AStA (2007) 91: 413–427
DOI 10.1007/s10182-007-0046-x

ORIGINAL PAPER

Statistics development: statistical methods meeting the user's needs

Jan Engel · Henriette (Jettie) C.M. Hoonhout

Received: 26 January 2007 / Accepted: 12 September 2007 / Published online: 31 October 2007

Summary Statistical methods have the potential of being effectively used by industrial practitioners if they satisfied two criteria: *functionality* and *usability*. Statistical methods are usually the product of statistical research activities of universities and other research organizations. Some already satisfy these criteria; however, many do not. The effect is that potentially relevant methods are not used in practice as often as they could be. In this paper we will present an approach regarding 'statistics development,' in which the end-user is given a central position, so that the results from statistical research aim to meet the needs and requirements of the practitioner. Examples of known and new methods will be presented, and we will discuss issues such as education in statistics, the link with statistical consultancy and publication of methods through various channels.

Keywords Statistics development · Statistical method · Functionality · Usability · User needs

1 Introduction

Statistical methods are potentially very effective when used by industrial practitioners, but only if they are functional and usable. Obviously, a method should be functional, i.e., it should contribute effectively to statistical data analysis and give answers to research questions. In addition, a method should be usable, or more specifically,

J. Engel (✉)
CQM, Vonderweg 16, P.O. Box 414, 5600 AK Eindhoven, The Netherlands
e-mail: Engel@cqm.nl

H. Hoonhout
Philips Research Europe, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
e-mail: jettie.hoonhout@philips.com

a method should be understandable, learnable, and easy to use in the perspective of the practitioner/user; software to support the application of the method should also be available.

Some of the methods that are the product of research activities at universities and other research organizations satisfy these criteria; however, many do not. Currently, statistics research usually results in the following: a journal paper and a conference presentation, all for peers, along with perhaps additional software that may support the practical use of the researched method. This is all very well and good when considering the statistics research community. However, the effect is also that potentially relevant methods often do not become known to industrial practitioners, and hence do not get used. Note that the meaning of 'industrial practitioners' is twofold. On one hand, these are the industrial researchers and developers that need statistical methods for daily use; on the other hand, these are the statistical consultants that give their advise to the before mentioned industrial community. When methods are not used by the potential target population, a chance to collect valuable feedback, which could lead to improvement of its methodology tool-box, or at least to check whether or not the method addresses a real need of users, is missed.

We will discuss in this paper the developmental process of industrial products, formalized by Design for Six Sigma, illustrating the creation process of products intended for use. We adopt this as a model for what we define as 'statistics development,' an approach that aims to develop statistical research results, which will reach the market place of the practitioner. Therefore we project the development of statistical methods onto this process development model and give suggestions for improving the R&D cycle of these methods so that their impact in industrial statistical practice will be stronger. Examples of known and new methods will be presented, and we will concentrate on industrial problems, although the applicability of this approach will likely be much wider.

This paper is organized as follows. In Sect. 2 we will introduce the current widely accepted view of modern industrial product development, a user-centered development, which emphasizes determining customer needs as a necessary starting point for successful product development. This view is also relevant and applicable to the development of statistical methods; here too, the needs and requirements of the intended users should be considered. Furthermore, we will present a systematic approach – Design for Six Sigma (DfSS) – adopted by many industries, which incorporates this customer-centered view, and aims to guide the development process by ensuring that important customer requirements are taken into account in a systematic, measurable and verifiable way. Then, in Sects. 3 and 4, we turn to statistics and usability, and discuss the user's needs regarding the functionality and usability of statistical methods, applying the ideas from this DfSS process development approach. We will conclude this paper with a discussion on the development of statistical methods in a wider perspective, including education, training, and other forms of 'knowledge transfer.'

2 Industrial product and process development

Modern approaches towards the development of industrial products aim to design products based on the customer's requirements and demands (e.g., Nielsen 1993, Beyer and Holzblatt 1997, Diederiks and Hoonhout 2007). This is generally called user-centered design or development. In addition, quality is made more explicit than ever and is specified in terms of customer needs. Instead of pushing technology into the market, customer needs are the central point for industrial development. One increasingly popular systematic approach that has its origin in US industries, with companies such as Motorola as the driving force, is what is called Design for Six Sigma.

Design for Six Sigma (DfSS; see Chowdhury 2002) is a philosophy and methodology for modern product development. There are two basic ideas in DfSS that are the key elements in this philosophy: (1) start with customer needs and not with technology (i.e., user pull, instead of technology push) as customers are the final users of the product, determining its success in the market, and (2) design quality into the product at an early stage of the development process to satisfy customer needs. This means that the product parameters should be at target value, with minimal variability, to be within the consumer-defined specification limits. Creating satisfactory products at the beginning is much less expensive than fixing products that are already in the market, and this consideration explains the success of DfSS in the industry.

Although there are different versions of the DfSS methodology, we think that the following variant, DMADV, is useful for our purposes. DMADV is the acronym for the five steps in DfSS that actually form the main body of the method.

These five DMADV steps are:

1. Define. Define the customer deliverables.
2. Measure. Measure and determine customer needs and specifications.
3. Analyze. Analyze the process options to meet customer needs.
4. Design. Design the process to meet customer needs.
5. Verify. Verify the design performance and the ability to meet customer needs.

It is very interesting to see in this five-step scheme that the first two steps are concerned with *customer* needs, and not with available or to be developed technologies. Technology considerations and technology development do not come into play until steps 3 and 4, where the process is designed in such a way that the outcomes will meet the needs of the customer. In the final step, the attention is again focused on the customer, checking how all efforts have worked out to satisfy his or her needs.

To continue this section, we will discuss DfSS a little further with the help of Fig. 1. This figure gives the main variables and factors that play a role in a process step.

The box in Fig. 1 represents the process step under consideration. As an example we give the process step for baking a cake in Fig. 2. The variables A, B and C are the measured inputs, such as the amount of eggs and butter in Fig. 2. The response is the output, and closely linked to customer needs: the taste and texture of the cake. The response should preferably be on target, and with a high chance of occuring between the specification limits, with which the customer agrees (e.g., the crust should opti-

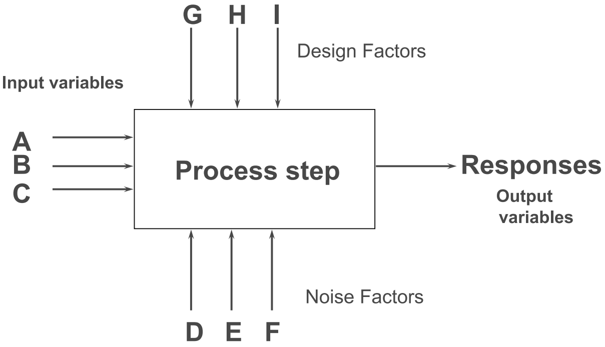

Fig. 1 Process step with input variables and responses, design factors and noise factors

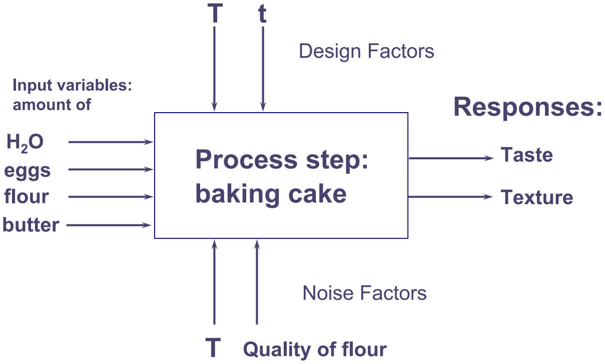

Fig. 2 Cake baking process step with input variables and responses, design factors and noise factors

mally be golden brown, but certainly not towards blond or dark brown). Finally, there are two types of factors: design factors and noise factors. Design factors represent all parameters in the process that can, in principle, be fixed at a certain value. Examples in the cake process step are baking temperature T and baking time t. The noise factors summarize all non-controllable random disturbances of the process. When the oven temperature T is not constant but varies over time, temperature T is a noise factor as well. Another noise factor is the quality of the flour that will vary over time because of batch-to-batch variations at the flour deliverer. Noise factors are a little inconvenient as they disturb the quality of the process: the customer targets need to be satisfied! Design factors and noise factors play an important role in what is called robust design (Taguchi 1986). Robust design teaches us to choose the design factors in such a way that the *effect* of all noise factors on the responses is considerably small. When this is not sufficiently successful, reduce the variability of the noise factors. The responses are then close to the target and the customer will be happy.

To conclude this point we consider DfSS from the point of view of the process step model of Fig. 1.

1. Steps D and M. Find targets and specifications for the process step responses from customer needs.
2. Steps A and D. Perform robust design to realize customer needs. This concerns the next three steps.

   - *Parameter design*. Set the design factors such that the mean of the responses is at the target. Also, set the design factors so that the responses have minimum *sensitivity* for (i.e., are most *robust* against) the variability in the noise factors. The responses are then close to the target.
   - Find *process window*. This is the window for allowed variability of the noise factors so that the responses are within the specification limits.
   - *Tolerance design*. When needed, reduce the variability in the noise factors so that this is limited to the process window, and the responses are within the specification limits.

How can DfSS play a role in statistics development? The following arguments seem to make sense:

1. A statistical method can be seen as a process step. We will explain this in the next section. This gives us a *first link* between statistics and industrial development with DfSS.
2. The starting point of DfSS is the customer, not the technology. As statistical methods have a large potential to be used by industrial practitioners when they have some level of utility and usability, a central position of the user in the development of statistics seems a fruitful *second link.*

We start with the methods produced by statistics research with a potential to answer specific research questions. There are many, so it makes sense to see what is available as a result of statistics development to solve the practitioner's problems. It is a known fact from industrial research that only a few research results will find their way into the market. The same may hold for statistics research. Important is that DfSS for statistics could help to establish a link between the user needs and these research results.

3 How DfSS can be of use in statistics development

Crucial for the next discussion is the following vision: a statistical method can be considered as a *process*, as well as a *product*. The argument for the process is as follows. Essentially, a statistical method transforms the input data into information that can be seen as the process response. An example is a statistical testing procedure that transforms input data into a decision (rejecting the hypothesis or not). Each statistical method is defined under a set of assumptions concerning data quality. For statistical testing this is, e.g., normality of the distribution of the data. However, the transformation should be done properly for various types of data, e.g., for non-normal data as well. This means that the principles of industrial process design (robust design) are relevant for the design of a statistical method: can we find a method that is robust against non-normality? Each statistical method can be seen as a product as well.

Indeed, it will be used as a tool by a practitioner and hopefully with success and some joy. Having said this, we will refer to the process view of statistics in the following section.

There we apply the DfSS principle of industrial process design to the design of statistical methods. Because this seems to be a new area we will only work out the main principles globally. In the following section we consider the five points from DfSS for statistics development, grouped in the following three clusters:

1. Define and measure. Determine the user's expectations and needs.
2. Analyze and design. Analyze and design statistical methods.
3. Verify. Evaluate whether the user's expectations are met.

We will reserve the phrase 'statistics development' for those activities that are needed to develop a statistical method from the research phase into a product that satisfies the user's needs.

So now we can easily interpret the process model from Fig. 1 in terms of statistics. The process step is the *statistical method*, the input variables are the *data* the method has to work and the response is the result of the method, for example the result of statistical testing. Then we have the design factors and noise factors. The design factors are concerned with the *choice* of method, the noise factors represent all *deviations from the assumptions* made on the method. Finally, the crucial point is to select the proper method (1) satisfying usability needs and (2) satisfying functional needs despite all *deviations* from the assumptions made on the method that always occur in practical situations.

4 DfSS for statistics

4.1 Define and measure

Contrary to statistics research, which is method oriented, users are problem oriented; they like to find answers to their research questions, and to them, the statistical method is merely a tool to reach that goal. For statistical methods to meet user's demands, statistics development should start with collecting and analyzing the user's questions and should aim to develop methods to support users in answering these questions. User needs of statistical methods include two main criteria: functionality and usability. Functionality is concerned with the technical aspects of the method: it should just do its job according to the needs of the user. Functionality is defined in statistical terms. This is different from the usability of the method that focuses on the usage aspects of the method, rather than on the technical utilitarian side. The International Organization for Standardization (ISO) defines usability as: 'the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' (ISO 9241-11 1998). Shackel (1991) provides a similar definition stating that 'usability of a system or equipment is the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfill the specified range of tasks, within the specified range of environmental scenarios' (p. 24 in Shackel 1991).

Furthermore, Nielsen (1993) provides the following criteria to define the concept of usability:

- Learnability. The ability to reach a reasonable level of performance, after an acceptable effort in training
- Memorability. The ability to remember how to use a product, after it has not been used for some time
- Efficiency. Trained user's level of performance
- Satisfaction. Subjective assessment of how pleasurable it is to use (which has strong motivational implications for the further use of a method)
- Errors. Number of errors, ability to recover from errors, existence of serious errors

When thinking about the usability of statistical methods, these criteria should be considered. Especially important aspects may be learnability and memorability of the method, effectiveness and efficiency, but also complexity of applying the method, understanding the basic operations of the data, and even enjoyment.

An example may be helpful to illustrate this idea.

Example 1: Two-sample testing problem A common problem in practical research is the question to verify whether the means of two populations, say μ_1 and μ_2, are different. We may need to prove that a new version of the product is of better quality than the version now on the market, or than the version from the competition. In statistics this research question is formulated in terms of a null hypothesis H_0: $\mu_1 = \mu_2$ and a test statistic is introduced for testing H_0. When the test statistic has outcomes in a specified area, H_0 will be rejected, and the means μ_1 and μ_2 will be stated as different. An example is the two-sample t-test. We will now try and find the consequences of *define and measure* for this testing problem.

Define and measure to meet user needs:

1. The *usability* needs will include an easy way to determine which test to use, i.e., being able to determine and understand the requirements and limitations of possible methods (e.g., depending on the properties of the data set, whether or not it is allowed to use a particular method), which operations are required on the data, and how the results should be interpreted. Research has shown that these are not trivial issues (e.g., Hoekstra et al. 2006a, b).
2. As a statistical test delivers a simple binary result ('to reject or not to reject'), it is suggested that the *functionality* needs are simple as well. This is not true. Primarily, they are specified by the errors of the first and second type. These are, respectively, the probability of rejecting H_0 when it is true (usually, it is limited to 5%) and the probability of not rejecting H_0 when it is false, where the latter is related to testing power (say, 90% at a specified difference of μ_1 and μ_2). A second and much more complex aspect, however, concerns the input data the method should be able to handle so that the user's needs are satisfied (data quality). This concerns the distribution of the data, outliers and data dependence.

To summarize, the functionality needs for the testing procedure are fixed by the user's specification of the error of the first and second type, and by the user-defined data quality under which the method should perform according to specification. It may sometimes be difficult to get the proper information from the users on functionality and usability. In testing, the user's needs are (often) not simply made explicit for the second type error. However, they will be for the first type error.

In this two-sample testing problem we consider (1) the Wilcoxon two-sample test, which is not much affected by non-normal distributions and outliers, and (2) the two-sample t-test. We return to this in the next section. Example 1 gives a simple illustration, whereas the following example is closer to our practice.

Example 2: Analysis of ordinal data Contrary to the physical sciences, measurements in the sociological and the psychological sciences are often indirect measurements of latent variables, and they are at the ordinal scale. For users who wish to do statistical testing on the data, methods for testing on ordinal data are needed. During the last 20 years many, sometimes complex, statistical methods have become available for the analysis of ordinal data (see Agresti 2002). The question is how to find functional *and* usable methods; complex methods are in most cases not *easy-to-use* methods.

The two cases that we will use to illustrate our point are described in Meerbeek et al. (2006), and Pelgrim et al. (2006), respectively. In both cases questionnaire data were obtained. Example 2 will also be re-addressed in the next section.

4.2 Analysis and design

When user needs have been specified, the analysis and design steps should help to find the proper methods to satisfy these needs.

1. The *usability* aspects concern ease of use of the method (e.g., no elaborate or complex data processing required), availability of support tools (e.g., software), accessibility of the method (e.g., where and how the user can obtain the method), clear specification of how data needs to be prepared, and ease of interpretation of the outcome. The final result is a ranking of the available methods on usability.
2. Consider the *rank-one* (most preferred) method; the *functionality* is decided as follows.

   - Find the process window for the rank-one method. Is it robust against deviations from the assumptions that are made on the method? What data quality (e.g., data distribution, outliers) is needed?
   - Compare data quality expected in the problem with the process window for the rank-one method. When there is no match, consider tolerance design. There are three options:

   1. Improve data quality so that it matches the process window for the rank-one method (e.g., eliminate outliers, improve experimental design, increase sample size). When unsuccessful go to next step.
   2. Re-define user's needs, and be happy with less functionality.
   3. Or move to the method next in rank.

A trade-off arises in the latter case: we can still apply the method that is preferred from a usability point of view, but it has less functionality, or reject this in favor of a method next in rank with higher functionality. As we mentioned above, note that in this analysis and design step sampling and experimental design will also be included. Improving data quality can sometimes be reached by improving the experimental design, training the panel, and so on.

Although each statistical method is supposed to perform well under a certain set of assumptions, reality forces us to also control the quality of the method under *deviations* from these assumptions. The question is what deviations from the assumptions are still acceptable, i.e., what is the *process window* of the method. Therefore we need to know: (1) what deviations from assumptions are still allowed to satisfy the functionality needs (the process window), and (2) what should we do when these deviations are exceedingly large?

Now we will reconsider our two examples.

Example 1: Two-sample testing problem (continued) For testing the two-sample case we assume the following two methods: (M1) two-sample t-test and (M2) two-sample Wilcoxon test.

Suppose that the user requires the proper control of the errors of type 1 and 2 for the analysis of maybe non-normal data, where outliers may be present. We will consider the consequences of the analysis and design steps. The analysis step analyzes the two methods available. The design step chooses what method to use.

1. The *usability*. Users would be helped when there would be support in determining what the properties of the data are, which methods can handle these properties, or what one needs to do to manage them. In practice, users often simply go for the method they are most familiar with, the t-test, ignoring possible properties of the data set that in principal would not allow its use. When ranking the above two methods, assume that method M1 comes out as rank-one, the most preferred method.
2. The *functionality*.

   (a) Find the process window for method M1. This gives an answer to the question to what extent the t-test will perform the testing properly. Note that the two-sample t-test makes the assumptions: (1) independence of the data, (2) normality of distributions, and (3) equality of standard deviations (homoscedasticity). The two-sample t-test should also preferably meet the functionality needs when these assumptions are not (fully) satisfied. For strong deviations from the assumptions this will be untrue. Within the process window, we can apply the t-test, outside this window we will loose functionality. For instance, the two-sample t-test is robust against deviations from normality (see Baker et al. 1966 and Zijlstra 2004) but it is sensitive to outliers. It is robust against heteroscedasticity but not to the dependence of the data.

   (b) Compare data quality expected in the problem with the process window for the t-test. When there is no match, consider tolerance design. There are the following three options.

- Improve data quality so that it matches the process window for the t-test as the rank-one method (e.g., data transformation to normality, eliminate outliers). When this is unsuccessful go to the following.
- Re-define user's needs. Accept large type 1 error or less power.
- Or move to the Wilcoxon test.

Example 2: Analysis of ordinal data (continued) For analyzing ordinal data from designed experiments, we will consider the following two methods for testing the main effects and interactions of the factors in the design: (M1) analysis by linear models (Anova) and (M2) analysis by ordinal models (see Agresti 2002). For this example we will consider again the two aspects usability and functionality.

1. The *usability.* Ordinal models are complex, they are quite useful for statisticians but not so much for the sociological and psychological communities. These, on the contrary, have a general preference to apply methods for *interval* scale data on the observed ordinal data, like Anova and principal components analysis. Therefore method M1 seems to be the method with the largest usability. For instance, Anova methods are relatively simple and well known to the mentioned communities. However, applying Anova on ordinal data maybe at the cost of the functionality of statistical testing.
2. The *functionality*.
   (a) Find the process window for method M1. This answers the question to what extent ordinal data can be analyzed as interval scale data.
   (b) Compare the data quality expected in the type of problem that has to be analyzed, with the process window for method M1. When there is no match, consider tolerance design. There are the following three options.

   - Improve data quality so that it matches the process window for M1. When this is unsuccessful go the next step.
   - Re-define user's needs, e.g., accept a larger type 1 and type 2 error.
   - Or move to method M2 and accept a lower usability of the method.

We will now further discuss this aspect of functionality in the context of the type of practical problems we meet. Indeed, many studies in the psychological sciences use rating scales, e.g., Likert scales in questionnaires, as the responses in experiments, and use interval scaled methods for analyzing these ordinal scale values. Two problems from our own environment will be presented, illustrating this issue.

The first example is a study in which different robot personalities (characters), and different levels of control for the user (i.e., to what extent the user is in control of the actions the robot will perform) were investigated (Meerbeek et al. 2006). In order to determine the preference of the participants in the test regarding the different combinations of personality and control, questionnaire data was collected: the participants were presented with statements, and asked to indicate their agreement on a 7-point scale ranging from totally disagree to totally agree. In total, 32 participants participated in this test.

The second example is a study in which combinations of ambient scent (fragrances dispersed into a room) and colored lighting settings were presented to participants, who were asked to assess the quality of the room on different aspects. Again, participants were asked to complete questionnaires, containing items addressing the evaluation of the room and the setting (Pelgrim et al. 2006). Participants could indicate their agreement with the items on a 7-point scale. In total, 90 participants took part in this test. In both examples, the data was analyzed by Anova (common practice in similar types of studies), although the measurement scale is ordinal, not interval.

Although method M1 is often used, as in the two presented problems, the process window for method M1 is not well determined. There are indeed some results from sociological and psychological literature, e.g., Baker et al. (1966) and Labovitz (1967), but no systematic study covering all relevant aspects seems to have been done. The question how to analyze ordinal data as interval scale data actually falls apart into three sub-questions that need an answer:

1. Can we *at all* assign interval scale numbers to the ordinal categories so that these interval scale numbers have a meaning for the problem?
2. If we can, what would the numbers be?
3. Can we analyze these interval scale numbers by Anova?

Note that the first and second question concern the measurement aspect of the problem, they are questions on type of scale and are related to the *measurement model* we assume for the data. The third question is a purely statistical one and it asks about the robustness of Anova against deviations from the standard Anova assumptions. It is thus closely related to example 1, as the two-sample t-test that is treated is a special case of an Anova test.

One systematic and practical approach to find the process window of method M1 is by Monte Carlo simulation, as an analytical study will likely be complex. Some of the above mentioned references use a Monte Carlo approach as well, but the following three steps seem to cover the problem more extensively:

1. Choose an ordinal model for a general practical situation, with some experimental factors affecting the ordinal response in an experimental design setting; a first study will be on a univariate response, although the problem is multivariate in nature.
2. Simulate data from this model by Monte Carlo simulation.
3. Analyze the data by methods M1 and M2 and compare the results that are found on testing the significance of the factorial effects.

When the testing results of M1 and M2 strongly agree, method M1 can be used instead of method M2. To find the process window effectively, remember from Sect. 3 that we have to investigate the noise factors where the noise factors represent all deviations from the assumptions made on the method. This means that all relevant uncertainties should be included in the Monte Carlo study to find their effect on the functionality of Anova. Certainly the following noise factors are important:

1. Number of categories at the ordinal scale.
2. Scale values that are assigned to the ordinal categories (e.g., linear, non-linear).
3. Size of factorial effects: zero effect or effect of practical relevant size.
4. Number of subjects in the experiment.

We need to analyze the effect of these noise factors on the functionality of method M1 and this will determine the process window. Then the value of Anova for ordinal data from designed experiments will become clear. This is a highly relevant topic for practitioners and definitely interesting for further research.

4.3 Verify

As discussed in Sect. 3, once a start has been made with the development of a product, and in the case of statistics with the development of a statistical method, it is vital that the result is evaluated. We need to check whether or not the requirements of the user determined in the first step, have been met. It is the final check on functionality and usability of statistical methods: what did we achieve? Are the methods actually used and what is the user's experience? Which of the usability criteria as defined in Sect. 4.1 have not been satisfied? What functionality is missing? We can learn a lot from user evaluation results in this step and these may be a starting point for further improvements.

Unfortunately, this is a step that is still missing in the domain of development of statistical methods. It is extremely rare that users and statisticians exchange experiences concerning the development and use of methods.

5 Discussion

Statistics research now usually is a 'technology push' and strongly method oriented. Like in industrial process design, a firm grasp on the technology will of course be a strength. However, we need to realize that ultimately the goal of statistics development is to provide users with an effective, useful, and usable new method. Often, the user is not a specialist and nevertheless likes to solve his or her problems, and here we have an issue. The user does not wait for methods but is waiting for help to answer his or her research questions. So this means that in statistics development we should change our focus from technology to the user, from methods as the product of statistics research to methods that satisfy the user's needs. If we manage to build a bridge between statistics researchers/developers and users, a significant step forward for statistics in practice will have been made.

We applied the DfSS approach on statistics, aiming to systematically consider the usability and functionality of methods, and thus coming to a rough selection of the best method. When none of the existing methods are useful, they will have to be extended, or even a new method will have to be found. We did not discuss this further in the paper, but this could be a topic for a further study. We recommend, however, that for finding (or developing) new methods, the DfSS approach is adopted as well. These considerations have an immediate link with statistical consultancy. Firstly, statistical consultants, just like industrial researchers, benefit from well developed, simple and robust statistical methods. Although consultants will have the ability to keep up with the latest developments in their field, they often simply lack the time. Secondly, statistical consultation is statistical education, a topic we will discuss later in this section. When simple methods have been used in the analysis of a dataset, the analysis results can be more easily explained by the statistical consultant to his or her client, and this indeed strongly increases the acceptability and the effect of the advise that has been given.

Finally, we would like to discuss some issues we observe in the use of statistics by users who are not trained as statisticians, but who did receive training in statistics, and who use statistical methods in their work.

Most users of statistical methods tend to stick to the methods they were taught during their training, and it is difficult for non-statisticians to learn about new developments in statistics, because information about these developments is usually published in journals and conference proceedings outside their own domain. A solution here could be that statisticians publish their methods not just in their own publication channels, but also in journals of other disciplines. Such publications should then focus on the application of their methods. Furthermore, one can think of handbooks and websites, targeting users in various domains, to which statisticians contribute. Some examples of such publications are already available (e.g., Bailey 1996, Friendly 2004, NIST/SEMATECH 2006).

What one also sees is that particular statistical methods tend to be used in one discipline, and not at all in others. In some cases disciplines could really benefit from the adoption of methods common in other disciplines. A classical example is design of experiments methodology, widely applied to industry that has its origin in agriculture. Other examples are generalized linear models from biostatistics, and multivariate statistics from psychometrics, both applied in industrial quality improvement.

In addition, software packages (like SPSS; see SPSS 2005) that support statistical analysis pose a particular problem . On the one hand, these tools claim to make it easier for users to conduct statistical analysis. On the other hand they are restrictive in the sense that users are not likely to look for alternatives, but stick with the methods offered in the package. In addition, for most users such packages will be like a black box: data goes in as input, and comes out as test results, but not much guidance is provided, so such packages will not help users in understanding exactly what is happening to their data (and whether or not the operations on the data are allowed).

At a more general level, the adequate use of statistical methods, or even the availability of suitable methods, strongly depends on the proper design of the study. In most cases, users do not give the statistical analysis of the results much thought until they actually have collected the data; sometimes this might result in data sets that in the end turn out to be very difficult to analyze.

In order to support users in the selection and correct use of statistical methods to analyze their data set, one could think of creating an aid, e.g., in the form of an expert system, which would provide more support than regular statistical software packages. Indeed, a number of efforts to develop such systems have been reported in the literature: HaKong and Hickman (1986), Sandals and Pyryt (1992), Capiluppi and Fabbris (1994), and Grabowski and Harkness (1996). It seems though that in more recent years interest in expert systems has dropped somewhat. In any case, with or without an expert system, the core demand of users will still be statistical methods that are usable and functional, and developed according to a user-centered approach, as described in this paper.

Apart from developing user-friendly methods, statistical education to teach these methods to students is an aspect that requires consideration. A proper education and training can encourage students to study this topic for at least the statistics they need for their research. The problem is, however, that two popular methods of statistics training are both wrong: drowning the student in hard-core mathematics and forcing him to understand information that is of little relevance to the way the student is going to use statistics, and, the opposite, only giving cookbook recipes and SPSS commands (the authors have seen examples of both approaches). Neither method is likely to leave warm feelings for statistics in the hearts of most of the students for whom statistics is not the major field of study. Viewpoints regarding teaching methods such as problem-based learning (Boyle 1999), where students are encouraged to understand and apply statistical methods in the context of real life problems, preferably with their own data sets, looks a much more effective way to create a basis for statistical thinking. Boyle (1999) showed that students in a statistics course that had adopted problem-based learning, and the use of real cases, improved their understanding of statistical methods, and were better able to apply the correct methods given a particular case.

The need for support in understanding and conducting statistical analysis is high, as is the demand for easy to use statistical methods – but to many, statistics still has a reputation of being overly complex, difficult, and even scary. This is made clear in numerous publications, with such titles as *Statistics without Tears* (Rowntree 1988), *SPSS Made Simple* (Kinnear and Gray 2006), and *Usable Statistics* (Sauro 2007). Such publications indicate that many practitioners feel overwhelmed by statistics and data analysis or confused by the manner in which results of tests are presented and interpreted. But these titles also indicate that there is a clear demand for statistical methods that are not primarily driven by technology, but that take the users and their needs as a starting point. This is exactly what a statistical consultant should do, carefully listen to his or her clients, determine their needs and create methods to fulfill those needs.

It is our experience that statistics development and use in industry highly benefit from close and frequent contacts between statisticians and users. A recent example is Rajae and Engel (2005), describing the development of a practical method for the analysis of paired comparison data, with an implementation in SPlus software, as a result of joint work of user and statistician.

Acknowledgement The authors wish to thank the two referees for their suggestions and remarks that improved this paper.

References

Agresti, A. (2002) Categorical data analysis, 2nd edn. Wiley, New York

Bailey, R.W. (1996) Human performance engineering: designing high quality professional user interfaces for computer products, applications and systems. Prentice Hall, Englewood Cliffs

Baker, B.O., Hardyck, C.D., Petrinovich, L.F. (1966) Weak measurements vs. strong statistics: an empirical critique of S.S. Stevens' proscriptions on statistics. Educational and Psychological Measurement **26**, 291–309

Beyer, H., Holtzblatt, K. (1997) Contextual design: a customer-centered approach to systems designs. Morgan Kaufmann, San Francisco

Boyle, C.R. (1999) A problem-based learning approach to teaching biostatistics. Journal of Statistics Education **7**. http://www.amstat.org/publications/jse/secure/v7n1/boyle.cfm. Accessed 31 Aug 2007

Capiluppi, C., Fabbris, L. (1994) STATREE: an expert system for choosing suitable statistical data processing techniques. In: Brunelli, L., Cicchitelli, G. (eds.) Proceedings of the 1st Scientific Meeting (of the IASE) at the University of Perugia. International Statistical Institute, Voorburg, The Netherlands, pp. 229–235

Chowdhury, S. (2002) Design for six sigma. Dearborn Trade, Chicago

Diederiks, E.M.A., Hoonhout, H.C.M. (2007) Radical innovation and end-user involvement: the Ambilight case. Journal of Knowledge, Technology & Policy **20**, 31–38

Friendly, M. (2004) Statistics and Statistical Graphics Resources. http://www.math.yorku.ca/SCS/StatResource.html. Accessed 31 Aug 2007

Grabowski, B.L., Harkness, W.L. (1996) Enhancing statistics education with expert systems: more than an advisory system. Journal of Statistics Education **4**. http://www.amstat.org/publications/jse/v4n3/grabowski.html. Accessed 31 Aug 2007

HaKong, L., Hickman, F.R. (1986) Expert systems techniques: an application in statistics. In: Proceedings of the 5th Technical Conference of the British Computer Society Specialist Group on Expert Systems, pp. 45–63. University of Warwick, Warwick, UK

Hoekstra, R., Finch, S., Kiers, H.A.L., Johnson, A. (2006a) Probability as certainty: dichotomous thinking and the misuse of p-values. Psychonomic Bulletin & Review **13**, 1033–1037

Hoekstra, R., Kiers, H.A.L., Johnson, A., Groenier, M. (2006b) Problems when interpreting research results using only p-value and sample size. In: ICOTS-7 International Conference on Teaching Statistics: Working Cooperatively in Statistics Education. Salvador, Bahia, Brazil

International Organization for Standardization (ISO) (1998) Ergonomic requirements for office work with visual display terminals (VDTs), part 11: Guidance on usability, ISO report 9241-11. ISO Central Secretariat, Geneva, Switzerland

Kinnear, P.R., Gray, C.D. (2006) SPSS 14 made simple. Psychology, Hove, UK

Labovitz, S. (1967) Some observations on measurement and statistics. Social Forces **46**, 151–160

Meerbeek, B., Hoonhout, H.C.M., Bingley, P., Terken, J. (2006) Investigating the relationship between the personality of a robotic TV assistant and the level of user control. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06), pp. 404–410. Hatfield, UK

Nielsen, J. (1993) Usability engineering. Morgan Kaufmann, New York

NIST/SEMATECH (2006) NIST/SEMATECH e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/. Accessed 31 Aug 2007

Pelgrim, P.H., Hoonhout, H.C.M., Lashina, T.A., Engel, J., IJsselsteijn, W.A., de Kort, Y.A.W. (2006) Creating atmospheres: the effects of ambient scent and coloured lighting on environmental assessment. In: Proceedings of the design & emotion conference. Chalmers University of Technology, Göteborg, Sweden

Rajae-Joordens, R., Engel, J. (2005) Paired comparisons in visual perception studies using small sample sizes. Displays **26**, 1–7

Rowntree, D. (1988) Statistics without tears: an introduction for non-mathematicians. Penguin Science, London

Sandals, L.H., Pyryt, M.C. (1992) New directions for teaching research methods and statistics: the development of a computer-based expert system. In: Annual conference of the american educational research association, San Francisco, CA

Sauro, J. (2007) Usable Statistics. http://www.measuringusability.com/stats/index.php. Accessed 31 Aug 2007

Shackel, B. (1991) Usability – context, framework, design and evaluation. In: Shackel, B., Richardson, S. (eds.) Human factors for informatics usability, pp. 21–38. Cambridge University Press, Cambridge

SPSS (2005) Statistical Package for the Social Sciences, release 14.0. SPSS Inc., Chicago

Taguchi, G. (1986) Introduction to quality engineering: designing quality into products and processes. Asian Productivity Organisation, Tokyo

Zijlstra, W. (2004) Comparing the Student's t and the ANOVA contrast procedure with five alternative procedures. Master's Thesis, University Groningen. http://www.ppsw.rug.nl/~kiers/ReportZijlstra.pdf. Accessed 31 Aug 2007