Statistics for Scientists: Just Another Toolbox?

Yves-Laurent Grize
Baloise Insurance
Aeschengraben 21
CH-4002 Basel, Switzerland
yves-laurent.grize@baloise.ch

1. The Perception of Statistics

What do we statisticians do? Here is a typical list of our activities, each one of them being more or less equally important to us: data collection and design, exploratory data analysis, inference, modeling of stochastic systems, data presentation and reporting.

But what do scientists think we are doing? First they perceived us as inference specialists, able to compute p-values and thereby deciding if a scientific contribution is publishable or not! We are also perceived as statistical software specialists able to adjust the parameters of complicated statistical methods and decipher the incoherent output of statistical packages. Often we may receive a huge amount of data to analyze, as it is clear that the more data we have, the better its statistical analysis will be. Rarely we will be consulted in advance on what data to collect or on what sample to use. Finally, very rarely, scientists will come to us and ask to build a model. But sometimes they do come to see us, like the following illustrative dialog shows:

Scientist: I will bother you just for a couple of minutes; I know that I need to do this analysis, and I simply wanted to ask you if you think that it is correct.

Statistician (*trying to be diplomatic*): Well, that looks pretty good, but let me ask you something: What is really the problem you want to solve?

Scientist (*frowning the eyebrows*): Aren't you a statistician? (*and starts to wonder if it was a good idea to come and ask for advice*).

Conclusion: Scientists don't want us statisticians to be „involved“ in their business!

This leads to a paradoxical situation: on the one hand, we know that statistics is about learning from Data. In some sense we sometimes even dare say that statistics is the science of doing Science! And yet on the other hand there is no appreciation by scientists of the role and power of statistics! Only as a tool is the usefulness of statistics recognized (e.g. for FDA approval, quality certification, journal publication, etc...).

Is there an explanation to this paradox?

J. Friedman has made the following affirmation, although in a slightly different context (see Friedman (2001)): "Statistics is perceived as a set of tools and not as a set of problems". I think that this is profoundly true and may be the main reason for our paradox. Indeed for scientists, statistics is just another toolbox among many others. The fact that it is the problem that defines the statistical tool, thereby the entire data based induction process, is simply not perceived. A good illustration of this point is my experience concerning the use of experimental design in different departments of large a pharmaceutical and chemical company: Because experimental design is based on the concept of experimentation itself, it is one of the most appropriate tool to understand and discover new relationships between parameters in a system. And yet research scientists used it much less than engineers (process optimization) and also less than laboratory scientists (development of analytical methods). May be statistics is seen as a threat to the researcher own creativity and problem solving and discovery skills!

Also the increase in computing power and the broad availability of software and PCs has probably accentuated the "tool-perception" of statistics over the last 25 years.

What can we do to try to change this misconception of our field? Two lines of actions appear possible. The first possibility is simply to refuse that conception and say out loud that statistics can be used to address and solve problems. This is usually what statisticians do, but too few scientists are listening and little progress has been made so far.

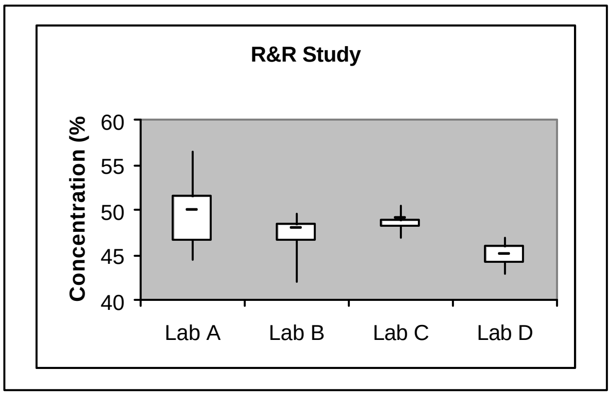
The second possibility is based on the belief that scientists are more likely to listen to other scientists than to statisticians. Therefore the more scientists practice statistics by themselves, the better they will eventually see and understand what statistics is really about. Hence we can hope that by providing the right tools to scientists, the "problem-perception" of statistics will diffuse through the scientific community and finally correct this fundamental misconception.

2. Statistical Tools for Scientists

Scientists will use statistical tools by themselves if they are easy to use, easy to understand and easy to communicate with. Scientists have an healthy mistrust of 'black-box' solutions. It is important that they understand how the statistical tools work. Furthermore it is also important that they can speak to other colleagues and convince them to also use these tools. Therefore the ease of use and of communication are crucial. Clearly this means that essentially graphical tools are required. In addition, these tools should promote the view that statistics is not just about p-values.

In the talk we shall expose and briefly discuss five tools, that, in our experience, qualify for our purpose. Certainly this list is not exhaustive, but we believe that at least the first four tools should be known and used by any scientist:

- Box-and-Whiskers Plots (or boxplot): surprisingly still very much unknown. A construction with Excel is possible (see e.g. www.mis.coventry.ac.uk/~nhunt/boxplot.htm).
- Contour Plots in Response Surface Analysis: probably the most used feature of experimental design software
- Biplot: Very compact representation of a fairly large number of variables and observations simultaneously. Draw-back: because it is based on a principal component analysis it may not be easy to explain. Its use is easy only with the appropriate code.
- Classification Tree: Extremely effective for communication; probably the tool where profound computational and statistical issues are best hidden; applicable to many situations (missing data).
- A-posteriori distribution and Bayesian inference: The Bayesian paradigm appeals to many scientists by allowing them to conveniently incorporate a-priori knowledge with information obtained from data. The use of an a-posteriori measure for describing uncertainty makes often more sense and is easier to interpret. Powerful Bayesian analyses have been implemented in the easy-to-use software BUGS: Bayesian inference Using Gibbs Sampling.



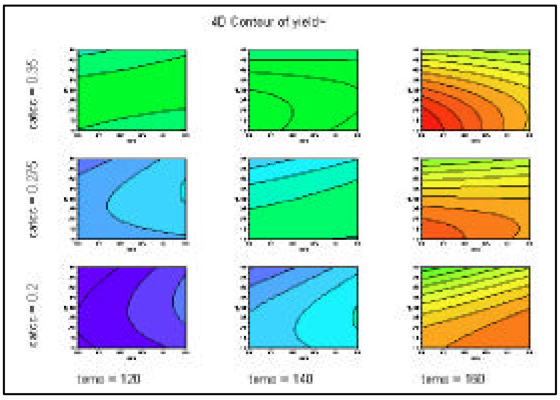
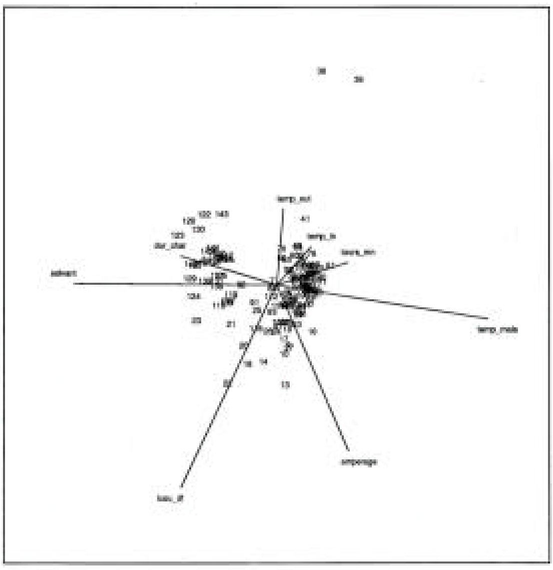
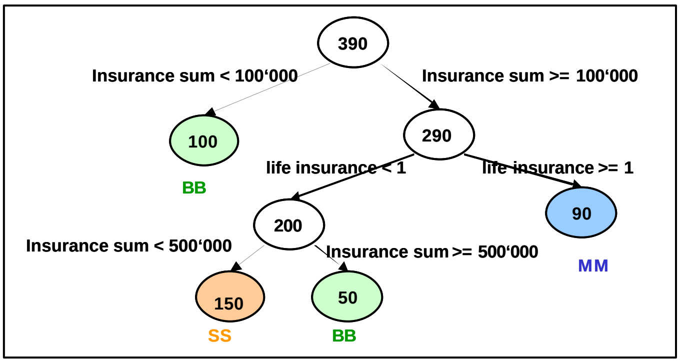


Figure 1 (left): Boxplots in a Repeatability and Reproducibility Study (EXCEL)
Figure 2 (right): Matrix Plot for Response Surface Analysis (MODDE)





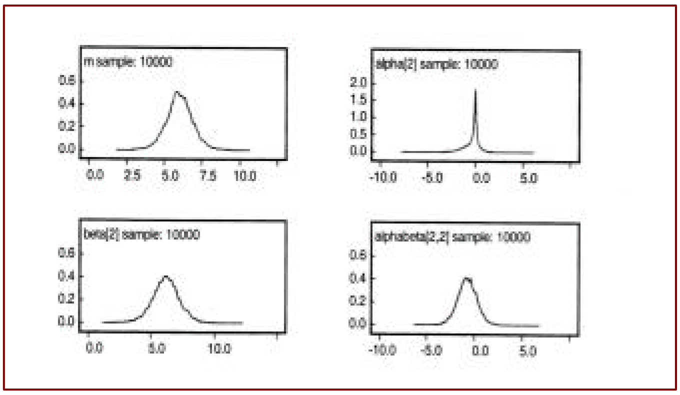


Figure 3(left): Biplot to summarize pigment production data in multivariate QC (Splus)
Figure 4 (top right): Classification Tree to Predict Insurance Customer Type (R)
Figure 5 (bottom right): A-Posteriori Distributions for selected parameters in a 2 way-ANOVA (BUGS) [from the MS Thesis of D. Spinnler, Univ. of Neuchâtel]

3. Statistical Education & Statistical Awareness

Finally, a second reason for the failure to perceive statistics as an integral part of the scientific induction process is the lack of understanding of what is called 'statistical thinking'. For me, Statistical Thinking is how to extract information from data when taking its variability into account. The recognition that data has not just one ‚dimension‘ (location) but two (variability) is here crucial! How can one possible correctly extract information from data without systematically taking its variability into account, for example when analyzing scientific experiments, making decision under uncertainty or, in general, learning from quantitative observations? And yet this obvious fact is so often overlooked, that there is probably a psychological reason for the failure of the human brain to do so. Therefore it must be a matter of education to explain the importance of the variability of data when making analyses.

The amount of data keeps increasing. The demand for statistical tools is on the rise (e.g. data mining, bioinformatics, etc…). Every day our society becomes more and more evidence-based. Statistical surveys especially reported via the media (e.g. Internet) are having more impact. All these reasons should make the case for pushing ahead with statistical education easier, and this not only for scientists, but for any educated citizen. It is a responsibility of statisticians to get more involved in managerial or leadership positions and use their obtained influence to increase statistical awareness in their company if they are working in industry or in the public (see Smith (2001)). Finally, because statistical thinking and science are so closely related, a better understanding of the former will also contribute to a better understanding of the later. According to C. Sagan, there actually is an urgent need to do so:

"If we can't think for ourselves, if we're unwilling to question authority, then we're just putty in the hands of those in power. But if the citizens are educated and form their own opinions, then those in power work for us. In every country, we should be teaching our children the scientific method [*or should we say "statistical thinking"; note from the author*]! and the reasons for a Bill of Rights. With it comes a certain decency, humility and community spirit. In the demon-haunted world that we inhabit by virtue of being human, this may be all that stands between us and the enveloping darkness." (Sagan (1997).

Statistics and Science have both much to gain from a better understanding of 'statistical thinking'.

REFERENCES

J.H. Friedman (2001), The Role of Statistics in the Data Revolution?, Int. Stat. Rev.69, 5-10.

N. Hunt (1999), Boxplot in Excels, http://www.mis.coventry.ac.uk/~nhunt/boxplot.htm, Coventry University.

C. Sagan (1997), The Demon-Haunted World: Science as a Candle in the Dark, Ballantine Books (Reprint edition).

A.F.M. Smith (2001), Public Policy Issues as a Route to Statistical Awareness, Int. Stat. Review 69 (2001), 17-20.

D. Spinnler (2001), Bayesian Methods in Practice: Application to the Swiss Canopy Crane Project, MS Thesis (supervisor YL Grize), University of Neuchâtel.

For information on the software products mentioned in the paper:

BUGS:	www.mrc-bsu.cam.ac.uk/bugs
EXCEL:	www.microsoft.com/office/excel
MODDE:	www.umetrics.com
R:	www.r-project.org
Splus:	www.insightful.com

ABSTRACT

The field of Statistics has changed in profound ways over the last 25 years. These changes have not only affected how Statistics is done but also how it is perceived. We will therefore examine what is in our opinion today the perception of Statistics by scientists and also by statisticians.

Based on various examples from our industrial consulting experience, we believe that for most scientists, Statistics is just another toolbox. This will lead us to make some suggestions on the requirements that statistical tools must fulfill to increase further their use in the scientific community, thereby changing the perception of what Statistics is really about.

Finally the importance of statistical education and statistical awareness will be stressed, as we think that this is an additional way to induce a change of perception and especially to make variability a better understood concept in general. Statistics and Science have both much to gain from a better understanding of 'statistical thinking'.

RÉSUMÉ

Le domaine de la Statistique a profondément changé durant ces 25 dernières années. Ces changements ont affecté comment la statistique s'applique mais aussi comment elle est perçue. Nous examinerons quelle est la perception que les scientifiques ont aujourd'hui de la statistique. A partir d'exemples tirés de notre expérience industrielle, nous pensons que la statistique est surtout perçue comme une boîte à outils. Ceci nous mène à faire des suggestions sur les critères que les outils statistiques doivent satisfaire pour être davantage utilisés au sein de la communauté scientifique et par là même pour contribuer à changer la perception de ce qu'est vraiment la statistique.

Enfin l'importance de l'éducation statistique et aussi de faire connaître la statistique est souligné. En particulier ceci devrait permettre de faire mieux comprendre le concept de variabilité qui est à la base de toute pensée statistique.

La Statistique mais aussi la Science ont toutes les deux beaucoup à gagner d'une meilleure compréhension de la pensée statistique.