Volume 25 (2), pp. 107–123
http://www.orssa.org.za

ORiON
ISSN 0529-191-X
©2009

# The role of statistics in operations research: Some personal reflections

L Paul Fatti[*]

*Received: 1 September 2009; Revised: 6 October 2009; Accepted: 8 October 2009*

**Abstract**

Statistics has a very important role to play in Operations Research (OR), yet many standard texts relegate it to a subordinate position behind, for example, deterministic optimisation techniques. I argue in this paper that statistics, and statistical thinking, are essential for the OR practitioner operating in the real, uncertain world. Using many examples from my own experience, I demonstrate that statistics is very useful, and indeed essential, when tackling real-world OR problems. Many of these examples have not been published before and will therefore hopefully be of interest in themselves, as well as illustrations of the points I wish to make.

**Key words:** Statistical thinking, problem formulation, Bayesian analysis.

## 1 Introduction

Most textbooks on Operations Research (OR) by which I include all synonyms such as Operational Research and Management Science, start with Linear Programming, its applications and extensions (Integer Programming, Nonlinear Programming, *etc.*) and only after that, usually about halfway through the book, get on to Probability and Statistics and their applications to OR. See, for example, Hillier and Lieberman [17], Taha [27], Wagner [28], and Winston [29]. (Exceptions to this are Ackoff and Sasieni [1] and Rivett [23, 24].) This approach sets the framework of OR firmly in the optimisation arena, within a deterministic space, and only later includes uncertainty as a possible aspect to take into account when approaching a problem.

Life, however, is not deterministic, and real problems mostly have to deal with uncertainty, especially as their solution invariably has to do with the future (why solve a problem that is in the past?). In this paper I will argue that in most problems tackled by operations researchers, uncertainty plays a major role. Even in cases where the system is essentially deterministic (as some engineering systems are) there will nevertheless be uncertainty about its parameter values, and parameter estimation falls firmly into the statistical arena.

[*] **(Fellow of the Operations Research Society of South Africa)**, School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, 2193, South Africa, email: fatti@stats.wits.ac.za

Statistics has developed tremendously over the past decade or two, and programs for implementing the new methods have followed in their wake, including those that require increasingly computer-intensive methods. Generalised Linear Models (GLIMs) have extended the Multiple Regression model to many other distributions besides the Normal, and corresponding link functions besides the linear. See, for example, McCulloch and Nelder [21]. Dobson [9] gives a good introduction to the subject and Agresti [2] discusses its application to count- and categorical data. The application of Bayesian methods has been considerably enhanced by the use of the computer-intensive Markov Chain Monte Carlo and related methods. See, for example, Press [22]. The Bootstrap, initially developed by Brad Efron, allows inference to be drawn by randomly re-sampling a random sample from the population a large number of times. Bootstrap inference can either be purely non-parametric or take place within a parametric framework. See, for example, Efron and Tibshirani [10]. Extreme value analysis for modelling extreme events, ranging from maximum wind-speeds to extreme movements on the stock market, has seen active development in recent years. Good references are Coles [8] and Beirlant, *et al.* [4]. A general reference to modern data mining methods and other recent statistical developments is Hastie, *et al.* [16]. While many of the new statistical methods will be useful additions to the Operations Researcher's toolkit, there is no substitute for careful consideration of the 'real' problem at hand and original thinking about its formulation.

This paper is a personal reflection on the important role that Statistics plays in OR, and indeed in problem solving in general. My ideas have developed from a long involvement in practical problems in the broad arena of OR and Statistics, and will be illustrated using case studies from my personal experience. Many of these studies have not been published before and will therefore hopefully be of interest in themselves, as well as illustrations of the points I wish to make.

## 2 Some of my earlier experiences in OR

Soon after completing a Masters degree in Statistics and OR in the late 1960s, I joined a consulting firm in London that was doing work on the application of OR in the then new areas of Health, Local Government and other areas that have subsequently become an arena for 'soft' OR. Equipped with my new toolkit of techniques, I looked around eagerly for problems to which I could apply them, but alas I found none, not even a simple LP. Much of my time there was spent discussing the structuring and formulation of problems [13], often developing special techniques for approaching them. In hindsight, being exposed to this environment was the best introduction to 'real' OR that I could have hoped for.

A problem area in which I got involved was that of planning for the recruitment and training of nurses for a hospital that was still being built. It was then that I became aware of the whole area of Manpower planning, and the book by Bartholomew [3], which approached the modelling of social systems through stochastic processes. This was a revelation to me, and opened the way for me to model the manpower system with new entrants, training, promotions and attritions being handled as a large interrelated stochastic process.

One of the problems that was doing the rounds of the firm at the time had to do with a large industrial company which had a regular monthly meeting for senior executives to discuss new projects and make decisions about them. In its simplest form, projects were discussed and then either accepted for implementation, held over for additional analysis and re-consideration at the next monthly meeting, or rejected. The company had noticed that some projects seemed never to leave the system but rolled around every month without ever being accepted for implementation or rejected. Some elementary data analysis by two colleagues seemed to confirm this and they were busy trying to analyse the decision processes at the monthly meeting in order to devise methods of cleaning up the projects that never left the system.

However, having been made aware of the application of stochastic processes to decision problems, it immediately struck me that the decision system could be modelled as a discrete-time Markov Chain, with two absorbing states (acceptance and rejection) and that such a system could not have projects that never left the system. A re-analysis of the data yielded estimates of the probabilities in the transition matrix and this allowed for a clearer understanding of the process and highlighted which aspects of the decision making system needed to be changed in order to improve its performance.

## 3 The importance of statistical thinking in problem formulation and data analysis

Statistical thinking may be loosely defined as the recognition that all, or mostly all, problems have aspects that are uncertain (or random) and that this uncertainty needs to be taken into account in any realistic solutions to these problems. Clearly, the nature of the problem and the environment in which it exists will determine the importance that statistics and probability should play in its solution.

Problem formulation requires original thinking, and the broader one's knowledge, the better able one should be at formulating OR problems, particularly difficult ones. It is my contention that a good knowledge of Statistics is an essential component of the OR practitioner's toolbox.

Some years ago, I was approached by a large distributor of hybrid seed maize which had recently installed a drying plant at their maize processing factory in order to bring the moisture content of 'wet' maize, delivered to the factory in very large trucks by independent producers, down to the required level of 12.5%. Producers were paid by the weight of maize delivered, reduced to this moisture content. Since the time that the new plant was installed, the factory had been experiencing a loss of maize, averaging at about 5%, but sometimes going up as high as 15%. My task was to find the source of this loss!

On my tour through the factory it soon became evident that losses could not be pinned on to individual consignments of maize, because after being weighed and its moisture measurement taken, a consignment entered the drying plant in a bin devoted to the particular hybrid type, and its identity was lost amongst the other maize consignments in the bin. The only link that was maintained with the individual consignments was through samples that were drawn randomly from them, which had their weight and moisture measured before going into the drying plant with the consignments, but kept separately and measured again at the end of the drying process.

Because of the losses they were experiencing, an investigation had already been conducted (as a result of which they increased the number of samples taken from each consignment from 1 to 4 and improved their record-keeping), but could not find any reason for these losses. On going through the company's records, the only useful data I could find was:

1. The reconciliation records for 15 production runs of different hybrid types, which recorded the total tonnages (corrected for moisture) which had been received by the processing factory from the drying plant and the tonnages (also corrected for moisture) for which the producers had been paid.
2. Data on 4 samples taken from each of 9 consignments, recording initial weight, initial moisture percentage before going into the drying plant, and the dry weight and dry moisture percentage immediately after coming out of the plant. Because the practice of taking four samples per consignment had only been instituted recently, none of these nine consignments had yet been processed through the plant.

## 3.1 Methodology

Before doing any data analysis, it is important that one thinks of the structure of the data and to take this into account in the analysis. For the 4×9 = 36 samples the obvious identity

$$\text{(initial weight)} \times (1 - \text{initial moisture}) = \text{(final weight)} \times (1 - \text{final moisture})\,.$$

holds amongst the four variables (assuming no losses occur). Therefore, for each sample, the relationship

$$\left(\frac{\text{final weight}}{\text{initial weight}}\right) \div \left(\frac{1-\text{initial moisture}}{1-\text{final moisture}}\right) = 1 \tag{1}$$

should hold, within the limits of sampling and measurement variability. For the 15 reconciliation results, the obvious result

$$\frac{\text{total kg of (dry) seed received from the drying plant}}{\text{total kg of seed, adjusting for moisture, for which the producers had been paid}} = 1 \tag{2}$$

should hold, if there were no losses (also within the limits of sampling and measurement variability). Summary statistics computed separately for these two ratios yielded the results in Table 1. The correspondence between the frequency distributions of these two ratios is remarkable, the more so since they refer to totally different sets of maize!

It is a known statistical fact that the ratio of two independent random variables is biased upwards. To take a specific example, if $X$ and $Y$ are independent random variables with the same mean $\mu$ and variances $\sigma_X^2$ and $\sigma_Y^2$, respectively, then a second-order Taylor expansion gives the approximate result

$$E\left[\frac{Y}{X}\right] \approx 1 + \frac{\sigma_x^2}{\mu^2} > 1.$$

Therefore, for both these ratios, one would expect that, if there were no losses, they would be greater than 1 on average. The fact that their sample means are both less than 1,

| Summary statistic | Ratio for samples | Reconciliation Ratio |
|---|---|---|
| Sample size | 36 | 15 |
| Mean | 0.9590 | 0.9508 |
| SD | 0.0404 | 0.0485 |
| Minimum | 0.8463 | 0.8557 |
| First quartile | 0.9422 | 0.9202 |
| Median | 0.9616 | 0.9515 |
| Third quartile | 0.9797 | 0.9896 |
| Maximum | 1.0316 | 1.0226 |

**Table 1:** *Summary of statistics for maize sample and reconciliation ratios.*

suggests either that there have been losses, as was claimed by the company, or that there were consistent biases in the measurements taken at the factory. Since the company had not found any source for the losses, I decided to pursue the second possibility.

### 3.2 A suggested cause for the downward biases

Because of the remarkable similarity between the frequency distributions of the two ratios (1) and (2) — not only are their respective means and standard deviations very close, but so also are their medians, quartiles, maxima and minima — it seems safe to assume that the downward biases in both these ratios do, in fact, come from the same source. Observations from the company's accounting system revealed that:

- The total mass of (dry) seed received from the plant was calculated from the input weight measurements, corrected using the moisture measurements.
- The total mass of seed for which the producers were paid, after adjusting for moisture, was calculated by multiplying the weight of each truck consignment by the quantity

  $$\text{final weight/initial weight} \tag{3}$$

  from the corresponding sample, including a possible minor moisture correction down to a dry moisture of 12.5%. The initial moisture measurement on the sample had no influence on this calculation.

This evidence seemed to point to the moisture measurements as being a likely cause of the downward biases in the ratios (1) and (2). The reasoning behind this conclusion is that the quantity in (3) appears in the numerator of (1), whereas it is involved in the denominator of (2). Therefore, if there was a systematic error in this quantity, it would affect the ratios (1) and (2) in opposite directions. However, since (1) and (2) have the same statistical distributions, it is unlikely that there could be anything but minor errors in this ratio. The measurement of the weight of the truck consignment is also unlikely to be a cause of the downward biases, since it does not play any part in the ratio (1).

It was therefore suggested to the company that the moisture measurements be investigated for possible biases, which could explain the downward biases in both ratios (1) and (2), and hence also the apparent losses experienced by the company. In other words, there was no loss of maize!

Shortly after the study was completed, the drying plant closed down as it was the end of the harvesting season. During the routine maintenance that was undertaken thereafter, special attention was paid to the moisture measuring equipment. After the plant re-opened, the losses seemed to have disappeared.

Lessons drawn from this study include:

- the value of drawing representative sample data from the system under investigation, even if the samples are small,
- using the sample distribution to draw inferences about the system, within the framework of an underlying model of the system, and
- the importance of statistical theory in drawing inferences (in this example, concerning the distribution of a ratio).

## 4 Problems where statistics provides the primary vehicle in the OR approach

By their nature, many problems are dominated by uncertainty, and as a result probability and statistics play a major role in their formulation and solution. A brief discussion on Decision Theory, followed by some examples from my personal experience, will illustrate the concepts involved.

### 4.1 Decision theory

The theory behind decision-making when outcomes are uncertain is couched within the Bayesian paradigm, which shows that optimal decisions are those that maximise expected utility [20]. Uncertainty is incorporated, initially via prior distributions, but which are generally updated using data, via Bayes' Theorem. Utility functions are usually estimated subjectively, although in many practical situations they can be assumed to be linear (at least over the range of financial consequences likely to be experienced).

Decision theory teaches us that when outcomes are uncertain, as is generally the case, then decisions that are likely to be best, on average, are the ones to search for. This requires an understanding of the statistical distribution of the outcomes and how this is affected by the different decisions. If there is more than one outcome of interest, then we need to understand the multivariate distribution of these outcomes and we should be able to find trade-offs between their different utilities [5].

### 4.2 Optimal strategies for search engine marketing

One of the ways in which the owners of search engines make their money is by offering sponsored links to websites via keywords, so that whenever someone uses the search engine to look up one of these keywords, the sponsored link comes up on the screen. If the user then clicks on this link to access the website, the owner of the website is charged a fee. In return, the user may buy something advertised on the website, in which case the owner receives a return from his investment in the form of a sale. The decision for the web-based advertiser is whether this form of advertising is worthwhile.

In order to analyse this decision, the advertiser needs the following information on the keyword:

- the cost per click (CC),
- the average (net) revenue per sale (R) — there is often more than one product advertised on a website,
- the conversion rate $p$, which is the proportion of clicks on the link that lead to a sale, and
- the earnings per click (EPC), which is the product of the conversion rate and the average revenue per sale.

A very simple profitability analysis for this keyword uses the following identities

$$\begin{aligned}\text{profit} &= \text{expected net sales revenue} - \text{click costs}\\ &= \text{no. sales} \times \text{expected revenue per sale (R)} \ - \text{no. clicks} \times \text{cost per click (CC)}\\ &= \text{no. clicks} \times (\text{conversion rate} \times \text{R} - \ \text{CC})\\ &> 0 \quad \text{if conversion rate} \times \text{R} > \text{CC}.\end{aligned}$$

A keyword is therefore expected to be profitable if conversion rate $p > CC/R$ or, equivalently, $EPC > CC$.

In order to estimate the conversion rate, one can count the number of clicks and the number of them that have led to a sale. The estimated conversion rate, call it $\hat{p}$, is then the ratio of the latter to the former. So the decision criterion for the advertiser is that the keyword is profitable if $\hat{p} > CC/R$; otherwise it is not.

This decision criterion is of rather limited use, as it depends on there having been a number of clicks and sales on that keyword — what if there have been no sales so far? A formal Bayesian approach would be to consider the number of conversions to be a Binomial$(p, n)$ random variable, where $n$ is the number of clicks on the website and $p$ the probability of a sale on each click (assuming independence between successive clicks). Since $p$ is an unknown random variable, we have to assume a prior distribution for it, and a common prior used in this situation (it is also the natural conjugate prior for the Binomial parameter $p$) is the Beta($\alpha$, $\beta$) distribution.

Now the expected value of the Beta($\alpha$, $\beta$) distribution is $\alpha/(\alpha+\beta)$, and if there have been $n_c$ clicks on this keyword and a resulting $n_s$ sales on them, then it is not difficult to show that the posterior distribution of $p$ is now a Beta($\alpha + n_s$, $\beta + n_c - n_s$) distribution with corresponding expected value (or posterior expectation) $(\alpha + n_s)/(\alpha + \beta + n_c)$. So the Bayesian decision criterion becomes:

> The keyword is profitable if the expected value of $p$, $E[p]$, is greater than $CC/R$. So, *á priori*, the keyword is profitable if $\alpha/(\alpha + \beta) > CC/R$, and *á posteriori*, (*i.e.* after the $n_s$ sales have been obtained from the $n_c$ clicks) the keyword is profitable if $(\alpha + n_s)/(\alpha + \beta + n_c) > CC/R$.

To put some (fairly realistic) numbers on this, suppose that the cost per click is R0.50 and that the net revenue per sale is R300. Thus the keyword is expected to be profitable if the conversion rate is at least 0.0017. Typical data from the search engine relating to all keywords in the particular marketing sector of interest, gave prior estimates of 0.2859 and 120 for $\alpha$ and $\beta$, respectively, with an expected conversion rate of 0.0024 or 0.24%. So, *á priori*, the keyword is expected to be (just) profitable. However, if, after 100 clicks on this website there have been two sales, then the posterior expected conversion rate would be 0.0104 or 1.04%, which would suggest that the keyword would be expected to be quite profitable.

This analysis represents only the start of the web-based advertiser's problem, because he has the option to change his 'bid price' on the keyword, which is the maximum price per click he is prepared to offer in an auction between other bidders for the same keyword. The higher the bid price, the higher up on the search page his advertisement is likely to appear, making it more visible to the user and thus increasing its 'click potential', and correspondingly decreasing it if the bid price is decreased. A full analysis of the web-based advertiser's problem (which will not be discussed further here) would seek to understand how the click potential responds to the bid price and hence to find the optimal bid price for the keyword.

This study illustrates how a simple analysis of a decision problem is enhanced by including a statistical model of the process and using Bayes' Theorem to incorporate data into the analysis towards finding an optimal decision strategy.

## 4.3 Inventory management

Whereas the OR student's first introduction to Inventory Theory is through the classic deterministic model, leading to the basic Economic Order Quantity and its variants, any realistic inventory system is heavily influenced by uncertainty (see, for example, Johnson and Montgomery [18]). Uncertainty enters through the demand process, which is generally random, as well as through the delivery lead-time.

An interesting practical case in my experience is the replenishment of Automatic Teller Machines (ATMs) by banks. In the situation where I was involved, the bank replenished its ATMs every night, but provision was made for additional 'emergency' replenishments in cases where an unusually high demand caused an ATM to run out of cash. The question from the inventory manager was how much cash to put in each ATM at each nightly replenishment. If the replenishment amount was too high, then the bank would incur excessive cash-holding and insurance costs, whereas too low an amount could lead to a costly emergency replenishment. The challenge was to determine an amount that balanced these two costs.

A simple probabilistic argument led to the expression

$$K = H\left(R - \frac{1}{2}E\left[X\right]\right) + V\left(1 - F\left(R\right)\right),$$

for the expected daily inventory cost at an ATM, where $H$ is the daily cash-holding cost, $V$ is the cost of an unscheduled replenishment, $R$ is the replenishment amount and $X$ is the random daily demand, with expected value $E[X]$, CDF $F(\cdot)$ and PDF $f(\cdot)$.

Elementary calculus then yielded the remarkably simple formula

$$R_{opt} = f^{-1}\left(\frac{H}{V}\right),$$

where $f^{-1}(\cdot)$ is the inverse function of the PDF, for the optimal replenishment amount for this ATM. Clearly, this formula will hold for any ATM experiencing the same distribution of daily demand (and cost structure). In practice, each ATM will have its own demand distribution, which will be influenced by the day of the week, day of the month and month of the year. So the next challenge was to find a distribution that could accommodate the variety of demand patterns experienced by the bank's many ATMs during the different times of the year.

The Gamma distribution, with probability density function

$$f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)},$$

where $\alpha$ is a shape parameter and $\beta$ is a scale parameter, is a flexible, right-tailed distribution which was found to fit the distributions of daily demand experienced by ATMs very well. For any ATM maximum likelihood estimates of the parameters $\alpha$ and $\beta$ could be obtained from the daily demand data over the previous month. However, these parameter estimates were found to be very erratic and resulted in highly variable estimates of the optimal replenishment amounts, so it was decided to model these parameters directly over the different ATMS and over the twelve months of the year, so as to obtain smoother estimates. The linear model $\ln(\alpha_{ij}) = \mu + r_i + c_j + \gamma_i \times j + \varepsilon_{ij}$ was therefore fitted to the logs of the individual estimates of the shape parameter, where $\alpha_{ij}$ is the shape parameter for the demand distribution at ATM $i$ during month $j$, $\mu$ is a general average, $r_i$ is the $i$th ATM effect, $c_j$ is the $j$th month effect, $\gamma_i$ is the monthly growth/shrinkage at ATM $i$, and $\varepsilon_{ij}$ is a random error term.

This model was applied separately to the different regions of the country, on the assumption that within a region the demand at different ATMs have similar monthly effects, except that they may have different growth (or shrinkage) patterns. Using logs makes this a multiplicative model, which was felt by the bank to be more realistic.

A similar model was fitted to the estimates of the scale parameter $\beta$. (Day of the week and day of the month factors were estimated from the raw demand data and these were used to 'de-seasonalise' the data before modelling. After modelling, these factors were re-applied to the distributions. 'Special days', such as those with sporting events, could be accommodated in a similar manner for the affected ATMs.) These models were refitted every month and the estimated parameters were then applied to the replenishment formula to give the optimal replenishment amounts for each ATM in the region for each day of the coming year.

This practical case study illustrates the application of statistical modelling to an optimal decision problem with uncertain outcomes. It required knowledge of distribution- and estimation theory as well as of linear modelling for the parameters.

# 5 Further problems where statistics has provided the primary vehicle for the OR solution

In my experience, it is always necessary, to keep one's 'statistical hat' firmly on when investigating OR-type problems. A few more examples in which I have been involved will illustrate the point in this section.

## 5.1 Valuating a property for housing bond purposes

One of the main bottle-necks in the provision of a home loan is the process of obtaining a value for the property. The traditional method employed by banks and other bond providers is to send an evaluator to value the property. This is both expensive and potentially slow, especially in times of high sales activity when valuators are busy. Since home buyers typically apply to more than one institution for a loan, and are often under time pressure to obtain one, the bank that is quickest at valuating properties enjoys a significant advantage over its competitors in securing customers.

Automated valuation is the process whereby such a value may be estimated on the basis of statistical models of house prices. This is generally much faster (and cheaper) than obtaining a value from a valuator, and can also be more accurate.

A very effective model for the purpose is the repeat-sale model, which estimates the house-price inflation index from the prices of houses that have sold twice (or more times) over, say, the last 15 years. It turns out that the logarithm of the ratio of the second price of a property to its previous selling price may be modelled using Multiple Regression, where the regression coefficients are functions of the annual inflation rates over the period between the first and the second sale.

Specifically,

$$Z_i = \log_e\left(\frac{y_{t_{i2}}}{y_{t_{i1}}}\right) = \sum_{k=1}^{T} a_k \delta_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $n$ is the number of houses that have been sold twice between the years of interest, 1 to $T$, where $0 \leq \delta_{ik} \leq 1$, and denotes the proportion of year $k$ lying between the first and second sales of house $i$, and where $y_{t_{i1}}$ is the first sale price of house $i$, which occurred at time $t_{i1}$ and similarly for $y_{t_{i2}}$. Clearly, $\delta_{ik} = 0$ if $k \notin \{t_{i1}, t_{i1}+1, \ldots, t_{i2}-1\}$, and it is not difficult to show that $a_k = \log_e(1 + r_k)$, where $r_k$ denotes the (unknown) house price inflation rate in year $k$.

Fitting this model to sales data obtained from the Deeds Office (and other sources) gives estimates of the regression parameters $a_k$ and hence of the annual house price inflation rates $r_1, \ldots, r_T$, between years 1 and $T$. Having estimates of the annual inflation rates allows for the current sale price of any property to be estimated, by applying these rates between its first sale date and the present time to its previous sale price. A 'confidence index' on this value may be estimated by logistic regression, which enables the bank to decide whether the predicted price can be accepted, or whether an independent evaluator needs to be called in.

## 5.2 Credit scoring and queueing theory

Credit scoring allows a bank or other credit granting institution to assess whether or not a customer is sufficiently creditworthy to be granted a loan. Being able to assess creditworthiness accurately is crucial for the financial health of a bank, as this allows it to maximise its loan book while keeping control of its default risk. Credit scoring models are generally based on the loan repayment records of previous customers, fitted to information supplied by the customer (such as salary, length of employment, residence type, *etc.*) as well as the customer's credit bureau record, using either Statistical Discriminant analysis or Logistic Regression.

Queueing theory has been applied successfully to modelling queues in banks, leading to rate-control policies in which the number of tellers is varied, depending the arrival rate of customers, so as to keep the probability of a person having to queue for more than a specified maximum time (say, 5 minutes) acceptably low. Tellers freed up during quiet periods can then be profitably occupied with back-office work. (See, for example, Gross and Harris [15].)

## 5.3 Optimal use of industrial gas

Many years ago I was approached by the owner of a small steel factory for advice about its usage of industrial gas for the heating of steel billets in the wire-making process. The tariff for gas varied considerably, depending on the smoothness of the usage of gas between and within days, the cheapest being for very smooth utilisation and the most expensive (by a factor of almost 2) when the utilisation was very variable. Since a single hour's (or day's) excess utilisation could affect the tariff for the whole of the next six-month billing period, it was imperative that the factory monitor its usage closely and closed down its process (with attendant high cost) as soon as the usage became too high.

Complex statistical modelling, based on the distribution of hourly and daily gas demand at the factory, was used to investigate the influence of different policies regarding switching off the process when the utilisation became too high, and enabled an optimal upper limit (as regards expected total cost) to be determined, above which the factory should be closed down. Additional modelling allowed for the effects of installing one or more large storage tanks to buffer the gas usage, from which the optimal number of tanks could be determined. (For more details, readers are referred to Fatti [11].)

## 5.4 Quality control in export coal

Quality management should be a central aspect of all business activity, be it in Industry, Commerce or Government, and the OR practitioner needs to be conversant with its basic philosophy and techniques. Control-charting is the most commonly used technique in the field, and is based on straightforward, yet powerful, procedures from statistical sampling- and distribution theory. However, many situations require more complex solutions, as illustrated in the following example from my personal experience.

The coal industry is one in which quality control is of crucial importance when producing coal for export purposes. The required quality of the coal is determined by the contact with the overseas buyer, which typically specifies minimum standards for quality variables (such as Net Calorific Value (NCV) or Ash Content) as well as the procedure by which conformance to these standards will be measured. Quality conformance is measured by an independent laboratory which randomly samples the coal as it is loaded onto the ship at the export harbour. Heavy penalties are levied on consignments that fall below the minimum standards, and if this happens frequently it could jeopardise the contract with the buyer.

Quality is managed by the coal producer via a gravity separation process in the wash-plant at the end of the coal mine's production process. Control at the wash-plant is exercised by varying the specific gravity (SG) of the heavy-liquid medium that is used to separate the lighter (higher quality) coal, which is exported, from the heavier (lower quality) coal, which is sent to the discard dump. Lowering the SG therefore improves the quality of the export coal and thus lowers the chance of incurring penalties, but at the cost of discarding a higher proportion of the mine's production. The control policy at the wash-plant has to strike a balance between these two criteria. However, after leaving the mine by rail to the export harbour, there are many sources of variability, outside the control of the mine, that influence the quality of the coal that is finally determined, not least of which is the sampling variability at the ship sampler and measurement variability at the independent laboratory. A further complication is the fact that, while the contract was specified in terms of NCV, it is the Gross Calorific Value (GCV) that was controlled at the wash-plant. The formula relating the NCV to the GCV includes a number of variables that themselves are subject to random variability.

By analysing all the sources of variability influencing the NCV as measured at the independent laboratory, and applying them to the NCV:GCV formula (for details, see Fatti and Stewart [12]) the target mean GCV for the wash-plant at the mine could be derived, so that the probability of any shipment from the export harbour incurring penalties would be limited to a specified value (typically 5%). The final step in the process was to derive a (Bayesian) control strategy for SG at the wash-plant, so as to maintain the GCV of the exported coal at its target mean, taking into account the variability in the GCV:SG relationship across the coal extracted from the different seams at the mine and blended before entering the wash-plant.

## 5.5 Experimental design towards optimising the performance of a hazardous waste plant

The manager at a recently commissioned plant for processing toxic water which leached off the dumps at a hazardous waste facility, wanted to find optimal performance conditions for the plant. Control over the plant's performance was exercised by varying the values of three process parameters. The investigation had to be carried out experimentally, as the influence of these parameters on the performance of the plant was unknown. Unfortunately, each experimental run, at fixed values of these parameters, took two weeks, as the plant's performance was very variable, so that meaningful readings required averaging over a relatively long period, and the plant took a long time to settle down when these parameters were changed at the beginning of a run. Furthermore, gradual fouling over time of the heat exchanger also affected the plant's performance, so this factor needed to be accounted for in the experimental design. The manager needed the results in six months, which effectively meant that only twelve experimental runs were possible!

After detailed discussion with the plant manager and the consulting engineer, it was decided to run the experiment as a $2^3$ factorial with centre-points at the start, two in the middle and one at the end of the experiment. The first half-fraction of the $2^3$ factorial was performed in runs 2 to 5 and the complementary half-fraction was performed in runs 8 to 11, with the remaining four runs all being centre-points — see, for example, Box, *et al.* [7].) This design allowed for possible linear effects of the three control parameters to be estimated, correcting for a possible decrease in performance due to the fouling of the heat exchanger. The centre points would also allow for the estimation of curvature, which in principle would enable the optimum operating conditions to be estimated. (In the end, the results of the experiment were not conclusive, but a tentative estimate of the optimal operating conditions could nevertheless be obtained.)

## 6 Problems where statistics provides a unique insight

Statistics can provide unique insights into situations and phenomena, which may not be intuitively evident, and sometimes may even be counter-intuitive. A well-known example is the solution to the Birthday Problem, from which it emerges that in a class of only 23 people the chance of at least two having the same birthday is 50%, and if the number goes up to 40 then the chance increases to 89%, despite there being very many more possible birthdays. Two further examples are briefly discussed here.

### 6.1 Testing whether an athlete has used a banned substance

Suppose that the test for a particular banned performance-enhancing substance used by athletes is 99% accurate, in that if an athlete has been using the substance, then there is a 99% chance of this being picked up by the test (sensitivity), and if he or she has not been using the substance, then there is also a 99% chance of this being confirmed by the test (specificity). Historically 3% of athletes have been guilty of using this substance.

Suppose that an athlete has tested positive for the substance. Then the probability of him actually being guilty may be obtained from Bayes' Theorem. More specifically,

$$\begin{aligned}\Pr[\text{guilty} \mid \text{positive result}] &\propto \Pr[\text{positive result} \mid \text{guilty}] \times \Pr[\text{guilty}]\\ &\propto 0.99 \times 0.03 = 0.0297,\end{aligned}$$

while

$$\begin{aligned}\Pr[\text{not guilty} \mid \text{positive result}] &\propto \Pr[\text{positive result} \mid \text{not guilty}] \times \Pr[\text{not guilty}]\\ &\propto 0.01 \times 0.97 = 0.0097.\end{aligned}$$

Therefore,

$$\Pr\left[\text{guilty} \mid \text{positive result}\right] = \frac{0.0297}{0.0297 + 0.0097} = 0.7538.$$

So despite the 99% accuracy of the test, the probability that an athlete who has tested positive for the drug is actually guilty of taking it, is only 75%!

This illustrates the fact that tests for rare conditions need to be extremely accurate to be of any value at all. It also emphasises the need for confirmatory testing if a positive drug result has been found — reporting the result without first confirming it with additional testing is irresponsible.

## 6.2 Regression to the mean

The regression to the mean phenomenon, first reported by Francis Galton [14] may be stated as follows, in terms of the heights of fathers and of their (adult) sons: "The sons of tall fathers tend to be tall, but not as tall as their fathers, and the sons of short fathers tend to be short, but not as short as their fathers." However, Galton wrongly termed this phenomenon as: "regression towards mediocrity," implying that the heights of successive generations of men would tend towards the mean. Instead, it is easy to show that this phenomenon is actually a necessary consequence of the distribution of heights being *stable* over successive generations.

This is easily demonstrated via the Bivariate Normal distribution (although it holds more generally). If random variables $X$ and $Y$ have a Bivariate Normal distribution with respective means and standard deviations $\mu_X, \mu_Y, \sigma_X$ and $\sigma_Y$, and correlation coefficient $\rho$, then the conditional mean of $Y$, given that $X$ has the value $x$, has the well-known linear form

$$E\left[Y|X=x\right] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\left(x-\mu_X\right).$$

Now consider the example where $X$ is the height of a father and $Y$ is the height of his (adult) son. Assume that their distributions are stable over generations, implying that the heights of fathers and sons have the same mean ($\mu$) as well as the same standard deviation ($\sigma$). Substituting this into the above formula and rearranging terms, yields the simple formula

$$E[Y|X=x] - \mu = \rho\left(x-\mu\right)$$

relating the son's expected height to the actual height of his father. Noting that the correlation coefficient $\rho$ between fathers' and sons' heights is positive (Galton estimated it to be about two-thirds), this formula implies that the son of a tall father (*i.e.* $x > \mu$) will also be expected to be tall (*i.e.* $E[y] > \mu$), but not as tall as his father, and correspondingly for the sons of short fathers.

The regression to the mean phenomenon is well-known in the mining industry, where the grade of ore is estimated by means of samples before it is extracted, and generally only the ore that is estimated to be richer is mined. Considering the estimated grade to be $X$ and the actual grade of the extracted ore to be $Y$, and that the ore is only extracted if $X$ is greater than some threshold value (which is generally higher than the average grade of the whole ore-body) Y would be expected to be lower than was predicted by the sample $X$. (This apparent loss is termed the 'mine call factor,' which presumably also includes actual losses suffered by the mine.)

Regression to the mean is also observed in sport, where the top performing players in one year generally do not perform as well during the next year (see, for example, Schall and Smith [25]). This is also seen in the performance of sports teams, which could partially explain why firing a coach after the team has had a poor season and appointing a new coach, generally seems to pay dividends. Regression to the mean suggests that this would tend to happen anyway!

# 7 Using statistics to enhance a deterministic model

There have been many attempts to improve optimisation models by incorporating uncertainty into them. Two examples will be discussed here, namely Stochastic Programming and Global Optimisation.

## 7.1 Stochastic programming

Stochastic programs are linear (or nonlinear) programs in which some, or all, of the coefficients are random variables. Ignoring this randomness and solving them in the usual deterministic manner could lead to solutions that are either not feasible, sub-optimal or both. In the first two special cases described below, stochastic programs can be solved with no more difficulty than their corresponding deterministic versions, whereas, in general, including randomness can result in large increases in problem size. For simplicity, I will consider only Linear programs:

- If only the objective function has random coefficients, then one can appeal to decision theory principles (assuming linear utilities) and replace the random coefficients by their expectations and solve the corresponding deterministic linear programming problem.
- If the right-hand sides of the constraints are random variables, then one can replace them by their probabilistic critical values and again solve the corresponding deterministic linear programming problem. Basing the critical values on the joint distribution of these variables will ensure that all the constraints are satisfied with a specified overall probability. Otherwise, Bonferroni's inequality can be used to specify a lower bound on this overall probability, using critical values based on the individual marginal distributions. This is a simple case of Chance-Constrained Programming.
- More generally, linear programs with elements in their coefficient matrices also being random, may be approached as recourse problems, which essentially look at each possible realisation of the random variables and the solution that minimises the expected costs, including those associated with possible violation of the constraints, is sought. As the number of possible realisations of the random variables increases, these problems can become very large. Chance constrained formulations of these problems usually end up with nonlinear constraints, even if the random coefficients are independent.

Kall and Wallace [19] and Birge and Louveaux [6] give good coverage of the field.

### 7.2 Global optimisation

Multi-start algorithms for finding the global minimum of a function that has many local minima often start at randomly selected points within the search region. Each (randomly started) iteration of the algorithm yields a (possibly local) minimum point and its corresponding minimum function value. The point with the lowest minimum value is the current best. The question is when to stop the process and use the current best point as the global minimum solution.

A feature of the particular algorithm in question (Snyman and Fatti [26]) was that the region of attraction of the global minimum (the region of starting points that would lead to this point) could be assumed to be larger than that of any other local minimum (however many local minima there may be). After $n$ iterations of the algorithm, the current best point will have been reached by $r \geq 1$ of them, and if $r$ is large enough, then it is likely that this point is the global minimum, in view of the fact that it has the largest region of attraction.

A Bayesian approach, based on this feature, using a prior distribution for the probability of the region of attraction of the current minimum, which is updated after every iteration, led to an expression for the posterior probability that the current minimum was indeed the global minimum. The stopping rule was based on this probability reaching a sufficiently high value (for example, 95% or 99%).

## 8 Final remarks

In this paper I have stressed the important role that Statistics should play in Operations Research. It follows, therefore, that in order to be effective as an OR practitioner, a good background in Statistics is essential. There is no rule that says which statistical techniques will necessarily be the most useful, but from the examples discussed in this paper, the following suggestions emerge as statistical *desiderata* for the well-rounded OR practitioner:

- Basic probability theory and stochastic processes
- Statistical distribution theory and Bayesian inference
- Basic data analysis, including graphical techniques
- Sampling and experimental design
- Regression modelling (including GLIM)

Clearly, there are many other techniques that are useful for operations researchers, including the most important of them all, namely statistical thinking!

## References

[1] Ackoff RL & Sasieni MW, 1968, *Fundamentals of operations research*, Wiley, New York (NY).

[2] Agresti A, 2002, *Categorical data analysis,* Wiley, Hoboken (NJ).

[3] Bartholomew DJ, 1967, *Stochastic models for social processes,* Wiley, London.

[4] Beirlant J, Goegebeur Y, Segers J & Teugels J, *Statistics of extremes: Theory and applications,* 2004, Wiley, Chichester.

[5] Belton V & Stewart TJ, 2002, *Multiple criteria decision analysis: An integrated approach*, Kluwer Academic Publishers, Boston (MA).

[6] Birge JR & Louveaux F, 1997, *Introduction to stochastic programming*, Springer, New York (NY).

[7] Box GEP, Hunter JS & Hunter WG, 2005, *Statistics for experimenters: Design, innovation and discovery*, 2$^{\text{nd}}$ Edition, Wiley, Hoboken (NJ).

[8] Coles S, 2001, *An introduction to statistical modelling of extreme values,* Springer, London.

[9] Dobson AJ, 1990, *An introduction to generalized linear models*, Chapman and Hall, London.

[10] Efron B & Tibshirani R, 1993, *An introduction to the bootstrap,* Chapman and Hall, London.

[11] Fatti LP, 1983, *Optimal smoothing of demand for industrial gas*, Journal of the Operational Research Society, **34(7)**, pp. 583–590.

[12] Fatti LP & Stewart TJ, 1986, *Quality control in export coal*, Journal of the Operational Research Society, **37(11)**, pp. 1073–1080.

[13] Friend JK and Jessop WN, 1969, *Local government and strategic choice: An operational research approach to the processes of public planning,* Tavistock, London.

[14] Galton F, 1886, *Regression towards mediocrity in hereditary stature,* The Journal of the Anthropological Institute of Great Britain and Ireland, **15**, pp. 246–263.

[15] Gross D & Harris CM, 1985, *Fundamentals of queueing theory*, 2$^{\text{nd}}$ Edition, Wiley, New York (NY).

[16] Hastie T, Tibshirani R & Friedman J, 2009, *The elements of statistical learning: Data mining, inference and prediction*, 2$^{\text{nd}}$ Edition, Springer, New York (NY).

[17] Hillier FS & Lieberman GJ, 1986, *Introduction to operations research*, 4$^{\text{th}}$ Edition, Holden-Day, Oakland (CA).

[18] Johnson LA & Montgomery DC, 1974, *Operations research in production planning, scheduling and inventory control,* Wiley, New York (NY).

[19] Kall P & Wallace SW, 1994, *Stochastic programming*, Wiley, Chichester.

[20] Lindley DV, 1992, *Making decisions*, 2$^{\text{nd}}$ Edition, Wiley, London.

[21] McCullagh P & Nelder JA, 1989, *Generalized linear models*, 2$^{\text{nd}}$ Edition, Chapman and Hall, London.

[22] Press SJ, 2003, *Subjective and objective Bayesian statistics: Principles, models and applications,* Wiley, Hoboken (NJ).

[23] Rivett P, 1972, *Principles of model building: The construction of models for decision analysis,* Wiley, London.

[24] Rivett P, 1994, *The craft of decision modelling,* Wiley, Chichester.

[25] Schall EM & Smith G, 2000, *Do baseball players regress toward the mean?* The American Statistician, **54**, pp. 231–235.

[26] Snyman JA & Fatti LP, 1987, *A multi-start global minimization algorithm with dynamic search trajectories*, Journal of Optimization Theory and Applications, **54(1)**, pp.121–141.

[27] Taha HA, 1976, *Operations research: An introduction*, 2$^{\text{nd}}$ Edition, Collier MacMillan, London.

[28] Wagner HM, 1969, *Principles of operations research,* Prentice-Hall, Englewood Cliffs (NJ).

[29] Winston, WL, 1994, *Operations research: Applications and algorithms*, 3$^{\text{rd}}$ Edition, Wadsworth, Belmont (CA).