Directions in Statistical Computing 2007

statistical software development, maintenance, and publishing

Outline

- Introduction
- Question & my Provocative Answer 😳
- Business Model
- R Framework (screenshots)
- Future work: Development of Compendium Publishing System (concept, screenshots)

Question

- Does it matter whether statistical software is created within a commercial or an academic environment?
- Yes it does...

On the other hand ...

- there is no reason why we should treat academic software any differently from commercial software.
- both types should:
 - comply to the same quality standards;
 - be easy to use;
 - be cost-effectively produced, maintained, and implemented;
 - be distributed using sound marketing principles;
 - offer a real, measurable "added value" resulting in positive returns for the consumer and the producer.

- Academic software should be managed as if it were a commercial product
- The observation that most academic software projects are free of charge, does not imply that the process of academic software development (incl. maintenance and publishing) is not bound to the basic laws of economics applying to both costs and benefits.
- The costs are mainly associated with production, implementation, and dissemination. The benefits of dissemination on the other hand, are not virtual but real and measurable: e.g. in terms of increased funding for future research and development.

This paradigm implies that academic software...

- should be "leading" (making sure that commercial software is improved and adopts new methods)
- development should make "business sense"
- production, maintenance, and publishing should be "creditable"

(Very Simple) Business Model

for statistical software development, maintenance, and distribution





Problems – part 1

- Do we want Papers to be part of the business cycle?
 - It is hard to compete with the publishing industry
 - Open-access publishing on the internet is 'disruptive'*
 - Cartel strategy employed by commercial publishers
- Cost of creating a paper* (published in an international peer-reviewed journal)
- Cost of dissemination of a paper (Ginsparg Index*)
- Cost per unit of impact (expected citation rates* SCI, SSCI, ...)

Example (closed-access world)

- Cost of producing paper: 15 000 €
- Dissemination cost: 15 000 €
- Expected citation rate (mathematics, 2004) = 0.06

 Total cost per unit of impact = (30 000 € / 0.06) = 500 000 €

Problems – part 2

- Agency problems
 - Individuals need positive incentives to do what is best for the community
 - Make software development & maintenance count
- Meta-data is required and must be maintained
 - technical, usage, educational, ...

R Framework

Building interfaces from meta-data in wessa.net

Example

 Let's create a Bivariate Kernel Density module (based on an R package) that makes Business sense.

Input of Data & Parameters

Specifica	ation of Data Input
Allow users to specify a subset of the dataset (from - to)	Ves Yes
Allow users to specify the width and height of charts	V Yes
Allow users to specify the minimum and maximum value of the y-axis	C Yes
Default Chart Title	Bivariate Kernel Density
Default X-axis Title	x
Default Y-axis Title	у
Number of datasets	2 datasets
Default dataset 1 (x)	88.6 71.6 93.3 84.3 80.6 75.2 69.7 82 69.4 83.3
Default dataset 2 (y)	20 16 19.8 18.4 17.1 15.5 14.7 17.1 15.4 16.2 15
Default dataset 3 (z)	
Number of parameters	7 parameters 💌
Default value (par1) or comma-separated list	50
Label of par1	xgridsize
ToolTip of par1	
Default value (par2) or comma-separated list	50
Label of par2	ygridsize
ToolTip of par2	
Default value (par3) or comma-separated list	0
Label of par3	xbandwidth (zero to use default)
ToolTip of par3	
Default value (par4) or comma-separated list	0
Label of par4	ybandwidth (zero to use default)
ToolTip of par4	
Default value (par5) or comma-separated list	0
Label of par5	correlation >(zero to use actual correlation)
ToolTip of par5	
Default value (par6) or comma-separated list	Y,N
Label of par6	display contours (Y/N)

R code with input validation



R code produces html



Descention (Tout onlu)*

Meta-tags, links, citations

Specification of Output Page					
Allow users to specify the type of output (html, Word, Excel,)	Ves Yes				
Name of page	bidensity				
Main Title	Bivariate Kernel Density Estimation				
Caption	Bivariate Kernel Density Estimation - Free Statisti				
This free online software (calculator) computes the Bivariate Kernel Density Estimates as proposed by Aykroyd et al (2002). The resolution of the image that generated is determined by xgridsize an ygridsize (the maximum value is 500 for axes). You may opt to have the contour					
Keywords	statistics software, free online calculator, kernel density, contour plot, regression, correlation, bivariate kernel density				
Menu Label This field is inactive because the software has been published on January 12 2006 16:27:26.	Bivariate Density				
Menu Description This field is inactive because the software has been published on January 12 2006 16:27:26.	computes Bivariate Kernel Density Estimates				
Citation of statistical methodology or applied work - 1.	Lucy, D. Aykroyd, R.G. & Pollard, A.M.(2002) Non-parametric calibration for age estimation. Applied Statistics 51(2): 183-196				

R Version, Category, Author

R version	R-2.4.1		
Category Descriptive Statistics - Ungrouped - Bivariate			
Command Build User Interface			
Send output to Browser Black/White 💌			
Source (publication) Lucy, D. Aykroyd, R.G. & Pollard, A.M.(2002) Nor			
Author	author's name goes here		
Submitted by*			
E-mail*	myemail@myhost.tld		
- 2zp j3p	2zpj3p		
Subm	nit Program		

Editor's screen

Recent Submissions waiting for approval

Author Home Page | Upload .RData | Request Library

	Name	Modified	Published	Action	
	3dplot.wasp	September 14 2006 03:10:00.	not published	view edit publish	
	JSE1_figure1.wasp	January 09 2006 03:27:00.	not published	view edit publish	
	JSE1_figure2.wasp	January 09 2006 02:46:31.	not published	view edit publish	
	NewRVersion.wasp	January 08 2007 06:39:17.	not published	view edit publish	
	QRW1.wasp	January 06 2006 10:00:50.	not published	view edit publish	
	QRW2.wasp	January 09 2006 05:07:30.	not published	view edit publish	
	RLibraries wasp	January 04 2006 09:01:42.	not published	view edit publish	
	SimTest1.wasp	January 15 2006 14:15:15.	not published	view edit publish	
	SimTest2.wasp	January 14 2006 17:16:18.	not published	view edit publish	
	SimTest3.wasp	January 14 2006 15:05:04.	not published	view edit publish	
	backtobackhist.wasp	January 08 2007 06:42:48.	October 16 2006 15:17:04.	view edit publish	
	bidensity.wasp	January 08 2007 06:47:37.	January 12 2006 16:27:26.	view edit publish	
	bootstrap.wasp	January 14 2006 17:30:38.	not published	view edit publish	
	bootstrapplot.wasp	January 08 2007 07:10:33.	October 10 2006 06:49:24.	view edit publish	I ne tramework creates
	bootstrapplot1.wasp	January 08 2007 07:12:15.	October 10 2006 06:49:18.	view edit publish	
	boxcoxlin.wasp	January 08 2007 07:16:04.	October 08 2006 14:15:02.	view edit publish	the Drandule when
	boxeexnorm.wasp	January 10 2007 03:38:21.	October 08 2006 16:12:00.	view edit publish	the Rmodule when
	cdff.wasp	January 02 2006 04:53:02.	not published	view edit publish	•••••
	cdfnorm.wasp	October 06 2006 10:34:14.	September 06 2006 15:44:33.	view edit publish	the aditor alight publich
	cdft.wasp	January 02 2006 04:54:30.	not published	view edit publish	
	cloud.wasp	January 02 2007 16:52:04.	October 05 2006 08:16:11.	view edit publish	
	density.wasp	October 06 2006 10:43:27.	January 08 2006 17:04:32.	view edit publish	
dum.wasp	November	01 2006 07:38:28	. not published		view edit publish
edauni.wasp	October 06	2006 10:45:59.	October 03 200	6 16:06:28.	view edit publish
	fitdistrexp wasp	October 06 2006 16:21:35.	October 07 2006 09:01:35.	view edit publish	
	fitdistrgamma.wasp	October 06 2006 13:30:33.	October 07 2006 09:01:48.	view edit publish	
	fitdistrinorm.wasp	October 06 2006 16:01:40.	October 07 2006 09:01:59.	view edit publish	
	fitdistrlogistic.wasp	October 06 2006 17:12:30.	October 07 2006 09:02:11.	view edit publish	
	fitdistmeabin.waso	October 06 2006 17:40:21.	October 07 2006 09:02:23.	view edit publish	
	fitdistmorm.wasp	October 25 2006 00:47:48.	September 06 2006 14:41:18.	view edit publish	
	fitdistrpoisson.wasp	October 06 2006 16:30:50.	October 07 2006 09:02:36.	view edit publish	
	fitdistrweibull.wasp	October 06 2006 10:21:03.	October 07 2006 09:02:48.	view edit publish	
	harrell_davies.wasp	October 22 2006 16:30:12.	January 21 2006 10:14:14.	view edit publish	
	hypothesismean1/wasp	November 16 2006 15:05:01.	September 28 2006 06:27:29.	view edit publish	
	hypothesismean2.wasp	October 06 2006 11:17:00.	September 28 2006 07:41:14.	view edit publish	
	hypothesismean3.wasp	October 06 2006 11:19:59.	September 28 2006 14:06:21.	view edit publish	
	hypothesismean4.wasp	October 06 2006 11:21:09.	September 29 2006 08:46:42.	view edit publish	
	hypothesismean5.wasp	November 02 2006 05:16:44.	November 02 2006 17:17:54.	view edit publish	
	hypothesismean6.wasp	November 02 2006 06:00:40.	November 02 2006 17:18:03.	view edit publish	
	hanaliha di sena ana masa	November 14 2006 14/26/04	November 14 2006 14:31:43	view Louist Louistics	

Homepage

		:: Free Statistics Software (Calculator) - Web-enabled scientific services & applications ::	[Select module] Load Module
	The non-commercia	I (academic) use of this software is free of charge. The only thing that is asked in return is to cite this software when results are used in publications.	Home Page Equation Plotter Time Socies Analys
		All rights reserved. Academic Scense for non-commercial use only.	Financial Database Multiple Regression Descriptive Statist
		Wessalnet offers these software applications free of charge:	Statistical Distribut
	Mathematical Equation Plotter	Plots mathematical equations, statistical distributions, derived functions, cumulative functions, and apply numerical integration. Equations can be formulated in algebraic form, for instance: y=exp(a*sin(x)+b). The equation is sent through an interpreter and the analysis and equation chart is generated (in .png format).	Academic citations Latest News Archive (old version FAQ
	Scientific Forecasting	Performs Univariate Box-Jenkins ARIMA modeling, forecasting, and various bootstrap simulation methods for the estimation of financial profit density functions according to the following strategies: BuybHold, Alexander's Fitterrule,	About Wessa.net Black&White
	sonsware	Tructors, Special and NHCL), window types or time series analysis tooli and available: provide the series of the Function, Special Analysis, Variance Reduction Matrix, Standard Devidion-Nean Plot, timmed Skewness & Kurtosis, supported rootsgram displays, percentiles, concentration, histograms, foreward & backward running autocorrelation, ARMA parameter estimation, and much more.	Blue Theme Normal Fontsize Increase Pontsize Decrease Fontsize Toggle underline
	Time Series Database	Use this Distabase to gain access to thousands of time series about financial markets. Note: this application is also evailable from within our Scientific Porecenting Software.	Server status page History list Illustrated history
tive cs	Nultiple Regression Software	Performs Multiple Regression Equation Nodeling with the following features: Ordinary Least Siguares Estimation, Historoxiediaticity batts, Autocorrelation batts, Musepechcabon feats, Nulticollinearity batts, Regression Charts, ANOUA tables, Goodness of Fit, and much more. Several types of regression will be available in future.	
	Descriptive Statistics Software	The Pres Statistics Calculator offers a vide range of descriptive and explorative types of statistical measures and analysis: Central Tendency, Average, Nean, Nedian, Variability, Interquartile Range, Concentration, Lorenz Curve, Gini Coefficient, Stewness, Kurbais, Quartiles, Percarbiles, Notched Bocpick, Histogram, Correlation, Partial Correlation, Rank Correlation (Spearman and Kendell), Simple Repression, Kernel Densky Estimation, Harnell-Davis Quantiles, Bivanide KDG, Correlation Matriz, Stem-and-leaf plot, Explorative Data Analysis	
	Statistical Distributions	Peatures Random Number Generators, PPCC Plots (ind. Tukes lambda), and Statistical Distribution Pitting Modules (Maximum Likelihood) for a series of important distributions: Beta, Inverted Beta, Cauchy 1, Cauchy (2 parameters), Chi, Chi Square (1 parameter), Chi Square (2 parameters), Erlang, Exponential, Fisher F, Garrma, Inverted Garrma, Gumbel, Laplace, Logistic, Lognormal, Normal, Pareto, Power, Rayleigh, «Olderbution, Restangular (Uniform), Student t, Triangular, and Weibull.	
	Statistical Hypothesis Testing Software	Offers statistical testing of a variety of hypotheses: Population Mean, Mean (critical value, p-value, type II error, sample size), Skewness/Kurtosis, Quasi Random-Walk	
	R modules	Registered authors can create, maintain, and publish 8-based modules and make them freely available within the Warrs net foremenced	

Load Module Home Page Equation Plotter Time Series Analysis Financial Databases Multiple Regression **Descriptive Statistics** Statistical Distributions Hypothesis Testing

~

[Select module]

5

Descriptive Statistics page

	::Free Statistics and Forecasting Software::	
	Etatistics Bertware (Calculator) - Web-enabled scientific services & applications ::	Seat Totue Loop Morule
The non-commercial (acad. Here you find a collect University, Bivers	anny use of this software is five of charge, the only thing that is usked in return is to ske this software when results are used in publications. Ion of Free Descriptive Statistics Software wondules (Calculators). The modules have been grouped in the and Tevanute Categories. All modules can be used with any dataset that contains ongrouped abservations.	Home Page Equation Biother Line Series Analysis Financial Databases Multure Regression Description Statistic Statistical Distribute Biothesis Fastion
Num Henu	recom to Main Neno	Academic citations
	Univariate Descriptive Statistics – Univoluted Data	Latest score
		Orchive (old version
Central Tendenry	arithments mean, geometric mean, harmonic mean, median, midirange, midmean, mbietness of cantral tendency (winstrived and trimmed mean), etc	Anchive joid version FAU About Wessenaet
Central Tandanry Variability	anthemetic mean, geometric mean, bermonic mean, median, multienge, mutmean, industriess of control tendency (winsuffeed and termined mean), etc range, senance, containd neuroticit, variation, NST, absolute deviation, interquentle ritterance, coefficient of quartic variation, Criti's mean difference, Leik's D, dispersion, diversity, qualitative variation, mean square neurotics, str	Archive (old version FAU About Wessenet Black@Winto Dive Theme Normal Fonteazo
Central Tendenry Variability Concentration	antimetric mean, geometric mean, harmonic mean, merian, merian, antimetric, mirmean, industriess of rentral Unions (Amazined and Kimmed mean), etc augusta variation, chin's mean difference, tesk's D, dispersion, diversity, qualitative variation, mean square destribution, etc.	Archive Join Versian FAU About Wesselowt Black&Winte Dine Theme Numial Punisize Increase Confeire
Central Tantanry Variability Concentration toments	antimeno mean, geometro: mean, harmonic mean, marian, multange, mutrange, mutranan, industriess of rentral Unders: (Amarica) and brimmed mean), etc range, vanance, dendard deviation, NST, absolute deviation, intergravitie difference, coefficient of quartit, variation, Ghi's mean difference, task's 0, dispersion, diversity, qualitative variation, mean square newton, etc concrust, economical index, Herfindahi, variation coefficient, Unit coefficient, Lorenc carve, etc general, particepted 0, carefored moments, immored moments	Archive Init Versian PAU Abaut Wesseanet Black&Winte Dim Theme Formass Fondate Decrease Pontate Tongla underline
Central Tandancy Variability Concentration homents Skewness/Kurtosis	Anthmenis mean, geometris mean, harmonis mean, median, multisona, mutrisona, mutrisona, robustness of rentral Unitoris (Antoniced and Veinmed mean), Vec Panja, Venanca, Bandard deviation, NST, absolute deviation, intergravitie difference, coefficient of Quartic verificity. Only mean difference, task's (b) dispression, diversity, qualitative verificition, mean square deviation, etc encross, economical index, Herfindahl, verificitien coefficient, Unit coefficient, Lorence carve, etc deviative, pan expressed & conserved moments, formand moments Histor and economical paths, German and Sample downeess. Prior diversity (occording to 6 different quarties), and harmonic, Histor beta 1:8 gamma 1, Pearson, Yald's skewness (occording to 6 different quarties 2, and harmonic katuras, curval, cample downeess, Driver diversity, Deter beta 2:6 gamma 2, and harmonic katuras.	nichtos foid vordan FAQ Abaut Wassa.ant Black@Wissa.ant Black@Wissa.ant Black@Wissa.ant Ding Varial Paularce Formass Fontstor Decrease Paularce Toggia underline Sarvar status page History list History list

Bivariate Descriptive Statistics - Ungrouped Data

Correlation	Pearson correlation, covariance, determination coefficient, scatter plot, etc			
Rank Correlation	Spearman Rank Order Correlation (corrected and non-corrected).			
Simple Regression	general linear model, mean and variances, covariance, correlation, least squares estimation, parameters, response, significance, determination coefficient, ANOVA, residuals, autocorrelation, model selection, mode			
Bivariate Density	computes Bivariate Kernel Density Estimates			
кениан канк соггенации	computes the Kenuali tau Kank Correlation between two data series			
Box-Cox Linearity Plot	computes the Box-Cox Linearity Plot			
Linear Regression Graphical Model Validation	computes the Simple Linear Regression model (Y = a + b X) and various diagnostic tools from the perspective of Explorative Data Analysis			
Back to Back Histogram	computes the Back to Back Histogram (sometimes called Bihistogram) for a bivariate dataset			

Bivariate KDE (R Module)



Handling Requests

- Find R module (Google)
- User submits request (html form, HTTP POST)
- Webserver loads R module
- R module creates pre-processed R code
- Webserver directs request to R server (callback)
- R server invokes R engine, stores output
- Webserver gets output parses through template
- Webserver sends html reply to user
- User fetches pictures/logfiles from R server

What do people say?

- "Fantastic, this is a great resource"
- "Very useful."
- "Lifesaver!"
- "I think that you have done an excellent work"
- "I think that having the R source code is a good idea."
- "Very simple in use, useful and nice-looking! Keep developin it!"
- "I think that a lot of help us to own job."
- "I think that...it is great. We (my 10 year old dtr. and I) used it for her science project--even she had an easy time with it. Thanks so much"
- many more....

What else do they say?

• Many useful suggestions:

– "There is a problem to open bitmap .png files".

- "I think that...it can be improved"
- "close it down" ⊗

R Modules ...

- are (almost) "user friendly"
 - fontsize, color, underline
 - simple submission form
 - output in html, Word, Excel
 - text-only input (do we need to import data files?)
- are "safe to use" unlike DIE software
- are compatible with any http-enabled user agent
- "firewall friendly"
- are used by many [statistics available, REPEC]

R Modules ...

- are "indexable", hence "findable"
 - hierarchy of hyperlinks
 - meta tags, titles, descriptions
 - archive of old versions
 - fast loading, pure html
- make "Business/Marketing Sense"

<u> </u>		Web	Images	Video	News	Maps	more »	
Googl	e	harrell	davis sof	tware			Search	Advanced Search Preferences

Web

Wessa.net - Free Statistics and Forecasting Software (Calculators ...

Wessa.net offers free educational forecasting software (time series analysis) and statistics software.

www.wessa.net/v1_1_20/stat.wasp - 38k - Supplemental Result -

Cached - Similar pages - Note this

Harrell-Davis Quantiles - Free Statistics and Forecasting Software ...

This free online software (calculator) computes the Harrell-Davis Quantiles and associated standard errors.

www.wessa.net/rwasp_harrell_davies.wasp - 34k - <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u> [<u>More results from www.wessa.net</u>]

EconPapers: HDQUANTILE: Stata module for Harrell-Davis estimator ...

HDQUANTILE: Stata module for **Harrell-Davis** estimator of quantiles. Nicholas Cox (). Statistical **Software** Components from Boston College Department of ... econpapers.repec.org/**software**/bocbocode/s449601.htm - 9k -Cached - Similar pages - Note this

EconPapers: Statistical Software Components

Statistical **Software** Components, from Boston College Department of Economics ... HDQUANTILE: Stata module for **Harrell-Davis** estimator of quantiles Downloads ... econpapers.repec.org/**software**/bocbocode/default7.htm - 22k -<u>Cached</u> - <u>Similar pages</u> - <u>Note this</u>

HDQUANTILE: Stata module for Harrell-Davis estimator of quantiles

HDQUANTILE: Stata module for **Harrell-Davis** estimator of quantiles ... **Software** component provided by Boston College Department of Economics in its series ... ideas.repec.org/c/boc/bocode/s449601.html - 8k - <u>Cached</u> - <u>Similar pages</u> - <u>Note this</u>

IDEAS: Statistical Software Components, Boston College Department ...

Statistical Software Components. Contact information of Boston College Department of ... HDQUANTILE: Stata module for Harrell-Davis estimator of quantiles ... ideas.repec.org/s/boc/bocode.html - 45k - Cached - Similar pages - Note this

R Modules have "impact"



R Modules ...



R Modules ...

- feature session management:
 - "remembers" datasets
 - maintains "history lists"

List of last 4 computations	Delete history		
Server Date	Module	Command	
Thu, 08 Feb 2007 08:53:45 -0600	Univariate Explorative Data Analysis	Print Word Excel Blog this Delete	
Thu, 08 Feb 2007 08:01:04 -0600	Random Number Generator - Log-Normal Distribution	Print Word Excel Blog this Delete	
Thu, 08 Feb 2007 07:55:43 -0600	Random Number Generator - Normal Distribution	Print Word Excel Blog this Delete	
Thu, 08 Feb 2007 07:43:05 -0600	Start of session		



Archive (old versions) FAQ About Wessa.net

Black&White Blue Theme Normal Fontsize Increase Fontsize Decrease Fontsize Toggle underline

Server status page History list Illustrated history list

In Future, R Modules will...

- allow your work to be archived*
- work like a blog (can be referenced in published articles)
- feature discussion
- collect additional meta-data
 - about usage
 - discussion
 - 'taxonomy'
- allow you to reproduce your computations (compendium)
- allow you to reuse the code and meta data to create derived works

Compendium

- An electronic <u>Article</u> that includes the Data and Software to reproduce the results?
- An electronically archived <u>Computation</u> that includes the Data, Meta-data, and Software to reproduce, and reuse it? Meta-data contains discussions, and information about usability & findability. It makes the compendium a tool of collaboration and dissemination.

Could we make computations reproducible for anyone?



Yes, we can...

This free online software (calculator) reproduces Example 6.5 (Prediction, Filtering, and Smoothing for the Local Level Model) in Shumway R.H., Stoffer D.S., Time Series Analysis and Its Applications (2nd ed. with R examples). The module includes all necessary functions to reproduce their results (Kfilter.R, Ksmooth.R, and ex65.txt). In addition, some hard-coded numbers have been replaced by user-defined parameters.



and when we Blog the computation...



...and submit meta information...



...we get a "citable computation"





http://moodle.org/mod/forum/discuss.php?d=44830

New types of MDB-driven applications

- Automatic creation of reference manuals (with cross-linking)
- On the fly Casebook creation
- Feedback administration system
 - Usability (GUI)
 - Bug reporting
 - Citation tracking
 - Refereeing
- MCMC Queuing (performance statistics)
- Collaborative Research & Learning (intelligent caching)
- Compendium Publishing

Future development

- Queue Management for MCMC
- Special apps require special UI:
 - Rich (Rpad-like) GUIs for graphics
 - Zelig
- Data import (SPSS, SAS, ...)
- Upload of large databases (=> dataframes always available)
- Sweave integration
- RSS/Atom support (keyword dependent)