Software for Statistics (for Physics)

Jim Linnemann, MSU

PHYSTAT2005, Oxford UK
September 13, 2005

Photo: Fermilab esd
What A Physicist Wants

1) A better environment for doing physics analysis using statistical tools
   Beyond what was in Root
   R envy: R Language used by statisticians

2) To know what’s available on the web
   software without writing it all from scratch
   Mark Twain & Louis Lyons’ advice

3) A physics-oriented repository
   Astronomers and Statisticians have them
   Improvement, not re-invention!
Steps Towards These Goals:

2004: PHYSTAT workshop at MSU
    (on statistical software, mainly Root and R)

2004: develop Page of Web pointers
    Shockingly, largest I currently know of in HEP

2005: Fermilab PHYSTAT workshop with Mark Fishler
    on statistical software repository for Physics

These were small invited-participation workshops, but intended to produce work
1) Root/R (mostly) Workshop

user.pa.msu.edu/linnemann/public/workshop

Luc Tierney, R core development team (thanks: J. Friedman, N. Reid)

Rene Brun, Root system

Astronomers:
   Eric Feigelson (StatCodes) and VOStat
   Tim Beers (MSU) Rostat (Robust statistics package—\textit{not} on web!)

Physicists:
   Harrison Prosper, Scott Snyder, Sherry Towers (TerraFerma),

Physicist R users from FNAL:
   Adam Lyon, Jim Kowalkowski, Marc Paterno

Outcomes:
   Adam Lyon wrote a Root Tree reader for R
   Eric Feigelson confirmed decision to implement VOSTATS in R
   Rene Brun: furthered interest in R, statistics, packages?
   \textit{Celebrating }\frac{1}{4} \textit{ century of encouraging Rene to do even better...}
Root: key features

GUI for graphics, “cuts”
(data subregion selection)

I/O for tree-structured data: scales to PetaBytes

Histogram as base metaphor

C++ at command line: macros, compiled macros

Extensible, though most wouldn’t say “easily”

See Anna Kreshuk talk Wed pm

Sophisticated nonlinear fitting and parameter errors

Statistical algorithms: few, but growing
R: key features

Elegant data manipulation language, S: \( (R = gnu S) \)
command prompt and macros
interpreted, heading to byte-compilation (often fast: wrapped C)
GUI: only now building hooks; most users satisfied with command line:

**QUICK & EASY interactive analysis**

Standard tool of professional research statisticians

**If they suggest a method, it’s probably in R**

bootstrap, lasso, bag, boost, cross-validate…

Sophisticated statistical tests and graphics (many not in HEP)
Links to further multidimensional graphics (Ggobi)

**Broad package library, with trivial download mechanism**

Straightforward extensibility to new packages; R or C code

Data in virtual memory

Data frames: vectors are a basic metaphor (cf. histogram in Root)
interfaces to databases (postgres; mysql)
Parallel computation under development
R graphics example: Scatterplot array

Can read Root Trees in R (Adam Lyon)

Many graphics, statistical tests available or downloadable trivially

Marc Paterno talk on R in HEP analysis: Wed pm
High Energy Physics

There are evolving plan for a HEP statistical software repository; your comments (and your collaboration's) would be welcome.

Particle Data Group Statistics Summary describes statistical methods (theory) on which there is consensus in HEP

Glen Cowan's statistical resources page (Royal Holloway physics); go up a link for some software associated with his book.

There are some statistical routines in Root (an interactive data analysis framework); and in cerlib, clhep and Fermilab's Zoom

FreeHep points to other HEP analysis software (including JAS, Java Analysis Studio), but does not have a specific statistics section.

CDF statistics committee a Tevatron experiment's statistics page: mostly methods discussion

A simple version of the D0 experiment's Bayesian limit calculator

Babar statistics working group a SLAC experiment's statistics page: methods and a few applets

Geant statistical packages, Maria Grazia Pia, HEP, INFN Genova, C++ library

Fermilab Advanced Analysis Group

TerraFerMA, Sherry Towers: a root-compatible package combining several classifiers and helping select candidate variables for multidimensional analysis.

gnu gsl (gnu scientific library) contains random number generators, as well as some histogramming, ntuples, moments for weighted events, and autocorrelation calculations.

sourceforge.net a broad repository of open source software. Basic browsing or search by name without subscribing.

You could troll about in the scientific/engineering section and find, for example, roofit.

The Computer Physics Communications program library contains a few items of interest; it requires a subscription to the journal.

Cedar is beginning a HEP archive.
More Resources: Astronomy, Statistics

Astrophysics

Statcodes Eric Feigelson et. al., Penn State: big collection, with commentary; see also his Astrostatistics book. Look here--much broader than astrophysics! Includes link to web-based VOSTAT (Virtual Observatory Statistics) project, largely implemented in R (see below).

Statistical Resources Eric Hooper, Harvard

StatPy Python interfaces to statistical software, Tom Loredo, Cornell; see also his Bayesian Inference in the Physical Sciences (Software Section) see especially the ominously-named BUGS (heavily used by statisticians), and BAYESPACK

Astrostatistics, Barry Madore, Cal Tech

Mutual translation glossaries for astronomers and statisticians

Statistics

http://www.galaxy.gmu.edu/papers/astr1.html George Mason Statistics
http://members.aol.com/ohnp71/avarta2.html free software and interactive pages from John Pezzullo (retired, Georgetown Statistics)
http://my.execpc.com/~helberg/statframes.html Clay Helberg of SPSS
http://www.stat.ufl.edu/vib/statistics.html Use your browser to search for Resources to get to the good stuff Journal of Statistical Software; in many programming languages.
http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml bugs Markov Chain MC package

From national labs:
http://my.execpc.com/~helberg/statframes.html Clay Helberg of SPSS
http://gams.nist.gov/ see Class L for a mixture of commercial and academic software NIST/SEMA TECH e-Handbook of Statistical Methods (Engineering Statistics: reference, but not much on multidimensional data, and little software under Tool and Aids)
StatCodes

Online statistical software for astronomy and related physical sciences

StatCodes is a Web metasite with links to public domain software implementing statistical methods. The codes are chosen for their potential utility for research in astronomy and other physical sciences. Some are obtained from large code archives (such as R, CRAN, StatLib, GAMS), while others are scattered throughout the Internet. It includes multi-purpose statistical computing systems, single-purpose packages, single-purpose source codes, and various Web resources and services. StatCodes listings are organized by statistical topic.

In mid-2005, StatCodes was updated and incorporated into Penn State's Center for Astrostatistics (CAST). CAST provides a range of statistical resources for the physical scientist including astrostatistical tutorials, annotated bibliographies, conferences and research in addition to StatCodes.
StatLib: Standard Source for Statistics Codes

StatLib -- Downloads

Search

[ Main download area | New | Popular | Top Rated ]

Category: Software Archive

Downloads also available in Software Archive Sub-categories:

- Applied Statistics algorithms (1)
- BLSS Archive (1)
- CMLIB archive (1)
- Designs Archive (13)
- DOS--General (29)
- DOS--S Index (41)
- FIASCO (1)
- General Archive (157)
- Glim Archive (5)
- Griffiths and Hill Archive (35)
- JASA Software Archive (11)
- JCGS Archive (26)
- Journal of Quality Technology Archive (42)
- MacAnova Archive (5)
- Maps (4)
- Matlab extensions and software (14)
- Minitab Macros (5)
- Multivariate Archive (14)
- Sapaclisp Archive (1)
- Software and extensions for the genstat language (9)
- Software and extensions for the S (Splus) language (284)
- The Comprehensive R Archive Network (CRAN) (2)
- Software for Graphical Modelling (5)
- XlispStat Archive (16)
- The American Statistician Software Archive (6)
- Sort downloads by: Title (A-Z) Date (A-D) Rating (A-D) Popularity (A-D)

Resources currently sorted by:Title (A-Z)
And yet more

Multivariate Analysis and Statistical Learning

Useful buzzwords to search on in **bold**; "statistics" will get you more data than methods.

**R** The R project for **Statistical Computing** GNU implementation of the S language

- Graphics, statistical algorithms, and a huge repository (CRAN) of R packages. Extensive online documentation. Published books include Introductory Statistics with R by Dalgaard; and Programming with Data: A Guide to the S Language by Chambers; Modern Applied Statistics with S-PLUS, by Venables & Ripley, and others; there's a very good [R tutorial](http://www.ggobi.org) here's another tutorial, but without graphics

**http://www.ggobi.org**/ Ggobi visualization package for multidimensional data.

- Includes dynamic graphics such as arrays of scatterplots, brushing techniques (highlighting groups of objects in one dimension and having their coordinates highlighted in other coordinates); parallel coordinate plots, and grand tours. Interfaces exist to R and Python front ends, and database back ends. I've skimped on Perl here and elsewhere but it often where you find Python interfaces, you'll also find Perl-though Ruby not as often.

**http://www.omegahat.org**/ The Omega project for Statistical Computing

- Interfaces between R, Python, XML, Java, databases, and other goodies. At this point, aimed more at developers than users.

Jerry Friedman (High Energy Physicist turned Statistician) has software for a number of multivariate techniques on the web, don't miss his book below.


- The best multivariate analysis and Statistical Learning textbook I know of; web site includes software. From a modern and sophisticated computational statistics viewpoint, but quite readable. Compares methods from trees to neural nets, kernel methods, and support vector machines, though nothing on genetic algorithms. You can even learn the meaning of useful things like bootstrapping and boosting and other post-1960's statistical jargon!

**http://a.fri.uni-lj.si/~aleks/orngp** Data mining in Python support vector machines, logistic regression, clustering, by Aleks Jakulin

**http://magis.fri.uni-lj.si/orange/** Orange

- a massive toolkit, including visualization, feature selection, many evaluation tools, including calibration curves and ROC (Receiver Operating Characteristic = efficiency for signal vs fraction of background: true positives vs. false positives). Practically all major algorithms from machine learning. Python is a popular interface to this library.

**http://www.pitt.edu/~csra/software.html** Multivariate Analysis Software

**http://www.ll.mit.edu/IST/Itnet/** Classification Software collection---easy to compare methods

**http://www.kdnuggets.com/software/classification.html** mixture of commercial and academic software links

**http://www.phi.tudelft.nl/PR/Info/software.html** Machine Learning Resources online

**http://www.ncrg.aston.ac.uk/NN/software.html** Neural Network Software list, including SNNS popular in Babar


- ROC curves and a critique of using "best accuracy" on test data sets as a comparison criterion across algorithms (and implicitly, perhaps, as a training
# Alphabetical R Package Listings

[cran.r-project.org](http://cran.r-project.org) | Packages

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bindata</td>
<td>Generation of Artificial Binary Data</td>
</tr>
<tr>
<td>Biodem</td>
<td>Biodemography functions</td>
</tr>
<tr>
<td>biopara</td>
<td>Self-contained parallel system for R</td>
</tr>
<tr>
<td>bitops</td>
<td>Functions for Bitwise operations</td>
</tr>
<tr>
<td>blighty</td>
<td>United Kingdom coastlines</td>
</tr>
<tr>
<td>BMA</td>
<td>Bayesian Model Averaging</td>
</tr>
<tr>
<td>boa</td>
<td>Bayesian Output Analysis Program (BOA) for MCMC</td>
</tr>
<tr>
<td>Bolstad</td>
<td>Bolstad functions</td>
</tr>
<tr>
<td>boolean</td>
<td>Boolean logit and probit</td>
</tr>
<tr>
<td>boost</td>
<td>Boosting Methods for Real and Simulated Data</td>
</tr>
<tr>
<td>boot</td>
<td>Bootstrap R (S-Plus) Functions (Canty)</td>
</tr>
<tr>
<td>bootstrap</td>
<td>Functions for the Book &quot;An Introduction to the Bootstrap&quot;</td>
</tr>
<tr>
<td>bvlr</td>
<td>Bayesian QTL mapping toolkit</td>
</tr>
<tr>
<td>BradleyTerry</td>
<td>Bradley-Terry models</td>
</tr>
<tr>
<td>brlr</td>
<td>Bias-reduced logistic regression</td>
</tr>
<tr>
<td>BRugs</td>
<td>OpenBUGS and its R interface BRugs</td>
</tr>
<tr>
<td>BSDA</td>
<td>Basic Statistics and Data Analysis</td>
</tr>
<tr>
<td>BsMD</td>
<td>Bayes Screening and Model Discrimination</td>
</tr>
<tr>
<td>caMassClass</td>
<td>Processing and Classification of Protein Mass Spectra (SELDI) Dat</td>
</tr>
<tr>
<td>car</td>
<td>Companion to Applied Regression</td>
</tr>
<tr>
<td>cat</td>
<td>Analysis of categorical-variable datasets with missing values</td>
</tr>
</tbody>
</table>
ASDS Astronomy Repository
asds.stsci.edu/packages.html

Software Packages, by category:

Top level links take you to a page listing packages in each category, along with their associated on-line documents. Links under each heading take you directly to the package description page.

- **Data reduction and analysis software**
  - AIPS++, AIPS, AstraImage, CoStat, eclipse, ESO-MIDAS, EXSAS, FIGARO, FTOOLS, GHRSDAF, GILDAS, GIPS, IDL, IDLastro, IRAF, LAIA, LinuxForAstro, MIRIAD, MR1, NEMO, Penelope, PROS, STARLINK, STARMAN, STSDAS, XANADU, YOMAMA

- **Document preparation software**
  - AASTeX

- **Graphics, plotting, image display, visualization ... software**
  - AIPSVIEW, CelestialMaps, CoPlot, DS9, FITSview, GGS, NICMOSlook, PGPERL, PGPLOT, PSPLOT, Radio-SkyPipe, SACimage, SAOng, SM, STAP, StarSplatter, VFleet, WebWinds, WIP

- **Modeling and simulation software**
  - Aberrator, adap, CLOUDY, COSMICS, CSUAPMM, DUSTY, FASTELL, FLASH, SPECTRUM, TIPSY

- **Math and statistics software**
  - Maple, StatCodes

- **Subroutine libraries**
  - FITSIO, NAg, NumericRecipes

- **Utility programs**
HEASARC Software

The following software is supported by the HEASARC. Most of this software is designed for professional researchers and advanced students for the analysis of scientific astronomical observations in FITS format. The 'f' program is also suitable for more general use by amateurs and educators for viewing astronomical image files. These software packages are to be downloaded to the users machine. Web-based tools are available on our utilities page.

HEASARC Software Packages

- **HEAsoft** - A unified release of the FTOOLS and XANADU packages
- **FTOOLS** - General FITS file utility programs and mission-specific data analysis tools
- **XANADU** - Suite of spectral (xspec), timing (xronos), and image (ximage) analysis programs
- **XSELECT** - Multipurpose tool for filtering event files and generating images, spectra, and light curves
- **XSTAR** - Program for calculating physical conditions and emission spectra in photoionized gases
- **f** - Interactive editor and viewer for astronomical data files in FITS format
- **Hera** - Enables complete interactive analysis over the Internet of data products retrieved from the Browse data archive
- **MAKI** - A multi-mission observation visualizer and planning utility
- **FITSIO** - A subroutine library for reading and writing FITS files for C and Fortran programmers
<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aCGH</td>
<td>1.1.7</td>
<td>Classes and functions for Array Comparative Genomic Hybridization data.</td>
</tr>
<tr>
<td>affy</td>
<td>1.6.7</td>
<td>Methods for Affymetrix Oligonucleotide Arrays</td>
</tr>
<tr>
<td>affycomp</td>
<td>1.4.3</td>
<td>Graphics Toolbox for Assessment of Affymetrix Expression Measures</td>
</tr>
<tr>
<td>affydata</td>
<td>1.4.1</td>
<td>Affymetrix Data for Demonstration Purpose</td>
</tr>
<tr>
<td>affyImGUI</td>
<td>1.2.7</td>
<td>GUI for affy analysis using limma package</td>
</tr>
<tr>
<td>affyPdnn</td>
<td>1.1.0</td>
<td>Probe Dependent Nearest Neighbours (PDNN) for the affy package</td>
</tr>
<tr>
<td>affyPLM</td>
<td>1.3.3</td>
<td>Methods for fitting probe-level models</td>
</tr>
<tr>
<td>altcdfenvs</td>
<td>1.1.11</td>
<td>alternative cdfenvs</td>
</tr>
<tr>
<td>annaffy</td>
<td>1.0.18</td>
<td>Annotation tools for Affymetrix biological metadata</td>
</tr>
<tr>
<td>AnnBuilder</td>
<td>1.5.31</td>
<td>Bioconductor annotation data package builder</td>
</tr>
<tr>
<td>annotate</td>
<td>1.5.16</td>
<td>Annotation for microarrays</td>
</tr>
<tr>
<td>apComplex</td>
<td>1.1.5</td>
<td>Estimate protein complex membership using AP-MS protein data</td>
</tr>
<tr>
<td>arrayMagic</td>
<td>1.5.8</td>
<td>two-colour cDNA array quality control and preprocessing</td>
</tr>
<tr>
<td>arrayQuality</td>
<td>1.0.11</td>
<td>Assessing array quality on spotted arrays</td>
</tr>
<tr>
<td>bim</td>
<td>1.0.3</td>
<td>Bayesian Interval Mapping Diagnostics</td>
</tr>
<tr>
<td>Biobase</td>
<td>1.5.12</td>
<td>Biobase: Base functions for Bioconductor</td>
</tr>
<tr>
<td>bioDist</td>
<td>1.0.0</td>
<td>Different distance measures</td>
</tr>
<tr>
<td>biomaRt</td>
<td>1.0.4</td>
<td>Interface to BioMart databases (e.g. Ensembl)</td>
</tr>
<tr>
<td>Biostrings</td>
<td>1.1.1</td>
<td>String objects repre.senting biological sequences</td>
</tr>
<tr>
<td>bridge</td>
<td>1.0.1</td>
<td>Bayesian Robust Inference for Differential Gene Expression</td>
</tr>
<tr>
<td>Category</td>
<td>1.0.1</td>
<td>Category Analysis</td>
</tr>
<tr>
<td>ChromoViz</td>
<td>1.0</td>
<td>Multimodal visualization of gene expression</td>
</tr>
<tr>
<td>CoCiteStats</td>
<td>1.0.0</td>
<td>Different test statistics based on co-citation.</td>
</tr>
<tr>
<td>convert</td>
<td>1.1.11</td>
<td>Convert Microarray Data Objects</td>
</tr>
<tr>
<td>ctc</td>
<td>1.2.7</td>
<td>Cluster and Tree Conversion.</td>
</tr>
<tr>
<td>ddmA</td>
<td>1.0.1</td>
<td>Efficient design and analysis of factorial two-colour microarray data</td>
</tr>
</tbody>
</table>
3) Repository Envy: Motivation for Fermilab Workshop

Where is the HEP contributed-package library?

Trying to start one at Fermilab

Wish I could find & install Root packages as easily as R packages

Design choices make it possible in R
Computing and Engineering for Physics Applications
Projects and Activities

Software for Physics Analysis
- Cernlib
- CLHEP
- Database Software
- BLASTMAN Accelerator Database
- ROOT
- Visualization
- ZOOM
- Minimization

Simulation
- Advanced Accelerator Simulation
- Geant4
- MCFast
- Monte Carlo Generators
- Patriot
- TeV4LHC
- Physics Projects
- StrHeP

Collaboration Tools
- CRL
- DocDB
- HEPBook

Electronics Systems Engineering
- ESE Hardware Projects
- DAQ R&D
- Pixel R&D
- Trigger R&D

Online
- DRIOD
- FISION
- Merlin
- ODE
- VxWorks

Accelerator Support
- Beam Document Database
Combined e-Science Data Analysis Resource for high-energy physics

The CEDAR project is constructing an e-Science resource for particle physics, combining the strengths of the well-established and widely used HepData database of high-energy physics data and the innovative JetWeb data comparison facility, and exploiting developing computing grid technology (Read more about CEDAR...)

News

First public CEDAR code release

posted 2005-07-26

We've just re-tagged the CEDAR code in selected CVS repositories for public release. This is CEDAR's first official product release, so get the champagne ready :-). The projects involved in this release are:

- LHAPDF version 4.1
- HZTool version 4.0
- HZSteer version 1.1
- KtJet version 1.5

You can download these new versions from the CEDAR downloads page. Have fun!
Some Repository Goals

**Sharing:** don’t reinvent the wheel, **improve** it

**Longevity:** organization as sponsor

Fermilab is interested

**Web Interface** needed
for upload, search, retrieval

**Hierarchy of Purposes:**

- **Archive** for software associated with papers
- **Packages**
- **Component library**
  - Web or Python interfaces to library
Level I: Archive Repository

compare: arxiv.org

Vision: Open Archive
Statistical calculations in refereed physics, or stat conferences
If you publish, put the code here and reference it

“guarantee” intentionally weak:
  once, it compiled and ran on some machine
  include data you used in paper

Possibly reuse of code with credit to authors
  Author   Title   Description   Experiment(?)
  keywords?? Later…

Your grad student can start here, compare methods…

Access goals:
  Documentation encouraged (not quite required?)
  Search/browse mechanisms
  Comments/ratings?

Auto versioning support by system
Candidates:
Small Statistical Packages

Calculation of significance, limits
Goodness of Fit Tests

Competing procedures exist
Some published (NIM; PHYSTAT)
Programs are very hard to find
know the Author; or at best, collaboration CVS
Only a few web interfaces (D0, Babar for example)

Programs not collected by Particle Data Group
publishes generally-recognized methods review
Level II: Package Repository
Re-use intended!

• Access:
  – Source code download
  – Install/run packages (standalone? In a framework?)

• Needs:
  – Conventions for upload/build: **design**
  – quality assessment/maintenance mechanisms
  – Documentation, version control
  – Multiple Platforms? starts to be an issue

• Highly Desirable:
  – Simple specifications to add a package to a framework: **design**
Candidates:  
**HEP Medium-sized Packages**

MC packages (Fortran or C++ typically)  
Jetset NN code part of Pythia  
TerraFerma (Towers)  
StatPatternRecognition (Narsky)  
Talk Wed pm; poster

Author-maintained; **hard to find**  
Not integrated with big packages/frameworks  
Not in central repository  
Not even a central link page to point to them
Level III: Executable Repository

Access Goals:
- Source code download
- Linkable library
- Python wrapper?
- Make it more straightforward to write “toy MC”
- Web execution interface?

Needs:
- Coherence criteria, scope/design
- Identification, supply of missing components
- Quality assessment/maintenance mechanisms
- Documentation, version control
This is the home of the software contributed to our budding repository.

If you use the software or analysis techniques from this page in your work, please give acknowledgement to the author of the particular item you used.

## 1 Bayesian Efficiencies and their Errors

The commonly used methods for the calculation of the statistical uncertainties in cut efficiencies ("Poisson" and "binomial" errors) are both defective in extreme cases. A method for the calculation of uncertainties based upon Bayes' Theorem is presented; this method has no problem with extreme cases. A program for the calculation of such uncertainties is also available. The method is explained in FERMILAB-TM-2286-CD, and the software is available here.

- [FERMILAB-TM-2286-CD.pdf](#)
- [effic2.tgz](#)
Discussing Now
with Fermilab Computing Division

Let us know if you think it’s worthwhile
  linnemann at pa.msu.edu ; mf at fnal.gov
Talk to me or Marc Paterno (poster Wednesday)
Marc Paterno here to assess interest level

  Workshop consensus statement:
  at poster; on stat resources page

Discuss with your collaboration and inform us
  email of support would be helpful

Thinking:
  Manpower, security, copyright/license (any repository)
The Reproducible Research Ideal
www.stat.washington.edu/jaw/jaw.research.reproducible.html

Save enough to re-produce the plot & tables in the paper. Really. Tomorrow or next year. A discipline!
“more benefits to authors than readers”
Assumes: stable environment, powerful script language
tar archive (data + script + paper text)
A natural thing to put in a statistical archive
R or Root scripting helps, but not enough
Design: Directory conventions, makefiles, ...
e.g. Arxiv.org you WILL be able to make pdf of paper
Achievable for most plots shown at this conference, and for most significance or limit calculations in our papers
problematic for large HEP data sets, long analyses…

Thanks: Tom Loredo
Summary

R has intrinsic attractions, and is a window to the statistical community

I would be delighted if everything in R appeared in Root, my everyday environment

I’ve started a link page: help me improve it

We are trying to start a HEP repository: encourage us if you think it’s a good idea