DAta Mining & Exploration Program

Statistical Calculations

User Manual

DAME-MAN-NA-0022

Issue: 1.0
Date: September 12, 2013
Author: M. Brescia, S. Cavuoti

Doc. : RegStatistics_UserManual_DAME-MAN-NA-0022-Rel1.0

Index

TABLE INDEX

FIGURE INDEX

1 Introduction

The present document is the user guide of the general statistics calculation service, as implemented and integrated into the DAMEWARE web application. It is a tool that can be used to execute statistical evaluation between two data vectors, useful for both regression post-processing and stand-alone statistics on a data table. The user data table could be formatted in one of the supported file types: ASCII (columns separated by spaces), CSV (comma separated values), FITS-Table (numerical columns embedded into the fits file) or VOTable.
This manual is one of the specific guides (one for each data mining model or services available in the webapp) having the main scope to help user to understand theoretical aspects of the tool, to make decisions about its practical use in problem solving cases and to use it to perform experiments through the webapp, by also being able to select the right functionality associated to the model, based upon the specific problem and related data to be explored, to select the use cases, to configure internal parameters, to launch experiments and to evaluate results.

The documentation package consists also of a general reference manual on the webapp (useful also to understand what we intend for association between functionality and data mining model) and a GUI user guide, providing detailed description on how to use all GUI features and options.
So far, we strongly suggest to read these two manuals and to take a little bit of practical experience with the webapp interface before to explore specific model features, by reading this and the other model guides.
All the cited documentation package is available from the address
http://dame.dsf.unina.it/dameware.html **, where there is also the direct gateway to the webapp.**

2 Statistical Calculations Theoretical Overview

The statistical indicators available through the web service are based on the following formulas.

$$\Delta z = (col1 - col2) \quad (1)$$

$$\Delta z_n = \frac{(col1-col2)}{(1+col1)} \quad (2)$$

$$bias = \frac{\sum_{i=1}^{N}(\Delta z_i)}{N} \quad (3)$$

$$bias_n = \frac{\sum_{i=1}^{N}(\Delta z_{n_i})}{N} \quad (4)$$

$$RMS = \sqrt{\frac{\sum_{i=1}^{N}(\Delta z_i)^2}{N}} \quad (5)$$

$$RMS_n = \sqrt{\frac{\sum_{i=1}^{N}(\Delta z_{n_i})^2}{N}} \quad (6)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z_i - \left(\frac{\sum_{i=1}^{N}\Delta z_i}{N}\right)\right]^2}{N}} \quad (7)$$

$$\sigma_n = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z_{n_i} - \left(\frac{\sum_{i=1}^{N}\Delta z_{n_i}}{N}\right)\right]^2}{N}} \quad (8)$$

$$MAD = Median(|\Delta z|) \quad (9)$$

$$MAD_n = Median(|\Delta z_n|) \quad (10)$$

$$NMAD = 1.4826 * Median(|\Delta z|) \quad (11)$$

$$\mathrm{N}MAD_n = 1.4826 * Median(|\Delta z_n|) \quad (12)$$

From a theoretical point of view there is a known relation between RMS (5) and standard deviation (7):

Given the RMS in (5), we have RMS $= \sqrt{mean^2 + \sigma^2}$. By knowing that σ^2 is the variance, we have that $RMS = \sqrt{mean^2 + variance}$.
Although the RMS and standard deviation are in principle different, sometimes they differ very little when the errors are sufficiently small but a little bit higher in some error bins.
MAD is less affected by outliers than standard deviation. NMAD is the normalized MAD.

For a direct comparison of results, present in literature, in terms of distance of mσ (with m = 1,2…) from the distribution of Δz, it is much more precise to use the standard deviation (7) as main indicator, rather than the simple RMS.

For the outliers evaluation, our point of view is that the statistical estimation would be always provided at all different multiples of the standard deviation (at least from 1σ to 4σ). This in order to give the possibility to evaluate the trend of the prediction scattering and to proceed with a deeper analysis of objects resulting as outliers at different degrees of scattering.

There is often a confusion about the relation between photometric and spectroscopic used to apply the statistical indicators. For instance, the performance could be very different if the simple Δz (col1-col2) is used instead of the normalized Δz_n (Δz/(1+col1)). Our concern is that the Δz cannot represent the best choice in the specific case of statistical evaluation. As known any not uniform distribution in a wide spread sample, and the related statistics is not able to give a consistent estimation at all ranges of bins. On the contrary, the normalized term Δz_n introduces a more uniform information, correlating in a more correct way the variation of the samples, and permitting a more consistent statistical evaluation at all ranges of the distribution.

More in detail, concerning the difference between terms Δz and Δz_n, from the mathematical point of view, the following considerations may be useful to understand the physical properties:

It is known that $z = {\Delta\lambda}/{\lambda} = \frac{\lambda obs - \lambda real}{\lambda real} = \frac{\lambda obs}{\lambda real} - 1$ hence:

$$1 + z = \frac{\lambda obs}{\lambda real} \qquad (9)$$

So far, by considering $z = \frac{\lambda obs - \lambda real}{\lambda real}$ and taking the (9) into account:

$$dz = d\left(\frac{\lambda obs - \lambda real}{\lambda real}\right) = \frac{d\lambda obs}{\lambda real} = \frac{d\lambda obs}{\lambda real}\frac{\lambda obs}{\lambda obs} = \frac{d\lambda obs}{\lambda obs}\frac{\lambda obs}{\lambda real} = \frac{d\lambda obs}{\lambda obs}(1 + z) \qquad (10)$$

From (10) it is easy to obtain:

$$\frac{dz}{(1+z)} = \frac{d\lambda obs}{\lambda obs} \qquad (11)$$

And the term $\frac{d\lambda obs}{\lambda obs}$ of equation (11) is exactly the variation between observed and real events, which is the main focus of any empirical estimation, especially for empirical data mining models which learn its prediction based on the observed information. And this result is invariant to the sample range considered. We conclude that the term $\frac{dz}{(1+z)}$ is the best choice on which to apply the statistical operators.

3 Use of the web statistical tool

The statistical tool needs information about the input data table, provided by the user (by uploading the input file within the web app), and about two columns of the input table to be used to evaluate their statistical correlation. The user can arbitrarily choose two columns among the available columns forming the input data table. Then, by applying calculations, the output will consist of a report file, with all statistical results correlating the two column data.

3.1 Use Cases

For the user the Statistics tool offers one use case only:

- ***Test***

As easy to understand, this use case enable user to execute all available statistical calculations between the two data vectors, selected during the setup phase.

3.2 Input

We also remark that in all DAMEWARE tools it is possible to use one of the following data types:

- ASCII (extension .dat or .txt): simple text file containing rows (patterns) and columns (features) separated by spaces, normally without header;
- CSV (extension .csv): Comma Separated Values files, where columns are separated by commas;
- FITS (extension .fits): fits files containing tables;
- VOTABLE (extension .votable): formatted files containing special fields separated by keywords coming from XML language, with more special keywords defined by VO data standards;

A correct dataset file must contain columns, among which the user can select the two involved in the statistics.

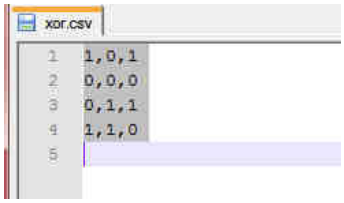


Fig. 1 – The content of a file sample used as input for statistics calculations

3.3 Output

In terms of output, different files are obtained, depending on the specific use case of the experiment.
In the case of **regression** functionality, the following output files are obtained in all use cases:

TEST	DESCRIPTION	REMARKS
Statistics_Test.log	Experiment report log	
Statistics_TEST_output.txt	Output of statistics	
Statistics_Test_params.xml	Experiment configuration	

Tab. 1 – output file list as will appear in the experiment tab of the web application

3.4 TEST Use case

The setup interface of test use case for the statistics tool, is shown in Fig. 7.

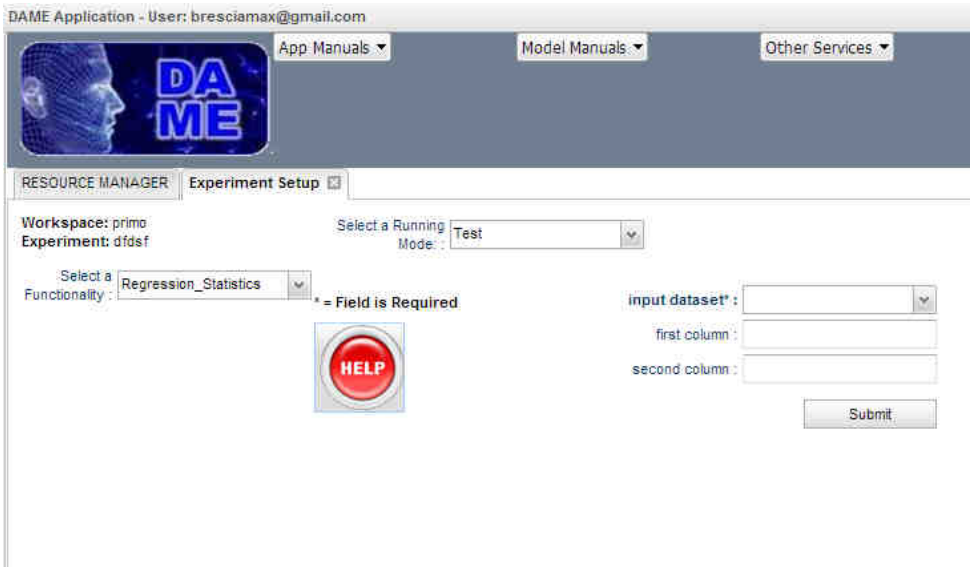


Fig. 2 – The setup tab for statistics test use case

The help page (available from Help button in the setup page) is at the address:
http://dame.dsf.unina.it/statistics.html#test

- **Input dataset**

  this parameter is a field required!

  This is the dataset file to be used as input for the test phase of the tool.

  The file must be already uploaded by user and available in the current workspace of the experiment. It can be the same input dataset already submitted during training phase.

  The format (hence its extension) must be one of the types allowed by the application (ASCII, FITS-table, CSV, VOTABLE).

- **First column**

  It is the first column index of the input data table to be specified as col1 parameter of all statistical formulas. If left empty, the first (from left to right) column will be considered. Be sure to specify an index > 0 and included in the range of existing number of columns within the input table.

- **Second column**

  It is the second column index of the input data table to be specified as col2 parameter of all statistical formulas. If left empty, the second (from left to right) column will be considered. Be sure to specify

a col1 < index < max_number_of_columns, as included in the range of existing number of columns within the input table.

4 Appendix – References and Acronyms

Abbreviations & Acronyms

A & A	Meaning	A & A	Meaning
AI	Artificial Intelligence	KDD	Knowledge Discovery in Databases
ANN	Artificial Neural Network	IEEE	Institute of Electrical and Electronic Engineers
ARFF	Attribute Relation File Format	INAF	Istituto Nazionale di Astrofisica
ASCII	American Standard Code for Information Interchange	JPEG	Joint Photographic Experts Group
BoK	Base of Knowledge	LAR	Layered Application Architecture
BP	Back Propagation	MDS	Massive Data Sets
BLL	Business Logic Layer	MLC	Multi Layer Clustering
CE	Cross Entropy	MLP	Multi Layer Perceptron
CSOM	Clustering SOM	MSE	Mean Square Error
CSV	Comma Separated Values	NN	Neural Network
DAL	Data Access Layer	OAC	Osservatorio Astronomico di Capodimonte
DAME	DAta Mining & Exploration	PC	Personal Computer
DAMEWARE	DAME Web Application REsource	PI	Principal Investigator
DAPL	Data Access & Process Layer	REDB	Registry & Database
DL	Data Layer	RIA	Rich Internet Application
DM	Data Mining	SDSS	Sloan Digital Sky Survey
DMM	Data Mining Model	SL	Service Layer
DMS	Data Mining Suite	SOFM	Self Organizing Feature Map
FITS	Flexible Image Transport System	SOM	Self Organizing Map
FL	Frontend Layer	SW	Software
FW	FrameWork	UI	User Interface
GRID	Global Resource Information Database	URI	Uniform Resource Indicator
GSOM	Gated SOM	VO	Virtual Observatory
GUI	Graphical User Interface	XML	eXtensible Markup Language
HW	Hardware		

Tab. 2 – Abbreviations and acronyms

Reference & Applicable Documents

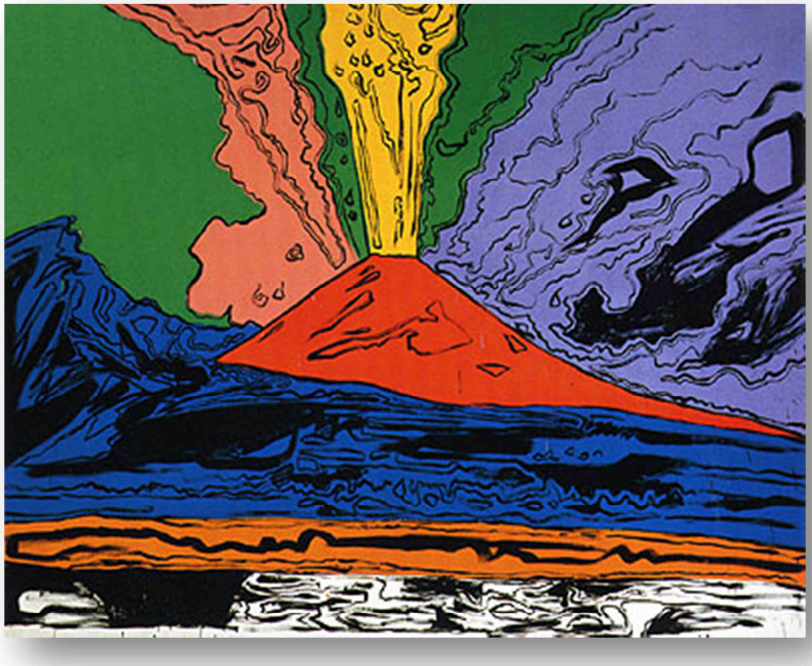
ID	Title / Code	Author	Date
R1	"The Use of Multiple Measurements in Taxonomic Problems", in Annals of Eugenics, 7, p. 179--188	Ronald Fisher	1936
R2	*Neural Networks for Pattern Recognition*. Oxford University Press, GB	Bishop, C. M.	1995
R3	*Neural Computation*	Bishop, C. M., Svensen, M. & Williams, C. K. I.	1998
R4	Data Mining Introductory and Advanced Topics, Prentice-Hall	Dunham, M.	2002
R5	*The Fourth Paradigm*. Microsoft research, Redmond Washington, USA	Hey, T. et al.	2009
R6	Artificial Intelligence, A modern Approach. Second ed. (Prentice Hall)	Russell, S., Norvig, P.	2003
R7	Neural Networks - A comprehensive Foundation, Second Edition, Prentice Hall	Haykin, S.,	1999
R8	*A practical application of simulated annealing to clustering.* Pattern Recognition 25(4): 401-412	Donald E. Brown D.E., Huntley, C. L.:	1991

Tab. 3 – Reference Documents

ID	Title / Code	Author	Date
A1	SuiteDesign_VONEURAL-PDD-NA-0001-Rel2.0	DAME Working Group	15/10/2008
A2	project_plan_VONEURAL-PLA-NA-0001-Rel2.0	Brescia	19/02/2008
A3	statement_of_work_VONEURAL-SOW-NA-0001-Rel1.0	Brescia	30/05/2007
A4	mlpGP_DAME-MAN-NA-0008-Rel2.0	Brescia	04/04/2011
A5	pipeline_test_VONEURAL-PRO-NA-0001-Rel.1.0	D'Abrusco	17/07/2007
A6	scientific_example_VONEURAL-PRO-NA-0002-Rel.1.1	D'Abrusco/Cavuoti	06/10/2007
A7	frontend_VONEURAL-SDD-NA-0004-Rel1.4	Manna	18/03/2009
A8	FW_VONEURAL-SDD-NA-0005-Rel2.0	Fiore	14/04/2010
A9	REDB_VONEURAL-SDD-NA-0006-Rel1.5	Nocella	29/03/2010
A10	driver_VONEURAL-SDD-NA-0007-Rel0.6	d'Angelo	03/06/2009
A11	dm-model_VONEURAL-SDD-NA-0008-Rel2.0	Cavuoti/Di Guido	22/03/2010
A12	ConfusionMatrixLib_VONEURAL-SPE-NA-0001-Rel1.0	Cavuoti	07/07/2007
A13	softmax_entropy_VONEURAL-SPE-NA-0004-Rel1.0	Skordovski	02/10/2007
A14	MLPQNA_DAME-SRS-NA-0009-Rel_1.0	Riccardi, Brescia	09/02/2011
A15	dm_model_VONEURAL-SRS-NA-0005-Rel0.4	Cavuoti	05/01/2009
A16	MLPQNA_DAME-SDD-NA-0015-Rel_1.0	Riccardi, Brescia	02/06/2011
A17	DMPlugins_DAME-TRE-NA-0016-Rel0.3	Di Guido, Brescia	14/04/2010
A18	BetaRelease_ReferenceGuide_DAME-MAN-NA-0009-Rel1.0	Brescia	28/10/2010
A19	BetaRelease_GUI_UserManual_DAME-MAN-NA-0010-Rel1.0	Brescia	03/12/2010

Tab. 4 – Applicable Documents

__oOo__



DAME Program
"we make science discovery happen"