Towards a Statistical System Analysis

Bernd Heidergott

Professor of Stochastic Optimization

Department of Econometrics and Operations Research

VU Amsterdam University, the Netherlands

Academic applied probability/operations research is mainly focused on the mathematical analysis of models that find their motivation in the outside (read, non-academic) world. In preparing a real-life problem for mathematical analysis, a "model" has to be distilled, and once this is done, reality is replaced by this model, which is subsequently analyzed with much energy and analytical rigor. However, hardly ever are the exact model specifications known, and defining parameters of the model under consideration, such as arrival rates in queueing networks, failure rates of servers in reliability models, or demand rates in inventory systems, are only revealed to the analyst by statistics. The classical approach for dealing with such parameter insecurity is to integrate out the system performance with respect to the assumed/estimated distribution of the unknown parameter.

We believe that in order to achieve a better understanding of model/parameter insecurity a closer look into the way "randomness" is used in the analysis of a given model is of importance. Randomness is an ubiquitous phenomenon. Without going too much into detail on the philosophical aspects of the concept of randomness and probability, one can loosely state that randomness is encountered as (1) lack of knowledge, or (2) variability in repeated realizations of a phenomenon. For example, (1) covers the so-called parameter insecurity and/or model insecurity. Indeed, often either the true distribution of a random variable used in a model is not known (=model insecurity) or the distributional parameters, such as mean, variance etc. are not known (=parameter insecurity). Statistics can then be used to narrow down the possible range of distribution models or range of parameter values, but reaching certainty is epistemologically impossible. This in contrast to (2), where in principle laws of large numbers and ergodic theorems are available that allow to produce reliable measurements for which mathematically supported quality assessments are possible. The concept described in (1) relates to subjective probabilities, whereas the (2) relates to the frequentialistic interpretation of probability. Consider, for the sake of exposition, the following simple problem. Let X_θ be random variable with cumulative distribution function F_θ, where θ denotes a parameter of the distribution, for example, the mean or the variance. Suppose we are interest in estimating the mean value of X_θ, denoted by μ_θ, and we perform a computer simulation to sample n independently and identically distributed realization of X_θ, denoted by $X_\theta(i)$, $1 \leq i \leq n$. Then, the sample

average

$$\bar{X}_\theta = \frac{1}{n} \sum_{i=1}^{n} X_\theta(i)$$

is a natural estimator for the mean. The above estimator is however deceivingly simple as it assumes that θ is known, i.e., we know the correct value of θ. Now assume that we do not know the exact value of θ. To formalize, let θ_0 denote the true value of θ and suppose that for the simulation we use, due to lack of better knowledge, θ_1. Then, the error we make in estimating μ_{θ_0} is

$$\mu_{\theta_0} - \bar{X}_\theta = \underbrace{\mu_{\theta_0} - \mu_\theta}_{(1)} + \underbrace{\mu_\theta - \bar{X}_\theta}_{(2)},$$

where the first error is due to our lack of knowledge and the second error can be controlled through the sample size. Much research in applied probability and statistics is targeted at reducing the second error. Although the presence of the type (1) error is acknowledged in the literature, how to deal with type (1) error is still an open question. The area of perfromability analysis is devoted to finding models for the lack of knowledge based on entropy. Starting point is expert knowledge about typical behavior of θ and taking the distributional model that maximizes the entropy with respect to the predefined characteristics provides a distributional model for θ, i.e., θ is now considered a random variable. Alternatively, a statistical estimator for θ may be available. Then, the sample-distribution of the estimator can be used as distributional model for θ. Consider, for example, the case where θ is estimated through a sample-mean of identical and independently distributed random variables, and assume that the sample-size is sufficient for the strong law of large numbers to hold, then we can model θ as $\theta(\omega) = \bar{\theta} + \mathcal{N}(\omega)$, where $\bar{\theta}$ is the sample average, i.e., point estimator, and $\mathcal{N}(\omega)$ is standard normal random variable. Put differently, statistics allows to build a distributional model for θ.

This lecture will elaborate on the above distributional model for parameter insecurity and is aimed at stimulating a discussion on the relation between statistics and applied probability/operations research. This lecture will advocate supporting the analyst by studying the risk incurred by parameter insecurity. Rather than taking an entirely statistical point of view by dismissing "model building" at all, we want to integrate the data-driven statistical nature of model building into the analytical analysis. We will discuss an analytical framework for doing so that allows for separating (i) the (analytical) analysis of the system from (ii) the statistical model for the parameter insecurity. We present a series of numerical examples illustrating our approach.