

## **Lecture 1: Introduction to statistics and astrostatistics**

**What is astronomy? What is statistics?**

**Overview of Statistics**

**Overview of Astrostatistics**

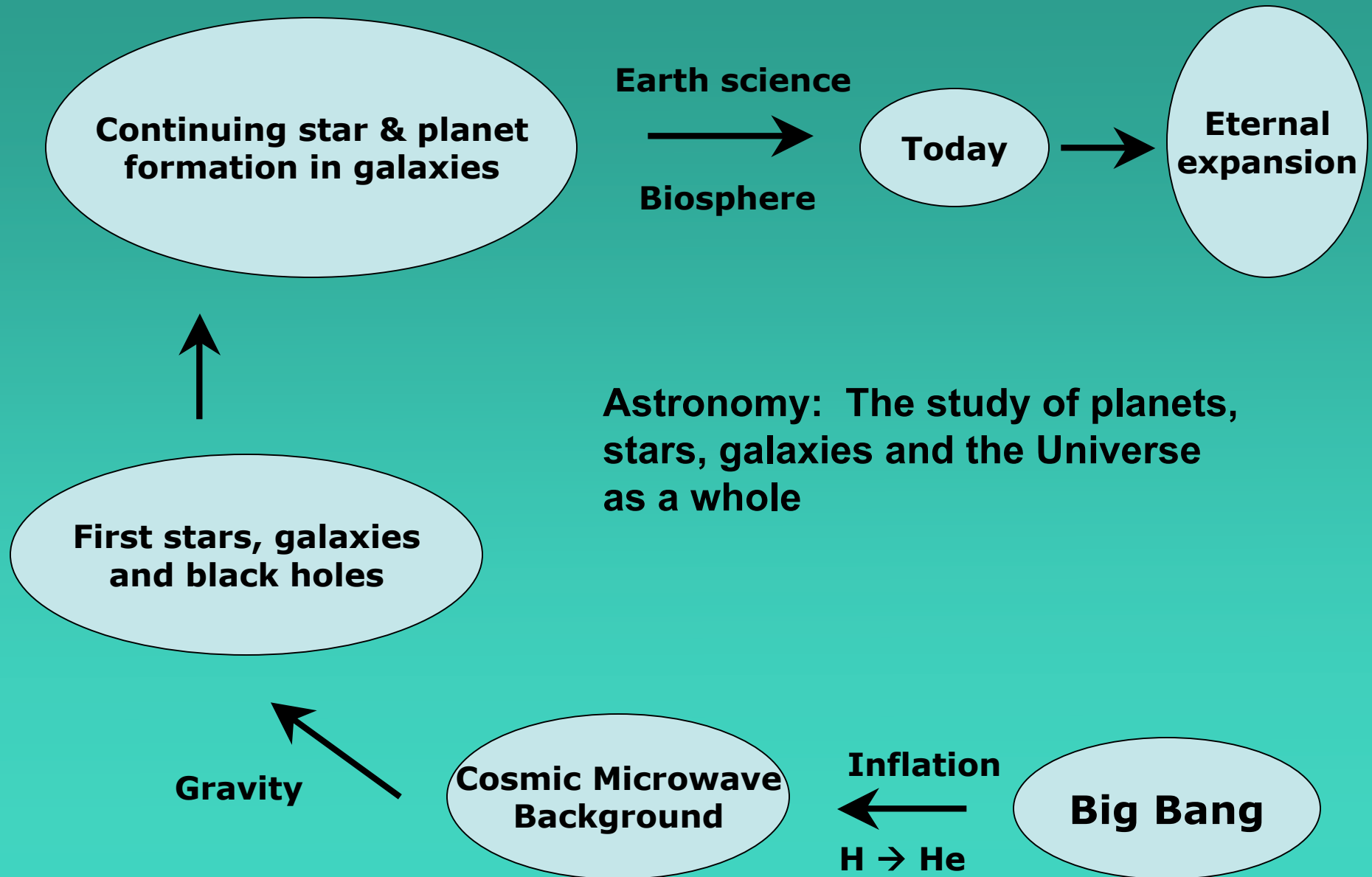
**Resources**

# What is astronomy?

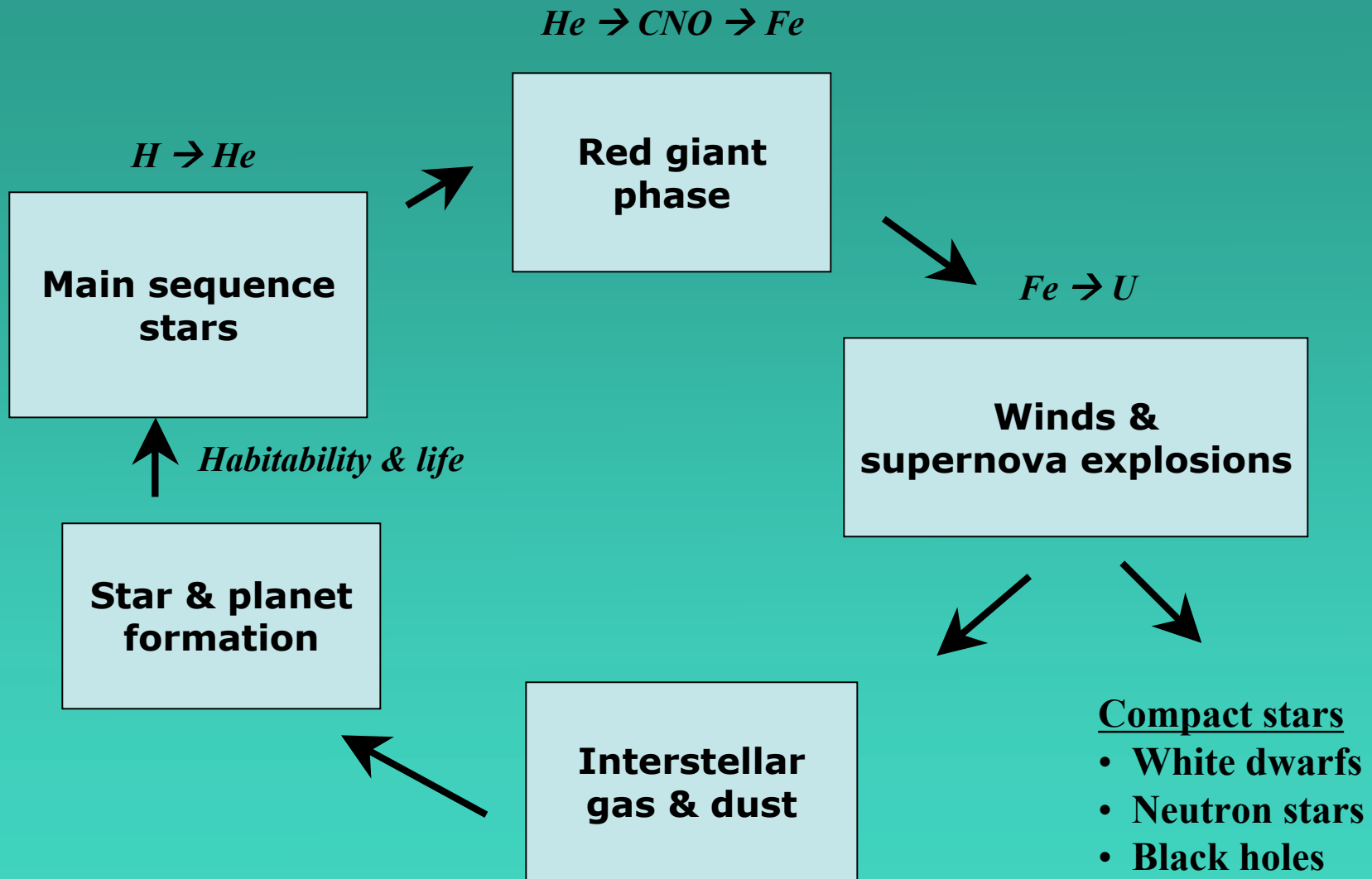
**Astronomy** (astro = star, nomen = name in Greek) is the observational study of matter beyond Earth – planets in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations. The perspective is rooted from our viewpoint on or near Earth using telescopes or robotic probes.

**Astrophysics** (astro = star, physis = nature) is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an indirect, inferential intellectual effort based on the assumption that gravity, electromagnetism, quantum mechanics, plasma physics, chemistry, and so forth – apply universally to distant cosmic phenomena.

# Overview of modern astronomy & astrophysics



# Lifecycle of the stars



# What is statistics? (*No consensus !!*)

## Definition 1: Statistics concerns data

- “The first task of a statistician is cross-examination of data” (R.A. Fisher)
- “[Statistics is] the study of algorithms for data analysis” (R. Beran)
- **Statistics** is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data  
(Wikipedia)

## Definition 2: Statistics links data to models ... which may or may not give correct interpretations of reality

- “A statistical inference carries us from observations to conclusions about the populations sampled” (D.R. Cox 1958)
- “Some statistical models are helpful in a given context, and some are not” (T. Speed, addressing astronomers)
- “There is no need for these hypotheses to be true, or even to be at all like the truth; rather ... they should yield calculations which agree with observations” (Osiander’s Preface to Copernicus’ *De Revolutionibus*, quoted by C.R. Rao)

## A pessimistic view of statistics and science:

“The object [of *statistical* inference] is to provide ideas and methods for the critical analysis and, as far as feasible, the interpretation of empirical data ... The extremely challenging issues of *scientific* inference [synthesizes] very different kinds of conclusions if possible into a coherent whole or theory ... The use, if any, of simple *quantitative* notions of probability and their numerical assessment [in scientific inference] is unclear”

(D.R. Cox *Principles of Statistical Inference*, 2006)

## An optimistic view of statistics and science:

“The goal of science is to unlock nature’s secrets. ... Our understanding comes through the development of theoretical models which are capable of explaining the existing observations as well as making testable predictions. ... Fortunately, a variety of sophisticated mathematical and computational approaches have been developed to help us through this interface, these go under the general heading of statistical inference.”

(P. C. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, 2005)

**My personal conclusions**  
**(X-ray astronomer with 25 yrs statistical experience)**

The application of statistics to scientific data is not a straightforward, mechanical enterprise. It requires careful statement of the problem, model formulation, choice of statistical method(s), calculation of statistical quantities, and judicious evaluation of the result.

Modern statistics is vast in its scope and methodology. It is difficult to find what may be useful (jargon problem!), and there are usually several ways to proceed. Some issues are debated among statisticians.

Statistics is based on mathematical proofs which limit the applicability of established results. It is easy to ignore these limits and make mistakes.

It can be difficult to interpret the meaning of a statistical result with respect to the scientific goal. We are scientists first! Statistics is only a tool. We should be knowledgeable in our use of statistics and judicious in its interpretation.

# Statistics: Some basic definitions

- Statistical inference
  - Seeking quantitative insight & interpretation of a dataset
- Hypothesis testing
  - To what confidence is a dataset consistent with a previously stated hypothesis?
- Estimation
  - Seeking the quantitative characteristics of a functional model designed to explain a dataset. An estimator seeks to approximate the unknown parameters based on the data
- Probability distribution
  - A parametric functional family describing the behavior of a parent distribution of a dataset (e.g. Gaussian = normal)
- Nonparametric statistics
  - Inference based directly on the dataset without parametric models
  - Independent & identically distributed (iid) data point
  - A sample of similarly but independently acquired quantitative measurements.

## Some basic definitions (cont.)

- Frequentist statistics
  - Suite of classical inference methods based on simple probability distributions. Parameters are fixed (though possibly unknown) while data are random variables.
- Bayesian statistics
  - Suite of inference methods based on Bayes' Theorem based on likelihoods and prior distributions. Data are fixed while parameters are random variables.
- $L_1$  and  $L_2$  methods
  - 19th century frequentist methods for estimation based on minimizing the absolute or squared deviations between a sample and a model
- Maximum likelihood methods
  - 20th century frequentist methods for parametric estimation based on the likelihood that a dataset fits the model (often like  $L_2$ )
- Gibbs sampling, MCMC, ...
  - New computational methods for integrations over hypothesis space in Bayesian statistics

## Some basic definitions (cont.)

- Robust methods
  - Statistical procedures that are insensitive to data outliers
- Model selection & validation
  - Procedures for estimating the goodness-of-fit and choice of parametric model. (Nested vs. non-nested models, model misspecification)
- Statistical power, efficiency & bias
  - Mathematical evaluation of the effectiveness of a statistical procedure to achieve its desired goals
- Two-sample & k-sample tests
  - Statistical tests giving probabilities that k samples are drawn from the same parent sample
- Independent & identically distributed (iid) data point
  - A sample of similarly but independently acquired quantitative measurements.
- Heteroscedasticity
  - A failure of iid due to differently weighted data points

# Some fields of applied statistics

- Multivariate analysis
  - Establishing the structure of a table of rows & columns
  - Analysis of variance, regression, principal component analysis, discriminant analysis, factor analysis
- Multivariate classification
  - Dividing a multivariate dataset into distinct classes
- Correlation & regression
  - Establishing the relationships between variables in a sample
- Time series analysis
  - Studying data measured along a time-like axis
- Spatial analysis
  - Studying point or continuous processes in 2-3-dimensions
- Survival analysis
  - Studying data subject to censoring (e.g. upper limits)
- Data mining
  - Studying structures in mega-datasets
- Biometrics, econometrics, psychometrics, chemometrics, quality assurance, geostatistics, astrostatistics, ..., ...

## Decisions needed by a practicing astronomer

- When little is known, exploratory data analysis and nonparametric inference may be best.
- When something is known about the underlying populations, then parametric inference can proceed. Please distinguish what is *truly established* vs. what is *habitually assumed* (e.g. distributions are Gaussian in logarithmic variables).
- When the goals are to establish the applicability of an astrophysical model, the choice between astrophysical models, and/or measurement of the parameters of models, then parametric inference is needed. Model validation and model selection methodology is still not well-established within statistics ... many debates continue.
- Within parametric inference, choices between frequentist (e.g. maximum likelihood) and Bayesian approaches.

# Astronomy & statistics: A glorious history

*Hipparchus (4th c. BC): Average via midrange of observations*

*Galileo (1572): Average via mean of observations*

*Halley (1693): Foundations of actuarial science*

*Legendre (1805): Cometary orbits via least squares regression*

*Gauss (1809): Normal distribution of errors in planetary orbits*

*Quetelet (1835): Statistics applied to human affairs*

***But the fields diverged in the late 19-20th centuries,  
astronomy → astrophysics (EM, QM)  
statistics → social sciences & industries***

## Do we need statistics in astronomy today?

- Are these stars/galaxies/sources an unbiased sample of the vast underlying population?
- When should these objects be divided into 2/3/... classes?
- What is the intrinsic relationship between two properties of a class (especially with confounding variables)?
- Can we answer such questions in the presence of observations with measurement errors & flux limits?

## Do we need statistics in astronomy today?

- Are these stars/galaxies/sources an unbiased sample of the vast underlying population? **Sampling**
- When should these objects be divided into 2/3/... classes? **Multivariate classification**
- What is the intrinsic relationship between two properties of a class (especially with confounding variables)? **Multivariate regression**
- Can we answer such questions in the presence of observations with measurement errors & flux limits? **Censoring, truncation & measurement errors**

- When is a blip in a spectrum, image or datastream a real signal? **Statistical inference**
- How do we model the vast range of variable objects (extrasolar planets, BH accretion, GRBs, ...)? **Time series analysis**
- How do we model the 2-6-dimensional points representing galaxies in the Universe or photons in a detector? **Spatial point processes & image processing**
- How do we model continuous structures (CMB fluctuations, interstellar/intergalactic media)? **Density estimation, regression**

# How often do astronomers need statistics? (a bibliometric measure)

Of ~15,000 refereed papers annually:

1% have *'statistics'* in title or keywords

5% have *'statistics'* in abstract

10% treat variable objects

5-10% (est) analyze data tables

5-10% (est) fit parametric models

# We use a unnecessarily narrow suite of statistical methods

Modern statistics is vast. Astronomy encounters problems in: image analysis, time series analysis, model selection, regression, nonparametrics, spatial point processes, multivariate analysis, survival analysis, ..., ... Monographs are published each year in each these fields.

The software situation is much improved: Since c.2002, **R** has emerged as the premier public-domain statistical software package. Similar to IDL in style, but with a huge range of built-in statistical functionalities.

See <http://r-project.org> and tutorials at <http://astrostatistics.psu.edu>

# A new imperative: Large-scale surveys, megadatasets & the Virtual Observatory

Huge, uniform, multivariate databases are emerging from specialized survey projects & telescopes:

- $10^9$ -object catalogs from USNO, 2MASS & SDSS opt/IR surveys
- $10^6$ - galaxy redshift catalogs from 2dF & SDSS
- $10^5$ -source radio/infrared/X-ray catalogs
- $10^{3-4}$ -object samples of well-characterized stars & galaxies with dozens of measured properties
- Many on-line collections of  $10^2$ - $10^6$  images & spectra
- Planned Large-aperture Synoptic Survey Telescope will generate ~10 Pby

*The Virtual Observatory is an international effort underway to federate these distributed on-line astronomical databases.*

Powerful statistical tools are needed to derive scientific insights from extracted VO datasets

# We are making mistakes!

- The **likelihood ratio test** for comparing two parametric models cannot be applied when a parameter is near zero (Protassov, van Dyk et al. 2002)
- Probabilities from the 1-sample **Kolmogorov-Smirnov test** comparing a univariate dataset to its best-fit model are incorrect (Lilliefors 1969; Babu & Feigelson ADASS XV 2006)
- The **Anderson-Darling test** is often more sensitive than the K-S test, and there is no valid 2-dimensional K-S test (Stephens 1974; Simpson 1951)
- **Power-law models** (= Pareto distribution) should not be fit to binned data, use the MLE on the original events (Crawford et al. 1970)

# We are making progress!

- Growth of cross-disciplinary research collaborations in astrostatistics: California-Harvard Astrostatistics Collaboration, groups at Carnegie-Mellon, Berkeley, Michigan, Penn State, Cornell, SAMSI, ...
- Growth of conference series (SCMA, ADA, PhysStat, SAMSI) and monographs for advanced statistical treatment of astronomical data
- Week-long Summer School in Statistics for Astronomers held at Penn State since 2005. In steady state, we are training ~10% of world's astronomy graduate students.
- Useful resources at <http://astrostatistics.psu.edu>

## Some methodological challenges for astrostatistics in the 2000s

- Simultaneous treatment of measurement errors and censoring (esp. multivariate). See B.C.Kelly (ApJ 2007)
- Statistical inference and visualization with very-large-N datasets too large for computer memories
- A user-friendly cookbook for construction of likelihoods & Bayesian computation of astronomical problems
- Links between astrophysical theory and wavelet coefficients (spatial & temporal)
- Rich families of time series models to treat accretion and explosive phenomena

## Some resources for learning & using statistics

These lectures assume familiarity with basic statistics for physical scientists at the level of:

- Data Reduction and Error Analysis for the Physical Sciences, Bevington & Robinson (2003)
- Practical Statistics for Astronomers, Wall & Jenkins (2003)
- Statistical Data Analysis, Cowan (1998)

Useful broad-scope volumes in statistics:

- Mathematical Statistics and Data Analysis, Rice (1995, undergraduate)
- Introduction to Mathematical Statistics, Hogg, McKean & Craig (2005, graduate)
- Statistical Models, Davison (2003, advanced with R code)
- Principles of Statistical Inference, Cox (2006, discursive)
- Statistical Analysis for the Physical Sciences, James (2006, graduate)

A selection of many topical volumes (see <http://astrostatistics.psu.edu> Bibliographies):

- Monte Carlo Statistical Methods, Robert & Casella (2004)
- Introduction to Modern Nonparametric Statistics, Higgins (2004)
- Applied Multivariate Statistical Analysis, Johnson & Wichern (2002)
- An R and S-PLUS Companion to Multivariate Analysis, Everitt (2005)
- Data Analysis: A Bayesian Tutorial, Sivia & Skilling (2006)
- Bayesian Data Analysis, Gelman, Carlin, Stern & Rubin (2003)
- The Analysis of Time Series: An Introduction, Chatfield (2003)
- A Wavelet Tour of Signal Processing, Mallat & Mallat (1999)
- Applied Measurement Error in Nonlinear Models, Carroll, Ruppert & Stefanski (2006)
- Astronomical Image and Data Analysis, Starck & Murtagh (2006)
- Statistical Analysis of Spatial Point Patterns, Diggle (2001)
- Visual Data Mining: Techniques and Tools for Data Visualization and Mining, Soukup & Davidson (2002)
- Elements of Statistical Learning: Data Mining, Inference, and Prediction, Hastie, Tibshirani & Friedman (2001)
- Survival Analysis: A Self-Learning Text, Kleinbaum & Klein (2005)
- Model Selection and Multimodel Inference, Burnham & Anderson (2002)

# The R Statistical Computing Package

## **Base R**

arithmetic & linear algebra, bootstrap resampling, empirical distribution tests, exploratory data analysis, generalized linear modeling, graphics, robust statistics, linear programming, local and ridge regression, maximum likelihood estimation, multivariate analysis, multivariate clustering, neural networks, smoothing, spatial point processes, statistical distributions & random deviates, statistical tests, survival analysis, time series analysis

## **Selected methods from Comprehensive R Archive Network (CRAN)**

***Bayesian computation & MCMC***, classification & regression trees, genetic algorithms, ***geostatistical modeling***, hidden Markov models, irregular time series, kernel-based machine learning, least-angle & lasso regression, likelihood ratios, map projections, ***mixture models & model-based clustering***, nonlinear least squares, multidimensional analysis, multimodality test, multivariate time series, ***multivariate outlier detection, neural networks***, non-linear time series analysis, nonparametric multiple comparisons, omnibus tests for normality, orientation data, ***parallel coordinates plots***, partial least squares, periodic autoregression analysis, principal curve fits, projection pursuit, ***quantile regression***, random fields, ***random forest classification***, ridge regression, robust regression, self-organizing maps, shape analysis, space-time ecological analysis, ***spatial analysis & kriging***, spline regressions (MARS, BRUTO), tessellations, ***three-dimensional visualization, wavelet toolbox***

## **Interfaces**

BUGS, C, C++, Fortran, Java, Perl, Python, Xlisp, XML

## **I/O**

ASCII, binary, bitmap, cgi, FITS, ftp, gzip, HTML, SOAP, URL

## **Graphics & emulators**

Grace, GRASS, Gtk, Matlab, OpenGL, Tcl/Tk, Xgobi

## **Math packages**

GSL, Isoda, LAPACK, PVM

## **Text processor**

LaTeX

*In recent years, R has become the premier public-domain statistical computing package. Easily downloaded from <http://www.r-project.org>. Tutorials available in recent books and from <http://astrostatistics.psu.edu/datasets>.*