Linköping University Postprint

Advanced statistics and data analysis in laboratory medicine: steep learning curve but substantial rewards

Elvar Theodorsson

N.B.: When citing this work, cite the original article.

Original publication:

Elvar Theodorsson, Advanced statistics and data analysis in laboratory medicine: steep learning curve but substantial rewards, 2008, Scandinavian Journal of Clinical and Laboratory Investigation, (68), 6, 434-436.

<http://dx.doi.org/10.1080/00365510701832245>.

Copyright © [Taylor & Francis Group](#), an informa business

Postprint available free at:

Linköping University E-Press: <http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-15178>

Advanced statistics and data analysis in laboratory medicine: steep learning curve but substantial rewards

Elvar Theodorsson
IBK/Clinical Chemistry
University Hospital
SE-581 85 Linköping
Sweden

e-mail: elvar.theodorsson@ibk.liu.se

Keywords: reference intervals, data analysis, statistics, computer programs

Running heading: Advanced statistics and data analysis in laboratory medicine

Reference intervals

The dependence of laboratory results on a multitude of different influence factors including age was evident long before the introduction of the concept of reference values by Gräsbeck & Saris in 1969 [1-5]. Comprehensive and advanced but also conveniently applicable methods for including these factors in the calculation of reference intervals have, however, until now been scarce. Age- and sex- specific reference intervals have been calculated using widely accepted univariate statistical methods [6-12] and are well accepted in patient care. A recent rich and updated resource on the evolving aspects of reference intervals is the whole issue volume 42, issue 7, 2004 of Clin Chem Lab Med.

The theory of dealing with multivariate normal distributions has been extensively developed during the last five decades resulting in a wealth of textbooks written from several practical angles [13-15]. The methods have been introduced into laboratory medicine [16-20], but have not found practical use or survived the test of time.

Very recently Hansen and co-workers [21] described an innovative method for estimating reference intervals in small sample sizes based on variance component analysis. They point out the value of analysing small (<120 observations) materials, fully aware of the substantial uncertainty. Furthermore, and perhaps even more importantly, they make an informed effort to investigate possible causes of the variation. The fundamental question is whether the observed variations represent stochastic variations in measurement results over time in homogenous populations or if the variations are caused by identifiable factors/causes. Statistical and mathematical techniques including variance component analysis are valuable tools for this end, but can in no way replace the informed study of the underlying biological processes explaining

the variation. It is possible and even probable that time and resources are better spent on detailed studies of these biological factors in smaller patient materials rather than on purely statistical and mathematical studies of large patient materials.

Variance component analysis has a long history with roots dating back to the 1860s, blooming in the mind and hands of Sir Ronald Fisher early in the twentieth century and having a strong boost in the period since the 1970s coinciding with the availability of ever better mathematical and computational methods. In addition to estimating reference intervals, this technique is e.g. highly valuable for identifying the most important sources of variation in a laboratory organisation consisting of several laboratories, instruments and methods used for measuring the same analyte. It may e.g. identify which method or analytical instrument contributes most to the total variation. Readers interested in this field are primarily directed to the classical text by Searle and co-workers [22] which is easily available in book stores reprinted as paperback in 2006, or the book of Cox and Solomon [23] which has half the number of pages, but is very much to the point. Mathematically more elaborate for medically trained persons is the book by Rao and Kleffe [24].

In the current issue of Scandinavian Journal of Clinical and Laboratory Investigation Johan Bjerner introduces novel and innovative techniques for partitioning reference intervals using alternative principles and novel computer- intensive approaches. The time seems ripe for a fresh look at techniques used for sampling, measuring and calculating reference intervals in order to take new bold steps in this field the results of which are so useful and widely used in patient care.

It is becoming increasingly difficult to refrain from understanding of calculus and advanced statistical models since recent standards/methods recommended in all measurement- related sciences (e.g. GUM [25, 26]) require knowledge of advanced statistical methods.

Understanding the principles and applying advanced mathematical and statistical methods in the daily tasks of the clinical laboratory may seem to be an impossible task for medically trained persons. However, the rewards are potentially substantial and the most important methods are conveniently available in software tools for personal computers.

Freely available tools for data analysis

There is an abundance of advanced statistical and mathematical software tools developed and marketed commercially. During recent years a number of high-end freely available software tools have become available including the **“R” data analysis tools**. The “Free Software”/GNU (<http://en.wikipedia.org/wiki/GNU>) “R” data/statistical/graphical analysis tools (<http://www.r-project.org/>) were initiated by R. Gentleman and R. Ithaka (R&R) at the University of Auckland, New Zealand. It was originally inspired by the commercially available tools “S” developed by John Chambers & collaborators at Bell Laboratories (formerly AT&T, now Lucent Technologies) (http://en.wikipedia.org/wiki/S_programming_language). R is the result of massive collaborative efforts with contributions from all over the world. Documentation of the R software project is i.a. found at R Wiki (<http://wiki.r-project.org/rwiki/doku.php>). Actually R may be seen as a different implementation of S. There are some crucial differences between R and S, but much code written for S runs unaltered under R.

Gnuplot (<http://www.gnuplot.info/>) is a command-line interface data and function plotting utility for several computing environments. It was originally developed in 1986, and is still going strong. **Dataplot** (<http://www.itl.nist.gov/div898/software/dataplot/>) is also available for several computing environments. In addition to several plotting options it has several routines for statistical analysis including non-linear modelling. **Gnumeric** (<http://www.gnome.org/projects/gnumeric/>) is a spreadsheet program with much more advanced

scientific functionality compared to traditional office- oriented spreadsheet programs. Its numerical functions have been shown to be more accurate than those used in Excel 2003 (<http://www.csdassn.org/reportdetail.cfm?ID=508>). **WinBUGS** (<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml> and <http://www.winbugs-development.org.uk/>) is an advanced interactive program for Bayesian analysis of statistical analysis using Markov Chain Monte Carlo (MCMC) approaches .

Choosing tools for data analysis is like choosing tools for other purposes. The highest probability for relevant results is usually obtained using tools you are thoroughly familiar with. Money spent on commercially available data analysis tools is usually very well spent, and the time taken learning their interface and functions will pay off, even in revised versions released over time. Choosing the tools that best suit your personal preferences and needs irrespective of whether the program happens to be commercial, shareware or freeware is commonly a prudent strategy.

Chemometrics and laboratory medicine

Svante Wold and colleagues at Umeå University in Sweden already in the early 1970's coined the term "Chemometrics" for describing their introduction of novel and advanced multivariate techniques e.g. PLS [27] for analysing chemical data and for experimental designs. They were joined by Bruce Kowalski in the U.S.A. who developed a public domain "toolbox" of data analysis tools focused on chemical applications ("ARTHUR"), by Harald Martens, Tormod Næss [28], Kim Espensen in Norway, Luc Massart [29] in Belgium just to mention only a few.

"Chemometrics is a chemical discipline that uses mathematics, statistics and formal logic (a) to design or select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems." [30]. Several of the techniques have become invaluable in several areas of analytical chemistry

and in the industry. Unfortunately their widespread use in laboratory medicine is lacking behind probably due to less familiarity with linear algebra [31, 32] statistics and the computational techniques needed.

The disciplines of laboratory medicine serve in the interface between basic biological and natural sciences (e.g. chemistry, molecular biology, microbiology, immunology, metrology) and the clinical disciplines. Historically laboratory medicine has been in the forefront of introducing new methods from information technology, statistics and data analysis into the medical sciences.

Developments in medicine have frequently been catalyzed by developments in the basic sciences.

However, as the scope of science expands, it becomes increasingly more difficult to acquire and maintain the knowledge and overview needed to break significant new ground. This is particularly evident when introducing new methods of data analysis based on advanced mathematics and statistics. The ever increasing wealth of data readily available in health care prompts us to use improved techniques of data mining and interpretation. As new, promising and readily available tools appear, they are sure to tempt a number of us to embark on the steep learning curve needed to truly understand multivariate analysis, linear algebra and the computer programs needed to make use of the new techniques in practical laboratory work and in patient care.

References

1. Gräsbeck R, Saris N-E. Establishment and use of normal values. *Scand J Clin Lab Invest* 1969;26, Suppl. 110:62-3.
2. Gräsbeck R. Reference values, why and how. *Scand J Clin Lab Invest Suppl* 1990;201:45-53.
3. Alstrom T, Grasbeck R, Hjelm M, Skandsen S. Recommendations concerning the collection of reference values in clinical chemistry and activity report by the Committee on Reference Values of the Scandinavian Society for Clinical Chemistry and Clinical Physiology. *Scandinavian journal of clinical and laboratory investigation* 1975;144:1-44.
4. Sunderman J, FW. Current concepts of "normal values," "reference values," and discrimination values" in clinical chemistry. *Clin Chem* 1975;21:1873-7.
5. Gräsbeck RA, T. Solberg, H E. *Reference Values in Laboratory Medicine*. Chichester: John Wiley & Sons; 1981.
6. Solberg H. The theory of reference values. Part 5. Statistical treatment of collected reference values determination of reference limits. *J Clin Chem Clin Biochem* 1983;21:749-60.
7. Solberg H. Statistical treatment of reference values. *Bull Mol Biol Med* 1983;8:13-9.
8. Solberg H. Statistical treatment of reference values in laboratory medicine: testing the goodness-of-fit of an observed distribution to the Gaussian distribution. *Scand J Clin Lab Invest* 1986;46:125-32.
9. Wade AM, Ades AE. Incorporating correlations between measurements into the estimation of age-related reference ranges. *Stat Med* 1998 Sep 15;17:1989-2002.
10. Wade AM, Ades AE. Age-related reference ranges: significance tests for models and confidence intervals for centiles. *Stat Med* 1994 Nov 30;13:2359-67.
11. Cole TJ, Green PJ. Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med* 1992 Jul;11:1305-19.
12. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. *Clin Chem* 2000;46:867-9.
13. Mardia KV, Kent JT, Bibby JM. *Multivariate analysis*; 1979.
14. Albert A, Harris E. *Multivariate interpretation of clinical laboratory data*. Marcel Dekker, Inc, New York and Basel 1985.
15. Martens H, Naes T. *Multivariate calibration*. John Wiley & Sons 1989.
16. Goldberg D, Ellis G. Multivariate analysis in the diagnosis of liver disease and as an aid to optimal selection of serum enzyme tests. *Scand J Clin Lab Invest* 1984;44:113-29.
17. Anderson J. Multivariate methods in clinical chemistry. *J Roy Statist Soc* 1976;139:161-82.
18. Cederblad G, Kågedal B, Norr A, Sörbo B, Tegler L, Wigertz O. Computer-assisted evaluation by multivariate reference region and cluster analysis as applied to serum thyroid hormones and thyrotropin. In: *Lecture Notes in Medical Informatics*, Vol 5, 1979:650-7.

19. Kågedal B. Screening for abnormal thyroid function in middle-aged women. With special reference to computer-assisted multivariate evaluation of laboratory results. Linköping: Linköping University; 1979.
20. Winkel P. Patterns and clusters-multivariate approach for interpreting clinical chemistry results. Clin Chem 1973;19:1329-38.
21. Hansen AM, Garde AH, Eller NH. Estimation of individual reference intervals in small sample sizes. Int J Hyg Environ Health 2007 May;210:471-8.
22. Searle SR, Casella G, McCulloch CE. Variance Components. New York: John Wiley & Sons, Inc.; 1992.
23. Cox DR, Solomon PJ. Components of Variance. Boca Raton: Chapman & Hall/CRC; 2003.
24. Rao CR, Kleffe J. Estimation of Variance Components and Applications. Amsterdam: Elsevier Science Publications; 1988.
25. ISO CD. Guide to the expression of uncertainty in measurement. Switzerland: ISO; 1995.
26. Ellison SLR, Rosslein M, Williams A. Eurachem/CITAC Guide Quantifying Uncertainty in Analytical Measurement. Second Edition: Eurachem; 2000.
27. Wold S, Martens H, Wold H. The multivariate problem in chemistry solved by the PLS method. In: Kågstöm BR, A editor. Matrix pencils - Lecture Notes in Mathematics. Berlin: Springer Verlag; 1983. p. 286-93.
28. Martens H, Næs T. Multivariate Calibration. New York: Wiley; 1992.
29. Massart DL. Chemometrics: A Textbook (Data handling in Science and technology): Butterworth Heinemann; 1990.
30. Massart DL, et al. Data Handling in Science and Technology 20A: Handbook of Chemometrics and Qualimetrics Part A: Butterworth Heinemann; 1997.
31. Searle S. Matrix algebra for the biological sciences. John Wiley & Sons, Inc 1966.
32. Basilevsky A. Applied matrix algebra in the statistical sciences. Elsevier, North-Holland 1983.