Statistics for Physical Scientists

Introduction

Welcome to this short course in statistics! It's designed to give researchers, particularly in the physical sciences, some practical background and guidance in applying common statistical tools. The course covers:

- basic summary statistics, probability distributions and data combinations,
- overview of the Frequentist and Bayesian frameworks,
- correlation testing and significance, and sample comparisons,
- hypothesis tests and p-values,
- model-fitting and hypothesis testing using the χ^2 statistic,
- regression analyses (including least-squares and Gaussian processes),
- principal component analysis,
- practical error estimates (jack-knife, bootstrap and Monte Carlo simulations),
- propagating errors and Fisher matrix,
- Bayesian likelihood methods (including MCMC) and model selection.

The course is structured in 6 classes, as described below, which are split into content presentation, worked examples and practical activities using the datasets provided. Each class comes with an accompanying **python Jupyter notebook**, which provides summary notes and code for all the worked examples. I am an astrophysicist, so many of the activities tend to utilise astronomical datasets, even though the techniques are general.

The course is planned to run in 6 two-hour workshops. Please don't hesitate to contact me if you have any questions about the course: Chris Blake, cblake@swin.edu.au, 03 9214 8624, room AR303.

Useful books

The following is an (incomplete!) list of books which contain a great deal of practical wisdom in using statistics.

- Practical Statistics for Astronomers (Wall & Jenkins)
- Statistics for Nuclear and Particle Physicists (Lyons)
- Practical Bayesian Inference: A Primer for Physical Scientists (Bailer-Jones)
- Modern Statistical Methods for Astronomy (Feigelson & Babu)
- Principles of Data Analysis (Saha)
- Bayesian Logical Data Analysis for the Physical Sciences (Gregory)
- Data Analysis: A Bayesian Tutorial (Sivia)
- Numerical Recipes: The Art of Scientific Computing (Press, Teukolsky, Vetterling, Flannery)

Content outline

Class 1: Probability & statistics

- The process of science
- Common uses of statistics
- Summary statistics and their errors
- Estimators and bias
- Optimal combination of data
- Probability distributions: Binomial, Poisson, Gaussian
- The Poisson error
- Special role of the Gaussian distribution, central limit theorem
- Confidence regions and tails
- Comparing the Frequentist and Bayesian frameworks
- Monte Carlo simulations

Class 2: Correlation testing

- Correlation versus independence
- Pitfalls when searching for correlations
- Correlation coefficient
- Pearson product-moment correlation
- Significance of correlation
- Hypothesis tests and p-values
- Student t-distribution, tailed tests
- Non-parametric correlation tests; Spearman rank correlation
- Bayesian correlation methods
- Are the means of two samples consistent?
- Kolmogorov-Smirnov test

Class 3: Model fitting

- Comparing data and models
- The χ^2 statistic
- χ^2 probability distribution, degrees of freedom
- Reduced χ^2
- Use of χ^2 statistic as hypothesis test
- Modification for correlated data
- Use of χ^2 statistic for parameter fitting
- Errors in parameters from $\Delta\chi^2$; joint confidence regions

Class 4: Regression

- Introduction to regression

- Least-squares linear regression
- Quantifying the regression fit
- The case of errors in both co-ordinates
- Principal component analysis, eigenvalues and eigenvectors
- Interpolation
- Gaussian processes (kriging)

Class 5: Error estimates

- What is an error?
- Statistical versus systematic errors
- Error estimation using approximate sampling procedures
- Jack-knife errors
- Bootstrap errors
- Combining and propagating errors, linear and non-linear cases
- Fisher matrix forecasts
- Monte Carlo simulations

Class 6: Bayesian likelihood methods

- Bayesian Methods: conditional probability, Bayes' Theorem
- Role of the prior and likelihood
- Posteriors and confidence limits
- Marginalization
- Use of likelihood for parameter fitting
- Monte Carlo Markov Chains
- Model selection; is adding another parameter justified?
- Bayes factor and Jeffreys scale
- Akaike information criteria