

## TEACHING STATISTICAL THINKING TO LIFE SCIENTISTS A CASE-BASED APPROACH

**Philippe Vandebroeck**

*WS CVBA, Brussels, Belgium*

**Luc Wouters**

*Barrier Therapeutics NV, Geel, Belgium*

**Geert Molenberghs**

*Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium*

**Jef Van Gestel and Luc Bijens**

*J&JPRD, Janssen Pharmaceutica, Beerse, Belgium*

*We describe a workshop on statistical thinking for scientists involved in pharmaceutical discovery research. The objectives were 1) to improve the quality of research data by developing a structured approach to bias and variability and 2) to establish a collaborative and informed relationship between scientists and statisticians by broadening their common basis. The corner stone was the introduction of statistical thinking and the didactical route to achieve this goal.*

**Key Words:** Bias; Case study; Learning process; Pharmacological research; Statistical education; Statistical thinking; Variability.

### 1. INTRODUCTION

This article describes the concept of an innovative workshop to improve statistical thinking skills of research scientists in pharmaceutical discovery research. The target audience of the workshop was pharmacologists, but we believe the concepts presented here are applicable to a broad range of life sciences settings in which interaction between substantive scientists and statisticians is mandatory. The workshop focuses on seven general principles of statistical thinking, brought to life in a case study: 1) Time spent thinking on the conceptualization and design of an experiment is time wisely spent; 2) The design of an experiment reflects the contributions from different sources of variability; 3) The design of an experiment balances between its internal validity (proper control of noise) and external validity (the experiment's generalizability); 4) Good experimental practice provides the clue to bias minimization; 5) Good experimental design is the clue

Received January 1, 2005; Accepted September 1, 2005

Address correspondence to Luc Wouters, Barrier Therapeutics NV, Ciplastraat 3 B2440, Geel, Belgium; E-mail: lwouters@barriertherapeutics.be and Philippe Vandebroeck, WS CVBA, Dikke Beuklann 92 B1090, Brussels, Belgium; E-mail: philippe.vandebroeck@ws-network.com

to the control of variability; 6) Experimental design integrates various disciplines; 7) A priori consideration of statistical power is an indispensable pillar of an effective experiment.

In Section 2 “statistical thinking” is introduced and contrasted with other ways of conceptualizing this skill. In Section 3, the workshop format is elucidated. Section 4 elaborates on the seven principles mentioned above. Section 5 is devoted to the case study. In Section 6, we present an assessment of this learning approach and our conclusions.

## 2. CONCEPTS OF STATISTICAL THINKING

We propose that the statistical thinker is a creative and confident professional, who is a diagnoser, intermediary, and communicator.

Although the concept of statistical thinking has seen increasing circulation over the last decade, there is no generally accepted definition. Chance (2002) states that “statistical thinking processes clearly involve, but move beyond, summarizing data, solving a particular problem, reasoning through a procedure and explaining the conclusion.” He states its defining characteristic “is the ability to see the process as a whole, including ‘why,’ to understand the relationship and meaning of variation, to have the ability to explore data in ways beyond what has been prescribed in texts.”

According to Garfield et al. (2003) “statistical thinking involves an understanding of why and how statistical investigations are conducted and the ‘big ideas’ that underlie statistical investigations. (...) Statistical thinking also includes being able to understand and utilize the context of a problem in forming investigations and drawing conclusions, and recognizing and understanding the entire process (from question posing, over data collection and choosing analyses, to testing assumptions, etc.).”

Following Chance and Garfield et al., the statistical thinker is a professional distinguishing himself by a sovereign command of the technical issues and a particular capacity for conceptual thinking. He is a confident, creative person, able to see the big picture, to outline research strategies, and to support them with a wide array of statistical methodologies, focused and effective. We expect statistical thinkers to be found in positions at all levels of research, including the managerial levels (Moore, 1990).

Our own conceptualization emerged from a decade of experience in persuading industrial and pharmaceutical researchers to be attentive to statistical principles (Vandenbroeck and Vandevyvere, 1996). The following definition is proposed: statistical thinking is a generalist skill that is focused on the application of nontechnical concepts and principles with the aim to better understand how statistical methods can contribute to finding answers to specific research problems and what the implications are in terms of data collection, experimental setup, data analysis, and reporting. From there, he is able to postulate which statistical expertise is needed to enhance the research project’s success. In this capacity, the statistical thinker acts as a *diagnoser*.

Although the “big picture” is important, it is not the only element needed to successfully act as an independent, confident source of professional expertise.

A research project's success is usually defined as a trade-off between various performance measures, e.g., resource economy, statistical power, pharmacological relevance. The statistical thinker has the ability to integrate these potentially competing priorities into a coherent and statistically underpinned research strategy. Statistical thinking is a practice fully integrated with the researcher's scientific field, not merely an autonomous science. Hence, the statistical thinker is deeply involved in applied research, with a good working knowledge of the substantive science.

As such, he is an *intermediary* between scientists and statisticians, which does not preclude that he belongs to either one of these professional disciplines, with sufficient background in the other.

These skills combined lead to efficiency, important to increase the speed with which research data, analyses, and conclusions become available, as well to enhance the associated quality, and to reduce the associated cost. Statistical thinking then helps the scientist to build a case and negotiate it on fair and objective grounds with those in the organization seeking to contribute to more business-oriented measures of performance. In that sense, the successful statistical thinker is a persuasive *communicator* (Abelson, 1995).

This definition is implemented through the principles and the workshop, which follows. These principles apply to the specific learning situation described here, but arguably also more broadly across a range of life sciences applications, pharmaceutical and other. The reader may wonder whether there is a connection to Six Sigma and Process Excellence and the concepts presented here. Although the former can be applied throughout all types of processes, but to a large extent to well-established ones, our approach is explicitly relevant in a research context, requiring tailor-made solutions.

### 3. DESIGN OF A LEARNING EVENT

In this section, we discuss our approach to designing the learning event. The aim was to help scientists expand their knowledge of the principles of statistical thinking. First, relevant characteristics of this particular research organization as they emerged from a thorough statistical needs assessment are discussed. This then leads to the workshop's design.

#### 3.1. The Discovery Environment

A series of 13 one-to-one interviews with researchers led to a better understanding of the organization and revealed the scientists' expectations of the role of statistics in their work. Our findings are rooted in this specific organization. It may be worthwhile to undertake their validation in similar but different organizational contexts. Four key points summarize our findings.

**3.1.1. Discovery is a complex research environment.** The drug discovery lab is the start of the pharmaceutical research chain (Spilker, 1989). Here, new chemical compounds are developed, screened, and tested for biological activity. When a compound is shown to have potential as a new drug, it is transferred to the drug development stage for in-depth analysis, first in additional animal studies,

later in humans. Drug discovery is not easy to formalize. A certain amount of “disorder” fosters creativity. Because of technological advances, the complexity of drug discovery has dramatically increased over the last 15 years. The result is an open and fluid research environment, with a great variety in research problems and strategies. This asks for a flexible and eclectic approach in statistical support strategies.

**3.1.2. Different strategies are used to control variability along the discovery chain.** There is a gradually changing perception of “effect” and “variability” along the discovery chain. On the one hand, researchers at the exploratory end of the chain look for strong effects only; it is important to discover potential early on. People working at the development side, on the other hand, deal with a variety of responses in a more complex experimental setting. They perceive their error margin as much smaller implying a more scrupulous data handling. Thus, Type I error tolerance is much larger at early stages: falsely identifying a compound as being active is less serious at that stage, unlike later in the chain, where continuing with a nonactive compound is much more serious.

As a consequence, variability (error) control is gradually more important at later stages. At such stages, there are two prevailing strategies: 1) controlling variability by (optimal) design and 2) accounting for variability by appropriate statistical analysis methods. (Of course, this is not the only problem to be tackled. For example, Type II error control is of the utmost importance as well. However, completeness is beyond the scope of this article.)

**3.1.3. A biological rather than a statistical rationale underpins research strategies.** Generally, scientists guided their experiments on biological rather than statistical grounds, the so-called “clinical model,” i.e., “a model in which normative biology or behavior can be studied, or in which a spontaneous or induced pathological process can be investigated, and in which the phenomenon resembles in at least one respect the same phenomenon in humans” (Van Zutphen et al., 1993, p. 191).

The development of an appropriate clinical model is art and science because “there are no rules regarding the choice of a proper animal model, nor are there rules for the extrapolation of results from the model to another animal species or man” (Van Zutphen et al., 1993, p. 191). Acceptable variability and clinical relevance are important, although not the only, desirable attributes. Researchers cited ease of induction, extent of documentation, sensitivity of targets, activity of reference compounds, and throughput as other requirements, although their relative importance may be graded differently, sometimes leading to fierce debate about a model’s validity and effectiveness. The process of model refinement is a multicriteria selection problem with several independent variables, e.g., source of raw materials and experimental animals, gender of animals, housing conditions, number of injections, dosage of compound, and type of cell line.

**3.1.4. Scientists may have polarized opinions on the benefits of statistics.** The use of statistics was a controversial issue in the lab, with strong and opposing opinions, ranging from pleas for a more widespread adoption

of statistical principles to the perspective that statistics merely drains time and resources. Regardless of their sentiment toward statistics, many held a restricted view on the role of statistics, defining it as a collection of techniques to confirm or reject hypotheses once the data were there. This confirmatory interpretation obscured other uses of statistical principles, such as graphing, exploring data, and experimental design. There was little interest for the principles of sound statistical design and analysis of experiments. For example, the concept of “experimental unit” was often left unspecified, as were statistical model assumptions and the balance between economy and information in experimentation.

### 3.2. Objectives of the Workshop

The interviews confirmed the need for and an interest in statistical thinking, underscoring that a statistical learning event would need to be carefully positioned in terms of target audience, format, and content, to minimize resistance. This called for an intervention that would lower barriers and remove unhelpful preconceptions, using an interactive, case-based and nontechnical route. The emphasis had to be on the thought process rather than on technicalities.

The primary objective of the workshop, therefore, was to improve the statistical thinking skills of researchers: 1) to enable assessment of the potential benefits and costs of having them guided by a statistical rationale in their work; 2) to help them maximize the benefits of the available expert statistical support in the laboratory; 3) to help them build and defend their case in shaping their research programs against the background of responsible financial resource management and business-oriented measures of performance. In addition, boundary conditions, as laid out in Table 1, were set.

Note that our definition states the statistical thinker should also have a sufficient working knowledge of statistical methodology to be effective. This is a process to be initiated, but clearly not completed by a single workshop.

**Table 1** Additional requirements for a workshop focusing on dealing with bias and variability in laboratory experiments

Condition	Remark	Implication
<i>Duration</i>	The event should not last longer than 2 days	Limited duration affects the learning objectives and the didactic approach adopted by the process
<i>Throughput</i>	At least 10 researchers should be able to participate in the event at the same time	Group dynamics and learning capacities are considered optimal in groups of 10–16 people
<i>High involvement</i>	The events needs to mobilize the energy and enthusiasm of the participants	
<i>Evolutionary</i>	The process needs to be able to evolve as a function of changing learning requirements	
<i>Hard deliverable</i>	The process needs to be supported by a creative tool kit	
<i>Budget</i>	The cost to design process and artefacts should not exceed the budgetary limits	

### **3.3. Format of the Learning Event**

The key workshop design challenge was to balance between sufficient possibilities for free-flowing and exploratory interaction between participants and adequate substance in terms of technical concepts to be transmitted. The process should not be too closed, with insufficient possibilities for participant interaction, but it should not be too open neither, because it would risk spilling over into softer issues such as organizational structures, relationships between managers and scientists, and strategic issues.

The learning event thus aimed for a format, sufficiently imaginative to generate participant energy, while getting the very specific learning messages across. This led to a 2-day course, combining a streamlined theoretical introduction to statistical thinking, with a more open and practical case, with interactive and game oriented components. We first discuss the key conceptual issues and then turn to the case study.

## **4. SEVEN PRINCIPLES OF STATISTICAL THINKING**

In selecting material for inclusion in the theoretical introduction of the course, we drew on our own experience as statistical practitioners supplemented with insights from key references (Abelson, 1995; Bechhofer et al., 1995; Box et al., 1978; Cox, 1958; Fisher, 1935; Hinkelmann and Kempthorne, 1994; Lindsey, 1999; Robinson, 2000; Ruxton and Colegrave, 2003; Santner and Tamhane, 1984; Selwyn, 1996). The course introduction focuses on the seven key principles outlined in the introduction. We discuss them in turn.

### **4.1. Time Spent on the Conceptualization and Design of an Experiment is Time Wisely Spent**

This principle, with its implications for cost and appropriateness of conclusions, is at the heart of the course. The statistical thinker is presented in the role model of smart researcher. Because we can structure the experimental nexus as a sequence of steps—definition, design, data collection, analysis and reporting—the smart researcher is someone who allocates a major part of his time budget to the first two steps.

With a good view on the architecture of the experiment, it is often possible to reduce the number of measurements to be taken, reducing the time spent in the lab and opposing the view of the “lab freak” that something interesting will come out if only lots of data are collected. The link between design and analysis is important too, opposing the view of the “data salvager” who believes that no matter how you collect the data, there is always a statistical fix-up at analysis time. Finally, the smart researcher can look ahead to the reporting phase with peace of mind, contrasting the view of the “novelist” who needs to spend a lot of time distilling a report from an ill-conceived experiment (Milliken and Johnson, 1984).

#### **4.2. The Design of an Experiment Reflects the Contributions From Different Sources of Variability**

One can get a grip on the thinking process surrounding the experimental structure by recognizing a fundamental similarity between experiments (Hinkelmann and Kempthorne, 1994, pp. 36–37):

$$\text{Response} = \text{Treatment Effect} + \text{Design Effect} + \text{Error},$$

which underscores that each experiment is based on a statistical model, implicitly or explicitly (Cochran and Cox, 1957).

The linear model decomposes the measured response to one or more sources of variability and random error. It confronts the scientist with the decision to either relegate sources of variability to the residual error term or to decompose the variability through treatment or other design factors. The treatment effects are associated to those experimental factors that the experimenter consciously manipulates to induce the response of the biological substrate or subject. The design effects, given the selected design, are determined by those sources of variability that can be identified but not consciously manipulated (Piergorsch and Bailer, 1997). Such elementary scaffolding plays a key role in making scientists more aware of the fact that thinking statistically is unavoidable, much less a choice. The model is the experimental setup's omnipresent statistical shadow.

#### **4.3. The Design of an Experiment Balances Between its Internal Validity (control of variability) and External Validity (generalizability)**

Each time a scientist develops a new experiment, he faces a dilemma. On the one hand, it is necessary to optimize the external validity of the experiment: conclusions need to be generalized from sample to population, calling for a sufficiently robust experimental setup, with respect to variation of the characteristics of the sample and settings. For example, the rationale behind an animal model is the ability to transpose experimental results from one biological system to another, a typical requirement in preclinical research. However, robustness with respect to populations and settings may jeopardize the researcher's ability to guarantee the internal validity of the experiment. An experiment is internally valid if, in the presence of treatment and/or design effects, observed response changes can be confidently attributed to these effects. It is only possible to guarantee this link if the "noise" or unfiltered random error in the experimental data is significantly smaller than the "signal," or variability due to treatment or design effects. In other words, to safeguard the internal validity of the experiment, the scientist needs to optimize the signal-to-noise ratio.

In most cases, balancing between external and internal validity will happen almost unconsciously. Pharmacologists have a natural leaning toward the biological relevance of their experimental layout. In adopting or updating an animal model, the requirement of generalizability to the target population is explicitly dealt with. However, scientists appear to be much less surefooted regarding internal validity, which is the next principle of statistical thinking.

#### **4.4. Good Experimental Practice Provides the Basis for Eliminating Bias**

A maximal signal-to-noise ratio is obviously desirable. Little can be done to enhance the signal apart from 1) an appropriate choice of the range of the experimental domain, neither too wide nor too narrow; 2) an adequate choice of the response metric, numerically if possible; and 3) the selection of precise measurement technology. Minimizing noise, made up of bias and variability, is more complex. Bias is a systematic deviation in observed measurements from the true value and variability refers to random fluctuation.

Emphasis is put on the fact that, compared with variability, bias is more damaging with respect to an experiment's ability to ensure internal validity. Therefore, it is essential to *minimize* and ultimately eliminate bias, whereas variability needs to be *controlled* and estimated (Montgomery, 1984). Fortunately, bias elimination is an issue of good experimental practice rather than of statistical technicalities. The correct application of blinding procedures, control measurements, standardization of experimental conditions, calibration of measurement instruments, and random sampling and randomization removes measurement, sampling, and investigator bias from the experimental data (Selwyn, 1996).

Most of these practices are well known to pharmacological researchers. However, the difference between random sampling and randomization is often not clear (Snedecor and Cochran, 1980). Random sampling from a population provides the foundation for the population model of inference. Although an easy concept, it is difficult to implement in practice, particularly in biomedical research. Randomization ensures that the effect of uncontrolled sources of variability has equal average effect in all treatment groups (Lehmann and D'Abrera, 1998). In other words, randomization is an operation that effectively turns lethal bias into more manageable random error (Lagakos and Pocock, 1990). It is important for scientists to understand that the only sensible location in the experimental nexus at which randomization needs to be applied is just prior to the administration of the treatment.

#### **4.5. Good Experimental Design Provides the Basis for Controlling Variability**

The experimental material is subject to various sources of variability. By carefully defining the experimental conditions, the researcher will be able to reduce the number of sources of variability without jeopardizing external validity (Piergorsch and Bailer, 1997). Some of the remaining variability will be associated with the treatment effects. Other sources, however, are not the result of the scientist's deliberate intervention strategy. By introducing design effects into the experimental model, the scientist is able to filter identifiable but uncontrollable sources of variability from the random error and hence to improve the internal validity (Montgomery, 1984). Assigning blocking factors (or covariates) is a way to do so. The basic idea behind blocking is to partition the total set of experimental units into homogeneous subsets (blocks), in that they respond more alike to uncontrollable source of variability (operator, batch, cage, etc.) This leads to a

better signal-to-noise ratio because the error is evaluated in each block so generated and then pooled over the whole experiment. There are various blocking configurations possible and following Hinkelmann and Kempthorne (1994, p. 54) we group them under the heading “error-control designs.”

Another conceptual distinction is the difference between experimental units and observational units. Initially, for many researchers, this is unclear and irrelevant (Hurlbert, 1984). However, this has major practical implications with respect to the experiment’s internal validity. The experimental unit corresponds to the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment. The observational unit is simply the unit on which the measurements are made. Often, but not always, they are identical. For example, in animal experiments the status of experimental unit is routinely assigned to individual animals, irrespective of the fact whether they were treated individually or not. When the experimental and observational units are different, proper reflection in the analysis strategy is in order. The situation is referred to as subsampling. An important feature of an error-control design with subsampling is that it allows for separation of experimental and observational error, which is useful when assessing the quality of the experimental and observational (measurement) procedures.

#### **4.6. Experimental Design Integrates Various Disciplines**

Researchers who have notions of experimental design tend to see it as a separate discipline. Many tend to approach the field in a “shopping” mode: from the range of typical designs that is presented in standard textbooks (Boniface, 1995; Montgomery, 1984; Sokal and Rohlf, 1981), they choose one as a template, for such reasons as convenience, accessibility, and experimental economy, in which the whole experimental setup is then indiscriminately squeezed. These kinds of intellectual shortcuts may lead to elegant experiments that often fail to answer the primary scientific question.

In this workshop, we clarify that experimental design can be viewed naturally as a synthetic approach to minimize bias and control variability. A researcher diligently applying the principles of statistical thinking will unwittingly and to their own surprise rediscover some classic designs without explicit recurrence to notions of experimental design.

An experimental design can be thought of as layered (Hinkelmann and Kempthorne, 1994): 1) the treatment design, determining the number and type of treatments; 2) the error-control design, separating sources of variability from random error; and 3) the subsampling design, allowing to distinguish between experimental and observational error. A completely randomized design is then simply a layout with a one-way treatment design. A randomized complete block design combines a one-way treatment design with a one-way block error-control design. A full factorial design combines a factorial treatment design with a completely randomized error-control design.

The choice of a particular design needs to be based on a variety of factors, prominently including resources and ability to establish a clear relationship between the measured response and relevant sources of variability.

#### **4.7. A Priori Consideration of Statistical Power is an Indispensable Pillar of an Effective Experiment**

Introducing an error-reduction design is a nontrivial step. As explained above, the purpose of adopting such factors is to separate identifiable but uncontrollable sources of variability from the error term. However, the relative efficiency of an error-reduction design to improve the signal-to-noise ratio depends on 1) the loss of degrees of freedom associated with the design factors and 2) the absolute size of the variability associated to these factors. Here, the researcher is faced with a potentially difficult question: Is it more advantageous to sacrifice degrees of freedom by using an error-reduction design compared to inflating the error term? This is the realm of statistical power calculations which opens up a new level of complexity in the practitioner's thinking (Barker-Bausell and Li, 2002; Cohen, 1988; Hintze, 1996; Machin and Campbell, 1987; Neter et al., 1990; Noether, 1987). Although conceptually simple, the notion of statistical power is rarely considered by most scientists. As a result, power calculation is often reduced to a technical exercise, ideally delegated to someone with appropriate expertise (Barker-Bausell and Li, 2002). Although scientists intuitively understand that power is predicated on the proposed sample size and the significance level which will be used to determine whether to accept the study's hypothesis, the link with a third basic parameter, the hypothesized effect size, is less well understood. This is nothing more than a standardized measure of the hypothesized mean difference(s) among the experiment's groups. The benefit of statistical power lies in its forcing researchers to think in terms of the strength of the effects their experiments are likely to produce. It is a crucial building block of the design process. An informed decision, for example, with respect to whether or not an error-control design is adopted can only be taken when it is clear what the resource constraints are for the experiment, what hypothesis it needs to confirm or not, at what significance level this has to happen and, not to be forgotten, how large the anticipated effect size is (Barker-Bausell and Li, 2002; Hintze, 1996; Machin and Campbell, 1987; Neter et al., 1990; Noether, 1987).

The above principles can be presented in a nontechnical way, with minimal reliance on technicalities and mathematical formalism. The seven principles do not exhibit a natural ordering. Researchers are likely to cycle through these principles during their design work in an iterative way, until they have been able to reconcile ambitions, resources and constraints.

In the workshop, the theoretical introduction is delivered by an external consultant who has played a leading role in the development of the learning event, to avoid opening the course with a long presentation delivered by a statistician in teaching mode. To create a more collaborative atmosphere between scientists and statisticians, a coaching mode is a key tool.

### **5. A CASE-BASED APPROACH**

The case study is the open part of the learning event, with opportunities for exploratory thinking and discussion. The challenge was to create a sense of balanced autonomy that would stimulate participants to think outside the box whilst providing enough support to work toward a viable solution strategy for the proposed research problem within the adopted time frame.

### 5.1. Balancing Openness vs. Guidance

The sense of openness was achieved in various ways. The case study avoided pharmacology-oriented research problems, to minimize the advantage of specialist biological knowledge and well-entrenched research practices. Hence, a case study was developed around a fictitious research problem in historically remote 17th century France. This engendered distance and unfamiliarity, allowing people to let go of their preconceptions. It also created a playful and relaxed atmosphere, reinforced by customized case materials, reflecting the historical period.

The problem is open-ended, with no single correct answer and various solution strategies possible. The challenge is to devise a persuasive solution strategy, rooted in the principles of statistical thinking. The result of the work is a process map, a high-level protocol, outlining the main steps in the experimental strategy and the logic behind them. Teams of five to seven researchers collaborate on the case for fairly long stretches at a time, enabling the team to settle down, find a working rhythm, and spend time digesting issues.

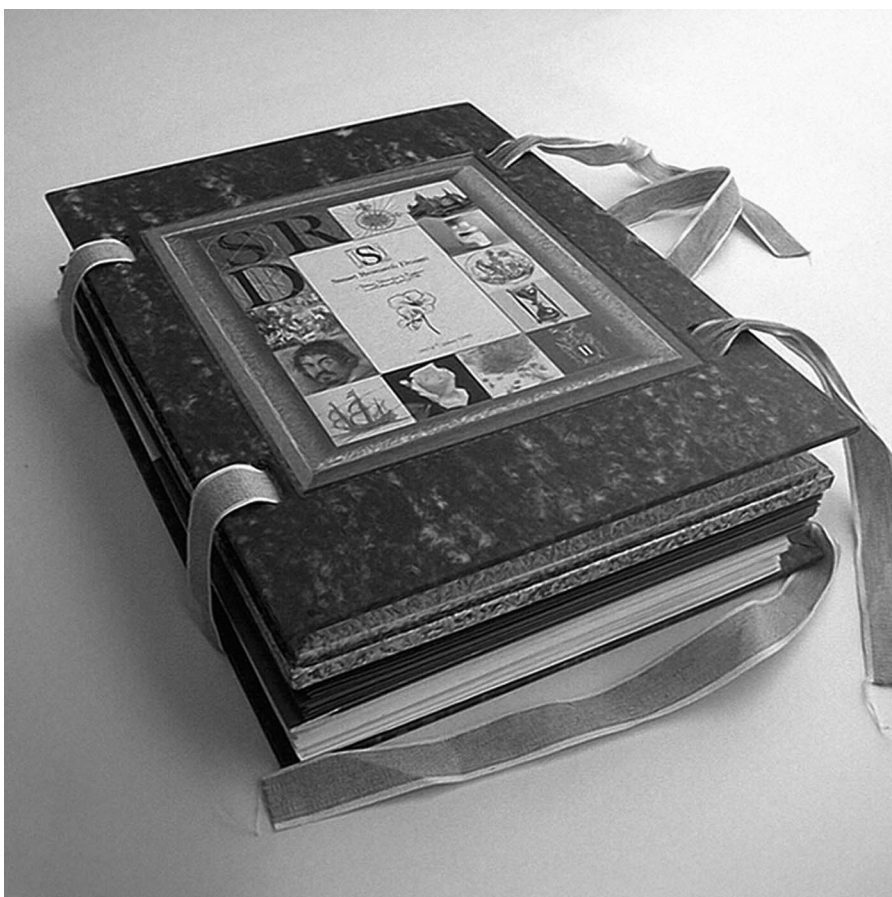
Openness is necessary to generate commitment, stimulate interaction and new ideas, but it needs to be balanced by a sense of direction to avoid energy diffusion and excessive divergence in solution strategies. This sense of direction is the result of various design decisions. 1) Researchers are stimulated to integrate principles of statistical thinking as much as possible in building their argument. So, there is a firm body of content, anchoring the teams' solutions strategies. 2) Each team receives a largely identical set of case materials. 3) Each team is supported by an expert from the laboratory's statistical support department, monitoring the discussion, challenging ideas, and helping the team out of blind alleys.

### 5.2. Description of the Case Study

The challenge posed by the case is to identify a chemical compound that will help in keeping cut roses fresh for as long as possible. To create a plausible context within which a harvested flower would be an interesting study object, the case is situated in the middle of the 17th century (Figure 1).

The background is the following. A young but wealthy nobleman invites a famous scholar and his assistant to his castle to experimentally determine how flowers can maximally be kept fresh, a problem particularly important because he had recently spent a significant part of his personal fortune on the purchase of a small collection of tulip bulbs (around the 1620s tulips commanded prices as high as 3000 guilders per bulb!). In particular, he wanted to impress and mollify the young Lady Y, daughter of one of his very influential peers, with an absolutely unique and long-lasting bouquet of flowers but, smartly enough, without completely surrendering his investment. For obvious reasons, the scholar is not allowed to experiment on the costly tulips themselves. Instead, the nobleman puts his castle gardens, dominated by roses, an earlier infatuation of his, at the disposal of the scientist (Figure 2).

In this apparently fanciful case, pharmacological scientists quite easily recognize a screening exercise to compare effectiveness of a range of chemical compounds in inducing a desired biological response. In addition, the need to extrapolate from the experimental organism (a rose) to a target organism (a tulip)



**Figure 1** The folder with the case materials.

is very similar to pharmacological research where results from cell, tissue, or animal experiments need to pave the way for tests on the human organism.

### **5.3. Working Through the Case**

Work on the case does not involve experimenting; hence no data are generated. The emphasis is on highlighting decision points in setting up the experiment and analyzing the resulting data. Typically, people spend about 40% of the total course duration (2 days) on case work. The case is documented by information sheets, which are collected in a handsomely designed folder. The sheets provide the team with information on the background of the research assignment, the composition of their compound library, preliminary indications on the potency of each of the library compounds, the layout of the castle grounds and the gardens, a rich source for blocking factors, and the available measurement instruments.

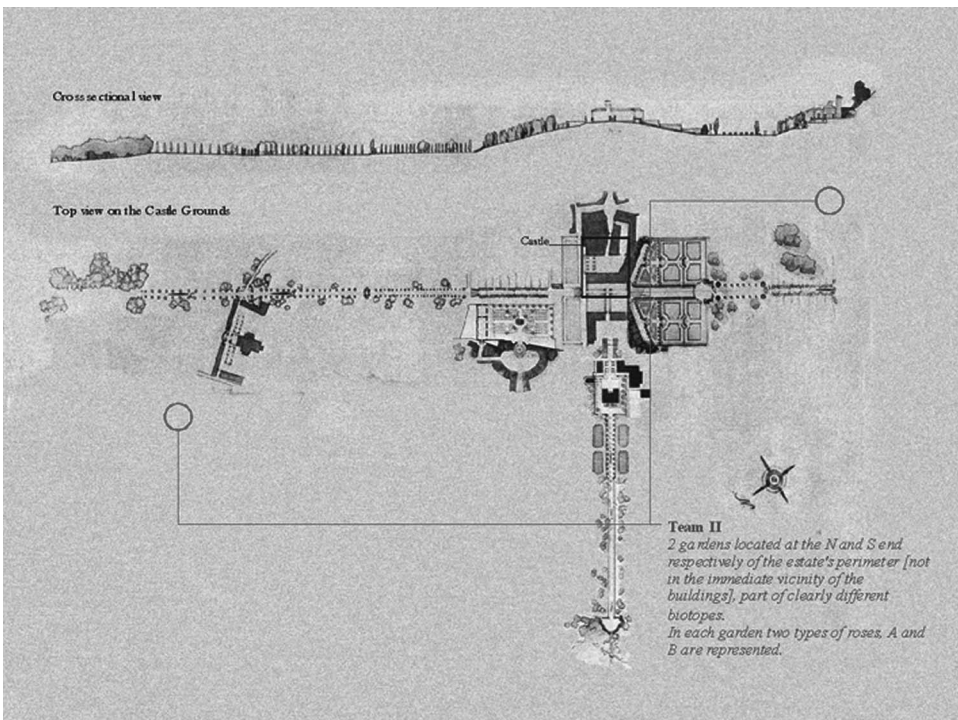
Case work is split into three assignments. The first assignment allows the teams to familiarize themselves with the case. The questions probe for a general

understanding of the context and force the participants to make explicit the sponsor's expectations and to examine the critical success factors for bringing the assignment to a successful conclusion. This assignment refers to the definition-phase of an experimental cycle.

A second assignment then focuses on key questions about the structure of the experiment: 1) formulating the hypothesis underlying the experiment; 2) operationalization of the response variable; 3) compiling an inventory of sources of variability; 4) basic structure of the experiment in terms of response and effects. Researchers need to indicate how they balance the dilemma between internal and external validity, relevant in this case because the target population (tulip) is a different biological organism than the experimental material (roses), against the background of resource constraints. To be sure, although it is clear from the assignment that resource economy is a desirable feature of the experiment, there is no rigid cap on the amount of experimental material used. Teams can, to an extent, negotiate for more resources with their sponsor in return for higher quality information.

A final assignment introduces the notion of experimental design and requests the teams to revisit their solution and indicate how the experiment is built up in terms of treatment, error control, and subsampling design.

The work on the case study is concluded by a presentation of each team's solution strategy with discussion.



**Figure 2** A map showing the location of the castle amidst the gardens.

## 6. ASSESSMENT AND CONCLUSION

The course “Smart Research Design” has now been taught to 75 researchers in seven sessions of 2 days to small groups of 10–12 participants, who have rated the learning experience very positively (an average overall satisfaction rate of 80% has been recorded in a survey involving 60 participants). The course has been perceived as innovative and relevant for their daily work. Apart from having acquired a new language within which to frame their experimental problems, participants report a renewed interest and confidence in the power of statistical methods to help achieve better results.

Improvements have been suggested as well. Not every participant appreciates the open character of the workshop. It remains a challenge to convince scientists that quality is not an objective, immutable characteristic of research data, but the result of contingent balance between resource requirements, physical boundary conditions, and a quest for a high signal-to-noise ratio, and needs to be reassessed whenever any of the constituting elements in the experimental context change.

In response to a desire for real-life evidence of the power of statistical thinking, the theoretical and general introduction is complemented with a shorter presentation in which typical research problems prominently feature, reassuring the participating scientists that the generic principles of statistical thinking can indeed be transposed to concrete laboratory contexts. In addition, it alerts them to the limits of statistical thinking and the need to interface at an early stage with a statistical expert to resolve the finer technical points.

From the vantage point of the statistical support unit, the course is a considerable success too. Many participants contact the unit for help. There is an increasing demand for more targeted statistical training. This evidence strengthens our confidence that, over time, a course like this provides a platform for building a research culture more receptive and less judgmental of the benefits of statistics in day-to-day discovery research and leading to a virtuous (as opposed to vicious) circle linking better cooperation between statistical experts and scientists to better designed experiments, to higher quality data, and ultimately to more confident researchers.

The interested reader can obtain more detail about the case study from the first author.

## REFERENCES

- Abelson, R. P. (1995). *Statistics as Principled Argument*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Barker-Bausell, R., Li, Y.-F. (2002). *Power Analysis for Experimental Research*. Cambridge: Cambridge University Press.
- Bechhofer, R. E., Santner, T. J., Goldsman, D. J. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. New York: John Wiley & Sons.
- Boniface, D. R. (1995). *Experiment Design and Statistical Methods for Behavioural and Social Research*. London: Chapman & Hall.
- Box, G. E. P., Hunter, W. G., Hunter, J. S. (1978). *Statistics for Experimenters*. New York: John Wiley & Sons.
- Chance, B. L. (2002). Components of statistical thinking and implications for instructions and assessment. *J. Stat. Educ.* 10(3).

- Cochran, W. G., Cox, G. M. (1957). *Experimental Designs*. 2nd ed. New York: John Wiley & Sons.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Cox, D. R. (1958). *Planning of Experiments*. New York: John Wiley & Sons.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Garfield, J., delMas, R., Chance, B. L. (2003). The web-based ARTIST: assessment resource tools for improving statistical thinking. Paper presented at AERA Annual Meeting, Chicago.
- Hinkelmann, K., Kempthorne, O. (1994). *Design and Analysis of Experiments*, vol. 1. New York: John Wiley & Sons.
- Hintze, J. L. (1996). *Pass User's Guide—Pass 6.0 Power Analysis and Sample Size for Windows*. Kaysville, Utah: NCSS.
- Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monographs* 54(2):187–211.
- Lagakos, S. W., Pocock, S. (1990). Randomization and stratification in cancer clinical trials: an international survey. In: Buyse, M., Staquet, M., Sylvester, R., eds. *Cancer Clinical Trials: Methods and Practice*. Oxford: Oxford Medical Publications.
- Lehmann, E. L., D'Abrera, H. J. M. (1998). *Nonparametrics—Statistical Methods Based on Ranks*. rev. 1st ed. Upper Saddle River, New Jersey: Prentice Hall.
- Lindsey, J. (1999). *Revealing Statistical Principles*. London: Arnold.
- Machin, D., Campbell, M. J. (1987). *Statistical Tables for the Design of Clinical Trials*. Boston: Blackwell Scientific Publications.
- Milliken, G. A., Johnson, D. E. (1984). *Analysis of Messy Data*. Vol. 1. London: Chapman & Hall.
- Montgomery, D. C. (1984). *Design and Analysis of Experiments*. New York: John Wiley & Sons.
- Moore, P. G. (1990). The skills challenge of the nineties. *J. R. Statist. Soc. A* 152:221–240.
- Neter, J., Wasserman, W., Kutner, M. H. (1990). *Applied Linear Statistical Models—Regression, Analysis of Variance and Experimental Designs*. 3rd ed. Homewood, IL: Irwin.
- Noether, G. E. (1987). Sample size determination for some common non-parametric tests. *J. Am. Stat. Assoc.* 82(398).
- Piergorsch, W. W., Bailer, A. J. (1997). *Statistics for Environmental Biology and Toxicology*. London: Chapman & Hall.
- Robinson, G. K. (2000). *Practical Strategies for Experimenting*. Chichester: John Wiley & Sons.
- Ruxton, G. D., Colegrave, N. (2003). *Experimental Design for the Life Sciences*. Oxford: Oxford University Press.
- Santner, T. J., Tamhane, A. C. (1984). *Design of Experiments—Ranking and Selection*. New York: Marcel Dekker.
- Selwyn, M. R. (1996). *Principles of Experimental Design for the Life Sciences*. Boca Raton, Florida: CRC Press.
- Snedecor, G. W., Cochran, W. G. (1980). *Statistical Methods*. 7th ed. Ames, Iowa: Iowa State University Press.
- Sokal, R. R., Rohlf, F. J. (1981). *Biometry—The Principles and Practice of Statistics in Biological Research*. 2nd ed. New York: W.H. Freeman and Company.
- Spilker, B. (1989). *Multinational Drug Companies—Issues in Drug Discovery and Development*. New York: Raven Press.
- Vandenbroeck, P., Vandevyvere, P. (1996). Statistics in a new business environment: An example. *Statistician* 45(3):287–292.
- Van Zutphen, L. F. M., Baumans, V., Beynen, A. C. (1993). *Principles of Animal Laboratory Animal Science—A Contribution to the Humane Use and Care of Animals and to the Quality of Experimental Results*. Amsterdam: Elsevier.