Title: "A New Family Of Generalized Power Transformations And The Educational Ambassador ASA Program"

Kelmansky, Diana
Universidad de Buenos Aires, Instituto de Cálculo
Intendente Guiraldes 2160, Pabellón II, 2do Piso.
Buenos Aires (1428), Argentina
E-mail: dkelman@ic.fcen.uba.ar

Martínez, Elena
Universidad de Buenos Aires, Instituto de Cálculo
Intendente Guiraldes 2160, Pabellón II, 2do Piso.
Buenos Aires (1428), Argentina
E-mail: emartin@ic.fcen.uba.ar

ABSTRACT

The Educational Ambassadorship is an initiative of the ASA that begun in 2005. This report describes the activities accomplished by Diana Kelmansky, the First Educational Ambassador from Argentina (years 2005 to 2010). Several major results of the initiative will be described. The first one: twenty new members were recruited by the ASA. The second one: students from Argentina, Ecuador, Mexico, Chile and Brazil were benefited by the semester and short courses together with lecture notes in Spanish on Microarray data analysis. The third one: a new research line is growing with several master students and senior researchers. Also consulting is being provided to biological researchers. Both consultants and researchers are regarding the course learned at JSM on microarrays data analysis as very valuable.

Finally there will be a description of the ongoing research: "A new family of generalized power transformations", a joint work with Elena Martinez. This family provides a framework for modeling not only microarray data (the subject on which the initial 2005 JSM course was held) but also oncoming massive DNA sequencing data.

1. Introduction

This paper gives a brief description, in chronological order, of the first author activities regarding the Educational Ambassadorship program, held from July 2005 to March 2011. These activities include current research work in collaboration with the second author, the original assignments and several additional ones that made the project even more challenging, exciting and useful.

2. First author activities regarding the Educational Ambassadorship Program

2.1. Minneapolis, USA, JSM 2005 two day intensive course

Attended the JSM 2005 two day (August 6-7) intensive course "Statistical Analysis of Gene Expression Data", given by Terrence P. Speed and collaborators. Study material for the following 6 months was selected.

2.2. Sierras de Córdoba, Argentina. First Argentine School of Mathematics and Biology (BIOMAT) 2005

Invited Conference: "Introduction to experiments and statistical analysis of microarrays". La Cumbre,

Sierras de Córdoba. December 1-10 2005.

2.3. Buenos Aires semester courses "Exploratory and confirmatory analysis of microarray's experiments data" March-July 2006, August -December 2007, March-July 2010

Elective doctoral and master course of the University of Buenos Aires (Master Program in Mathematical Statistics of the FCEN UBA). It took place at the Instituto de Cálculo and included lectures, discussion of current papers and lab sessions using the R environment. Lecture notes and lab guides were prepared and can be downloaded from the website: www.dm.uba.ar/ materias/ analisis_expl_y_conf_de_datos_de_exp_de_marrays_Mae/2006/1/

An additional consequence of this course was the R introductory guide in Spanish that has been recently linked to two pages:

- http://knuth.uca.es/R/doku.php?id=documentacion (R project of Cádiz University)
- http://ocw.um.es/gat/contenidos/ldaniel/ipu_docs/calculo_estadistico/ipu_calculo_estadistico.html (OpenCourseWare of Murcia University)

2.4. Buenos Aires workshop on microarrays for invited biologists. A new learning experience. July 2006

As a final assignment of 2006 "Exploratory and confirmatory analysis of microarray's experiments data" course, the students agreed to participate as teaching assistants of a 4 hours workshop were biologists from various institutions were invited. This enabled the students to practice with the just learned methodology and also to interchange experience with the biologists, some of which worked with microarray technology.

2.5. Seattle, USA, invited lecture "Statistics in Argentina" on JSM 2006

The educational ambassador 2005-2006 experiences were presented at the JSM session organized by the Committee on International Relations in Statistics. A discussion on the current situation of statistics in Argentina was held with the participants.

2.6. Seattle, USA, JSM 2006 one day intensive course

Attended the course "Methods and Computational Tools for the Screening and Classification of Microarray Gene Expression Data", McLachlan G. University of Queensland, Kim-Anh Do, M. D. Anderson Cancer Center. JSM 2006

2.7. Quito, Ecuador, Escuela Politécnica Nacional Master in Applied Statistics intensive course August 2006 "Introduction to the analysis of microarray experiment data"

An intensive course with final exam: "Introduction to the analysis of microarray experiment data" as part of the VII Applied Statistics Seminar, was given in Quito. Labs with R as well as lectures were delivered to an audience of 25 students. Lecture notes, selected papers and lab guides were provided. Ecuador August 23 - 31 2006.

2.8. Madrid, Spain, Lecture "Microarray Experiments: where statistics and molecular biology meet" October 2006

A lecture: "Microarray Experiments: where statistics and molecular biology meet", was presented at the Workshop on Robustness and Statistical Inference in honour of Víctor Yohai. Dr. Yohai was given the title of "Doctor Honoris Causa" at the Universidad Carlos III. Madrid. España. Oct 4-5. 2006.

2.9. Acapulco, Mexico, short course "Statistical Issues with Microarray Processing and Analysis", October 2006

26 students attended the course "Statistical Issues with Microarray Processing and Analysis" jointly given with Terry Speed at the XXI Foro Nacional de Estadística. Continuing Education Program. American Statistical Association. Asociación Mexicana de Estadística. Acapulco, Mexico. October 9 - 10. 2006.

2.10. Rosario, Argentina, mini course "Microarray experiments, from molecular biology to statistics" , October 2006

The mini course "Microarray experiments, from molecular biology to statistics" (Experimentos de microarreglos, desde la biología molecular a la estadística) was given at the joint meeting of SOCHE, SAE, GAB, IASI statistical societies. There were students from Argentina, Chile, Uruguay and Brazil. JIE 2006. Rosario Argentina. Octuber 9-13, 2006.

2.11. La Falda, Córdoba, Argentina, short course. Second Argentine School of Mathematics and Biology . "Microarray experiments data analysis". June - July 2007

The course "Microarray experiments data analysis" (Análisis de datos de experimentos de microarreglos) was given at the Second Argentine School of Mathematics and Biology. It was a one week intensive course where mathematics and biology students from all the country interacted. Córdoba - Argentina. June 28 -July 7 2007.

2.12. National University of Córdoba - Argentina, intensive course "Exploratory and Confirmatory Anaysis of Micoroarrays Experimental Data". August 2007

The course with final exam "Análisis exploratorio y confirmatorio de Datos de Experimentos de Microarrays" was included in the Magister of Applied Statistics program of the Universidad Nacional de Córdoba, as part of the "Applied Regression and ANOVA course" was given in an intensive schedule of 6 hours a day with labs. August 13-17 2007.

2.13. Antofagasta, Chile. October 17-19 2007. XXXIV Jornadas Nacionales de Estadística SOCHE.

- Short course "Microarray experiments, a challenge for statistical analysis. Current methods and new proposals".
- Plenary lecture "Generalized power transformations".

2.14. Mendoza, Argentina, LVIII Reunión Reunión anual de la UMA. July 2008

Invited lecture "Genomic data analysis, a challenge for mathematicians and statisticians". Mendoza, Argentina. 24- 27 Sep 2008.

2.15. La Falda, Córdoba, Argentina, Third Argentine School of Mathematics and Biology (BIOMAT III)

July 23-28, 2008. Invited conference " Models and transformations in microarray experiments data"

2.16. Invited paper ISI 2009 presentation

"Transferring Knowledge in South and Central America -Microarray data analysis-" Kelmansky Diana. 57th Session of the International Statistical Institute Invited paper presentation: The American Statistical Association Educational Ambassador (EA). Durban South Africa 16-22 August 2009.	 Closing the American Statistical Association Educational Ambassador (EA) Session.

2.17. XXXVII Argentine Statistical Association 2009 meeting. Coloquio Argentino de Estadística SAE.

Opening conference: "Epigenetics, genomics, microarrays' revolution and statistics" (Epigenética, genómica, la revolución de los microarreglos y la estadística). Catamarca. Argentina. October 7, 2009.	 D. Kelmansky with the local press

2.18. Franco-Argentine Centre of Science, Information and Systems (Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas) CIFASIS-CONICET. 2010

Invited Conference "Genomics and statistics". Rosario Argentina. 17 de Mayo 2010.

2.19. Terry Speed Conference at FCEN- UBA. Buenos Aires March 10 2010.

The Instituto de Cálculo of University of Buenos Aires was pleased to receive Dr. Terry Speed. for a friendly meeting with the Institute' statistical researchers and also for his conference "Statistical challenges with next-generation DNA sequencing". The conference was held at the Faculty of Exact and Natural Sciences of Buenos Aires University.

2.20. Rosario, Argentina, 40 hours course *"Statistical aspects of Microarray's Data Analysis"* May-June 2010

"Statistical aspects of Microarray's Data Analysis". Elective doctoral course of the National Rosario University. 25 Students passed the final exam.

3. Other results

3.1. Papers and book chapter

- *Microarray Experiments: where statistics and molecular biology meet*. Kelmansky, D. Chapter of Statistical Methods for Microarray Data Analysis Series: Methods in Molecular Biology Yakovlev, Andrei Y.; Klebanov, Lev; Gaile, Daniel (Eds.) ISBN: 978-1-60327-336-7. Humana Press. In preparation. 2011.

- "On the glog-normal distribution and its association with the gene expression problem." Leiva, V., Sanhueza, A. Kelmansky, D., Martinez, E. 2009. *Computational Statistics and Data Analysis*, 53:1613–1621 *ISSN* 0167-9473, http://dx.doi.org/10.1016/j.csda.2008.04.012.
- "Introducción a los Experimentos y Análisis de Datos de Microarreglos" Diana Mabel Kelmansky - Elena Julia Martínez Acta Nº 13 de la Academia Nacional de Ciencias. ISSN-0325-7533. 2007. págs 85-95.
- "Experimentos de Microarreglos: El Debate" Diana Mabel Kelmansky Acta Nº 14 de la Academia Nacional de Ciencias. ISSN-0325-7533. 2008. págs 171-182.

3.2. Advisorship

A graduate thesis on mathematics was successfully completed (October 2008) and 3 master theses are ongoing as a consequence of the program. Also short statistical advisory services on microarray data analysis were provided to biological researchers.

3.3. Increase ASA membership

10 new members were recruited from the Buenos Aires 2006 semester course and 10 more during 2006 Quito Ecuador intensive course.

4. Ongoing research

"A new family of generalized power transformations", joint work with Elena Martinez. This family provides a framework for modeling not only microarray data (the subject on which the initial 2005 JSM course was held) but also oncoming massive DNA sequencing data.

Gene expression microarray intensity data and massive DNA sequencing data show high heteroskedasticity, i..e. higher intensity measures show higher variability. To cope with this drawback, standard microarray data analysis is based upon a log2 transformation of the intensity data (Speed 2003, Smyth et al. 2003), however it can inflate the variance of observations with intensities near the background values (Durbin et al. 2002). Rocke and Durbin (2001), Durbin et al. (2002) and Huber et al. (2002) have simultaneously proposed a variance stabilizing transformation, the generalized logarithm: glog (x) = $\operatorname{arsinh}(x) = \log(x + \sqrt{x^2 + 1})$. This proposal has been criticized (Speed 2003) for its similar behaviour to the log transformation for high intensity values, as there is a body of evidence that this is a too severe transformation in that range. The need of power transformations other than the log transformation has also been emphasized by other authors (Berg et al 2006).

To overcome the above mentioned drawbacks we introduce and study the analytical properties of a generalized power transformation family, gpower(x,p), which has arsinh(x) as one of its members, in the same continuous manner as the natural log belongs to the Box-Cox power transformation:

$$\text{gpower}(x,p) = \begin{cases} \dfrac{(x + \sqrt{x^2 + 1})^p - 1}{p} & \text{if } p \neq 0 \\ \log(x + \sqrt{x^2 + 1}) & \text{if } p = 0 \end{cases}$$

A simple graphical procedure for the estimation of the p parameter is being evaluated for different simulated scenarios in a Monte Carlo study and also for real data.

5. Final Comments

During the activities held as an Educational Ambassador, biologists as well as statisticians, confirmed the difficulties that both face regarding interdisciplinary work. Many of them were thankful for clarifying concepts that are usually taken for granted in either specialty, others pointed out that they faced greater difficulties as most of the references were in English. In this respect the Spanish lecture notes gave them an important tool for their studies. There is much more work left for interdisciplinary activities and teaching regarding molecular biology and specially microarrays.

REFERENCES

Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) "A variance-stabilizing transformation for gene-expression microarray data". *Bioinformatics*, **18** (Suppl. 1), S105–S110.

Gordon K. Smyth , Yee Hwa Yang and Terry Speed. " Statistical Issues in cDNA Microarray Data Analysis". *Methods in Molecular Biology*, **224**, Humana Press, Totowa, NJ, 2003, pages 111-136.

Huber,W., Von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

T. Speed - Editor. "Statistical Analysis of Gene Expression Data". 2003 . Chapman&Hall