cobe/firas_monopole_get.html

## Part III — Modeling Methods (Weeks 8–10)

### Week 8: Linear and Nonlinear Regression

- Simple and multiple linear regression, OLS/weighted least squares (accounting for heteroscedastic measurement errors)
- Nonlinear least-squares fitting (e.g., `scipy.optimize.curve_fit`) for non-linear physical models
- Model diagnostics: residual analysis, parameter covariance matrix, correlated fit parameters
- Lab: fitting a nonlinear physical model with full covariance analysis. **Dataset:** published radial-velocity measurements for a known exoplanet host star (e.g., 51 Pegasi, or a system from the NASA Exoplanet Archive) — fit a Keplerian orbit model via nonlinear least squares to recover period, amplitude, and eccentricity with uncertainties. **Access:** exoplanetarchive.ipac.caltech.edu
- *Final project proposal due*

### Week 9: Model Comparison and Experimental Design

- ANOVA and its connection to regression (brief, as a bridge concept)
- Model selection criteria: AIC, BIC, likelihood-ratio tests, Bayes factors (preview)
- Principles of experimental/observational design: randomization, blocking, systematic vs. statistical uncertainty
- Lab: comparing nested physical models. **Dataset:** the Pantheon+ Type Ia supernova compilation (distance modulus vs. redshift) — compare a matter-only cosmological model against a model with dark energy (extra parameter) using AIC/BIC and a likelihood-ratio test. **Access:** github.com/PantheonPlusSH0ES/DataRelease

### Week 10: Generalized Linear Models and Time-Series/Spectral Analysis

- Logistic and Poisson regression (count-rate modeling, e.g. source detection rates)
- Introduction to time-series analysis: autocorrelation, periodograms, Lomb–Scargle for unevenly sampled astronomical data
- Power spectra and signal detection in noisy time series
- Lab: period-finding in a real light curve. **Dataset:** a *Kepler* or *TESS* light curve for a known variable star or transiting exoplanet host (retrieved via `lightkurve`) — recover the period with a Lomb–Scargle periodogram and compare to the published value. **Access:** mast.stsci.edu (data), lightkurve.github.io/lightkurve (tool)

## Part IV — Modern and Computational Methods (Weeks 11–12)

### Week 11: Resampling Methods and Bayesian Inference / MCMC

- Bootstrap and jackknife methods for error estimation when analytic errors are hard to derive
- Cross-validation; permutation tests
- Bayesian inference: priors, posteriors, credible intervals; Markov Chain Monte Carlo (Metropolis–Hastings, brief intro to `emcee`/`PyMC`)
- Lab: Bayesian parameter estimation via MCMC for a physical model, compared against the Week 8 frequentist fit. **Dataset:** the same exoplanet radial-velocity system used in Week 8 — refit the Keplerian orbit with `emcee`, compare posterior credible intervals to the earlier frequentist confidence intervals. **Access:** exoplanetarchive.ipac.caltech.edu

### Week 12: Multivariate Methods, Model Selection, and Project Presentations

- Principal component analysis (e.g., spectral classification), basic clustering
- Model selection revisited: AIC/BIC/cross-validated error and Bayesian evidence
- Brief survey of machine learning in physics/astronomy (classification of transients, anomaly detection) as a natural extension
- In-class demonstration dataset: Gaia DR3 color-magnitude diagram data for a nearby star cluster (e.g., the Pleiades) — PCA and clustering to identify cluster membership. **Access:** gea.esac.esa.int/archive
- Final project presentations

---

## Final Project

Students propose a research question tied to a physical or astronomical dataset (real, public, or simulated — e.g., an open catalog, detector data, or simulated experiment), and conduct an independent analysis applying at least two methods from the course, ideally contrasting a frequentist (likelihood/chi-square) and Bayesian (MCMC) approach where applicable. Deliverables: a short proposal (Week 8), a written report in the style of a short paper (~10–15 pages), and a 10-minute presentation (Week 12).

---

## Appendix: Dataset & Archive Access

| Week(s) | Source | Link |
|---|---|---|
| 1 | Gravity Spy (LIGO/Virgo glitch classification) | www.gravityspy.org |
| 2 | Fermi GBM Burst Catalog (HEASARC) | heasarc.gsfc.nasa.gov/W3Browse/fermi/fermigbrst.html |
| 3 | NIST/CODATA Fundamental Physical Constants | physics.nist.gov/cuu/Constants/ |
| 4 | SDSS DR17 (SkyServer / CasJobs) | skyserver.sdss.org/dr17/ |
| 5 | CosmicWatch Desktop Muon Detector | www.cosmicwatch.lns.mit.edu |
| 7 | COBE/FIRAS CMB Monopole Spectrum (NASA LAMBDA) | lambda.gsfc.nasa.gov/product/cobe/firas_monopole_get.html |
| 8, 11 | NASA Exoplanet Archive | exoplanetarchive.ipac.caltech.edu |
| 9 | Pantheon+SH0ES Data Release (GitHub) | github.com/PantheonPlusSH0ES/DataRelease |

| | | |
|---|---|---|
| 10 | MAST Archive / Lightkurve | mast.stsci.edu |
| 12 | Gaia Archive (ESA) | gea.esac.esa.int/archive/ |

*Note: archive interfaces and access methods (e.g., astroquery module APIs) occasionally change between releases — verify each link and query syntax before the term starts and again before each relevant session.*