Contributions Towards the Operationalisation of Empirical Bayes for Discrete Event Simulation

Shona Blair
Department of Management Science
University of Strathclyde
Glasgow, UK
2025

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy.

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Shona Blair Date: 12/12/2025

Acknowledgements

I would like to express my sincere thanks to my PhD supervisors Professor John Quigley and Professor Tim Bedford, for their insight, expertise and guidance over the years. Their ongoing support has been greatly appreciated.

I would like to thank the EPSRC research council for their funding of this project by means of the CASE studentship they awarded, and Dr Mark Elder of Simul8 for his support of the project and the access provided to observe some of the work that is undertaken by his company. I'd also like to thank Professor Stephen Chick of INSEAD for his contributions as external project supervisor and for several fruitful discussions during the research process.

I am also immensely grateful to Kerem Akartunali for his continuous support and his comments on earlier versions of the thesis, and to my brother, Graham Baird, for his helpful guidance and fruitful discussions.

Above all, I would like to thank my mother, Anne Baird, for her love, belief, encouragement and support. I deeply wish that she were here to witness this submission.

Abstract

Discrete event simulation (DES) is a well-established methodology in operational research and management science, facilitating the design, analysis and improvement of complex real-world systems. In many cases the nature of the underlying system results in DES models which are large in scale, complex, and computationally expensive to run. These attributes complicate their use in practice and necessitate the careful design and analysis of DES experiments to ensure valid and efficient inference concerning DES model quantities of interest.

It is envisaged that empirical Bayes (EB) methods could provide a welcome addition to the DES practitioner's toolkit of experimentation methods. EB methods have been much used in recent years to exploit the parallel structures of the large-scale data sets arising through the use of modern scientific technologies such as microarrays and FMRI. Whilst the structural similarities between problem settings lends intuitive support to the idea that EB methods may be of benefit, implementation is unlikely to be trivial.

This thesis presents an investigation into the utility of adopting an EB approach to DES model experimentation. The specific contributions are discussed in more detail next.

We first present a computational study examining the application of EB to DES model experimentation. This study establishes proof of concept by demonstrating that substantial efficiency gains are possible, while also identifying key practical challenges. These challenges motivate the need for greater operationalisation to support practitioners applying EB in DES contexts.

The first contribution of this thesis is the design and development of a decision

support tool to guide practitioners on the suitability of EB. This involves the determination of the factors of relevance in measuring EB suitability, and the development of a predictive statistic for the relative performance of EB versus traditional approaches in the DES experimentation context.

Towards the second contribution, this thesis investigates adapted EB procedures, designed and developed to overcome issues and challenges identified in the DES experimentation context. This work involves the novel use of a classical weighting mechanism, allowing the data from model scenarios to be weighted according to their similarity, limiting unhelpful bias and ensuring robustness for DES practice.

The final contribution of the thesis is the presentation of a numerical study demonstrating the use of the methods developed on a large-scale, industrial DES simulation model. Whilst the earlier numerical testing of the methods developed demonstrates their statistical efficiency, this study allows us to demonstrate their applicability to and practical value within a real-world DES context.

Taken together, these contributions demonstrate the potential for EB to support effective and efficient DES model experimentation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The purpose of this chapter is to introduce the topic of the thesis to the reader, providing background and setting expectations for what follows in later chapters. To begin with, the motivation for the study is discussed, through which a high-level overview of the study emerges. Following this, a formal description of the research design is presented, providing the reader with an understanding of its aims and objectives, and the methodology employed in pursuing them. The chapter concludes with an outline of the thesis structure, clarifying how the content of the study aligns with the chapters of this document.

1.1 Research Motivation

Discrete event simulation (DES) is a well-established methodology in the field of operational research and management science (OR & MS) facilitating the design, analysis and improvement of complex real-world systems [213]. At its core, DES involves abstracting the fundamental structure of the system of interest and using this information to construct a stochastic and dynamic computer-based model of the system.[1] A process of experimentation is conducted with the computer model in order to gain insight into, and understanding of, the performance of the computer model. It is hoped that the careful application of appropriate modelling and experimentation techniques will ensure that the insights gained concerning model behaviour are generalisable to the behaviour of the real system [154].

[1] DES differs from other forms of modelling, such as system dynamics (deterministic) or spreadsheet-based models (no time component). A fuller characterisation of DES as compared with other simulation approaches is provided in Subsection 2.1.1.

This research explores the potential benefits of applying empirical Bayes (EB) methods to DES model experimentation: an idea that, to the best of our knowledge, has not been widely examined in the simulation literature. To motivate this exploration, we begin with a discussion of simulation experimentation and its varying levels of computational intensity, before turning to the structure of EB methods and why they may offer an attractive, and as yet underutilised, approach.

The process of simulation typically involves running the model multiple times and statistically analysing the resulting data. The time and resources required to do this, however, vary considerably and are dependent on a number of interrelated factors. These include the variability in the data generated, the level of accuracy needed in the analysis, the run time of the model itself, and the computational resources available. Greater data variability or a higher standard of accuracy typically call for more runs of the model, which naturally increases the length of time needed for experimentation. Meanwhile, whilst some DES models may run in seconds, others, especially larger-scale models with detailed logic, may take hours or even days for a single run.

Even in relatively straightforward applications, the computational demands of simulation can escalate quickly. Consider, for example, a hospital evaluating its triage procedures in its accident and emergency department. The objective might be to evaluate three different triage protocols across five staffing levels, producing 15 distinct simulation scenarios. If each scenario requires 50 runs to achieve a desired level of statistical precision, the total amount of simulation needed already reaches 750 runs. Assuming each run takes just five minutes, the full experiment demands over two and a half continuous days of computation, even before accounting for further exploration or model refinement.

This simple illustration highlights a broader point: the demands of simulation can grow rapidly as studies become more ambitious. Beyond individual examples, recent Winter Simulation Conference themes and keynotes (Table 1.1) reflect this trend, showing how the field's focus has shifted toward issues of scale: extending simulation across larger systems, richer datasets, and more demanding environments. The recurring emphasis on complexity, data intensity (e.g. social,

behavioural, and scientific contexts), and methodological sophistication (e.g. distributed and agent-based simulation) in these themes illustrates how simulation is now expected to operate at ever broader and deeper levels of analysis.

Year	*Theme* and Relevant Keynotes
2016	*Simulating Complex Service Systems*
	S.E. Page - "Many Model Thinking"
	S. Sanchez - "A Data Farmers Almanac"
2015	*Social and Behavioural Simulation*
	J. Epstein - "Agent-Zero and Generative Social Science"
	A. Law - "Discrete-Event and Agent-Based Simulation and Where to Use Each"
	P. L'Ecuyer - "Imitation Challenges: From Uniform Random Variables to Complex Systems"
2014	*Exploring Big Data through Simulation*
	R. Roser - "The Higgs Boson - the Search for the Particle and the Role of Simulation"
	R.M. Fujimoto - "Parallel and Distributed Simulation"
2013	*Making Decisions in a Complex World*
	E. Bonabeau - "Big Data and the Bright Future of Simulation (the Case of Agent-Based Modelling)"
	B. Nelson - "The Simulation Curmudgeon"
2012	*WSC Goes Europe*
	T.A. Henzinger - "The Propagation Approach for Computing Biochemical Reaction Networks"
	C. Balbo - "Modelling and Simulation of Complex Systems: Are Petri Nets Useful?"

Table 1.1: Recent Winter Simulation Conference themes and complex-system related keynotes.

For practitioners, these broader trends translate into concrete operational pressures. As simulation studies grow in ambition and scope, the time and resources needed for experimentation can become a limiting factor. Analysts may be forced to limit the number of scenarios examined, reduce the number of runs per scenario, or simplify model structure in ways that reduce fidelity. Each of these compromises risks weakening the reliability or scope of insights available to decision makers. This is especially problematic in contexts where simulation supports operational or strategic decisions under tight time or resource constraints, since poor evidence can carry significant practical costs.

Addressing this need for experimentation to scale efficiently without prohibitive increases in computational cost has prompted the development of a variety of methods for DES experimentation. Some build on classical statistical design, focusing on how scenarios are chosen and runs allocated. Others use more adaptive or computational strategies to reduce the number of runs required or to approximate model behaviour. Each has strengths, but each also carries trade-offs, and none fully resolves the burden in all contexts. Notably, conventional approaches typically treat scenarios independently, overlooking the information that might be

shared across them.

It is at this point that EB methods suggest themselves as a promising addition. By allowing data from related scenarios to be combined in a principled way, EB procedures have the potential to increase the efficiency of estimation while still preserving scenario-specific insight, provided they are calibrated to balance these two aims. In this context, each DES scenario may be viewed analogously to a population in EB settings, with individual simulation runs corresponding to observations; this conceptual mapping underlies the relevance of EB to DES experimentation. The current study therefore explores how such methods, suitably adapted, might be used to support more effective and efficient DES experimentation.

At their core, EB methods offer a statistical framework for making better use of information when many related estimation problems are addressed simultaneously. Rather than analysing each population in isolation, EB draws strength from the totality of data observed across the entire collection of populations, recognising patterns that recur across them. This shared information can improve the stability and reliability of results, particularly when the information for any single population is weak.

A classic illustration is Efron and Morris's baseball study [80]. The focus of the study is not sport itself, but on the statistical challenge of estimating performance from sparse data. Here, batting averages for individual players are estimated not just from each player's own limited number of at-bats, but also by pooling information across the entire league. The effect is that extreme batting averages, otherwise unstable due to small sample sizes, are "shrunk" toward the overall mean. This shrinkage effect is illustrated in Figure 1.1, a scatterplot where the raw player averages are contrasted with their EB-adjusted predictions.

The study neatly demonstrates both the promise and the pitfalls of the EB approach. For the majority of players, "borrowing strength" across related but weakly supported populations yields more stable and accurate estimates than treating each population in isolation. Yet this mechanism relies on an assumption of similarity among the populations being pooled, namely that they are genuinely

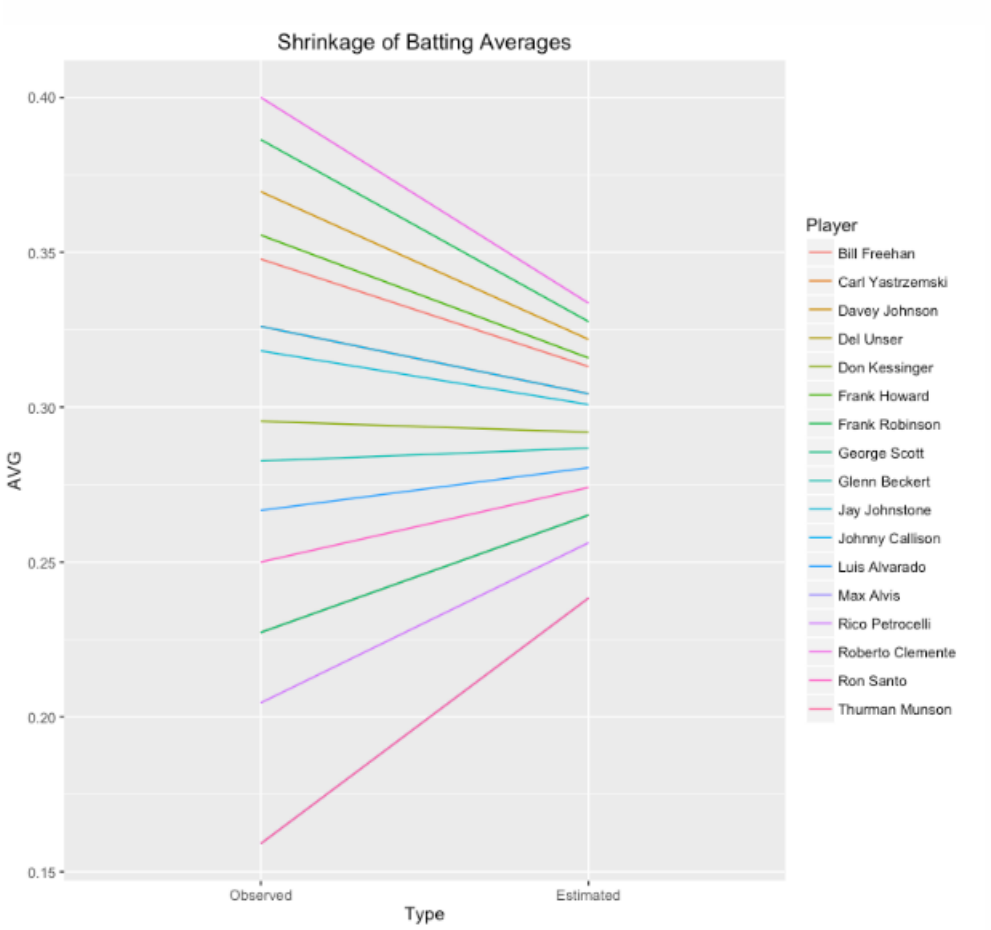

Figure 1.1: Illustration of the shrinkage in batting averages (reproduced from Albert [3] in relation to Efron and Morris's baseball example [80]).

related. This requirement, that the populations form a sufficiently coherent *family*, is central to the success of EB in practice. When this assumption is violated, and the pooled data are insufficiently homogeneous, the resulting shrinkage can introduce unwelcome bias into the estimates. Efron and Morris illustrated this issue by discussing the case of Roberto Clemente, an exceptional hitter whose EB-adjusted prediction was pulled too far towards the overall league mean because he was not typical of the rest of the players.

The effectiveness of this balance has been demonstrated in several domains. In genomics, for example, the populations are genes, and in brain imaging, they are voxels, each case involving many related but noisy estimation problems. The sig-

nificant success of EB in such applications shows that the approach can deliver substantial practical gains when the underlying structure genuinely supports information sharing. Given the structural parallels between such contexts and DES experimentation, where multiple, closely related scenarios are explored at the same time, it is intuitively reasonable to expect that EB could, in principle, yield similar benefits. Whether those gains can be achieved in practice, and under what conditions, are central questions motivating the present research.

Like these other domains, simulation experimentation exhibits a naturally parallel structure. An analyst might increase the number of servers in a queue, remove a ward from a hospital model, or adjust the staffing mix on a production line. In each case, the scenarios share most of their underlying logic, resources, and constraints, and yet they are usually analysed in isolation. This practice overlooks a defining feature of DES: the degree to which its experimental design already links model scenarios through shared structural elements.

As noted, each model scenario may be viewed analogously to a population, while a run of a scenario corresponds to an observation. The connection is immediate: the data generated in one scenario may carry information that can help interpret others. Results from a four-server queue, for instance, are not irrelevant to those from a five-server queue; performance in one hospital ward may offer context for another. In principle, the family of scenarios may be seen as forming a structured ensemble, much like the parallel populations encountered in genomics or brain imaging research.

Yet, despite these clear structural parallels and apparent potential, EB has received little explicit attention in the simulation literature. Adapting it for use in DES, however, is unlikely to be trivial. This form of "large-scale" inference features intrinsic challenges related to data heterogeneity and bias, alongside other conceptual and computational difficulties, which differ from those likely to be encountered in DES experimentation. Their intersection adds a further layer of complexity to practical application. Addressing this intersection, and providing a level of operationalisation that enables practitioners to realise the benefits of EB in DES, is therefore the central concern of the present research.

1.2 Research Design

It is helpful to distill the broad themes emerging from the foregoing discussion into a more focused description of the overall research design. We begin, in Subsection 1.2.1, with a brief discussion of the philosophical positioning of the research project. We continue, in Subsection 1.2.2, with the presentation of the overall research aim, followed by a set of more specific research objectives. Finally, in Subsection 1.2.3, we outline and discuss the research methodology applied in the thesis.

1.2.1 Research Philosophy

In this subsection, we briefly discuss the philosophical positioning of the research project. We do so by considering its ontological and epistemological foundations: that is, the assumptions underlying the work outlined and motivated in the preceding section.

To begin, we recognise the focus of the current research is the use of DES models. Given we are concerned with the use as opposed to the construction of DES models, we may consider DES models as artificial, concrete and well-defined microcosms, isolated from the rest of reality, whether natural or social. As such, there seems little difficulty in adopting an ontological position positing a fixed external reality.

Returning once more to reflect on the focus of the research, we note the importance of empirical observation in the use of DES models. In DES experimentation, the primary means of learning about the model (and so inferring about the world) is through adaptation of the model and observation of the results. Furthermore, we note the primary role played by the data values obtained through experimentation, and the secondary role played by their real-world significance and interpretation. These reflections highlight the need to adopt an epistemological position emphasising the centrality of empirical observation in the pursuit of knowledge within the project.

The above considerations reveal the positivist orientation of the current research project. It is, of course, fully acknowledged that a more nuanced discussion of research philosophy would possibly bring to light certain points of divergence

between the research project and the positivist paradigm. However, a detailed examination of alternative research paradigms early in the study indicated that positivism aligns most closely with the empirical, model-based nature of the proposed research.[2]

1.2.2 Research Aims and Objectives

In this section, we present the aims and objectives of the research, solidifying the preceding introductory discussion and equipping the reader with a map of the work presented in the remainder of the thesis.

Overall Research Aim

The aim of the current research is the investigation of the utility of adopting an EB approach to DES model experimentation, exploring the potential for material gains in statistical efficiency. Further, it aims to identify and/or develop EB-appropriate DES experimental contexts and DES-appropriate EB inferential procedures, so as to operationalise EB for use in DES model experimentation, facilitating its fruitful application in this previously unexplored context.

Research Objectives

To operationalise a broad research aim such as this, it is necessary to develop a number of more narrowly focused research objectives, enabling the overall research effort to be partitioned into its various component tasks. As such, consideration has been given to developing a suitable set of research objectives for the current project. The results of this process are outlined as follows.

R1 Given the variety of approaches to DES experimentation and EB inference, the first research objective involves the identification of the most promising overlap between the two fields, or the first logical step in the exploration of EB in DES. This will include:

(a) Confirmation of the lack of traction of EB in DES experimentation through a thorough review of literature. Any applications of EB to DES should be identified and discussed within the literature review.

[2]The discussion of such a detailed philosophical examination is beyond the scope of the thesis.

(b) A review of the different goals of and approaches to DES experimentation with a view to identifying the most promising area(s) in terms of the potential to apply EB.

(c) A review of the different approaches to EB inference with a view to identifying the most promising procedure(s) in terms of their potential applicability to DES experimentation.

R2 The second research objective involves a preliminary computational study to evaluate the applicability of EB to DES experimentation, informed by the results from R1. More specifically, by way of research sub-objectives, we have the following:

(a) The computational study will aim to provide proof of concept. That is, it will aim to demonstrate the potential for material gains in statistical efficiency with the application of EB to DES experimentation.

(b) The study will aim to highlight the possible issues and challenges in the application of EB to DES experimentation. In particular, it will investigate data heterogeneity, and the possible negative impact of bias from applying EB in DES experimentation.

(c) The study should also highlight opportunities for the operationalisation of EB in DES experimentation, e.g. in terms of providing practitioner support.

R3 The third research objective investigates the possibility of a decision support tool to guide practitioners on the suitability of EB given a particular DES experimentation context. This work necessitates the following research sub-objectives:

(a) Preliminary experimentation and discussion to determine the factors of relevance (in particular, relating to data heterogeneity) in measuring EB suitability given a particular DES experimentation context.

(b) The design and development of a tool utilising the aforementioned factors that supports practitioners deciding whether or not to apply EB in their particular DES experimentation context.

(c) The development of an appropriate evaluation approach to measure the effectiveness of the decision support tool (discussed further in R5).

R4 The fourth research objective investigates the possibility of adapted EB procedure(s), designed and developed to address the issues and challenges identified in the DES experimentation context (as outlined in R2). This work necessitates the following research sub-objectives:

(a) Preliminary experimentation and discussion to identify how data heterogeneity measurement can be taken into account to limit the negative impact of bias in the application of EB to DES experimentation.

(b) The design and development of adapted EB procedure(s) that limit the negative impact of bias and so ensure robustness (in particular, in relation to data heterogeneity) for DES experimentation practice.

(c) The development of an appropriate evaluation approach to measure the effectiveness of the adapted procedure(s) (discussed further in R5).

R5 The fifth research objective relates to the mechanics of implementation of the four research objectives outlined above (R1–R4). This involves the identification and/or development of fundamentals necessary for the programme of work as follows:

(a) The identification of DES models and simulated development/test environments to support the computational work of the thesis. In particular, this relates to the implementation of the proof of concept study of R2, the development and evaluation of both the decision support tool of R3 and the adapted EB procedure(s) of R4.

(b) The development of a set of error measures providing a multi-dimensional picture of the relative performance of the different inferential approaches (e.g. EB and traditional methods) in a DES experimentation context.

Having presented the research aims and objectives, we now turn our attention to discussing the research methodology involved in pursuing them.

1.2.3 Research Methodology

The purpose of this section is to present and discuss the methodological approach that will be implemented to deliver the research aims and objectives outlined in the previous section. This is an important step as it helps better acquaint the

reader with the research as a whole, clarifying and justifying the methodological choices made in undertaking it. As such, it enhances the validity and credibility of the eventual research findings.

We begin with a discussion of the two methodological traditions underlying the current project, namely statistical analysis and experimental mathematics. This is followed by a more detailed discussion concerning the specific methods applied in the course of the research, namely the use of Monte Carlo simulated data sets, toy simulation models and real simulation models. Finally, links between the research methods and the research aims and objectives are highlighted, creating a useful methodological map to guide the reader.

1.2.3.1 Methodological Positioning

In this subsection, we briefly outline and discuss the two broad methodological traditions, statistical analysis and experimental mathematics, underpinning the current research project.

Given the focus of this project, the exploration of EB in DES model experimentation, it is natural that such work will necessitate substantial use of statistical analysis. To begin, the fundamental role of empirical observation within this project was highlighted in Subsection 1.2.1 in discussing the philosophical positioning of the work. In this discussion, DES experimentation was characterised as a process of adapting and observing the DES model to learn about its operation and thereby infer about the real world. Such an experimental process generates substantial amounts of quantitative data, and a key challenge in DES practice is the selection and application of appropriate analytical procedures to make sense of this data. Statistics, described as "the practice or science of collecting and analysing numerical data in large quantities" by Oxford English Dictionary [175], provides us with an extensive toolkit of approaches to make sense of data and form meaningful conclusions from it. Indeed, the relevance of statistical analysis to DES experimentation may be appreciated from a perusal of Section 2.2 of the literature review, summarising both DES experimentation goals and the current analytical, and frequently statistically oriented, approaches to achieving them.

Returning to reflect on the nature of the project, as outlined in the overall research aim, our focus is the investigation of the utility of adopting an EB approach to DES experimentation, as compared with more traditional statistical approaches. This relative evaluation of different approaches to the implementation of DES experimentation adds another dimension to the required statistical analysis and can be viewed through the lens of statistical decision theory [24]. A more detailed discussion of statistical decision theory is deferred until Section 3.3 of Chapter 3. However, here we acknowledge the need to compare different analytical approaches (or decision rules) through an error measurement strategy (or loss function) [24]. An appropriate error measurement strategy will provide a comprehensive assessment of the effectiveness of the different analytical approaches, despite the presence of stochastic variation in the results obtained. The topic of error measurement will be revisited multiple times, not only in Chapter 3, but also in the further relative evaluations of EB and traditional approaches presented later in the thesis.

Another broad methodological tradition underpinning the current research project is experimental mathematics. Experimental mathematics is a type of mathematical investigation in which computation and heuristic arguments are used to investigate mathematical structures and identify their fundamental properties and patterns. Often, these approaches provide insights which may later lead to a full proof. More precisely, Bailey and Borwein [30] give experimental mathematics as a methodology of mathematical practice that includes the use of computations in the following contexts:

1. Gaining insight and intuition.
2. Discovering new patterns and relationships.
3. Using graphical displays to suggest underlying mathematical principles.
4. Testing and especially falsifying conjectures.
5. Exploring a possible result to see if it is worth formal proof.
6. Suggesting approaches for formal proof.
7. Replacing lengthy hand derivations with computer-based derivations.
8. Confirming analytically derived results.

Mathematics has a long history as an experimental endeavour: for example, extant records of early mathematics, such as those from the Mesopotamian era, frequently

involve lists of numerical examples rather than algebraic proofs [30]. Furthermore, it is often the case that mathematicians "know" their theories to be "true" long before a proof is attained. However, this approach to mathematics largely fell out of favour in the nineteenth century when the field became increasingly concerned with putting the subject on entirely firm foundations, with fully rigorous proofs becoming the gold standard [30]. The burden of this standard often restricts the application of mathematics from dealing with complex real-world problems and recently, especially in light of the increasing availability of computational power since the mid-twentieth century, experimental mathematics is enjoying an upswing in its popularity.

In terms of the applicability of experimental mathematics to the current research, we first note that the application of EB in DES experimentation appears relatively underexplored. Whilst the immaturity of the field likely indicates a great scope for contribution, it also implies that results are more likely to be computational or approximate in nature than theoretically derived or exact. We also recognise the integral role played by experimentation and computation in the current research. As discussed above in relation to statistical analysis, there are several experimental dimensions to the current work: firstly, experimentation with the DES model, and secondly, experimentation with the analytical approaches. Lastly, we note the practical, results driven field of OR & MS as the context for the research, emphasising the importance of impact over elegant formulation. Together, these considerations highlight a clear role for an experimental mathematics inspired methodological approach.

1.2.3.2 Research Methods

Building on the methodological foundations outlined above, this section describes the research methods used in the thesis for the evaluation of EB in DES experimentation, and for the development of novel approaches that operationalise EB for use in DES practice.

Before proceeding, we note the useful role of the framework by Hoad et al. [119] in finalising, presenting and discussing our choice of methods. In their article, Hoad et al. present a classification of methods for the evaluation of DES analysis ap-

proaches. Their classification is based upon a review of literature and investigation of simulation practice. The creation of this classification supports an associated attempt by the researchers to compile a sufficient repository of data sets to serve as a testbed for the evaluation of proposed DES analysis approaches.[3]

Several reasons exist for the use of this framework to guide the methodological choices made in the current research project. First and foremost, to the best of our knowledge, this work is the first of its kind, neatly surveying and summarising the different methods applied to such evaluation problems. We also note that Hoad et al. are well-established researchers in the field of DES experimentation and analysis, and further that their findings align closely with our own observations of this field of research.

We next turn our attention to briefly characterising and discussing the three key evaluation methods identified by Hoad et al., and which underpin much of the analysis presented in the thesis. These are the use of Monte Carlo simulated data sets, toy simulation models, and real simulation models.

M1 - Monte Carlo simulated data sets

- Hoad et al. define this method as the creation of data sets from "known equations, ... with a known value for some specific attribute."
- They use the term "artificial data" to refer to this type of data set.
- Some examples of the use of this type of data generation in the evaluation of DES experimentation approaches include:
  - Ockerman and Goldsman [172] use data from random walks, MA(1) and AR(1) processes to evaluate variance estimators;[4]
  - White et al. [245] use data from an AR(2) process to evaluate a variety of truncation heuristics for simulation initialisation.

[3]The completion of the data set repository did not align with the timelines for this research project, and thus unfortunately it could not be utilised.

[4]Here, MA(x) refers to a moving average process with window x, and AR(y) refers to an autoregressive process of order y. The interested reader is referred to Box et al. [35] for more details concerning such modelling approaches.

M2 - Toy simulation models

- Toy simulation models are very simple models, not necessarily created as real models of a real system, but often for educational or explanatory purposes.
- Hoad et al. use the term "artificial model" to refer to the type of model. They distinguish between two variants of such models:
  - Those for which the response parameters are known e.g. waiting time in an M/M/1 queuing model;[5]
  - Those for which the response parameters are not known but can be estimated/controlled e.g. inventory level in a model featuring a single item (s, S) inventory management policy.[6]
- Some examples of the use of this type of model in the evaluation of DES experimentation approaches include:
  - Law and Kelton [139] use data from various queuing models (some with theoretical results available, some without) to evaluate approaches for determining appropriate initialisation conditions;
  - Hsieh et al. [122] use data from various queuing and inventory models to evaluate approaches to response parameter estimation.

M3 - Real simulation models

- Hoad et al. use the same terminology in their framework, defining real models as "discrete event simulation models of real existing systems, created in 'real circumstances' (e.g. in business, academia, etc)."
- Some examples of the use of this type of model in the evaluation of DES experimentation approaches include:
  - Yaesoubi and Roberts [249] use a DES model of the history of colorectal cancer, called the Vanderbilt/NC State model, to evaluate factor screening approaches;

[5] An M/M/1 queuing model represents a single-server queue with exponential waiting and service times. The interested reader is referred to Gross and Harris [109] for more information on queuing theory models.

[6] The single item (s, S) inventory model is described in detail in Chapter 3 of the thesis.

  - Chwif et al. [57] use a DES model of the production of content for Panorama Studios (a television company) to evaluate an approach to validation and verification in the absence of data.

Before considering the criteria upon which such methods may be compared, we first note they are very much complementary in nature. That is, a strength of one method is very likely to be a weakness of another, and a well-rounded methodological approach would likely encompass balanced use of all three methods.

To better appreciate the profile of each of these different evaluation methods, we consider their performance with respect to three different criteria. These reflect the key practical concerns in the evaluation of DES experimentation approaches, namely:

- Efficiency of data generation
- Knowledge of model behaviour
- Real-world relevance

In considering the first criterion, efficiency of data generation, we note the decreasing performance of the methods as we move from the first (M1) to the last (M3).

Thinking first about Monte Carlo simulation, essentially the simulation of random variables, the situation is simple, and computation fast. In considering a toy simulation model, however, say a single server queue, such a model features items flowing through a system, resulting in more variables to simulate, and more quantities to monitor. Such a situation is, therefore, a little more complex, and so the computation is a little slower. Finally, we consider a real simulation model, a potentially complicated, interconnected network of queues and activities, with multiple inputs and outputs. Clearly, this situation is far more complex, and as a result the computation is far slower. Put simply, the greater the complexity of operations involved, the longer the time required to generate the necessary data. This constraint on effective DES model experimentation was highlighted in the introduction, and analogous comments apply here to the evaluation of approaches to DES experimentation.

In considering the second criterion, knowledge of model behaviour, we again note the decreasing performance of the methods from the first (M1) to the last (M3).

Again, considering Monte Carlo simulation first, we typically know a great deal about, and have full control over, both the data generating process, and the distribution obtained. In considering a toy simulation model, however, we may know the model outputs of interest theoretically, or we may only know these approximately from data. Finally, in considering a real simulation model, it is highly unlikely we will know the model outputs of interest. Instead, these will need to be approximated through the generation of large quantities of data, a potentially costly affair given the possibly complex nature of the model and its behaviour.

There are clear advantages to knowledge of the simulation model, whether complete or partial, in undertaking the current research. Put simply, the purpose of DES experimentation approaches is to learn about the underlying simulation model, and it is much easier to evaluate their efficacy and efficiency if we have good knowledge of the model's behaviour to compare with the results obtained.

Finally, in considering the third and last criterion, real-world relevance, we note the increasing performance of the methods from the first (M1) to the last (M3).

Naturally, the Monte Carlo simulated data sets are the furthest removed from real DES practice, and are the most "artificial" in nature. Toy models represent an interesting intermediate step, offering at least some practical applicability: for example, some real businesses manage their inventories using the single item (s, S) inventory policy. Unsurprisingly, the use of a real simulation model from a practical case study is the most valid in terms of real-world relevance. Of course, the scale and complexity of real simulation models varies a great deal, thus the variation in their distinguishability from toy models is worth acknowledging.

Taken together, the three methods reveal complementary strengths and weaknesses. Monte Carlo data sets offer efficiency and full control, toy models provide partial realism and interpretability, and real models contribute practical relevance. These complementary characteristics highlight the value of incorporating all three methods within a comprehensive evaluation strategy, ensuring that the advantages

of one can offset the limitations of another.

1.2.3.3 Research Method Mapping

Having characterised the three research methods, and discussed their relative strengths and weaknesses, we now discuss their application in the work of the thesis to address the research aims and objectives. Thus, we map the methods as follows:

R1 This research objective relates to a review of literature, and so does not involve the use of any of the research methods.

R2 This research objective relates to the preliminary computational study applying EB to DES to establish proof of concept. Research method M2, or the use of a toy simulation model, will be used in the delivery of this objective.

- It is envisaged that the intermediate nature of this model will enable the delivery of the research objective, without the additional complexity of a large-scale real simulation model, unnecessary at this initial stage of the research.

R3 This research objective relates to the design and development of a decision tool regarding the EB suitability of a given DES experimental context. Here:

- M1, or the use of Monte Carlo simulated data sets, will be used to design the decision tool.
- M1 will also be used for the preliminary testing of the decision tool.
- M3, or the use of a real simulation model, will be used for further testing, and to evaluate the real-world applicability of the decision tool.

R4 This research objective relates to the design and development of adapted EB procedure(s) to limit the impact of bias in DES experimentation contexts. Here:

- M1, or the use of Monte Carlo simulated data sets, will be used to design the adapted EB procedure(s).
- M1 will also be used for the preliminary testing of the adapted EB procedure(s).

  - M3, or the use of a real simulation model, will be used for further testing, and to evaluate the real-world applicability of the adapted EB procedure(s).

R5 This research objective relates exclusively to the mechanics of implementation, and cross-references the other research objectives already discussed above. In this sense, it does not require the application of any research methods beyond those already outlined above.

Having presented and discussed the research design underlying the thesis, the next subsection provides an outline of its remaining structure.

1.3 Outline

The purpose of this section is to outline how the remainder of the thesis is organised and to clarify how its content, as outlined in the preceding sections, aligns with the chapters of the document. More specifically, it will highlight how the research objectives will be accomplished via the work of the different chapters. The thesis is organised as follows.

Chapter 2 concerns the review of literature, presenting and discussing preliminary material underpinning the work of the thesis. It comprises three main sections: Foundations of DES, DES Model Experimentation, and EB Methodology. This chapter therefore primarily relates to Research Objective R1.

In Chapter 3, a computational study into the application of EB to DES experimentation is presented. This involves the identification of a simple DES test model, a DES-appropriate EB estimation procedure, and a set of error measures to provide an overall picture of statistical performance. It further provides proof of concept concerning the utility of EB in DES experimentation, and highlights opportunities for its operationalisation. As such, Chapter 3 relates to Research Objectives R2 and R5.

Chapter 4 concerns the design and development of a decision support tool to guide practitioners on the suitability of EB to a given DES experimentation context. This contribution involves identifying the factors of relevance in measuring

EB suitability, and their subsequent development into a predictive statistic regarding relative performance in the DES experimentation context. It further involves the formalisation of the decision support tool, using logistic regression, to provide objective classification outcomes regarding the EB suitability of the DES experimentation context in question. The work of this chapter therefore primarily relates to Research Objectives R3 and R5.

Chapter 5 investigates adapted EB procedures, designed and developed to overcome issues and challenges identified in the DES experimentation context. The chapter begins with an investigation of pooling, in which only data from 'similar' scenarios are included in the EB analysis for a given DES model scenario. However, issues with this approach prompt the investigation, design and development of an adapted EB estimator. This work is based on the novel use of a classic weighting mechanism, allowing the data from model scenarios to be weighted according to their observed similarity, limiting unhelpful bias and ensuring robustness for DES practice. Thus, this chapter primarily relates to Research Objectives R4 and R5.

The focus of Chapter 6 is the demonstration of the practical value of the methods developed in Chapters 4 and 5, exploring their applicability in the context of a large-scale, industrial DES case study. This involves an initial overview of the problem context of the case study in question, followed by discussion of the experimentation conducted and the results obtained. The work of this chapter therefore relates to Research Objectives R3(c), R4(c), and R5.

In Chapter 7, concluding thoughts are presented. The chapter begins with a review of the research aims and objectives and a discussion of the progress made in accomplishing them, emphasising the contributions made to both theory and practice. The chapter concludes with a discussion of the project's limitations and possible avenues for future research.

Chapter 2

Literature Review

In this chapter, important literature and preliminary material supporting the thesis is presented and discussed. We begin in Section 2.1 by outlining the fundamentals of DES, describing the intrinsic nature of the approach and discussing the stages typically involved in a study. In Section 2.2, DES experimentation is discussed, with a view to characterising its specific nature, and providing an overview of the methods currently found in the literature for its accomplishment. Finally, in Section 2.3 we seek to provide an overview of the EB approach to statistical inference, discussing how it differs from alternative approaches, its historical development, and its application to date in DES experimentation contexts.

2.1 Foundations of DES

In this section, an introduction to DES is presented. To begin, a high-level sketch of the nature of DES is provided. This overview encompasses both the types of systems commonly studied, and the specific characteristics of the models used. This is followed by a very simple example DES model to illustrate the foregoing discussion. To provide greater context and background, the different stages comprising a typical DES project are outlined, assisting the reader in better understanding where EB might find useful application. The section concludes with a brief discussion of the varying levels of complexity of DES models, of direct relevance to the following section on DES experimentation, as such complexity often complicates the experimentation process.

2.1.1 An Introduction to DES

Simulation is a term that is used in a variety of different ways in a variety of different contexts [178]. Simultaneously describing the use of physical, scale models in engineering contexts, "human in the loop" simulation for training of power plant operators and airline pilots, video games offering the experience of navigating any imaginable alternate reality, and computer representations of the laws that govern scientific phenomena, the scope of simulation is extremely broad [144, 246]. In light of such diversity, it is important to provide a clear characterisation of the nature of DES, the simulation of interest in this research.

To begin, as with White and Ingalls [246], the distinction between experiential and experimental simulation is noted. Classifying the above examples, game consoles and "human in the loop" training environments exemplify experiential simulation, whilst physical engineering models and computer models of scientific phenomena represent experimental simulation. White and Ingalls explain that in the first category, learning, or indeed entertainment, occurs through the user's experience of being immersed in and engaged with the simulation. In the second category, however, they note emphasis lies on the construction of a model of a system, and subsequent experimentation with this model, for the purpose of better understanding the underlying system. Applying this distinction to better characterise DES, we note DES represents a form of experimental simulation.

Whilst this provides some clarification, it also immediately raises questions concerning the types of systems and models involved in DES studies. Even more fundamentally, it is important to define the terms "system" and "model" themselves. Different definitions abound, however as with Sànchez [195], we shall define a system to be "a set of elements which interact or interrelate in some fashion," and a model to be "a system which we use as a surrogate for another system." Any definition of such terms will necessarily be broad and somewhat abstract, but in light of the importance of these terms, it is nonetheless helpful to provide some clarity.

Returning to our question on the types of systems typically studied using DES, we begin by noting that DES is, broadly speaking, an OR & MS methodology.

The field of OR & MS is concerned with the application of analytical approaches in organisations, be they public or private, to generate some improvement in their operations [127]. DES represents just one such analytical approach; others include mathematical optimisation, queuing theory, statistical modelling and forecasting, strategic options development and analysis (SODA), and soft systems methodology (SSM) [183, 248]. Whilst DES may at times find application in other fields, it has been developed, used and studied by management scientists and operational researchers [107], and is widely regarded as constituting a core component of the OR & MS analyst's toolkit [37]. Moreover, as this research is proposed as a contribution to the field of OR & MS, attention will be restricted to discussing DES as it relates to this field.

The focus in OR & MS on understanding and improving organisations sheds some light on the types of systems commonly studied through the lens of DES. Here, it is useful to consider the classification of systems proposed by Checkland [49]. This framework identifies four fundamental classes: natural systems (e.g. our solar system), designed physical systems (e.g. a manufacturing plant), designed abstract systems (e.g. mathematics) and human activity systems (e.g. international politics). As with Robinson [190], who also discusses this classification for similar purposes, we note that the classes of most relevance presently are designed physical systems and human activity systems. This follows from the fact that these classes of systems, or hybrids of both, are those most relevant to the study of organisations. Therefore, DES may be used to understand the operation of a manufacturing plant or warehouse (a designed physical system), or to improve the delivery of health care within a given region (a human activity system).

The points made by Robinson on the types of systems typically simulated in OR & MS applications are confirmed by the perusal of a list of core DES application areas. The following list has been compiled from similar lists in Law [153] and Pidd [184]. It highlights some key areas in which DES has a well-established, successful track record, and provides examples of DES use within each such area:

- Manufacturing: designing plant facilities, optimising shop floor layouts and re-engineering assembly lines.
- Healthcare: organising and improving facility layouts, and managing supply

chains for efficient delivery of resources

- Business: re-designing a firm's financial and accounting systems, optimising staffing levels and designing service facilities
- Logistics: designing distribution centres and improving fleet management
- Defence: efficiently allocating military resources, and forming effective defensive and offensive battle strategies
- Transportation: improving road traffic systems, shipping operations, airport terminal operations and air traffic control systems

Further information on DES application areas may be obtained from the past proceedings of the Winter Simulation Conference (WSC), the pre-eminent global DES conference. This conference is generally regarded as being representative of the state of the art in terms of DES theory and practice [11], and is an excellent source of information on all aspects of DES.

Having now discussed the types of systems typically analysed using DES, we now turn our attention to better understanding DES models. To begin, it should be noted that, in the main, DES is a computer simulation approach. That is, in the vast majority of situations, the model constructed to represent the system of interest is computer-based. As authors Law [154] and Pidd [184] amongst others note, whilst it is technically possible to carry out DES without the use of a computer, this rarely, if ever, occurs in practice. This is due to the obvious advantages offered by computer implementation (namely, vastly increased speed and the ability to tackle far more complex problems). As such, we discuss only computer-based DES models.

Computer simulation is itself a broad field, and so it is helpful, as with Law [154], to consider certain dimensions along which computer simulation models may be classified. Firstly, Law distinguishes between static and dynamic models. A dynamic model is one whose state evolves through time, whilst a static model is one which represents only a single point in time, or which has no time element at all. Secondly, Law makes the distinction between discrete and continuous models. A discrete model is one in which the model state changes instantaneously at distinct

points in time, whereas a continuous model is one in which the model state changes continuously with respect to time. Finally, Law considers stochastic versus deterministic models. A stochastic model is one in which input quantities are allowed to vary randomly according to assigned probability distributions. This in turn results in random variation in the simulation's output quantities. In a deterministic model, no such variation occurs.

Having discussed these distinctions, it is now possible to provide a more specific characterisation of the models involved in a DES approach. DES models evolve through time, their state changing instantaneously at distinct time points. That is, they are necessarily dynamic and discrete [154]. The manner in which time evolves in dynamic, discrete simulation models requires some discussion. Two main approaches have emerged for advancing simulation time in such models, namely next-event time advance and fixed-increment time advance [153]. In next-event time advance, simulation time only moves forward when an event occurs that precipitates a change in the state of the model. However, in fixed-increment time advance, simulation time moves forward at regular fixed intervals, regardless of whether or not an event triggers a change in model state. Whilst both approaches are valid, it is clear that the next-event approach is likely to result in greater computational efficiency, avoiding the potential for redundant model re-evaluation inherent in the fixed-increment approach [184]. It is therefore not surprising that the next-event approach has proven more popular in practice. DES models, as their name suggests, feature the next-event approach to time advance. Finally, we note that DES models are also stochastic in nature [154] (this point already discussed in some detail in the preceding chapter). Thus, in summary, we have that DES models are dynamic, discrete and stochastic in nature, with time handled via a next-event time advance mechanism.

Having introduced DES in a very general manner, we next present a simple, example DES model. Firstly, this example illustrates the preceding discussion, providing a practical example to solidify the theoretical concepts and ideas covered. Secondly, it highlights some key points concerning DES not emphasised in the preceding discussion. Lastly, it allows for the introduction of some useful DES terminology in a practical context.

Example: Simple DES Call Centre Model

Here, we present and discuss a simple DES call centre model, created in Simul8, a popular DES software.[1] An image of the model is provided in Figure 2.1 below. The model mimics the operation of a call centre over a period of 40 hours, or one working week.

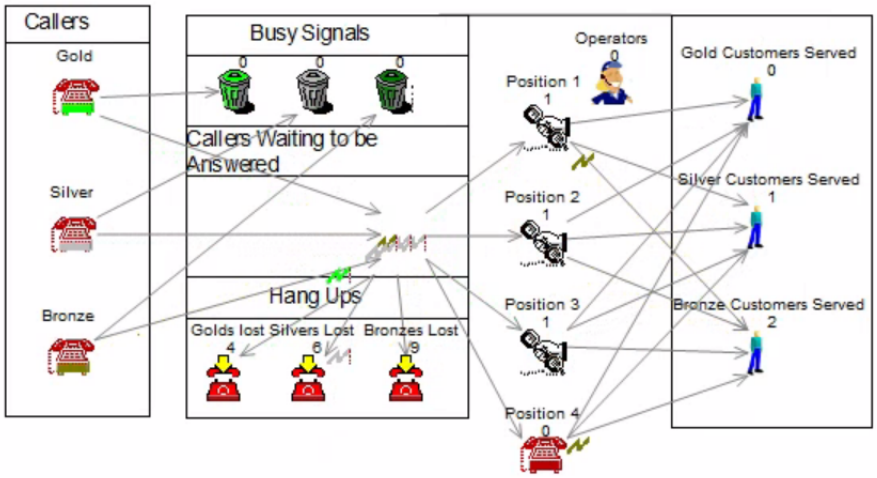

Figure 2.1: Template model "Simple Call Centre" available within Simul8 software.

Three types of customers visit the call centre: gold (top priority customers), silver (important, but not top priority customers), and bronze (low priority customers). The three telephone icons on the left hand side represent model entry points for each customer type. By clicking on each icon, we can view information about the arrival pattern of each type of customer. For example, the inter-arrival time for silver customers is modelled as an exponential distribution with a mean of 0.5 minutes.

Upon calling, customers typically wait in a queue for service from an operator. Gold customers may skip ahead of all other customer types, and silver customers may skip ahead of bronze customers. The queue, however, has a maximum capacity of ten customers. If a customer calls when the queue is already full, they will hear a busy signal and will not be able to join. Such customers are considered

[1] We refer the interested reader to Subsection 2.1.2.2 for a discussion of different types of DES software.

lost, as they then directly leave the call centre. Some queuing customers tire of waiting and leave before receiving service: these are termed reneging customers. Information on the renege time distribution for a particular type of customer may be found by clicking on the appropriate entry point. For example, the renege time of silver customers is modelled using a normal distribution with a mean of 0.5 minutes and a standard deviation of 0.1 minutes. The busy signal and hang up icons show the cumulative numbers of each type of customer affected by each of these issues.

Service from an operator requires two components, an available "position" (hardware, such as a desk and/or headset) and an available operator. In Figure 2.1, we can see four operator positions represented by the four telephone icons. Here, positions one to three are occupied, whilst position four is empty, a bronze customer having just finished service. Above these positions, we can see an operator icon. This represents the pool of available operators. Here, the adjacent zero indicates there are currently no available operators (also indicating the call centre has a total of three operators). Service, of course, takes time, and clicking on any of the positions reveals service time is modelled using a normal distribution with a mean of 2 minutes and a standard deviation of 0.5 minutes, regardless of customer type.

Once customers receive service, they leave the call centre via the exit points on the right hand side, represented by the three blue customer icons. The numbers adjacent to each of these icons show the numbers of gold, silver and bronze customers who have left the system after service since the beginning of the week. By clicking on a particular exit point icon, we can view information on the distribution of the time spent in the call centre for the given customer type. After running the model, we might be interested in performance measures such as the mean time spent in the call centre, or the percentage of gold customers served in ten minutes or less.

Having briefly outlined the operation of the model, we now discuss its use in practical decision making. In the previous discussion, the number of operators and positions were straightforwardly assumed to be three and four, respectively. In reality, a decision maker might want to experiment with these values to observe the impact on, for example, the percentage of gold customers served in under ten minutes. Alternatively, they may wish to run the model with four operators and

five positions to assess whether or not these changes result in a significant enough increase in the aforementioned percentage to consider investment.

Modelling assumptions (such as the service distribution or routing rules) or input quantities (such as the mean service time or number of operators) varied as part of experimentation are termed *factors*. The different settings or values assumed by a factor during experimentation are termed its *levels*. Performance measures of experimental interest, such as the percentage of gold customers served in under ten minutes, or the number of customers lost to reneging, are called *responses*. Here, changing the number of operators from three to four, and the number of positions from four to five, defines a new *model scenario*. Multiple, distinct model scenarios might be included in an analyst's *experimental design*. Model scenarios included in an experimental design may be termed *design points*.

In light of the inherent variability, i.e. through the modelling of inter-arrival, service and renege times with probability distributions, the analyst would most likely execute each design point a number of times. This allows the collection of a sample of data, and so more reliable inference regarding each response of interest. Each execution of the model is termed a *run* of the model, and a run of each design point included in an experimental design is termed a *replication*. The length of simulated time of each run, here 40 hours, is called the *run-length*.

The examination and discussion of this simple DES model has provided a concrete example to make real our more theoretical, initial discussion of DES. It also highlights two important points concerning the nature of DES not emphasised through our initial discussion. Firstly, DES models typically involve the flow of entities through a network of queues and activities. Whilst this is a very simple example, this description of our model also holds true for DES models of far greater complexity, and so provides a useful characterisation of DES models in general. Secondly, a key advantage of the DES approach is its ability to model the interaction of system components under uncertainty, equipping the OR & MS analyst with a useful and widely applicable tool.

2.1.2 DES Project Life Cycle

Having introduced DES, we now provide an overview of the key steps involved in a typical DES project. The purpose of this subsection is twofold. Firstly, it provides additional background and context, equipping the reader with a mental map of the various processes involved in DES, and how they relate to one another. Secondly, it also provides the reader with a better idea of where the proposed research contributions might find useful application within the DES project life cycle.

Structures for sound simulation projects feature frequently in DES literature, with contributions appearing from authors such as Shannon [212], Law [154], Pidd [184] and Banks et al. [11], amongst others. Whilst superficially different, the labelling and level of division changing from case to case, there is in fact a great deal of overlap. Here, we use one such contribution, proposed by Robinson [190] and based in part on the work of Landry et al. [152], as a vehicle for discussion. The reasons for the selection of this particular framework are that it is clear, concise, and of sufficient generality to be realistic. Whilst the basic structure of this discussion follows Robinson [190], other sources have been incorporated as relevant.

Robinson [190] identifies the fundamental processes involved in DES as conceptual modelling, model coding, experimentation and implementation. This framework is visually represented in Figure 2.2, taken from Robinson [190]. Whilst these activities are introduced in a linear fashion, Robinson highlights that in reality much iteration and repetition occurs, as indicated by the bi-directional arrows. The rectangles in Figure 2.2 represent the deliverables (a conceptual model, a computer model, solutions and/or understanding, and an improvement in the real world) associated with the 'completion' of each of the four key processes.

The remainder of this subsection will focus on providing a brief outline of each of the four key processes identified by Robinson [190]. Before proceeding, however, we will discuss two further important activities, namely verification and validation. Verification entails making certain the computer model accurately represents the conceptual model, whilst validation describes the broader aim of ensuring the computer model is sufficiently representative of the system it aims to model [47]. Robinson notes the slightly different status of these activities, as compared with

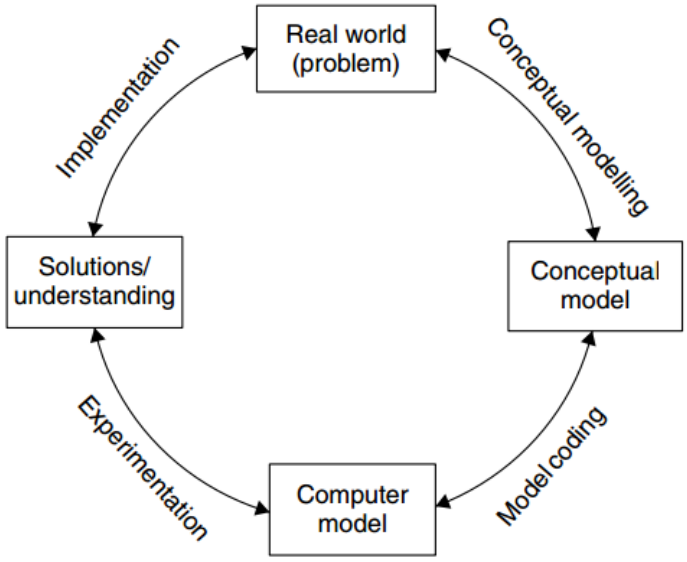

Figure 2.2: A visual representation of the framework proposed by Robinson [190], illustrating the fundamental processes and stages of a typical simulation study.

the processes mentioned previously, suggesting they should be viewed as continuous efforts on the part of the modeller, and not as processes occurring at specific points in the DES project life cycle. Several forms of validation are identified by Robinson, including conceptual model validation, data validation, white-box and black-box validation, experimentation validation and solution validation. Furthermore, he suggests that verification may in fact be considered alongside these other forms of validation, as a result of its relatively narrow scope and aims. For more information on verification and validation in DES, we refer the interested reader to the comprehensive tutorial by Sargent [204]. Further discussion of validation may also be found in Subsection 2.2.3.1, where it is reviewed part of a taxonomy of DES experimentation goals.

2.1.2.1 Conceptual modelling

Robinson [189] introduces conceptual modelling as "the process of abstracting a model from a real or proposed system." Pidd [184], meanwhile, describes conceptual modelling as "an activity in which the analyst tries to capture the essential features of the system being modelled." These views give us some sense of the nature and purpose of conceptual modelling. To gain a clearer picture, however,

we discuss certain aspects of DES conceptual modelling in a little more detail.

Seeking to provide greater specificity, Robinson [189] defines a conceptual model as " ... a non-software specific description of the computer simulation model (that will be, is or has been developed), describing the objectives, inputs, outputs, content, assumptions and simplifications of the model." In discussing the process of conceptual modelling, a variety of individual tasks and sub-processes are identified by Robinson [190]. These include: gaining sufficient familiarity with and knowledge of the problem situation; discussion and elicitation of the objectives of the study; abstraction of the appropriate inputs, outputs and content for the conceptual model; and the collection and analysis of the data required. In examining an alternative framework proposed by Banks et al. [11], clear parallels with Robinson's programme of tasks may be observed upon noting the preliminary steps of "problem formulation," "setting of objectives and overall project plan," "model conceptualisation' and "data collection."

Pidd [184] also discusses conceptual modelling, suggesting that the particular subset of features that ought to be deemed essential (and so included in the conceptual model) depends greatly on the purpose of the project. He further argues that a model may not be considered universally valid, but rather it may only be considered valid in connection with a specific purpose. Another issue discussed by Pidd concerns the form of the specification of the conceptual model, noting that "it may exist as a set of flow diagrams, as a textual description or as a mix of the two." Pidd also points out that in some cases (e.g. small-scale projects) the conceptual model may not exist in any concrete form, but instead remain solely in the mind of the analyst.

2.1.2.2 Model Coding

Robinson [190] describes model coding, the second process identified in his framework, as "the conversion of the conceptual model into a computer model." He is quick to highlight that coding is meant in a general sense, as the translation of the non-software specific conceptual model into a software specific computer model, rather than the use of a programming language, per se. This notion is reflected in the use of "model translation" by Banks et al. [11], and "computer implementa-

tion" by Pidd [184], to describe the same process.

Banks et al. [11] distinguish between three different categories of DES software. Firstly, they mention the use of general-purpose programming languages, citing C/C++ and Java as examples. Secondly, they refer to simulation programming languages, referencing GPSS/H and SIMAN V as examples. Thirdly and lastly, they mention simulation environments. The authors note that the products they term simulation environments differ in many aspects, e.g. in terms of cost and application area. However, these products routinely feature characteristics such as a graphical user interface (GUI), animation and "an environment that supports all (or most) aspects of a simulation study." Packages such as Arena, Flexsim, ProModel and Simul8 (amongst others) are given as examples of simulation environments. Banks et al. also mention the possibility of using spreadsheet software for certain classes of simulation problems, such as "risk analysis, financial analysis, some reliability problems, and others." However, they also note that "the spreadsheet has severe limitations for most complex real-world dynamic, event-based simulations." These views are much in line with Pidd's [184], where he distinguishes between general purpose programming languages, simulation languages and GUI-based VIMS (Visual Interactive Modelling System) packages. He acknowledges the value of spreadsheet simulation in certain simple problem situations, but contraindicates its use in more complex dynamical system modelling applications.

Law [154] and Robinson [190] make similar points regarding spreadsheet simulation, but instead distinguish between just two categories of software: general purpose programming languages and specialist simulation software. In terms of the advantages offered by general purpose programming languages, they identify: increased modelling flexibility, more efficient codes (and so shorter required simulation time), increased model portability, and lower cost. For specialist simulation software, they identify: a framework more naturally suited to the task at hand, a lower required coding skill-level, and less required coding time. In terms of software choice, Robinson [190] provides a detailed guide to software selection, but suggests that "as long as the software suffices, the expertise of the modeller ... is probably of far greater importance." Law [154] meanwhile suggests the software choice should ideally determined by the needs of the project, however in practice,

availability and analyst skill set are often the determining factors. We refer the interested reader to Swain [223] for an extensive list of commercial software and their characteristics and capabilities.

2.1.2.3 Experimentation

Pidd [184] introduces simulation experimentation, the third of Robinson's processes [190], in a very general way, describing it simply as "the use of the model," and noting the entire point of constructing a simulation model is to perform experiments with it. Robinson [190] highlights the purpose of such experimentation is "to obtain a better understanding of the real world and/or find solutions to real-world problems." He describes simulation experimentation as "a process of 'what-if' analysis," in which model factors are altered, the model is executed and model responses are observed. He suggests that the knowledge gained from each iteration of the process should be used to guide future iterations, with the process continuing until sufficient insight has been obtained. These views provide a very quick overview of the nature and purpose of DES model experimentation.

Given the detailed exposition of DES model experimentation presented later in Section 2.2.1, we defer further discussion of the topic here. This later material draws on Barton's taxonomy [16], which distinguishes several broad classes of experimentation activities: including validation, factor screening, sensitivity analysis and gaining understanding, predictive modelling, choosing the best design, and optimisation and robust design. This taxonomy provides a useful organising framework for appreciating the breadth of aims that DES experimentation may serve, and it helps clarify where different methodological contributions, including those developed in the current thesis, naturally sit within that broader landscape. Accordingly, we conclude this brief introductory treatment and proceed to discuss implementation, the final stage in Robinson's DES project framework.

2.1.2.4 Implementation

In discussing implementation, his fourth key process, Robinson [190] reflects on the general meaning of the term as "putting something into effect or carrying something out" to shed light on how it applies in a DES context. He also notes that "it is in this part of a simulation study that the modelling effort actually has an

effect on the real world." Pidd [184], recognising the practical focus of OR & MS applications, remarks that the topic of implementation should not be neglected in discussions of DES best practice.

Robinson [190] distinguishes between three forms of implementation that may accompany a simulation project, namely "the implementation of the findings, the implementation of the model and implementation as learning." Implementation of the findings refers to putting the proposed solution into practice in the real system, whereas implementation of the model relates to the application of a DES model developed for on-going organisational use. Implementation as learning is less explicit, but can essentially be described as the process by which the increased knowledge of system behaviour on the part of the client leads to improved decision-making and management in the real world.

Pidd [184], meanwhile, describes two types of implementation. The first type is termed "a tangible product" and refers to a set of recommendations based on the analysis carried out in the simulation study. Pidd notes that "these benefits are usually the official reason for conducting the study in the first place." The second type Pidd refers to is "improved knowledge and insight," which he claims occurs when "models and model-building are used as tools for thinking." Pidd asserts that this latter type of implementation occurs mainly at two points in the project: during conceptual modelling (through asking pertinent questions and collecting relevant data) and during experimentation (through counter-intuitive results leading to the correction of faulty mental models of the system).

Banks et al. [11] highlight two further important points concerning implementation in DES projects. Firstly, they claim successful implementation relies upon the successful execution of earlier processes, with validation particularly emphasised. Secondly, they highlight that successful implementation is also dependent upon good communication, and in particular "how thoroughly the analyst has involved the ultimate model user during the entire simulation process."

Having discussed each of the four key processes, and in doing so obtained a better picture of how DES is applied in practice, we will now turn our attention to DES model complexity.

2.1.3 DES Model Complexity

In this section, we turn our attention to the discussion of the varying degrees of complexity of DES models. Initially, we discuss the concept of complexity itself, relating it to the DES context. This is followed by a summary of the different contributions from DES researchers on the topic of model complexity. Next, we examine the relationship between complexity and modelling practice. We finish this section with a discussion of "necessary" complexity, and some of the resulting practical complications. As previously mentioned, model complexity can complicate the process of DES experimentation. As such, this section provides both useful context for Subsection 2.1.3 on DES experimentation, and motivation for the current research as a whole, with its focus on increasing the efficiency of DES experimentation.

Complexity is a broadly defined concept, with a wide variety of interpretations across different contexts. In general parlance, complexity is often cast as "the state or quality of being intricate or complicated," in reference to concepts, ideas and situations [175]. In various fields of inquiry, however, complexity has a specific, technical meaning. For example, in computational science, complexity concerns the scale of resources required for the execution of algorithms [174]. Commonly used measures of problem complexity are: time complexity, the length of time required for solution; arithmetic complexity, the number of arithmetic operations required for solution; and space complexity, the volume of memory required for solution. In software engineering, programming complexity relates to the number of interactions between the different elements of the software program, and is a measure of the complexity or intricacy of the software design [138]. In business, complexity management is an approach that aims to study, evaluate and control complexity in organisations [5]. Complexity, from this viewpoint, stems from factors such as the diversity of an organisation's product and customer portfolios, the intricacy of its hierarchy, and the variety and types of materials, processes, and technologies, involved in its operation. In cognitive psychology, the Hrair limit asserts that a human being cannot deal with more than 7 ± 2 activities simultaneously, suggesting fundamental limitations regarding the complexity of concepts which may be comprehended [256]. Each of these notions of complexity bears some relevance to our present discussion, as we shall see below.

In the context of DES, there appears to be no single, clear-cut definition of model complexity, nor generally accepted way of measuring it [56, 39]. Some contributions to the debate relate the complexity of a model to structural aspects (e.g. the number of levels, components, and interconnections); others focus on cognitive aspects (e.g. model transparency and comprehensibility); others still include the notion of calculational complexity (e.g. the extent of the calculational operations required for model execution).

Ward [238] discusses and attempts to decompose model complexity by differentiating between two key concepts: transparency and constructive simplicity. Transparency relates to the level of "user-comprehension of a model," whilst constructive simplicity is defined as "a primarily objective concept ... describing the form and level of detail of the model." Here, Ward contends that if a model is constructively simple, then it is also, necessarily, transparent. Meanwhile, Brooks [39] highlights that model complexity is frequently conflated with level of modelling detail, but that subtle distinctions exist between the two concepts. Brooks and Tobias [40] suggest that three sub-components are useful in decomposing model complexity: the number of model elements, the number of relationships between the elements, and the nature of the relationships between the elements.

Some progress has been made concerning the development of objective, quantitative metrics of the complexity of DES models. Wallace [235] defines the control and transformation metric, CAT, which measures the complexity of a simulation model represented using condition specification, a model specification language developed by Overstreet and Nance [179]. Schruben and Yücesan [258] propose complexity measures based on event graph representations of DES models [210], following from work on the measurement of complexity within software engineering. They also provide several specific use cases for their complexity measurement tools. They note that a priori assessment of complexity can be useful in DES project planning to help ensure budget limitations are not exceeded. Further, they also highlight complexity measures may be useful in classifying DES models to ensure that model test beds are sufficiently comprehensive for use in DES methodological research.

Despite the fact that no precise definition of complexity exists, much appears to have been written in the DES literature attesting to the superiority of simple models. In his DES textbook [184], Pidd outlines six principles of modelling in management science, several of which focus on different aspects of keeping models simple. Many other authors, including Zeigler [259], Law et al. [155], Salt [192], Pedgren, Shannon and Sadowski [181], Paul [228], Brooks [39], and Robinson [189], express similar views and set out the advantages of simple DES models. These include shorter development times, shorter run times, increased comprehension of model structure and content, and more straightforward interpretation of model results.

Against this, however, several authors point out the dangers inherent in model oversimplification. Alan and Pritsker [2] reflect upon experience in model development, noting that simpler models are not always easier to update and develop as modelling requirements change. Yücesan and Schruben [258] contend that simpler models are not always as easy to understand, code and debug. In the context of simulation in healthcare, Davies et al. [67] point out that simpler models require more extensive assumptions to be made about the operation of the underlying system. They highlight the potential danger of setting the system boundary too narrowly, and in doing so, missing important aspects of the underlying system. In discussion of agent-based simulation, Edmonds and Moss [73] argue there is a need to move away from the concept of KISS (keep it simple, stupid) to a new mantra: keep it descriptive, stupid (KIDS).

Overall, it seems sensible to make models as simple as possible, but no simpler. Occam's razor, frequently referenced in discussions of modelling best practice, states "the best explanation of an event is the one that is the simplest, using the fewest assumptions or hypotheses" [173]. Thus, in a DES context, given two models, each equally representative of the underlying system, we would opt to use the simpler of the two. In doing so, we would be favouring simplicity, but not at the cost model validity.

Models, of course, should not be unnecessarily complex, rather the complexity should be driven by context and purpose. Large-scale, complex models are frequently necessary in the simulation modelling of military systems [43, 59, 157], bio-

logical systems [86, 207], social systems [28] and telecommunications networks [70, 240], amongst other areas. Typically, issues of modelling, validation and experimentation are more challenging in such cases, and are referred to as "problems of scale" by Nicol et al. [170]. Simplification methods such as those proposed in Zeigler [259], Innis and Rexstad [128], Yin and Zhou [253], Robinson [190] and Chwif et al. [58], are valid and useful contributions to DES methodology, but they provide only a partial answer to the problem of complexity. Efforts must also be made to develop techniques enabling analysts to deal effectively with the large-scale, complicated and computationally-intensive models necessitated by many real systems.

2.2 DES Model Experimentation

Having introduced DES in the preceding subsections, we now turn our attention to the exploration of DES model experimentation. This is an important part of the literature review, as the proposed contributions of the thesis lie, largely, within the area of DES experimentation.

To begin with, a broad sketch of experimentation in DES is provided. This aims to highlight the identifying characteristics of DES model experimentation, as compared with experimentation in other fields. Following this, a taxonomy of DES experimental goals is presented. This is useful in light of the great diversity that appears in DES model experimentation. Lastly, a literature review of the existing approaches to DES experimentation is provided, structured with respect to the aforementioned taxonomy of goals.

2.2.1 Experimentation in DES

Experimentation is an intrinsic element of the DES approach. Indeed, as noted in relation to Robinson's framework [190] , experimentation with the simulation model constitutes the primary means by which knowledge and understanding concerning the system of interest is obtained [184]. Whilst some knowledge is gained at other stages of the DES life cycle, such as in conceptual modelling or implementation, the core purpose of the design and development of a DES model is to conduct a process of experimentation with it, in order to gain insight. Expressing

this slightly differently, Law [154] highlights that a DES model cannot be solved analytically, rather it must be "solved" numerically, by way of a process of experimentation. Experimentation, therefore, is the means by which we leverage the simulation model to learn about the underlying system of interest.

Looking a little more broadly, experimentation itself is an extremely useful, intuitive and widely applied means of learning, intrinsically linked to the scientific method [34]. The Britannica Dictionary defines an experiment as "a scientific test in which you perform a series of actions and carefully observe their effects in order to learn about something" [176]. Thus, it may be seen that experimentation constitutes a fundamental, empirical learning mechanism.

Experimentation has itself been the subject of much academic study, and an extensive body of literature exists concerning appropriate and indeed optimal methods for the design and analysis of experiments. Such methods are statistical in nature: classical experimental design is a well-established field of mathematical statistics, with its origins in the agricultural experiments of Fisher [88] in the 1920s and 30s. This foundational work was subsequently developed for use in industrial experiments through the work of Box and Wilson [36] in the early 1950s, and Taguchi et al. [226] in the late 1970s and 80s. The techniques of design of experiments (DOE) developed in such settings are now applied in fields as diverse as psychology, business, biology and astronomy [164, 32].

Returning to the context of DES model experimentation, we briefly discuss a number of its hallmark characteristics, as an awareness of these DES-specific characteristics is helpful for the development of effective DES model experimentation methods [145]. Essentially, these points may be divided into two categories: those stemming from the greater degree of control typically available in DES model experimentation and those relating to the nature of the problem settings typically tackled in such a context. We discuss these issues in turn in the following subsections. Before proceeding, however, we briefly discuss some pertinent DOE terminology, building upon (and, where appropriate, translating from) the DES terminology outlined previously in Subsection 2.1.1. Firstly, we note that we will use *factor*, *response*, *level*, *experimental design* and *design point* as introduced in Subsection 2.1.1. Alongside *model scenario* and *run*, however, we have *experimen-*

tal scenario and *trial*, as the closest to equivalent terms in general DOE parlance. Lastly, we introduce *design space*[2] and *response space*, the multidimensional spaces implicitly defined by the factors and responses, respectively, of interest.

2.2.1.1 Large-Scale Nature of DES Experimentation

Typically in physical experiments, there are many factors potentially of interest to the experimenter, each with a range of possible levels. This could equate to an extremely large number of experimental scenarios, if each possible combination were to be examined. Usually, however, the experimenter is faced with a number of constraints which limit the scale of the experimentation possible in practice.

Often, a subset of the potentially interesting factors will be difficult or impossible to control, and therefore cannot be varied during experimentation. Further, each additional trial may require significant additional resource, placing yet more constraints on the scale of experimentation. This may be compounded by the fact that additional resources may be required for the measurement of each response of interest, thereby further limiting the dimensionality of the experimentation. These constraints, common to physical experimentation, have led to traditional DOE's focus on smaller-scale experimentation.

In DES model experimentation, however, the situation is quite different. To begin, the DES experimenter has far greater control over the experimental environment. They have the opportunity to treat any model assumption or quantity as an experimental factor, varying its level simply by changing a few digits on a computer program. Further, the resource cost of a simulation run is, usually, lower than that of a trial of a physical experiment. Once completed, the only additional cost incurred in running a DES model is time, and often this can be scaled, with weeks or months of simulated time passing in minutes of real time. In DES experimentation, data on many different responses of interest can be collected simultaneously, usually without incurring any additional cost.

[2]Here, we note the distinction between the experimental design and the design space, with the experimental design a specifically selected subset of the design space to be evaluated during the course of experimentation.

The preceding discussion might lead the reader to believe that anything is possible in the world of DES experimentation. However, there are still constraints in this context. Whilst the typical reduction in resource required for DES experimentation leads to a greater budget for experimentation, the increased control over the environment leads to a vastly greater design space to be explored. This is due to the curse of dimensionality, a concept first introduced by Bellman [21] in the area of adaptive control processes. In this context, the issue is stated as follows: "the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with respect to the number of input variables (i.e. dimensionality) of the function" [50].

Let us illustrate the curse of dimensionality as it applies to DES by means of a simple example. If we consider a DES model with just two factors, each with just two levels, this results in $2^2 = 4$ scenarios. If, however, we let the number of factors increase, e.g. with k factors, letting $k = 1, 2, 3, \ldots$, we see that $2^3 = 8,\ 2^4 = 16,\ 2^5 = 32,\ 2^6 = 64, \ldots$. That is, a linear increase in the number of factors leads to an exponential increase in the number of scenarios. Generally speaking, this exponential increase in model scenarios cannot be offset by the relative reduction in the cost of experimentation in a DES environment. That is, when the number of scenarios to be examined grows exponentially, experimentation can quickly become impractical, even impossible, no matter how quickly or cheaply the model can be executed. This constitutes a significant constraint on the effective use of DES models in practice [200].

We also note that with the ability to collect data on any response of interest without incurring additional cost, the dimensions of the response space can also grow very large. This can lead to potential complications in the analysis and interpretation of the experimental data obtained. Such ideas demonstrate the potentially large-scale nature of DES model experimentation. This characteristic of DES model experimentation presents many opportunities, but also a number of challenges [198]. It implies that a change of perspective may prove beneficial in approaching the design and analysis of DES experiments.

2.2.1.2 Sequential versus Simultaneous DES Experimentation

In classic experimentation contexts, trials are, as far as possible, carried out simultaneously, with sequential designs viewed as being susceptible to validity issues. When data must be collected sequentially, the order in which samples are collected is typically randomised to ensure that changes in the experimental conditions over time do not unduly bias the results [164]. As such, many classic experimental designs are so-called 'one-shot' designs. Such issues are not relevant in the DES context, owing to the increased degree of control over the experimental environment [144]. Indeed, the vast majority of simulation experiments are carried out sequentially.

This difference between contexts may be exploited to facilitate more efficient experimentation. An example involves a simple two-stage design, where the first enables the construction of a first-order metamodel, and the second augments the experimental design with additional design points, allowing the construction of a second-order metamodel.[3] Here, we note that a *metamodel* is a model of the simulation model, approximating the relationship it defines between its input factors and output responses [144]. With this sequential approach, the first stage is implemented, and the model constructed and evaluated for fit. At this point, the experiment may be terminated or continued according to the adequacy of the degree of fit obtained in the first stage [200]. On a similar note, ranking methods sometimes feature a two-stage structure, with an initial number of simulation runs executed for each scenario.[4] The data obtained is analysed and used to determine an appropriate number of subsequent runs so as to provide sufficient inference on the ranking problem at hand [154].

Of course, such ideas are not universally applicable, however, this additional degree of freedom is a notable feature, and so is worth taking into account in considering new methods for DES experimentation.

[3]The use of linear regression models as simulation metamodels is discussed in further detail in Subsection 2.2.3.4.

[4]Ranking and selection methods in DES are discussed in further detail in Subsection 2.2.3.5.

2.2.1.3 Complexity of DES Response Surfaces

DES is frequently applied in more complex problem contexts. Indeed, Law [153] notes that DES is often referred to as a "method of last resort," to be applied when direct experimentation and analytical modelling do not offer sufficient flexibility to validly represent the system of interest.[5] Such problem contexts are a long way from the typical settings of the agricultural and chemical experiments leading to the development of classic DOE [164]. The complexity inherent in DES problem context soften translates to a complex *response surface* [144]. A response surface represents the relationship, here defined by the DES model itself, between the input factors and the output responses.

Traditionally, linear regression models have commonly been used in the analysis of experimental data [164]. This involves the fitting of a linear model to the response surface implicitly defined by the experimental process. Valid use of such models requires a number of assumptions, namely: the response surface is adequately represented by a low-order polynomial model; and the errors are normally, independently and identically distributed (NIID), with zero mean [164]. In DES experimentation, however, the adoption of such assumptions is not straightforward. Whilst the normality assumption may be justified, to some extent, through an appeal to the central limit theorem (CLT) [25], we note, however, that DES responses are not always the equivalent of simple sums of independent factors. Therefore, as highlighted by Kleijnen [143], in practice, this assumption should be checked. Kleijnen also discusses the assumption of identically distributed errors, noting that it is common to find differences in the variance of errors between different parts of the design space. This phenomenon is termed *heteroscedasticity*, and it is commonly encountered in queuing theory models [109], with the variability of the response increasing sharply as the traffic intensity approaches 100%.

Additionally, the response surfaces of simulation models may, in some cases, be highly nonlinear. They may feature spikes (localised regions differing dramatically from the immediate vicinity) and/or thresholds (contours of discontinuity). Some regions of the response surface may even display chaotic behaviour, that is the response appears entirely unpredictable, owing its extreme sensitivity to changes

[5] Law also reminds readers that DES is a very frequently necessary method of last resort, owing to the complexity of systems commonly studied.

in factor levels. It is noted that such issues are more common in agent-based simulation models (as opposed to DES models), and the interested reader is referred to references [43, 199, 234] for a more detailed discussion of these aspects of agent-based simulation. These characteristics of DES experimentation are also worthy of consideration when proposing alternative methods for its accomplishment.

2.2.1.4 Formal and Informal Methods in DES Experimentation

An extensive range of tools and techniques supporting DES experimentation may be found in the literature (with a review of these approaches provided in upcoming Subsection 2.2.3). In light of the diversity of available approaches, it might be expected that DES practice would reflect academic theory. That is, we might expect DES practitioners to make use of well-researched and mathematically optimal experimentation methods in their DES studies. However, it has been noted by many experts that, in practice, much DES experimentation is carried out in an informal and ad-hoc manner. Mark Elder, founder of Simul8, notes that practical DES experimentation tends to be conducted informally, through a process of asking pertinent what-if questions, rather than formally, through the application of mathematically optimal experimentation methods.[6]

Law [153] and Sanchez [200] also note the informal nature of practical DES experimentation. Law simply highlights that much DES experimentation is carried out in a less than optimal manner, whilst Sanchez notes the prevalent use of the so-called trial-and-error and one-factor-at-a-time (OFAT) approaches. Whilst such methods have intuitive appeal, and can at times prove successful, with knowledge gained in model building used to guide the experimentation process, they are less than optimal and can lead to erroneous conclusions. The trial-and-error approach is by nature subject to much inefficiency, whilst the OFAT approach always fails to capture information on interaction effects present between the factors [164]. Therefore, as with Law and Sanchez, we note that better and more reliable methods exist.

The disconnect between the formal approaches available in the literature and the informal methods applied in DES practice warrants further attention [117]. One

[6] Personal communication with Mark Elder.

explanation might be that the formal, more technical methods found in the literature are inaccessible to DES practitioners, especially those lacking a strong mathematical background. It is also possible that the theoretical advantages offered by these methods do not translate into significant enough practical advantages to justify changes in practice. Finally, the benefits of these more formals methods may not be communicated with sufficient clarity, leading to a lack of awareness on the part of DES practitioners. These points highlight areas where care should be taken in the design and dissemination of new DES experimentation methods.

2.2.2 A Taxonomy of DES Experimentation Goals

The high degree of flexibility offered by DES leads to a huge variety of different applications areas. Whilst the systems studied and models created differ significantly from one application area to another, the steps involved in applying the DES methodology do not, as discussed in the preceding subsections. This point also holds true in relation to DES model experimentation. It is found in practice that even across diverse application areas, DES experimentation is carried out in support of similar goals, and tends to progress through similar stages in a similar manner [16]. Here, we discuss these common goals of DES experimentation.

Initially, experimentation is conducted to support the validation of the simulation model. This is an essential step to help ensure that the conclusions drawn from the simulation project are meaningful and reliable. Following this, a common preliminary stage is factor screening. The general idea is that many model inputs (assumptions or quantities) are initially considered as factors to be varied during subsequent experimentation. Factor screening experiments are then performed to reduce this set of possible factors. Ideally, only a relatively small subset of truly important factors will remain after this process.

Having reduced the dimensionality of the design space to a manageable level, attention often turns to gaining a greater understanding of the behaviour of the responses of interest. This may involve obtaining a better picture of how changes to the factor levels impact the responses of interest, usually termed sensitivity analysis. Alternatively, it may involve attempting to predict the responses of interest for new model scenarios, not just those examined during the course of experimen-

tation. This usually involves the construction of a simulation metamodel.

The final stages of the experimentation process often centre around the identification of the model scenario(s) maximising (or minimising) the responses of experimental interest. This may involve selecting the best from a set of model scenarios, or optimising across a continuous design space. Alternatively, the prime concern may be the selection of a model scenario minimally affected by variations in environmental conditions (i.e. model aspects normally considered constant in the experimental process). Such a model scenario is termed robust.

The foregoing discussion is summarised in the taxonomy of DES experimentation goals proposed by Barton [16], summarised in Table 2.1. We note that very similar classifications of experimental goals are presented by Kleijnen and Sargent [146] in the context of computer experimentation, and by Montgomery [164] in the wider context of scientific and engineering experimentation. Barton's taxonomy provides not only a useful structure for the review of DES experimentation literature in the following subsection, but also a useful framework to better understand where the contributions of the thesis might find application in the DES experimentation process.

Lifecycle stage	Simulation goal
Early	Validation
Early	Screening variables
Middle	Sensitivity analysis / understanding
Middle	Predictive modelling
Late	Selecting the best configuration
Late	Optimisation / robust design

Table 2.1: Simulation goals by lifecycle, as proposed by Barton [16].

2.2.3 Existing Approaches in DES Literature

In this section, we seek to provide an overview of the approaches found in the literature corresponding to each of the DES experimentation goals identified in Barton's taxonomy [16]. That is, in the following subsections, we discuss in turn

validation, factor screening, sensitivity analysis / understanding, predictive modelling, selection of the best scenario, and optimisation / robust design.

2.2.3.1 Validation

In any real-world modelling context, questions concerning the validity of the model will likely arise, owing to the potential impacts of the implementation of the insights obtained. Here, we briefly discuss the methods found in the literature for assessing the validity of DES models.

Distinct from the narrower aim of verification (primarily concerned with ensuring the fidelity of the computer-based implementation of the conceptual model), validation has the broader scope of addressing the general suitability of a DES model for a particular domain of application. Referring to the classical definition of Schlessinger et al. [208], we define model validation as the "substantiation that a computerised model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model."

In the context of DES experimentation, validation is an important step, affecting any subsequent experimental goals of the study. It is therefore identified as an "early" goal in Barton's taxonomy [16], even though it is often considered a continuous process, with activities continuing until the end of the project [190]. As such, there are various forms of validation discussed in the literature, including but not limited to conceptual model validation, data validation, experimentation validation and solution validation, as noted by Robinson [190]. Sargent [203] also distinguishes between validation for observable systems and validation for non-observable systems, as well as between subjective and objective approaches.

Perhaps more concretely, the validation process may be seen to govern the development of model logic and the collection of modelling data. This early validation-focused experimentation often leads to the adjustment of modelling assumptions, in particular those associated with model logic or data, so as to ensure a good fit between model performance and observed real-world data. Whilst conceptually simple, some key challenges exist, such as the lack of a one-to-one pairing between model inputs and responses (for example, several different sets of model inputs

may generate the same model responses), and the lack of a unique model response for a given set of modelling inputs and assumptions. The process of DES model validation can be both difficult and costly: whilst the failure of a DES model to provide accurate results for a single model scenario invalidates the model, it is not always clear whether a broad range of scenarios with successful results is sufficient to 'validate' a model. The need for extensive model testing to address this issue, particularly for models of complex systems, may require substantial computational investment.

There has been a substantial amount of research in the area of validation, where many tools and techniques have been developed and evaluated. In particular, we note the broad range of statistical methods employed in validation. These range from classical methods, such as Balci and Sargent [9, 10] (based on two-sample t-tests and t-confidence intervals, respectively) to more recent developments such as the semantic composability focused work of Szabo and Teo [225], and the multivariate hypothesis testing based work of Sargent et al. [205]. The interested reader is referred to Sargent [203] for a thorough review of simulation model validation, as well as to Tolk et al. [229] for a thorough discussion of validation in military applications: a domain frequently featuring large-scale and complex systems, exacerbating the standard issued faced in DES model validation. Finally, we note in summary that whilst a range of different approaches to validation exist, with the ever-increasing scale of DES experimentation, there still exists a need for more rigorous experimentation.

2.2.3.2 Factor Screening

DES models frequently feature a large number of potential experimental factors. As discussed in some detail in Subsection 2.2.1.1, this characteristic of DES experimentation follows from the fact that in a simulation experiment essentially everything is controllable. That is, any and all structural assumptions and input quantities can in theory be varied, and so may be added to the list of possible experimental factors.

Case studies involving large numbers of possible experimental factors are prevalent in the DES literature. In one example, Yaesoubi and Roberts [249] examine

72 factors in a medical simulation model called the Vanderbilt/NC State model, designed and constructed for the purposes of investigating the natural history of colorectal cancer. Even allowing each factor to assume just two levels, if we desired to run just a single replication of each possible combination, this would lead to a total of $2^{72} \approx 5 \times 10^{21}$ simulation runs, clearly infeasible regardless of the amount of computing power available.

Whilst it is entirely possible to, a priori, designate a large number of factors as constant environmental factors, and to then experiment with just a few, a more considered approach involves performing a factor screening experiment. Such experiments are typically conducted in the preliminary stages of the experimental process and involve the investigation of a large number of controllable factors, with a view to eliminating the unimportant ones. By the sparsity of effects principle, most often only a few factors are responsible for the majority of the variation in the response(s) [164]. An effective screening procedure should correctly and efficiently identify this small subset of important factors. The factors identified as important may then be further explored in later phases of the experimental process, perhaps used to construct a metamodel to understand and optimise the DES model response.

Beyond its core purpose, factor screening may also be used to confirm (or disconfirm) prior expectation with regards to the important factors, helpful in validating the simulation model. Moreover, factor screening may actually assist in the data collection required for DES input modelling. By performing a preliminary data collection exercise to construct a working model, factor screening may be performed to help identify the influential factors, and so enable the informed prioritisation of further data collection efforts [250].

A range of different approaches to factor screening have been proposed in the DES literature. A very popular approach to factor screening in classic experimentation is the use of two-level factorial and fractional factorial designs (typically of resolution III and IV) [164]. Such designs have been used in DES experimentation, however they are generally not well-suited for use in situations involving more than around twenty factors [230]. More complex approaches such as Latin hypercube and supersaturated designs [115] are also classically used in factor screening, yet

do not appear to be greatly used in DES experimentation. This is perhaps owing to the relatively complex nature and derivation of such designs, and the limitations placed upon the number of factors which may be assessed [250]. The interested reader is referred to the classic texts by Montgomery [164] and Hinkelmann and Kempthorne [115] for more information on such experimental designs.

The majority of recent developments in factor screening in DES model experimentation have been sequential methods, in particular those based upon sequential bifurcation (SB). This approach, based on binary search technique, was originally proposed in relation to deterministic computer simulation experimentation by Bettonvil and Kleijnen [26, 27] and subsequently adapted for use in stochastic computer simulation by Cheng [52]. Its popularity stems from its increased efficiency over approaches used in classic experimental design.

The general approach of SB involves the initial aggregation of all factors into a single group, with this group tested to assess whether or not it contains an important factor effect. If so, the next step splits the group into two subgroups and tests each of these for important factor effects. The process continues in a similar manner, where subgroups deemed important are split for further testing and those without important factor effects discarded. In the final stage, the effects of all individual factors not in subgroups deemed unimportant are estimated and tested. The typical assumptions made in SB are that a first-order model adequately represents the simulation response surface and that the signs of all main effects are known and non-negative [144].

A number of developments of the SB approach have been proposed in the DES literature. Two important advances are controlled sequential bifurcation (CSB) [236] and a variant called CSB-X [237]. CSB incorporates a two-stage hypothesis testing approach into SB to control error and power in factor assessment, providing the user with more certainty regarding the results. CSB-X builds on the former approach through the relaxation of certain assumptions concerning interaction effects, leading to an increase in performance in relevant situations. Also of interest are fractional factorial controlled sequential bifurcation (FF-CSB) and a variant FFCSB-X, both proposed by Sanchez et al. [201], which utilise an initial fractional factorial design to relax certain assumptions made in the CSB and CSB-X

approaches, again providing more efficient inference in relevant situations. Finally, MCh-X [250] is a modification of Cheng's method which is able to perform well when interaction or quadratic effects are present, increasing the range of situations in which the approach is applicable.

Although differing approaches have been proposed, as is usual in experimentation, a gain in computational efficiency comes with either a reduction in the information obtained, or with the additional requirement of more stringent and limiting assumptions. It also appears that most currently available approaches are based on the estimation of first- or second-order linear regression models. Given the scale and complexity of DES models, a clear need exists for more flexible and rigorous approaches to such experimentation.

2.2.3.3 Sensitivity Analysis and Understanding

Given validation assesses the simulation model's representation of reality, and factor screening identifies its truly important factors for later experimentation, it may be argued that the sensitivity analysis and understanding stage in Barton's taxonomy [16] corresponds to the start of 'true' DES model experimentation. After all, most often, it is only at this point does the analyst begin to use the model to ask the questions the model was created to answer. Whilst Barton [16] does not provide great detail as to what is meant by "Sensitivity Analysis, Understanding," from the structure of the taxonomy it is clear this stage relates to increasing understanding of the model's response surface. This may relate to increasing understanding of a point in the design-response space, a localised area of this space or its entirety, depending on the context. Frequently, this relates to the asking of what-if questions, without the imposition of a formal experimental design or predictive model [144].

The point made around the wide breadth of possible scope of understanding sought represents a convenient juncture at which to discuss some important DES experimentation terminology. In his classic paper [61], Conway introduces the distinction between "tactical" and "strategic" simulation experimentation. He defines strategic planning as "the design of an experiment that will yield the desired information," thus including aspects such as the selection of factors and/or model

scenarios of interest. Tactical planning, on the other hand, is defined as "the determination of how each of the test runs specified in the experimental design is to be executed," and so includes practical issues such as the run-length and initial condition specifications. As such, in this subsection, we firstly discuss tactical aspects of simulation, so as to provide a more complete characterisation of DES experimentation practice, before moving on to introduce strategic aspects of simulation: the focus for the remainder of the review of DES experimentation.

To begin, we note that the methods proposed in the tactical DES experimentation literature tend to focus on the creation of appropriate samples of response data from a single long model run. They seek to increase the efficiency in estimating the response(s) of interest through the creation of multiple, approximately independent and identically distributed (IID) observations from this single long run of the model. The idea behind such an approach is to avoid the many shorter runs that would normally be required to achieve such a quantity of data. This enables computational savings to be made through the omission of the time normally spent warming-up the model on each of these runs. However, two key challenges in this context are determining an appropriate warm-up period for the single, long simulation run, and contending with the presence of autocorrelation between successive observations of many DES responses of interest [144].

A variety of approaches have been offered to assist the DES practitioner in determining an appropriate warm-up period, that is, in determining a point at which the model has reached a steady-state [190]. A classical approach to this issue, and perhaps the most frequently used in practice [154], is Welch's method [243]. This is a graphical method, involving the averaging of the response time series across a number of different model runs, followed by the use of different moving averages to evaluate when the model has reached its steady state. Aiming to provide a comprehensive review of the area, Hoad et al. [118] survey the approaches found in the literature, grouping them into five categories: graphical methods; heuristic approaches; statistical methods; initialisation bias tests; and hybrid methods. Hoad et al. [118] also proposed their contribution to the area, the MSER-5 heuristic, based on the MSER heuristic introduced by White [244]. The interested reader is directed to this survey as a means of navigating this area of the DES literature.

Surveying the methods available in the literature to make use of data from a single, long simulation run, several categories exist: the batch means method, spectral analysis, autoregressive analysis, the standardised time series method, and the regenerative method [154]. The first four approaches stem from the assumption of (at least some form of) stationarity.

The batch means method assumes the simulation is covariance-stationary and involves determining an appropriate batch length such that the means from each of the batches are essentially uncorrelated. It was first discussed the context of simulation by Conway [61], with subsequent derivatives such as consistent batch means [66] and overlapping batch means [163] introduced later. Spectral analysis, discussed by Fishman [89], likewise relies on stationarity (so that a spectral density is well defined) and uses Fourier Transform theory to estimate the variability of the batched sample means in support of confidence interval creation. Autoregressive analysis also assumes covariance-stationarity, but also assumes the simulation can be adequately represented by a p-th order autoregressive model, with output depending linearly on previous values and an additional stochastic term. This approach was introduced by Fishman [90, 94, 93], with subsequent development by Yuan and Nelson [257] amongst others. The standardised time series approach, introduced by Schruben [209] makes stronger assumptions in calculating confidence intervals for the mean, namely that the process is strictly stationary and *phi*-mixing. In contrast to all of these, the regenerative method is based on the notion that the simulation probabilistically "starts over" at random points; these regeneration points induce cycles that can be treated as (approximately) independent observations for inference. This approach was developed simultaneously by Crane and Inglehart [63, 64, 65] and Fishman [91, 92]. The interested reader is referred to Alexopoulos [4] for a comprehensive review of such methods.

Having now addressed two of the key tactical aspects of DES experimentation, namely how to determine appropriate warm-up periods and how to obtain approximately IID samples from a single long run, it is now appropriate to turn attention towards what Conway [61] characterised as the "strategic" aspects of simulation experimentation. Whereas tactical planning is concerned with how a given scenario is executed, strategic planning focuses on selecting which scenarios to investigate and determining the breadth of the design–response space to be ex-

plored. In this sense, the focus shifts from the reliable execution of single scenarios to the broader task of accumulating evidence across multiple scenarios to support broader inferences about system behaviour.

In the overall process of DES experimentation, these early-stage strategic issues can be understood as occupying a position between the initial tasks of validation and factor screening on the one hand, and the more sophisticated inferential goals of predictive modelling, selection of the best scenario, and optimisation on the other [16]. For instance, if a simulation model of a proposed warehouse layout were built to confirm expectations concerning its operations, the analyst would be concerned only with tactical aspects of the model's use. By contrast, if questions remained about the suitability of the layout, the analyst would also need to consider strategic aspects, in particular, which layout variants to explore during experimentation. The range of scenarios to be investigated, whether just two or three alternatives or a continuous multi-dimensional design space, would depend greatly on the context in question, as would the way in which evidence across these scenarios is utilised to inform inferences about system behaviour. These decisions about which scenarios to run and how to interpret results across them constitute the bridge into navigating the design–response space [154, 144]. The importance of efficient experimental design in enabling such exploration, while avoiding unnecessary computational effort, has been strongly emphasised in the literature [200].

The difficulty of this stage has been described as a 'chicken-and-egg problem' by Kleijnen et al. [145]: the analyst's choice of experimental design directly influences the analytical methods that can subsequently be applied, while at the same time the choice of analytical approach should ideally inform the construction of the experimental design in the first place. This interdependence makes early-stage strategic experimentation particularly challenging, since the analyst must balance the need for flexibility in design with the constraints imposed by their intended inferential goals. Care is therefore required to ensure that decisions taken at this stage do not preclude the use of more sophisticated analytical techniques later in the study [145].

As was discussed in Subsection 2.2.1.4, much practical DES experimentation continues to be conducted in an informal manner, with analysts often relying on trial-

and-error or one-factor-at-a-time investigations, rather than formal experimental designs [154, 200]. This tendency is especially pronounced at the early strategic stage, where the natural inclination is to explore a few plausible scenarios and observe their outcomes in order to build intuition about system behaviour [190]. While such practices can yield useful insights, they lack the structure required to ensure efficient coverage of the design–response space or to support reliable inferences across scenarios. Consequently, this stage of experimentation remains underdeveloped in the literature [144, 200], despite its importance in guiding subsequent analysis and, for many practitioners, representing the effective end point of a simulation study.

Although factor screening is usually framed in terms of identifying the most influential factors from a large set, several of the designs employed in screening can also contribute to early-stage strategic experimentation. Fractional factorial designs and Latin hypercube sampling, for instance, provide structured ways of spanning the design space, moving beyond one-factor-at-a-time perturbations and enabling a more representative picture of system behaviour [200]. In principle, such designs can support broader exploration at the strategic stage, but their explicit use for this purpose has received relatively little attention in the DES literature. Their limited uptake may reflect both the way these designs are typically framed in the literature and the fact that, while useful for mapping, they do not in themselves provide a framework for drawing inferences across multiple scenarios.

Simple sensitivity analysis techniques represent another strand of methods relevant to early-stage strategic work. Graphical approaches, such as scatterplots of responses against factors, together with correlation or regression-based measures, are widely used by practitioners because they are straightforward to apply and easy to interpret [154]. These methods can reveal broad relationships, thresholds, or regions of instability in system behaviour, even if they lack formal inferential guarantees. However, such graphical approaches become increasingly difficult to apply in higher dimensional design–response spaces, where interactions between factors cannot easily be visualised. More rigorous variance-based sensitivity analysis methods, such as Sobol indices or ANOVA type decompositions [193], are available and offer a systematic decomposition of response variance. However, their uptake in DES has been limited, owing to their computational demands and

the need for large numbers of replications across the design space.

Taken together, these approaches show that while some tools for exploration of the design–response space are available, their application to early-stage, understanding-oriented experimentation in DES remains limited. The methods either do not scale effectively to higher dimensional problems, do not address the central challenge of combining evidence across scenarios, or are not practically feasible given typical resource constraints. As a result, early-stage, understanding-oriented experimentation remains underdeveloped both in practice and in the methodological literature [144, 154]. This gap highlights a clear need for more rigorous yet practically applicable methods to support broader and more efficient exploration of the simulation design–response space.

2.2.3.4 Predictive Modelling

Moving beyond early-stage strategic approaches, Barton's taxonomy [16] next identifies predictive modelling as a more formal means of characterising the relationship between factors and responses in DES experimentation. Predictive modelling encompasses a range of surrogate or metamodel approaches, but within the DES literature it has been dominated by response surface methodology (RSM). Response surface methodology involves the use of mathematical and statistical techniques to form approximating, functional relationships between an output response of interest and a number of associated input factors. Originally devised by Box and Wilson for the "exploration and exploitation" of relationships between process yield and associated process variables in chemical experiments [31, 36], RSM has been very widely used and studied in DES experimentation [68, 130, 171], and has been referred to as the 'practical state of the art' in terms of DES metamodelling by Law [153]. Although other predictive approaches, such as Kriging and Gaussian process surrogates have been studied, their uptake in DES has been limited compared to RSM, partly because of their computational complexity and implementation challenges [144].

Response Surfaces and Functional Approximations

The general purpose of experimentation is to better understand the relationship between input factors and output responses. Suppose that there are k factors,

denoted $\xi_1, \xi_2, \ldots, \xi_k$, which are thought to affect a single response, y. In response surface methodology, we assume:

$$y = f(\xi) + \varepsilon,$$

where $\eta = E(y) = f(\xi)$ represents the mean response as a deterministic function f of the factors ξ, and ε represents the stochastic deviation of an individual response observation about its mean value [144].

The true nature of the underlying functional relationship between the factors ξ and the mean response η is unknown. We seek to gain insight by constructing local, low-order polynomial approximations using data obtained from designed experiments on the simulated system of interest. Such polynomials may be considered as truncated Taylor series expansions, and thus the validity of such an approach is supported by Taylor's theorem in mathematics [33].

It is often convenient to standardise the experimental factors $\xi_1, \xi_2, \ldots, \xi_k$ as follows. If factor ξ_i varies over the interval $\xi_{io} \pm \Delta_i$, then we may consider the transformation:

$$x_i = \frac{\xi_i - \xi_{io}}{\Delta_i},$$

where x_i instead varies over the canonical interval $[-1, 1]$. This is simple linear transformation of original factor ξ_i is readily reversed if required [154].

As mentioned above, the form of the approximating function is generally a low-order polynomial. Although globally the response surface may be complex, we often find that polynomials of only first- or second-order provide adequate representations over limited regions of the design space [33].

The general form of a first-order (or main effects) model is:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon.$$

In a DES context, this corresponds to estimating the direction of change in the response with respect to each factor. For example, if staffing levels increase in-

crementally, how much does mean waiting time decrease? With two factors, this model represents a plane in three dimensions: β_1 and β_2 capture how sensitive the output is to each factor while holding the other constant. Such models are useful for screening factors and identifying broad trends, but cannot capture curvature or interaction effects [154, 144].

To capture curvature and interactions, the model can be extended to second-order:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon.$$

In DES experimentation, quadratic models provide more freedom than main effects models, allowing us to approximate more complex system behaviour:

- Curvature: diminishing returns from adding more staff: the first additional server might reduce delays sharply, but further increases would likely yield smaller gains.
- Interactions: joint effects, such as how additional staff and variability in arrival patterns together affect performance.

With $k = 2$, this resulting surface is curved in three-dimensional space, representing effects such as "bowls," "ridges," or "saddles." These models are more flexible than first-order models but they require more design points to estimate the parameters β reliably [154, 144].

More generally, for any order d, we have:

$$y = g^T(x)\beta + \varepsilon, \tag{2.1}$$

where $x = (x_1, x_2, \ldots, x_k)^T$, $g(x)$ is a vector of dimension p consisting of powers and cross-products of powers of the factors $x_1, x_2, \ldots, x_k$ up to order d, β is the parameter vector, and ε is the random error term.

These polynomial models rely on statistical assumptions that can be fragile in a DES experimentation context. As discussed earlier in Section 2.2.1.3 , the adequacy of low-order polynomial approximations can be undermined by nonlinearities, thresholds, or discontinuities, and the standard "white noise" assumptions

of assumptions of regression models (errors with zero mean, constant variance, and independence) are not always defensible in simulation [143]. In particular, DES implementations of queueing models frequently exhibit heteroscedasticity, with error variance increasing sharply as traffic intensity approaches system capacity [109]. While RSM provides a practical and widely used framework, its validity is contingent on these assumptions holding at least locally, reinforcing the importance of careful design, diagnostic checking, and interpretive caution in DES experimentation.

Least-Squares Estimation

To fit the models, we require experimental data. Suppose that a series of n trials are executed, in each of which the response y is observed for specified settings of the k factors. The totality of these settings comprises what is termed the experimental design, compactly represented by the design matrix $\mathcal{D}$:

$$\mathcal{D} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

Each row of $\mathcal{D}$ corresponds to a single design point, one combination of factor levels tested in the experiment, and can be viewed as a coordinate in the k-dimensional design space [164]. In DES experimentation with a call centre model, a row might represent, for example, a particular combination of arrival rates, staffing levels, and resource priorities, with the corresponding output being mean waiting time.

Substituting this data into the general model yields:

$$y = X\beta + \varepsilon,$$

where $y = (y_1, y_2, \dots, y_n)^T$, X is an $n \times p$ matrix with $g^T(x_u)$ as u^{th} row, and vector of random errors $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ [164].

The least-squares estimator $\hat{\beta}$ is derived by minimising the sum of squared resid-

uals, leading to the normal equations:

$$X^T X\hat{\beta} = X^T y,$$

which, provided $X^T X$ is invertible, yields:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \tag{2.2}$$

the least squares estimator of β [164].

Thus, the fitted model is given as:

$$\hat{y} = g^T(x)\hat{\beta} \quad \text{for } x \in \mathcal{S},$$

with $\mathcal{S}$ denoting the design space [164].

This highlights the critical dependency on the design matrix: if the design points do not adequately span the design space, $X^T X$ may be singular or ill-conditioned, making estimation unstable [164]. In DES experimentation, this translates into the need for careful design selection, as too few runs, or poorly chosen factor levels, can leave the fitted model unstable and its predictions unreliable [144, 154].

Despite their intuitive appeal and practicality, response surface methods have seen little substantive development in DES over the past few decades [117, 153]. Indeed, more advanced variants of RSM are not widely used in routine DES applications, reflecting a broader gap between methodological literature and practical uptake.

Alternative predictive approaches based on Gaussian processes, with Kriging as the most widely used form, offer more flexible ways to approximate more complex factor–response relationships [144]. These methods are non-parametric and capable of capturing highly nonlinear behaviour, making them attractive where low-order polynomial surfaces prove too restrictive. In practice, however, their adoption in DES has been limited [117]. Kriging, originating in geostatistics, requires specification of a covariance kernel and estimation of hyperparameters, both of which demand technical expertise and substantial computational effort [144]. More general Gaussian process models share these challenges and, while conceptu-

ally elegant, become difficult to scale as the number of simulation runs grows [144]. For most DES practitioners, this means the additional burden of mastering and applying such methods often outweighs their benefits [153].

There is therefore no "free lunch" in this setting: while alternatives such as Kriging and Gaussian processes expand the modelling toolkit, they do so at the cost of accessibility and computational efficiency [131]. Their limited use to date reflects the same barriers that constrain the adoption of more advanced response surface methods. Nonetheless, there remains scope for approaches that combine efficiency with greater accessibility, both conceptually and practically, offering DES practitioners more powerful yet usable predictive modelling tools.

2.2.3.5 Comparison and Selection of Scenarios

In this subsection, we turn from predictive modelling across continuous design spaces to problems of comparison and selection among discrete model scenarios. Such problems are interesting, and important in the DES context, as many contend the real value of simulation lies in the comparison of different model scenarios to make purposeful and positive changes to the real system [154]. Careful statistical analysis is required to ensure reliable judgements in such problem settings, in light of the stochastic nature of DES model response(s) and the need to draw multiple inferences simultaneously [154]. To begin, we briefly discuss the comparison of just two simulated scenarios, before moving on to survey a range of classical comparison and selection problems for small, given subsets of model scenarios.

Comparison of Two Scenarios

With just two model scenarios to consider, the appropriate analytical approach depends on a number of points, namely whether the data samples are independent, homoscedastic (or, of equal variance), and of equal size [154].

Supposing the two samples are equal in size and so may be paired, then a paired t-confidence interval may be used to assess the size of any difference in the mean response between the two model scenarios. In such a context, there is no need to assume independence, and we note the use of common random numbers (CRN) between the model scenarios can lead to positive correlation in the observed data,

and so shorter confidence intervals in the estimation of the mean response. Likewise, there is no need to assume homoscedasticity with this method. We also note that the confidence intervals obtained are exact if the observed data are normal, and approximate otherwise.

Supposing, on the other hand, the two samples are different in size, then a different approach is required. In this case, we could calculate a two-sample t-confidence interval to evaluate any difference in the mean response between model scenarios. This, however, would require an assumption not only of independence between the samples, but also of strict homoscedasticity. As such, given the context of DES experimentation, a better approach would be to use Welch's approximate t-confidence interval [241], which does not require equal variance across the model scenarios. Law [154] notes this accommodation of unequal sample sizes allows for the approach to be used in validation, with one model scenario representing the simulation model (typically with more available data) and the other representing the real system (typically with fewer available data).

Comparison and Selection Across Multiple Scenarios

These techniques aim to, in some sense, optimise the model response of interest over a relatively small subset of model scenarios. Typically, these model scenarios are taken as given, and classically, no more than twenty are allowed, however, more recent developments allow for a larger number to be analysed [224]. As with Swisher et al. [224], we will distinguish between multiple comparison procedures (MCPs) and ranking and selection procedures (R&S). The former aims to understand the differences between pairs of model scenarios, whilst the latter has the more ambitious goal of honing in on the model scenario deemed "best" in a particular sense [154]. Of particular importance in many such procedures is the ability to efficiently discard any model scenarios clearly inferior to the best [141]. Typically, there is a set budget available for computation, and so it is clearly advantageous to focus as much of this budget as possible on obtaining more precise inference on model scenarios more likely to rank highly. This notion has given rise to either partly or fully sequential procedures [154].

Looking at MCPs, there are a range of standard problems examined in the liter-

ature, namely: all pairwise comparisons, comparisons with a control (MCC), and multiple comparisons with the best (MCB) [154].

Looking first at the problem of all pairwise comparisons, Swisher et al. [224] distinguish between two categories of approach: (1) the brute-force approach; and (2) all pairwise multiple comparison analysis (MCA). In the brute-force approach, all possible pairwise confidence intervals are constructed, with a confidence level as directed by Bonferroni's inequality. With this approach, the larger the number of model scenarios, the higher the required confidence level, and so the greater the width of the individual comparison intervals. As has been well-noted in the DES literature, this can lead to problems in practice, and unless there are clear differences between the model scenarios, it may not yield helpful information in terms of drawing informative comparisons [224]. In MCA, on the other hand, rather than creating individual confidence intervals, each with a different half-width, a simultaneous set of confidence intervals is created, each with the same half-width, and based on the pooled standard deviation and appropriate Student's t distribution quantile. This approach, attributed to Tukey [231], avoids the aforementioned issue with the brute-force approach, and is usually preferred [224]. Yang and Nelson [252] provide an adaptation of MCA that allows for the use of CRN.

Turning attention to MCCs, oftentimes the control model scenario represents the current, real system, with the remainder of the model scenarios representing possible alternatives to it [154]. Law [154] presents an approach similar to the brute-force approach above, where individual confidence intervals are created for each necessary comparison, and confidence levels are similarly determined by Bonferroni's inequality. Dunnett [72], meanwhile, offers a procedure for the creation of simultaneous confidence intervals, and Yang and Nelson present a revision that allows for CRN [252].

Finally, in discussing MCBs, we first note their connection to R&S problems, as these approaches necessarily attempt to identify the "best" model scenario [154]. Hsu [123] is credited as the first to provide a procedure for addressing the MCB problem, with Hsu and Nelson [124] demonstrating the procedure in the context of simulation. In terms of developments, Yang and Nelson [251, 252], Nelson and Hsu [169], Nelson [168], and Nakayama [166] offer procedures that may be used

alongside variance-reduction techniques such as CRN.

Focusing instead on R&S, again, there are standard approaches routinely discussed in the literature, namely: indifference zone procedures (selection of the best model scenario); subset selection procedures (selection of a subset containing the best model scenario); and alternative procedures outside of these approaches [224].

Looking first at indifference zone procedures, their goal is to select the best model scenario in terms of a particular response, for example, to select the model scenario with the highest mean profit. The indifference aspect stems from the setting of an indifference zone, a numerical value such that we are indifferent if the difference between two model scenarios is less than this threshold value. The first indifference zone procedure is credited to Bechhofer [19], with this procedure assuming unknown means but a known common variance. An important step forward came with the work of Dudewicz and Dalal [71] presenting a procedure for unknown and/or unequal variances, and so operationalising the approach for DES experimentation. Rinott [186] presents an adaption of the procedure providing a higher probability of correct selection, but necessitating a greater number of data observations. Further work in this area includes Koenig and Law's [148] formulation of Dudewicz and Dalal's [71] work to create a screening procedure, and Goldsman's [105] use of standardised time theory to support variance estimation in R&S approaches.

There is no specification of an indifference zone in subset selection procedures, rather these approaches stem from the original work of Gupta [110] providing a subset of random size with a pre-specified probability of containing the best model scenario. This work was extended by Gupta and Santner [111] and Santner [202] to permit the specification of the maximum subset size. These approaches traditionally assume equal and known variances, however a development by Sullivan and Wilson [221] presented an approach allowing for unknown and/or unequal variances. Despite this adaptation, subset selection procedures are typically less popular in the DES community, as practitioners typically like to determine the best model scenario, and not a subset containing the best model scenario [224].

Finally, we examine alternative approaches in the R&S space. The modelling of the

R&S problem as a multinomial selection problem has been studied in the literature, with Chen [51] adopting a subset selection framework, and Goldsman [103, 104] adopting an indifference zone formulation. Fully sequential approaches have been offered by Kim and Nelson [140] and Goldsman et al. [106], however, it is noted that such approaches typically involve a substantial computational overhead. Finally, Chick [53] employs a decision theoretic approach to the selection of the best model scenario. Further work on this topic was carried out by Chick and Inoue comparing Bayesian and frequentist approaches [129], and incorporating sample costs and value of information reasoning to increase efficiency [54, 55].

Comparison and selection procedures represent one of the most mature and extensively studied areas of simulation methodology, with decades of research producing statistically rigorous techniques [224]. At the same time, practical application often involves trade-offs: methods that maximise rigour can be computationally demanding, while those that prioritise efficiency may sacrifice guarantees of correctness [141]. Moreover, these approaches remain tied to comparisons across predefined model scenarios, limiting their adaptability in more exploratory or iterative study designs. In this sense, while the field is well established, there is no universally optimal solution, and scope remains for approaches that better balance statistical assurance, computational efficiency, and practical usability in simulation experimentation.

2.2.3.6 Optimisation and Robust Design

In this subsection, we continue our exploration of Barton's taxonomy [16] by considering the late-stage simulation goals of optimisation and robust design. Optimisation dominates both the literature and practice in DES, reflecting the frequent desire to identify model scenarios that maximise or minimise performance measures [6, 48, 95]. Robust design, while far less widely applied, offers a complementary perspective by focusing on solutions that remain effective under variability and uncertainty in system inputs [197, 226]. Accordingly, we devote most attention to optimisation, before briefly addressing robust design. This completes our discussion of Barton's taxonomy, after which we synthesise insights across the reviewed areas of DES experimentation.

The process of identifying the configuration of model factors giving rise to a maximum (or minimum) output response, without the explicit evaluation of each possibility, is termed optimisation [48]. Naturally, optimisation is a very common goal in practice, with practitioners frequently desiring to optimise, rather than simply improve, their system of interest. It is particularly common in engineering contexts, for example, in process design and reliability [164]. Figure 2.3 from Carson and Maria [48] highlights that any simulation optimisation strategy must be compatible with the what-if nature of simulation experimentation, where the information obtained from initial runs of the model is used to guide the choice of subsequent runs in the search for optimum.

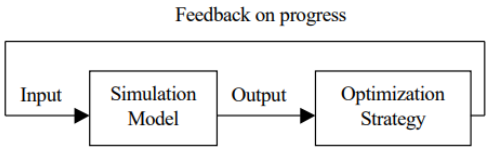

Figure 2.3: A simulation optimisation schematic, from Carson and Maria [48].

A variety of different approaches have been developed to support simulation optimisation, as may be seen from Figure 2.4, also from Carson and Maria [48]. The figure emphasises six main areas of the field, namely: gradient-based search methods, stochastic optimisation, response surface methodology (RSM), heuristic methods, a-teams, and statistical methods. We discuss each of these briefly in turn below. However, before proceeding, it is helpful to first introduce a little optimisation-specific terminology (see, e.g. Law [154]).

In optimisation, the objective function is the quantity to be optimised, parameterised by the function's input variables. In a simulation context, we can consider the response surface to be our objective function, and the model factors our input variables. A working solution is the current "best" configuration of input variables with regards to the objective function. In a simulation context, we can interpret this as the current best model scenario in terms of the response(s) of interest. The feasible region is the area of the input variable space considered in a given optimisation problem, and is analogous to the design space in a simulation context. Finally, we distinguish between local and global optima. A local optimum is an

extreme point where the objective function is more extreme than at neighbouring points within a region, but not necessarily more extreme than at any point across the feasible region (or design space, in simulation terminology). No such distinction need be made regarding a global optimum, an extreme point across the entire feasible region (or design space). Having introduced all useful terminology, we now discuss each branch of Carson and Maria's [48] framework in turn.

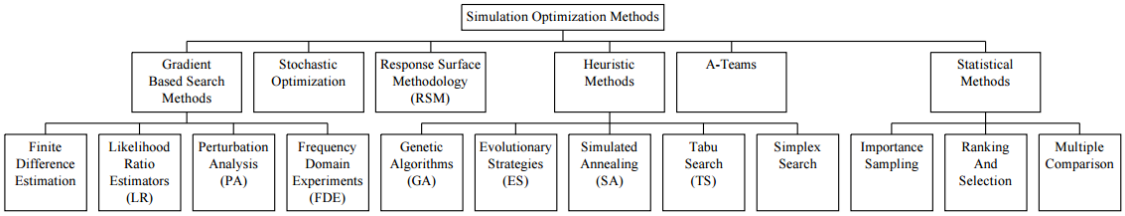

Figure 2.4: Carson and Maria's classification of simulation optimisation methods [48].

Examining Figure 2.4, we can see that the right-most branch is labelled statistical methods, with sub-branches including ranking and selection and multiple comparisons [20]. Recalling the discussion of the previous subsection, we do not cover these topics further, but rather note their ambiguous nature, classified separately from optimisation in Barton's taxonomy [16], yet treated as part of optimisation in Carson and Maria's classification [48]. The remaining sub-branch relates to importance sampling. While often viewed primarily as a variance-reduction technique, importance sampling can be integrated into optimisation frameworks to improve efficiency, particularly when systems involve rare but critical events. Its effectiveness lies in reweighting the probability distribution so that rare events occur more frequently in simulation, thereby yielding more accurate performance estimates with fewer replications [7, 113].

Gradient based search methods involve the estimation of the gradient or shape of the objective function to allow the use of classical, deterministic mathematical programming approaches. Commonly used techniques include finite differences, likelihood ratio (LR), perturbation analysis (PA) and frequency domain experimentation (FDE) [48]. Finite difference methods are simple to implement, approximating gradients by perturbing inputs and comparing outputs, but they are computationally expensive and can often produce noisy estimates in stochas-

tic settings [100]. LR estimators instead exploit the probabilistic structure of the model to derive analytic gradient estimates; these can be more efficient but require strong mathematical tractability, which can be difficult to achieve in complex DES models [100]. PA provides another approach, tracking how infinitesimal changes in inputs affect outputs within a single run. When applicable, it can be highly efficient, though it requires smoothness and differentiability, mathematical assumptions that are not always met in practice [116]. Finally, FDE involves introducing sinusoidal perturbations into inputs and estimating gradients from the output, a technique with strong theoretical foundations but limited use in DES practice [116]. Despite their theoretical appeal, these methods are often computationally intensive and can face difficulties in practical application to complex DES models [95].

Stochastic optimisation presents the problem of finding a local optimum given that the objective function is not known, but can be estimated or evaluated using data. Classical stochastic optimisation schemes, such as stochastic approximation methods, are iterative in nature and involve the estimation of the gradient of the objective function [188]. While foundational, they are often computationally inefficient and practically problematic in DES contexts [150]. More recent developments emphasise stochastic programming, for example through the stochastic counterpart method, which applies statistical inference over multiple scenarios to improve performance [214]. Despite their complementary strengths, both streams face significant challenges in DES, including the computational burden of repeated simulation and the difficulty of achieving reliable convergence in high variability, high-dimensional settings [48].

A popular approach in DES practice is the optimisation of the response surface via RSM, for which we introduced the necessary modelling fundamentals in Subsection 2.2.3.4. Starting from an initial point in the design space, RSM constructs an experimental design locally around this point and then uses the data obtained to fit a first- or second-order approximation, geometrically interpreted as a planar or curved surface, respectively [33]. Although the true response surface may be highly complicated globally, it is commonly assumed that a first- or second-order model will provide sufficient information to guide the search. Oftentimes, first-order models are used to guide the search towards the optimum, whilst a

second-order model is used to estimate the response surface around the optimum, allowing for curvature. In RSM, frequently first- and second-order models are employed iteratively, where mechanisms such as a goodness-of-fit test may be used to assess the adequacy of the fitted models [33].

As Barton [15] argued in detail in his review of metamodelling, RSM belongs to the broader domain of metamodel-based optimisation, where RSM is deemed a "local" approach, owing to its tendency to get stuck in local optima. As such, we briefly discuss the more general concept of metamodel-based optimisation. As highlighted previously, simulation models are frequently large-scale and structurally complex, difficult to interpret and computationally expensive to run. A common approach in such cases is the construction of a simulation metamodel to approximate the response surface [144]. The simulation metamodelling task involves the construction of a functional approximation to the mean response(s), and the most commonly proposed metamodels include low-order polynomial models and Kriging models, each discussed in Subsection 2.2.3.4. The construction of a simulation metamodel offers a number of advantages [15], such as the explicit relationship between factors and responses, the increase in transparency and understanding, its deterministic nature and the computational efficiency in its evaluation. We also note Barton's metamodel-based optimisation framework [15], and the associated schematic detailing both local and global metamodel-based optimisation strategies as detailed in Figure 2.5, and refer the interested reader to Barton and Meckesheimer [17] for a comprehensive discussion of this topic.

More recent developments extend this framework further, particularly through the use of Gaussian process metamodels, of which Kriging is the most widely applied form. When combined with adaptive sampling strategies, these models provide a principled way of balancing exploration and exploitation in the search for optima. Efficient Global Optimisation [134] is a prominent example, combining Kriging predictions with uncertainty estimates to guide the sequential placement of new simulation runs. Such approaches extend RSM's traditionally local focus into genuinely global search strategies, making them potentially valuable for computationally expensive DES optimisation problems. At the same time, Gaussian process-based methods are technically demanding and computationally costly to scale, limiting their practical uptake in DES [144]. Thus, while they offer powerful

extensions to classical RSM, they do not represent a panacea, and scope remains for approaches that combine accessibility, scalability, and robustness in DES optimisation.

We now turn our attention to the discussion of heuristic methods, a diverse range of approaches featuring direct search in which, unlike in previously discussed methods, no attempt is made to estimate or evaluate the gradient of the objective function. Genetic algorithms combine stochastic operators with selection, aiming to strike a balance between exploration (of the feasible region) and exploitation (of "good" solutions). The name comes from the similarity of the process to natural selection, where the fittest individuals are selected to reproduce, leading to a fitter overall population in the next generation [120, 102]. Evolutionary strategies likewise draw inspiration from biological evolution, using mutation, recombination and selection, along with a set of stopping criteria, to iteratively refine and improve the population of considered solutions [185, 211].

Simulated annealing is a stochastic global search technique, frequently used in cases of nonlinear objective functions problematic for local search techniques. It can accept "worse" solutions as the current working solution, with the probability of this occurrence decreasing as the search progresses, and so allows location of the global optimum [142]. Tabu search, first applied in integer programming, integrates memory into a local search approach. In doing so, it helps avoid many of the problems classically associated with local search methods, such as getting stuck in local optima [101]. Finally, the Nelder-Mead simplex method is a classic direct search technique, based around a polygon with $n + 1$ vertices in the case of a feasible region featuring n variables. The method involves evaluating the objective function at the vertex points, and then moving away from the vertex with the least optimal value [167]. Overall, heuristic methods offer flexibility and conceptual simplicity, but often require large numbers of simulation runs and provide no guarantees of convergence to a global optimum.

Finally, we discuss Asynchronous Teams (A-Teams). A-Teams combine multiple optimisation approaches within a cooperative framework, where different agents apply heterogeneous search heuristics (e.g. genetic algorithms, simulated annealing, tabu search) and exchange information asynchronously to accelerate conver-

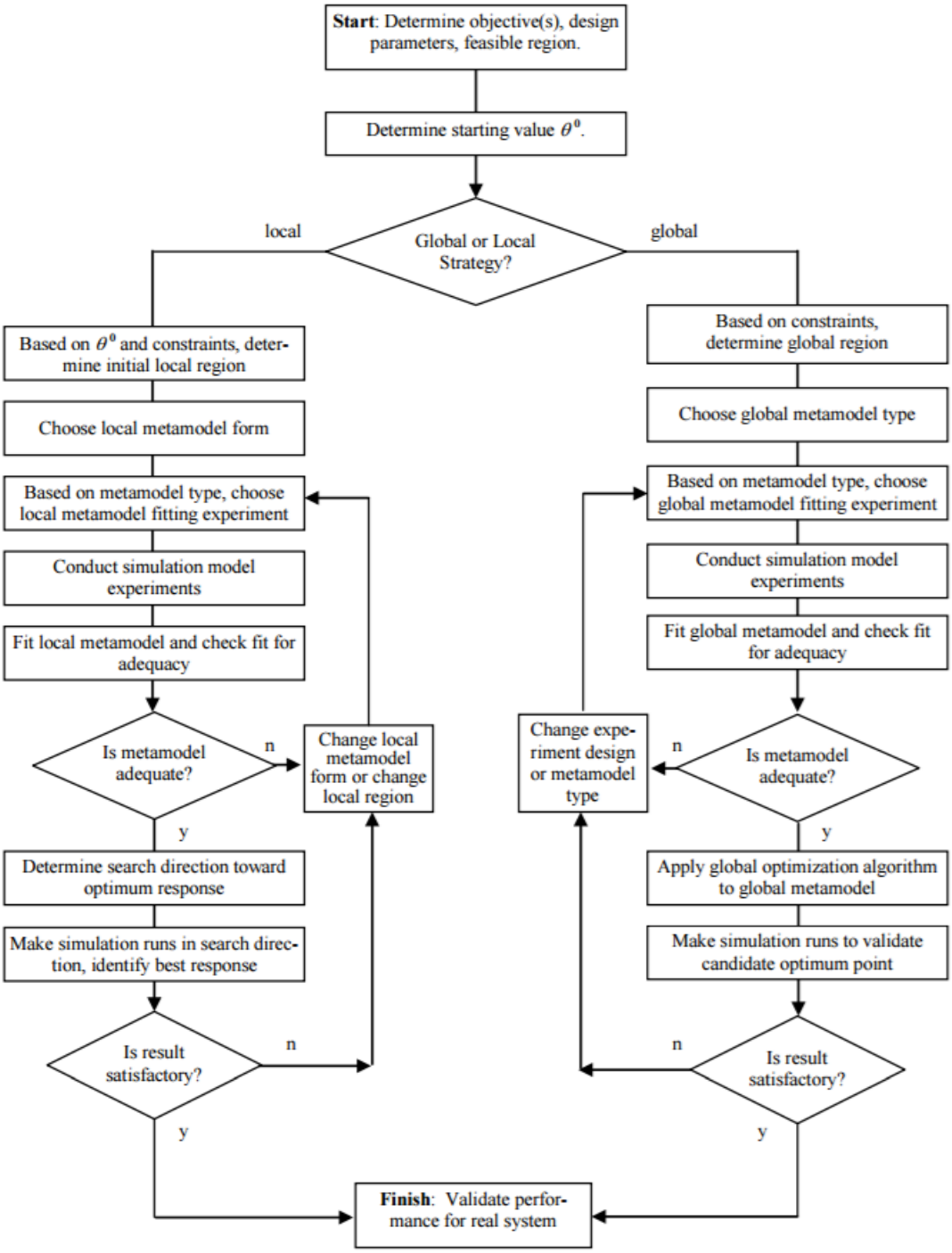

Figure 2.5: Barton's schematic of metamodel based optimisation [15].

gence and improve solution quality. This structure aims to exploit the strengths of complementary methods, making A-Teams particularly attractive in complex or multi-objective optimisation contexts [48, 227]. While they represent a flexible

and innovative approach, their application in DES optimisation has so far been limited, reflecting both their computational demands and the greater maturity of more established heuristics.

Clearly, a wide variety of approaches have been contributed towards simulation optimisation, many adapted from other, more established fields. Given the challenge of optimising across a model's entire design–response space, each method carries its own strengths and limitations, with no single approach emerging as universally superior in DES contexts [6]. This suggests both the maturity of the field and the continuing scope for contributions that improve efficiency, enhance robustness, or lower the technical barriers to practical application.

2.2.4 Conclusions

Having now provided an overview of DES and reviewed the approaches found in the literature for the accomplishment of DES experimentation, with goals as characterised in Barton's framework [16], it is helpful to make several comments.

To begin, we first note the purpose of Section 2.2.1 in relation to the research objectives, namely the accomplishment of Research Objective R1(b). That is, to review the different goals of and approaches to DES experimentation with a view to identifying the most promising area(s) in terms of the potential to apply EB.

In reflection, the preceding subsections have emphasised not only the wide variety of goals of DES experimentation, but also the great range of approaches found in the literature in support of such goals. In general, there is no single approach that dominates across the different goals. Even in predictive modelling, where RSM-type approaches have long been the most widely used, other methods such as Kriging have also received significant attention. At the same time, a clear gap persists between theory and practice, with much DES experimentation in practice continuing to rely on informal, ad hoc methods. A further observation relates to the uneven maturity of the areas of Barton's taxonomy. For example, "gaining understanding" (in Barton's taxonomy) or early-stage, understanding-oriented experimentation remains relatively underdeveloped, whereas later stages such as multiple comparison procedures and ranking and selection are more mature and

well-developed in the literature. This imbalance highlights both the challenges and opportunities for methodological contributions across the taxonomy.

We may finally note, in relation to Research Objective R1(a) around the confirmation of lack of traction of EB in DES experimentation, the absence of mention of applications of EB to DES in the preceding subsections. This is indeed due to a lack of significant application of EB in DES experimentation literature. The few, isolated results found in the literature, and the searches conducted to provide assurance regarding such claims, will be discussed in the final section of the literature review on the EB methodology, completing the necessary work for Research Objective R1(a).

2.3 EB Methodology

Having reviewed both DES and DES model experimentation in some detail, this section now turns to the EB methodology, proposed as holding particular value in this context. First, an overview of statistical paradigms is provided, focusing on the frequentist and Bayesian approaches to statistical inference, the two classical paradigms. Next, EB is introduced through the presentation of a simple Bayesian model and discussion of how an EB approach to such a modelling context would differ. These practical examples provide the reader with a clearer sense of the underlying nature of the EB approach and the mechanics involved in its application. Then, a brief survey of the historical development of EB is given, followed by an overview of more recent advances in high-dimensional statistics. Finally, the section concludes with a discussion of EB in DES experimentation, in particular highlighting the very limited application of EB in DES to date, and demonstrating the broad scope for methodological contribution: a gap which provides the central motivation for the present thesis.

2.3.1 Statistical Paradigms and EB

As noted by Carlin and Louis [45], two important researchers in the field of EB inference, the term EB "refers to a class of models to some statisticians; to others, a style of analysis; to still others, a philosophy for screening statistical procedures." In light of such complexity, it is useful to first situate EB within the two classi-

cal paradigms of statistical inference: the frequentist and Bayesian approaches.[7] These paradigms differ fundamentally in how they view parameters and how inference is drawn, and EB may be regarded as drawing on elements of both.

Before turning to these paradigms, it is worth clarifying scope. Statistical inference can be broadly divided into parametric and non-parametric approaches [239]. Parametric approaches make use of the established theory of mathematical statistics, rely on stronger assumptions, but are able to achieve inference with relatively small amounts of data. Non-parametric approaches, by contrast, avoid such assumptions but require much larger data volumes [239]. Relating this distinction back to the current study, the central motivation behind the proposed use of EB in DES experimentation is the increased statistical efficiency it potentially offers. For this reason, the discussion here, and the subsequent development of EB methods, will necessarily focus on parametric inference. We also note that more time will therefore be devoted to discussion of the Bayesian approach than to the frequentist. This is, firstly, because the majority of statistical education focuses on frequentism, and thus the basics require less explanation [76]. Secondly, the structure of an EB analysis is built on that of a fully Bayesian analysis, and so a good understanding of the Bayesian approach is essential for the work that follows. At this stage, however, it is helpful to outline the structure of a typical data analytic setting, which provides both a framework and the necessary terminology for further discussion.

Thus, adopting a parametric approach, we typically have a sample of data $\mathbf{y} = (y_1, \ldots, y_k)$, from which we wish to make inferences about the broader population or process of interest, following the general formulation set out by Young and Smith [255]. To begin, a sampling model is assumed, $f(\mathbf{y}|\theta)$, where θ denotes the parameter (or vector of parameters) governing the distribution. Both y_i and θ may in general be vector-valued. When viewed as a function of θ for fixed $\mathbf{y}$, the distribution $f(\mathbf{y}|\theta)$ is called the likelihood function, often denoted $L(\theta|\mathbf{y})$. The aim of inference is then to use the observed data to draw conclusions about the unknown parameter θ. In the context of simulation, for instance, θ could represent

[7] Other schools of thought exist, such as likelihoodist, fiducial, or design-based approaches [62], but the frequentist and Bayesian paradigms dominate both theory and practice and are the most relevant for situating EB.

a performance measure of interest, and the observed data $\mathbf{y}$ arising from repeated runs of the model [144].

Within the frequentist paradigm, the parameter θ is regarded as a fixed but unknown constant, while the observed data $\mathbf{y}$ are treated as random, generated under repeated sampling from the model [25]. The uncertainty in inference therefore arises entirely from the randomness of the data, not from the parameter θ itself. This distinction is fundamental: because θ is fixed, inference must proceed by studying the behaviour of estimation and testing procedures across hypothetical repetitions of the sampling process [25]. In this view, a statistical procedure is considered good if it performs well under repeated sampling across a range of possible parameter values. This is an important point, as much of frequentist methodology flows directly from this stance.

Building on this stance, we next consider how inferential procedures are constructed in practice. Within the frequentist school, a variety of approaches exist, reflecting different underlying principles. One classical approach is the method of moments, attributed to Pearson [180]. This approach equates population and sample moments and solves, providing a procedure for drawing inference about the unknown parameter values. So if $\theta_k = E[X^k]$ is the k^{th} population moment, whilst $\hat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ represents the k^{th} sample moment, then solving the equations $\theta_k = \hat{\theta}_k$ for as many moments as there are unknown parameters results in an estimation procedure for the parameter θ.

More influential, however, is Fisher's method of maximum likelihood [87]. In this approach, the parameter θ is estimated by finding the value $\hat{\theta}$ that maximises the likelihood $L(\theta|\mathbf{y})$. That is, finding the value $\hat{\theta}$ such that:

$$\hat{\theta} = \arg\max_{\theta}[L(\theta|\mathbf{y})].$$

The resultant value, $\hat{\theta}$, is termed the maximum likelihood estimate (MLE). This principle provides a systematic way of generating inference procedures and serves as one of the central operating mechanisms of frequentist inference [25, 77]. Alongside these, other guiding principles such as least squares estimation and resampling methods (e.g. the bootstrap) further illustrate the diversity of approaches within

the frequentist tradition [112].

Having described the principles by which frequentist procedures are constructed, we turn to the ways in which these procedures are typically deployed in practice. The most common task is estimation, where the aim is to provide a numerical value, or estimator, for the unknown parameter θ. Point estimators, such as those obtained by the method of moments or maximum likelihood, yield single-valued estimates, while interval estimators are designed to provide a range of plausible values for θ. The latter are formalised as confidence intervals, which are constructed so that, under repeated sampling, they contain the true parameter θ with a specified long-run frequency (for example, 95% of the time). Importantly, this confidence level is not the probability that a particular interval from a given dataset contains the true parameter, but rather a statement about the procedure's long-run performance across repeated samples [25].

Another central component of frequentist inference is hypothesis testing, where procedures are used to evaluate competing claims about the parameter θ. A null hypothesis H_0 is specified, often representing a baseline or default position, and is tested against an alternative H_1. The test statistic is chosen according to the model and the hypothesis under consideration, and its sampling distribution under H_0 provides the reference for decision making. The outcome is typically summarised through a p-value, which measures the extremity of the observed data under the assumption that H_0 is true. As with confidence intervals, the interpretation of the p-value lies in the repeated-sampling framework: it is not the probability that the null hypothesis is true, but rather the probability of obtaining data as or more extreme than that observed, given that H_0 holds [25].

Together, estimation and hypothesis testing form the core inferential tools of the frequentist paradigm, while methods such as regression and time-series analysis extend these principles to more complex modelling tasks [255].

In the frequentist framework, procedures are evaluated through their repeated-sampling properties. A central concern is bias, the extent to which an estimator systematically deviates from the true parameter θ, and variance, which measures its stability across samples. These two aspects combine in the mean squared

error (MSE), a common overall measure of performance. Beyond finite-sample behaviour, frequentist assessment also considers asymptotic properties: an estimator is consistent if it converges to the true parameter θ as the sample size n grows, and efficient if it achieves the smallest possible variance among unbiased estimators. Ultimately, a procedure is judged good if, across repeated sampling, it performs reliably, balancing accuracy and precision in the long run [46].

In contrast, the Bayesian paradigm treats the parameter θ itself as uncertain, representing this uncertainty through a probability distribution on θ [46]. Once the data $\mathbf{y}$ are observed, they are considered fixed, and all uncertainty is attributed to the parameter θ. This inversion of roles has profound implications: inference is framed as updating beliefs about the parameter θ in light of new evidence, rather than assessing the long-run properties of procedures. From this starting point arises the entire Bayesian framework for statistical inference [46].

Having established this perspective, we now turn to the process by which beliefs about the parameter θ are updated in light of observed data. The Bayesian framework starts with a familiar element: a sampling model $f(\mathbf{y}|\theta)$ relating the data to the parameter of interest. A probability distribution describing θ is specified, reflecting the totality of information available before observing the data $\mathbf{y}$. This distribution is termed the *prior* distribution. The information contained in the data $\mathbf{y}$ is used to update the prior distribution, resulting in a *posterior* distribution. The revision process, in which prior becomes posterior, relies on Bayes theorem (Equation (2.3)) [98].

Letting $\pi(\theta)$ denote the prior distribution, and $f(\mathbf{y}|\theta)$ denote the likelihood function, the posterior distribution is given by:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{m(\mathbf{y})} = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)d\theta}. \tag{2.3}$$

Originally demonstrated in the 1763 publication by Thomas Bayes [18], an English clergyman and amateur mathematician, Bayes theorem has come to form the cornerstone underlying the entire Bayesian statistical framework [135].

The final expression in Equation (2.3) reveals that the posterior distribution $\pi(\theta|\mathbf{y})$

is simply the product of the prior distribution $\pi(\theta)$ and the likelihood $f(\mathbf{y}|\theta)$, renormalised to integrate to one. The integral, which ensures this normalisation, is given by:

$$m(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta, \tag{2.4}$$

and is known as the *marginal likelihood* [98]. In this sense, it may be regarded as a normalisation constant.

Often, the greatest difficulty encountered in obtaining the posterior distribution lies in the evaluation of the marginal likelihood integral (2.4), which is generally intractable unless specific, computationally convenient pairings of prior and likelihood are selected to permit a closed form solution. Such pairings are termed *conjugate* [46].

As a final point, the process of updating the prior to obtain the posterior may also be carried out sequentially. Suppose the dataset $\mathbf{y}$ is formed of two subsets $\mathbf{y}_1$ and $\mathbf{y}_2$. Then:

$$\begin{aligned} \pi(\theta|\mathbf{y}_1, \mathbf{y}_2) &\propto f(\mathbf{y}_1, \mathbf{y}_2|\theta)\pi(\theta) \\ &= f_2(\mathbf{y}_2|\theta) f_1(\mathbf{y}_1|\theta)\pi(\theta), \end{aligned}$$

where $f_j(\mathbf{y}_j|\theta)$ denotes the likelihood contribution of subset j. Equivalently, this can be written as:

$$\pi(\theta|\mathbf{y}_1, \mathbf{y}_2) \propto f_2(\mathbf{y}_2|\theta)\pi(\theta|\mathbf{y}_1).$$

That is, the posterior based on the full dataset $\mathbf{y}$ is the same as that obtained by first updating with $\mathbf{y}_1$, and then again with $\mathbf{y}_2$. This concept can prove useful in contexts where data arrive sequentially over time, and it highlights the flexibility of Bayes' theorem as a general mechanism for inference [46].

Bayes' theorem thus serves not only as the basis for updating prior to posterior, but also as the foundation for the entire Bayesian framework. This simple result, relating the unconditional and conditional distributions of random variables, enables prior information and data information to be combined in a formal and systematic manner to support inference regarding the parameter.

In the Bayesian paradigm, inference is encapsulated in the posterior distribution $\pi(\theta \mid \mathbf{y})$ itself. This distribution represents the complete state of knowledge about the parameter θ after observing the data $\mathbf{y}$ and combining it with the prior $\pi(\theta)$. In practice, however, it is often useful to distil this information into more concise forms. Point summaries such as the posterior mean, median, or mode provide single-valued estimates, while interval summaries take the form of credible intervals, which specify ranges of parameter values that together contain a stated proportion of the posterior probability. Unlike frequentist confidence intervals, these have a direct probabilistic interpretation: for example, a 95% credible interval indicates a 95% posterior probability that θ lies within the specified bounds [98].

Beyond estimation, the posterior also provides a natural basis for hypothesis testing and model comparison. Rather than relying on test statistics and long-run error rates, Bayesian analysis evaluates competing claims by comparing the posterior support each receives. The most widely used formal tool for this purpose is the Bayes factor BF_{10}, which quantifies how much more strongly the data support one model over another, with the marginal likelihood $m(\mathbf{y})$ acting as the normalising quantity [137]. In this way, the probabilities assigned to models or hypotheses reflect both prior assumptions and the evidence provided by the observed data.

Together, posterior summaries and Bayes factors form the core inferential tools of the Bayesian paradigm, while applications such as regression and time-series modelling extend these principles to more complex data structures. In such contexts, the posterior distribution serves as the basis from which estimates, uncertainty statements, and model comparisons are derived [46].

In the Bayesian framework, procedures are evaluated through the coherence and properties of the posterior distribution $\pi(\theta|\mathbf{y})$ [46]. Coherence requires that probabilities combine consistently according to the rules of probability, providing a principled foundation for inference. Beyond this, assessment considers how the posterior behaves: as the sample size n increases, it should concentrate around the parameter θ, a property known as posterior consistency. Calibration also plays a role, linking Bayesian probability statements to observed frequencies, e.g. 95% credible intervals should contain the true value about 95% of the time under re-

peated sampling. Ultimately, a Bayesian procedure is judged good if it coherently synthesises prior information $\pi(\theta)$ with data $\mathbf{y}$ into a posterior that yields stable and interpretable probability statements about the parameter [46].

The two paradigms present contrasting but complementary virtues. Frequentist methods benefit from a long tradition of development and a well-established toolkit, with procedures such as the method of moments, maximum likelihood, and regression now widely applied across disciplines. Much of their appeal lies in their perceived objectivity, as no priors are required, and in their mathematical tractability, often translating into computational efficiency. Their key evaluation criteria (namely bias, variance and efficiency), are precisely defined, making their performance easy to judge under the repeated sampling framework. Yet these same features also reveal their limitations. The reliance on hypothetical repetitions can lead to less intuition in terms of practical decision making, and confidence intervals and p-values are easily misinterpreted as probability statements about parameters. Moreover, frequentist procedures often evolve in an ad-hoc manner without a unifying principle, and the framework provides limited scope for incorporating prior information or modelling complex structures [46, 76].

The Bayesian framework, by contrast, gains strength from its coherence: all inference flows from the posterior distribution, which directly represents uncertainty about the parameter θ. This permits probability statements that align naturally with how results are often understood, and provides a principled means of incorporating prior information into the analysis. In the past, practical analyses often relied on conjugate prior–likelihood pairings chosen for their computational convenience. However, more recent advances have broadened the scope of models that can be fitted in practice, making Bayesian methods increasingly feasible in applied settings. Nonetheless, the approach has vulnerabilities of its own. It depends on the specification of a prior distribution, which may be subjective or controversial, and results can be especially sensitive to this choice in small samples. Complex models may also be computationally demanding, and historically the Bayesian approach has been less familiar in some fields, though this is gradually changing [46, 76].

Placed side by side, the two perspectives show why EB has proved appealing.

EB seeks to retain the coherence of Bayesian updating while reducing dependence on fully specified priors, by estimating aspects of the prior distribution directly from the data. In this sense, it offers a bridge between the paradigms: preserving the interpretability and flexibility of Bayesian inference while drawing on the efficiency and operational objectivity associated with frequentist approaches [45]. To illustrate this synthesis in practice, the next section turns to a simple example: the normal–normal model. This example shows, in a transparent way, how the Bayesian framework can give rise to the EB approach, which forms the focus of the thesis.

2.3.2 EB in Practice: A Worked Normal–Normal Example

To illustrate the EB approach, this section presents a worked example based on the standard normal–normal model. This serves two purposes. Firstly, it makes explicit the mechanics of Bayesian inference in a setting where exact posterior derivation is feasible. Secondly, it demonstrates how the EB approach modifies this structure, highlighting key conceptual and practical differences.

The example is presented in three parts. Section 2.3.2.1 outlines the Bayesian formulation, including the full conjugate derivation of the posterior distribution. Section 2.3.2.2 then interprets the resulting posterior parameters, emphasising the roles of shrinkage and precision. Finally, Section 2.3.2.3 provides the EB formulation, showing how hyperparameters may be estimated directly from data and discussing the implications of this shift.

2.3.2.1 Bayesian Formulation

We begin with the standard conjugate case of a normal likelihood with a normal prior. This example provides a transparent setting in which the role of the prior, likelihood, and resulting posterior can be seen explicitly.

Suppose we have a sample of n observations, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ from a normal distribution $N(\mu, \sigma^2)$. Here, we assume that mean μ is unknown, but that variance σ^2 is known. Further, we assume the observations y_i are conditionally independent given μ. As before, we seek to use the sample of data $\mathbf{y}$ to make some inference about the unknown parameter μ.

Given these assumptions, the sampling distribution is:

$$f(y_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right),$$

and so, the likelihood function for the sample is given by:

$$f(\mathbf{y}|\mu) = \prod_{i=1}^{n} f(y_i|\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right).$$

Making use of the identity:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \mu)^2 &= \sum_{i=1}^{n}((y_i - \bar{y}) - (\mu - \bar{y}))^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(\mu - \bar{y}) + \sum_{i=1}^{n}(\mu - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\mu - \bar{y})^2,
\end{aligned}$$

allows us to express the likelihood function as:

$$f(\mathbf{y}|\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{n}{2\sigma^2}(\mu - \bar{y})^2\right).$$

On omitting factors constant with regard to μ, and reversing the terms in the remaining bracket, we obtain:

$$f(\mathbf{y}|\mu) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right), \tag{2.5}$$

which is proportional to a normal density with mean μ and variance σ^2/n.

Expressed in this way, we observe that the likelihood $f(\mathbf{y}|\mu)$, as a function of μ, is dependent on the data $\mathbf{y}$ only through $\bar{y}$ and n.

Reversing the bracket again (for later algebraic convenience), and making use of

the proportional form of Bayes' theorem (as discussed in Section 2.3.1), we obtain:

$$\pi(\mu|\mathbf{y}) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right)\pi(\mu). \tag{2.6}$$

From Equation (2.6), it is clear that the posterior distribution $\pi(\mu|\mathbf{y})$ will include factors of the form $\exp(c_1(\mu-c_2)^2)$, for certain constants c_1 and c_2. To simplify the resulting algebra and ensure a closed-form solution, it is natural to choose a prior $\pi(\mu)$ that is also of this functional form. The simplest family of probability distributions with densities of this type is, as might be expected, the normal family. We therefore assume:

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right),$$

with prior mean and variance μ_0 and σ_0^2, respectively. We also note that the parameters of the prior distribution, here μ_0 and σ_0^2, are generally termed *hyper-parameters*.

Combining the prior and likelihood using Bayes theorem yields:

$$\begin{aligned}\pi(\theta|\mathbf{y}) &\propto \exp\left(-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)\\ &\propto \exp\left(-\frac{(\sigma_0^2 n+\sigma^2)\mu^2 - 2(\sigma^2\mu_0+\sigma_0^2 n\bar{y})\mu}{2\sigma_0^2\sigma^2}\right).\end{aligned}$$

Hence, upon completing the square and omitting factors constant in μ, we obtain:

$$\pi(\theta|\mathbf{y}) \propto \exp\left(-\left(\mu-\left(\frac{\sigma^2\mu_0+\sigma_0^2 n\bar{y}}{\sigma_0^2 n+\sigma^2}\right)\right)^2 \Big/ \left(\frac{2\sigma_0^2\sigma^2}{\sigma_0^2 n+\sigma^2}\right)\right),$$

which is recognisable as the kernel of a normal density.

This uniquely identifies the posterior distribution as normal, with mean μ_n and variance σ_n^2 as given by:

$$\mu_n = \frac{\sigma^2\mu_0+\sigma_0^2 n\bar{y}}{\sigma_0^2 n+\sigma^2} \qquad \text{and} \qquad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2 n+\sigma^2}.$$

We note that, through the use of the normal–normal conjugate pairing, we are

able to obtain the exact form of the posterior distribution without the need to perform any integration. This analytical tractability is one of the main advantages of conjugate prior-likelihood pairs [46]. More importantly for our purposes, the resulting expressions for the posterior mean and variance allow us to examine how prior and data-based information are combined. The next subsection on explores these ideas further through interpretation of the posterior parameters.

2.3.2.2 Parameter Interpretation: Shrinkage and Precision

The previous derivation yielded explicit expressions for the posterior mean and variance under the normal–normal conjugate model. We now examine these parameters more closely to understand how they reflect the influence of prior beliefs and observed data, introducing the key ideas of shrinkage and precision, both central to the EB approach. For a more detailed discussion of these topics, see Carlin and Louis [46].

The posterior mean may be re-expressed as follows:

$$\mu_n = \frac{\sigma^2\mu_0 + \sigma_0^2 n\bar{y}}{\sigma_0^2 n + \sigma^2} = \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\bar{y} = w\mu_0 + (1-w)\bar{y},$$

where $w = \frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}$.

As $0 < w < 1$, it can be seen that μ_n is a weighted average of the prior mean μ_0 and the sample mean $\bar{y}$. When the prior variance σ_0^2 is large relative to the variance of the sample mean σ^2/n, the weight w is close to 0, and so the posterior mean μ_n is close to the sample mean $\bar{y}$. Conversely, when σ_0^2 is small relative to σ^2/n, then w is close to 1, and the posterior mean μ_n is pulled towards the prior mean μ_0.

This behaviour is often described as *shrinkage*: the posterior mean μ_n is "shrunk back" from the sample-based estimate $\bar{y}$ toward the prior mean μ_0, as the weight or shrinkage factor w scales from 0 to 1. The extent of this shrinkage depends on relative uncertainty, allowing the posterior to automatically adjust depending on the degree of certainty associated with each information source. Shrinkage provides a principled way to blend prior beliefs with empirical evidence, and is a hallmark of Bayesian and EB approaches alike.

The posterior variance is similarly informative. It can be seen that:

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2 n + \sigma^2} = \frac{\sigma_0^2(\sigma^2/n)}{\sigma_0^2 + (\sigma^2/n)} = w\sigma_0^2 = (1-w)\frac{\sigma^2}{n},$$

from which we can see that the posterior variance σ_n^2 is smaller than either the prior variance σ_0^2 or the variance of the sample mean σ^2/n.

Indeed, working in terms of the *precision*, the reciprocal of the variance, we set:

$$\lambda = n/\sigma^2, \quad \lambda_0 = 1/\sigma_0^2 \quad \text{and} \quad \lambda_n = 1/\sigma_n^2.$$

We then obtain:

$$\lambda_n = \frac{\sigma_0^2 n + \sigma^2}{\sigma_0^2\sigma^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = \lambda_0 + \lambda,$$

showing that the posterior precision λ_n is simply the sum of the prior precision λ_0 and the precision of the data λ. This reflects the fact that in combining both prior and data sources of information in the posterior, we have greater certainty in our inference than from either source alone.

Finally, we note as $n \to \infty$, we have:

$$w \searrow 0, \quad (1-w) \nearrow 1, \quad \text{so that} \quad \mu_n \to \bar{y}, \quad \sigma_n^2 \to 0.$$

This makes intuitive sense; as our sample size increases, the data accounts for a larger proportion of the overall information, and as such should dominate the resulting inference. This is reflected by the posterior mean μ_n tending to the sample mean $\bar{y}$. Further, with the increased information available from a larger sample size, our uncertainty decreases, and this is reflected in the increased precision of the posterior distribution.

Indeed, this increasing dependence on the data over the prior is a common feature in Bayesian inference and is formalised in the fundamental Bernstein-von Mises theorem [232]. The theorem states that regardless of the choice of prior, assuming that the true parameter θ is contained within the prior distribution's support, the posterior distribution will asymptotically converge to a normal distribution cen-

tred on the frequentist MLE of θ. Practically speaking, this result implies that for large sample sizes, and assuming a reasonable prior, the frequentist and Bayesian estimates will increasingly agree [46].

Having now derived the posterior distribution and examined its interpretation, we next turn to the EB formulation. In particular, we highlight how the current Bayesian formulation of the normal–normal example would differ if undertaken from an EB perspective, thereby enabling a direct comparison of the two paradigms. This shift in perspective retains the Bayesian logic of updating, but removes the need to fully specify a prior distribution, a trade-off with important conceptual and practical implications.

2.3.2.3 EB Formulation and Discussion

We now return to the normal–normal model and consider its EB formulation. To begin, we introduce the hierarchical Bayes (HB) approach, which provides a natural framework for modelling problems featuring a parallel structure [46]. We proceed to discuss how the fully Bayesian hierarchical and EB approaches differ, before explicitly working through the EB version of the normal-normal model. Finally, we consider key conceptual and practical challenges related to applying EB in practice, including computational challenges and the double use of data.

An HB model extends the standard Bayesian framework by introducing an additional level of uncertainty modelling. At the first level, we have, as before, a likelihood $f(\mathbf{y}|\theta)$, where the parameter θ governs the observed data $\mathbf{y}$. At the second level, rather than fixing parameter θ, we treat it as random and model it with a prior distribution $\pi(\theta \mid \eta)$ dependent on a hyperparameter η. In the HB approach, however, we extend this model to a third level of uncertainty, where the hyperparameter η is itself assigned a so-called *hyperprior* distribution $h(\eta)$. In this way, the model forms a three-level hierarchy: likelihood for the data, prior for the parameter, and hyperprior for the hyperparameter.[8]

Whilst HB offers a flexible approach to modelling problem contexts where uncer-

[8] Whilst three levels have been discussed here, in theory the HB approach can extend further dependent on the nature of the problem context and the computational limits of the implementation.

tainty exists at multiple levels, in practice, its use is often constrained by the need to specify hyperpriors [46]. EB methods, however, offer an alternative approach: they retain the same hierarchical structure as a fully Bayesian approach, but allow the estimation of hyperpriors directly from the observed data. This reduces elicitation demands while preserving Bayesian updating, and is useful where many related inferences are sought in parallel. For contexts featuring parallel structures, potentially including those encountered in DES experimentation, EB therefore enables information sharing across populations without necessitating full hyperprior specification [46].

To better appreciate the distinction between HB and EB approaches, it is helpful to consider how the posterior distribution $\pi(\theta|\mathbf{y})$ is derived in each case.

In the HB approach, a hyperprior $h(\eta)$ distribution must be specified for the hyperparameter η. So, to obtain the posterior distribution for parameter θ, we then need to integrate over the uncertainty in hyperparameter η:

$$\begin{aligned}\pi(\theta|\mathbf{y}) &= \int \pi(\theta|\mathbf{y},\eta)h(\eta|\mathbf{y})\,d\eta \\ &= \frac{\int f(\mathbf{y}|\theta)\pi(\theta|\eta)h(\eta)\,d\eta}{\int\int f(\mathbf{y}|\theta)\pi(u|\eta)h(\eta)\,du\,d\eta}.\end{aligned} \tag{2.7}$$

That is, inference proceeds by weighting the conditional posterior $\pi(\theta|\mathbf{y},\eta)$ by the posterior distribution $h(\eta|\mathbf{y})$ (sometimes termed the hyperposterior), and averaging over all possible values of hyperparameter η [46].

In EB, however, the posterior distribution is obtained differently. Instead, the marginal distribution of the observed data $m(\mathbf{y}|\eta)$ is used to obtain an estimate for the hyperparameter η:

$$m(\mathbf{y}|\eta) = \int f(\mathbf{y}|\theta)\pi(\theta|\eta)d\theta. \tag{2.8}$$

In this way, the marginal likelihood is treated as a likelihood for the hyperparameter η. Inference usually proceeds then by maximising it, that is, finding the value of η that makes the observed data most probable. This results in the marginal maximum likelihood estimator (MMLE), denoted $\hat{\eta}(\mathbf{y})$ [46].

This empirical estimate $\hat{\eta}(\mathbf{y})$ is then substituted into the prior distribution $\pi(\theta|\eta)$, and the posterior distribution is then given as:

$$\pi(\theta|\mathbf{y}) \approx \pi(\theta|\mathbf{y}, \hat{\eta}(\mathbf{y})). \tag{2.9}$$

So, whilst HB averages over uncertainty in η to derive the posterior, EB obtains the posterior by plugging in an estimate of η derived from the observed data. In this way, the integration problem in Equation (2.7) is replaced by the maximisation of Equation (2.8). In thise sense, EB often represents a simplification: locating an optimum of the marginal likelihood is often more straightforward than evaluating a potentially high dimensional integral. In some models closed-form updating is feasible, but where it is not, hyperparameters are typically estimated through approximation or numerical optimisation. Three widely used strategies include Laplace approximation methods for evaluating the marginal likelihood, EM-based marginal maximum likelihood estimation (commonly employed in normal–normal EB settings), and Monte Carlo approaches for cases where neither optimisation nor approximation is tractable [46].[9]

Having distinguished HB and EB inference procedurally and computationally, we now demonstrate the EB mechanism more concretely by returning to the normal-normal model and showing how hyperparameters are estimated from the data. therefore, let us assume the following two-level model:

$$\begin{aligned} Y_i|\mu_i &\overset{iid}{\sim} N(\mu_i, \sigma^2), \qquad i = 1, \ldots, k \\ \mu_i|\mu_0 &\overset{iid}{\sim} N(\mu_0, \sigma_0^2), \qquad i = 1, \ldots, k \end{aligned}$$

In this case, we assume k normally distributed populations, with each mean modelled as an independent draw from a common distribution $N(\mu_0, \sigma_0^2)$. We also assume that both σ^2 and σ_0^2 are known, and so we only seek to estimate μ_0.

Standard results show that Y_i are marginally IID normal random variables with mean μ_0 and variance $\sigma^2 + \sigma_0^2$. The joint marginal density of the observed data

[9] For a more comprehensive overview of computational approaches for EB estimation, including optimisation and approximation frameworks in hierarchical settings, the interested reader is referred to Carlin and Louis [46].

$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is therefore the product of the individual marginal densities $m(y_i|\mu_0)$, giving:

$$m(\mathbf{y}|\mu_0) = \frac{1}{(2\pi(\sigma^2 + \sigma_0^2))^{k/2}} \exp\left(-\frac{1}{(2\pi(\sigma^2 + \sigma_0^2))} \sum_{i=1}^{k} (y_i - \mu_0)^2\right).$$

Standard results pertaining to the normal distribution yield:

$$\hat{\mu}_0 = \bar{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

as the marginal maximum likelihood estimate of μ_0.

Substituting $\hat{\mu}_0$ into the prior distribution, and carrying out analogous steps to those of the Bayesian derivation, we obtain the following EB estimate of the posterior distribution for each component population as:

$$\pi(\mu_i|y_i, \hat{\mu}_0) = N(B\hat{\mu}_0 + (1 - B)y_i, (1 - B)\sigma^2), \qquad i = 1, \ldots, k,$$

where $B = \sigma^2/(\sigma^2 + \sigma_0^2)$.

Thus, the EB point estimate for μ_i is:

$$\hat{\mu}_{i,EB} = B\bar{y} + (1 - B)y_i. \tag{2.10}$$

Here, we note that this formula is the same as that obtained under the Bayesian formulation (see Subsection 2.3.2.2), with the exception that the prior mean μ_0 is replaced by the sample mean $\bar{y}$ of all observed data values. Thus our estimate is simply a weighted average of the estimated prior mean $\bar{y}$ and the data value y_i. As such, it may be seen that the inference obtained regarding a single population parameter μ_i depends on data obtained across all of the populations. This idea is often referred to as 'borrowing strength' [45].

Re-expressing Equation (2.10), we obtain:

$$\hat{\mu}_i = \bar{y} + (1 - B)(y_i - \bar{y}),$$

which clearly shows that the standard estimate y_i is shrunk back toward the common mean $\bar{y}$ a fraction $(1 - B)$ of the distance between the two. Setting $B = 0$ yields no shrinkage, as the expression reduces to the maximum likelihood estimate y_i. At the other extreme, setting $B = 1$ corresponds to full pooling, where all group means collapse to the common mean estimate $\bar{y}$. This behaviour illustrates the characteristic EB balance between individual and pooled information.

Strictly speaking, using the data in such a way is a violation of a core principle of the Bayesian approach, namely that the prior distribution must be specified independently of the data. In EB, the observed data are used twice, first in constructing the empirical prior and then again in posterior updating, and the resulting inference is therefore characterised by some as 'overconfident' [46]. Adjustments are often made to account for this issue, particularly in relation to the construction of interval estimators, such as those proposed by Carlin and Gelfand [44], Farrell et al. [85], Laird and Louis [151], and Morris [165]. Although this double use of data departs from the fully Bayesian ideal, contemporary EB methodology explicitly accounts for such behaviour, with extensive empirical validation demonstrating strong performance across diverse applications. As such, the concern is largely philosophical rather than operational for most practical purposes.

2.3.3 Review of EB Literature

Having introduced the broad theory and practice of EB, in the following subsections we discuss the history of the EB approach, briefly surveying its development and general applications.

EB methods have a rich history with roots going back to the work by von Mises on credibility theory in the 1940s [160], which introduced ideas later formalised within EB frameworks. The first major work, however, is attributed to Robbins [187], whose approach to EB inference is frequently termed *nonparametric empirical Bayes* (NPEB). In this approach, the entire prior distribution is estimated from the available data. This is in contrast to the approach outlined in preceding Subsections 2.3.2.1, 2.3.2.2, and 2.3.2.3 in which only the hyperparameters of the prior distribution are estimated from the data. Such an approach is frequently called *parametric empirical Bayes* (PEB). The reduction in dimensionality associated

with PEB makes EB practically easier to compute and implement, with closed-form updates in many standard models. PEB was greatly developed in a series of papers by Efron and Morris in the early 1970s [78, 79, 80], and these contributions helped establish EB as a practical inferential strategy capable of outperforming standard estimators in multigroup problems.

To contextualise subsequent developments, we distinguish classic EB (from the early 1970s to early 2000s) from a second period beginning in the early 2000s, initiated by Efron's reinterpretation of the false discovery rate (FDR) through an EB lens. This EB–FDR connection catalysed a rapid expansion of EB methodology into large-scale inference problems. We discuss each of these periods of EB development in turn, beginning with classical EB.

A vast literature exists regarding classical methods and applications of EB analysis. Morris [165] presents the table of Figure 2.6 summarising a selection of the diverse applications seen in the literature at that time. These include areas as diverse as quality control in industrial settings, admissions procedures in educational institutions, false alarm analysis for emergency response, and epidemiological prevalence estimation. In later years, application areas only continued to grow. By way of example, EB has been used widely in environmental and natural resource contexts, including ecology, as summarised by Ver Hoef [233], water resource modelling and management by Madsen and Rosbjerg [159] and Smith et al. [215], and forest inventory management by Burk and Ek [42].

EB has also been applied extensively in safety and reliability domains, such as road safety measurement, as summarised over a twenty year period by Persaud and Lyon [182], and nuclear reliability modelling by Hill et al. [114], Gaver and Worledge [97] and Martz et al. [161], indicating its acceptance and reliability in high-stakes analytical settings. In such reliability settings, EB is often preferred because it stabilises inference when failures are rare or unobserved: a scenario in which classical estimators can perform poorly [162]. Its continued use in these high-stakes applications reflects the degree of confidence placed in EB as a dependable inferential tool.

Extensive coverage of classical EB applications may be found in texts such as

Application	*Author*	*i*	Y_i, θ_i	*Technical Features*
Revenue Sharing—Census Bureau	Fay and Herriot (1979)	Small areas	Per capita census income	a, b(REG: Housing value, IRS Income per exemption).
Quality Assurance—Bell Labs.	Hoadley (1981)	Time periods	# failures,	a, b(GM), c(Poisson), d (to determine probability system is out-of-control.)
Insurance Rate-making	Many, eg. Buhlmann (1970), Morris and VanSlyke (1978)	Group of insureds or territory	Insurer's risk per unit exposure	a, b(GM) Credibility ≡ $1 - B_i$
Law School Admissions—Educ. Test. Serv.	Rubin (1981)	Law school	Weight for LSAT score relative to GPA.	a, b(GM)
Fire Alarms—New York City	Carter and Rolph (1974)	Alarm box locations	False alarm rate	a, b(GM), c(Poisson).
Epidemiology—El Salvador	Efron and Morris (1975)	City	Toxoplasomosis prevalence rate	a

Note: Key to "Technical Features": Letters indicate presence of the following concerns. a: Unequal variances ($V_i \neq V_j$). b: Shrinking to estimated points (GM = grand mean, REG = regression surface). c: Nonnormal sampling distribution (mentioned only if alternative distribution explicitly acknowledged). d: Interval estimates.

Figure 2.6: Overview of documented applications of EB from Morris [165].

Carlin and Louis [46], and Gelman [98]. Collectively these works demonstrate the methodological breadth and practical maturity of classical EB, well before the emergence of modern large-scale inferential problems.

As discussed elsewhere in this thesis, EB provides a principled means of borrowing strength across related inferential problems. While the classical applications reviewed above primarily concerned low- to moderate-dimensional multigroup problems, the approach became especially attractive as new scientific technologies began producing data in thousands of parallel streams, for example, microarray gene-expression profiles and voxel-level fMRI measurements,[10] where pooling became an intrinsic feature of the analysis, as opposed to an optional efficiency enhancement. The scale and dimensionality of these problems exceeded the scope of classical EB formulations and motivated a new wave of methods focused on inference at scale, particularly those linked to and inspired by the FDR procedure. The resulting literature is extensive, and only a selective review is possible here; however, several developments are sufficiently influential to warrant emphasis.

Given the centrality of FDR ideas to this modern wave of EB methods, it is natural to begin with the seminal work of Benjamini and Hochberg [23], which introduced

[10]Microarrays measure gene-expression levels across thousands of genes simultaneously, while fMRI records voxel-level brain activity over time by tracking changes in blood oxygenation. However, we note that no detailed understanding of these technologies is needed for the purposes of this thesis, they are referenced only as key producers of this type of large-scale parallel data set.

the FDR procedure. This novel approach to the simultaneous testing of a large number of related statistical hypotheses involves controlling the expected proportion of incorrectly rejected null hypotheses. This lies somewhere between the naive use of individual test p-values and the classical and ultra-conservative Bonferroni correction, which controls the possibility of reporting a single false positive across all tests. It is thus well-suited to applications in which factors are many but effects are sparse, as is the case in the testing of genomic microarray data. Many variants have since been proposed and are the subject of a comprehensive review by Benjamini [22].

Whilst the FDR technique was originally developed from a frequentist perspective, its subsequent re-casting in an EB framework by Efron et al. [82] in the early 2000s marked a pivotal shift toward high-dimensional EB inference. The two-groups EB model, in which test statistics are viewed as arising from a mixture of null and non-null populations, provided a new conceptual template for large-scale testing problems. Subsequent developments by Efron [75] introduced empirical estimation of the null distribution itself allowing inference without strict reliance on theoretical null assumptions, and also formalised the local false discovery rate (local fdr), which provides the posterior probability that a given test statistic arises from the null component of the mixture, offering a direct EB-based measure of significance at the level of an individual test.

Following Efron's reinterpretation through the two-groups model, several influential contributions expanded the reach of EB methods in high-dimensional hypothesis testing. For example, Storey's work [219, 220] reframed multiple testing as an estimation problem by introducing the q-value and the positive FDR (pFDR), providing a practical means of estimating the expected proportion of false positives among the rejected hypotheses and offering a more interpretable alternative to p-values in settings with thousands of parallel tests.

Subsequent refinements strengthened both the theoretical and methodological foundations of EB-based hypothesis testing. Genovese and Wasserman [99] analysed FDR behaviour under a stochastic-process framework, clarifying its asymptotic properties and reinforcing the interpretation of FDR as an estimable quantity. Sun and Cai [222] advanced EB ideas to settings where hypotheses are organised

into meaningful scientific groups, such as specific biological pathways or spatial clusters. They demonstrated that by exploiting this inherent structure, rather than treating all hypotheses as exchangeable, substantial efficiency gains could be obtained.

A further important development was introduced by Efron [74], who examined the impact of correlation among test statistics and proposed EB-motivated adjustments to account for such dependence, addressing a pervasive practical challenge in genomic and neuro-imaging applications. His analysis demonstrated how dependence aware corrections can restore more reliable significance assessment in genuinely high-dimensional datasets. Taken together, these contributions established a broad methodological foundation for modern EB approaches to high-dimensional testing.

Important developments were also made on the estimation side of the EB methodology, where the main objective was to stabilise large numbers of parallel parameter estimates through appropriate use of shrinkage. For problems requiring the simultaneous estimation of many normal-means parameters, Johnstone and Silverman [132, 133] showed that prior information could be estimated from the data to support adaptive shrinkage and thresholding, providing a principled alternative to fixed or heuristic rules. Practically, their work demonstrated that these data-adaptive procedures can achieve performance close to optimal in settings with many weak or noisy effects. Brown and Greenshtein [41] strengthened the theoretical footing of this work by demonstrating that EB shrinkage estimators can attain, asymptotically, the same level of accuracy as the best procedures available when each parameter is estimated separately. Related contributions extended similar principles to regression and other multivariate settings, where empirical estimation of prior structure helped to stabilise inference when many coefficients were estimated together [76]. Taken as a whole, this body of work consolidated EB shrinkage as a central component of modern high-dimensional inference, parallel to the developments in EB-based testing.

A common theme emerging from these developments is the way large-scale inference blurs the conventional boundaries between estimation and testing. As Efron observes [76], procedures designed for testing often rely on the accurate

estimation of underlying distributions, while estimation methods frequently derive their practical value from the stability they provide in relation to subsequent tests. The widely used approach of Smyth [216] exemplifies this interdependence: gene-specific variances, which in isolation tend to be estimated with limited precision, are stabilised through an EB shrinkage step before being incorporated into moderated test statistics. Such interactions between estimation and testing are increasingly typical of EB approaches in large-scale settings.

While earlier contributions addressed shrinkage of means and variances in isolation, later developments by Zhao [260] and Hwang et al. [126, 125] built on a shared normal–lognormal framework to introduce procedures in which both components are shrunk simultaneously. Zhao's contribution focused on point estimation, while the accompanying work by Hwang and colleagues extended the same formulation to interval estimation and hypothesis testing, yielding closed-form EB expressions for both location and scale parameters. These methods form part of a broader movement within modern EB to stabilise multiple aspects of the inferential model, particularly in settings where variability differs across groups. Such features may be of relevance in contexts like DES model experimentation, where data arising from different model scenarios often exhibit heterogeneous levels of variability.

These strands of work reflect the substantial growth and diversification of EB methodology over recent decades, from early shrinkage rules to modern procedures capable of addressing large-scale, heterogeneous problems. Together, they illustrate the breadth and maturity of the EB toolkit now available. In the following subsection, we consider the extent to which such methods have been applied within DES model experimentation.

2.3.4 EB in DES Model Experimentation

To the best of our knowledge, literature examining the use of EB methods in DES experimentation contexts appears to be extremely limited. This claim is based primarily upon a thorough search of the past proceedings of the Winter Simulation Conference (WSC). As previously discussed, WSC is the pre-eminent global simulation conference, and is widely regarded as representative of state of the art in the area. Additionally, searches of Google Scholar and arXiv were undertaken

to provide a comprehensive review of literature. This section aims to provide an overview of the sparse literature that does exist to establish the state of the art in this domain.

A few initial comments may be made about the literature identified before reviewing each contribution in turn. One common feature of these studies is that they tend to feature EB only in a limited capacity, often only in connection with a small part of their overall framework. For example, a Bayesian approach may be proposed for the specification of DES model input quantities, with EB invoked solely to specify only a small subset of these quantities where no prior information is available. Furthermore, the use of EB is frequently not elaborated in sufficient depth, making it difficult to obtain a clear understanding of what is being proposed. A final observation is that this literature is focused largely on simulation input analysis, leaving the area of simulation output analysis largely unexplored.

One exploration of the application of EB in a DES context is the approach proposed by Jun and Hui Ng [136] for stochastic computer model calibration. Calibration refers to "adjusting the unknown input parameters of the computer model to fit the real observed data," and their approach attempts to account for a variety of uncertainties, including uncertainty in the calibration parameters themselves. They propose an entropy based criterion, arguing that it improves the efficiency of the estimation of the calibration parameter. Their study examines the effects of noise and design size on predictive performance and notes that the EB component performs well, particularly when sufficient simulation replications are available and stochastic noise is small.

The study of Zouaoui and Wilson [261] proposes a Bayesian approach for probabilistic input modelling. In this study, EB is mentioned only very briefly as an option to support the specification of the prior hyperparameters supposing they are not known a priori from existing expert judgement. The remainder of the framework presents a classical Bayesian approach to the problem of accounting for input parameter uncertainty in DES modelling. Whilst the study incorporates empirical testing demonstrating a good performance, the approach itself is clearly more Bayesian than EB in nature.

The metamodelling approach of Yin et al. [254], similar to the previous article, also proposes an overall Bayesian approach. EB is only involved in the estimation of a subset of the unknown input parameters, with these quanities initially assumed known in order to develop and derive the proposed framework. The authors note the key reason supporting the use of EB in the estimation is to achieve better computational efficiency, however, the lack of empirical testing, or even further detail concerning the practical implementation of EB, restrict the conclusions that can be drawn.

Overall, the discussion of these few applications of EB in DES indicates a clear lack of detailed investigation into the potential of EB for DES model experimentation. It is evident that current work employs EB only in narrow ways and primarily within the context of input modelling. This absence of sustained investigation leaves considerable scope for further work on the potential role of EB within DES experimentation.

2.4 Conclusions

Having now reviewed the literature on both DES model experimentation and the EB methodology, it is possible to draw together several conclusions relevant to Research Objective R1. This chapter has mapped the landscape on both sides: the goals and challenges of DES experimentation, and the development of EB as a framework for increasing the efficiency of inference across related estimation problems. Bringing these literatures together enables a clearer understanding of the gaps that remain and the areas in which EB may have particular relevance for DES.

The discussion in Section 2.2 highlighted the variety of goals and approaches associated with DES experimentation, as structured by Barton's taxonomy, and the uneven maturity of methods across these goals. In particular, the discussion suggested that early-stage, understanding-oriented experimentation may offer especially promising opportunities for EB, given both the need to synthesise information across related scenarios in a statistically efficient manner and the relative lack of maturity of this area. More generally, contexts involving multiple related model scenarios, heterogeneous experimental data, or limited computational budgets were identified as areas where EB-style borrowing of strength could, in theory,

be beneficial, thereby addressing R1(b).

Turning to the evidence for existing EB usage in DES, the review in Section 2.3.4 found almost no sustained application of EB in this domain. Where EB is mentioned, it is typically confined to a narrow role within broader Bayesian frameworks, most often in input modelling, and is rarely described in sufficient detail to support careful assessment or reproduction. This review, taken alongside the absence of EB approaches within the broader literature on DES model experimentation, confirm that EB has gained little traction in DES model experimentation, thereby completing the work required for R1(a).

At the same time, the review of EB methodology undertaken in Sections 2.3.1–2.3.3 shows that a rich and mature body of EB techniques already exists in the wider statistical literature. Classical parametric EB, large-scale EB testing procedures, and modern shrinkage methods all illustrate how information can be borrowed across related populations to stabilise inference in high-dimensional or data-limited settings. Particularly relevant strands include EB shrinkage for many means, EB approaches to variance stabilisation such as Smyth's moderated statistics, and more recent "double shrinkage" procedures in which both means and variances are shrunk within a common hierarchical formulation, as in the work of Zhao [260] and Hwang et al. [126, 125]. When viewed alongside the parallel scenario structures, heterogeneous variances, and limited computational budget often encountered in DES experimentation contexts, these developments suggest that there is a non-trivial overlap between the kinds of problems for which EB methods have been designed and the challenges faced in DES. This addresses R1(c) by identifying a set of EB approaches with plausible relevance to DES contexts.

The review has identified both the DES settings in which EB may hold promise and the EB techniques with potential relevance to those settings. The next step, pursued in Chapter 3, is to assess whether this proposed intersection genuinely has merit in practice and to clarify what further methodological development or operationalisation would be required for EB to be used effectively by DES practitioners.

Chapter 3

An Application of EB to DES Model Experimentation

In this chapter, we present a numerical study on the application of EB in the context of DES experimentation. Building directly on the discussion in Chapter 2, the study brings together the insights gained there regarding EB-appropriate DES settings and DES-appropriate EB procedures, initiating a programme of evaluation into the potential overlap between these two methodological areas.

This is an important chapter for a number of reasons, offering contributions towards the accomplishment of several of the research objectives (specifically, R2 and R5). To begin, it instills the reader with a better grasp of the intrinsic nature of the research, presenting a concrete application of EB within DES model experimentation. Indeed, it is only at this point that we move beyond speculative discussion concerning the apparent dovetailing of the two research areas.

More specifically, the chapter provides a convenient point at which to present a number of concepts fundamental to both this study and the wider research programme in general. We first outline the DES model used as a test model. Consistent with the discussion in Chapter 1, a toy simulation model is used, an intermediate choice between Monte Carlo simulated data sets and the complexity of a full-scale real model, making it well suited for an initial exploratory study. We then introduce the EB estimation procedure, together with its frequentist counterpart, judged most appropriate in light of the insights from Chapter 2, particularly

the relevance of shrinkage-based methods in settings involving multiple related scenarios. A further component is the development of a set of error measures designed to provide a comprehensive picture of the relative performance of the two estimation approaches. As such, this chapter contributes towards the accomplishment of Research Objective R5.

These elements form the basis for a structured experimental design intended to evaluate EB performance across a range of practically relevant conditions. Using the DES test model, we construct a range of experimental configurations that vary in the number and similarity of parallel scenarios, as well as in the available computational budget. These configurations are chosen to reflect early-stage, understanding-oriented experimentation settings identified in Chapter 2 as especially promising for EB.

With this experimental structure in place, the study provides proof of concept, demonstrating that material gains in efficiency are possible with the application of EB to DES model experimentation. Further, we identify potential pitfalls for the DES practitioner, and highlight areas where some degree of operationalisation is likely required to support the practitioner in their attempts to realise the benefits of this approach. These challenges, arising for example from scenario heterogeneity and the resultant potential for bias as highlighted in Chapter 1, motivate the further methodological work undertaken later in the thesis. As such, this work fulfils Research Objective R2.

The remainder of the chapter proceeds as follows. Section 3.1 introduces the (s, S) inventory model that serves as the test model for the study. Section 3.2 presents the EB estimation procedure and its frequentist counterpart, while Section 3.3 outlines the statistical decision-theoretic framework used to evaluate their performance. Section 3.4 then details the experimental design of the numerical study, followed in Section 3.5 by the presentation and interpretation of the results. Section 3.6 concludes with a summary of the main findings and their implications for the subsequent chapters.

3.1 DES (s, S) Inventory Model

The purpose of this section is to provide the reader with a clear understanding of the DES model that will be used as a test model in the numerical study presented in this chapter. The test model selected for use is an (s, S) inventory model, based on an example DES model presented in Chapter 1 of Law [154]. This model represents a single-product inventory system, employing a classical (s, S) inventory policy. The model features a stationary approach to inventory management, with periodic review over a finite time horizon, that is widely used in practice [8]. Due to their widespread practical applicability, there has been substantial interest within the DES community towards the study of such models [96, 194].

For present purposes, we note that this model is classified as a "toy" model in the taxonomy presented in Subsection 1.2.3. With the strengths and weaknesses of each approach discussed previously in Subsection 1.2.3, here we discuss the reasons for the selection of this particular "toy model," aside from its aforementioned ubiquity in DES modelling literature. Firstly, Law's model is programmed in C code and relies upon an open source simulation library, SIMLIB [154]. As such, it is readily and freely available for any interested reader wishing to replicate the results or indeed build upon the research. Further, the model is experimentally convenient, running quickly enough to allow the generation of sufficient quantities of experimental data within reasonable timescales. Additionally, we note Law's stature within the DES community and his authorship of the classic DES text "Simulation Modeling and Analysis", with more than $175,000$ copies in print and over $20,000$ citations. Finally, the model has also been used as a DES test model in a number of other articles proposing new approaches to DES model experimentation and analysis, such as those by Koenig and Law [148] and by Luo and Hong [158].

The foregoing discussion highlights the model is well suited for use as a test model in the current study. Having briefly motivated its use, in Subsections 3.1.1 to 3.1.3, we turn our attention to describing the characteristics of the model that are relevant to our numerical study, including its conceptual structure, the specification of input quantities, and the definition of output quantities of interest.

3.1.1 Conceptual Model

The (s, S) inventory policy implies a simple approach to inventory control, where s denotes the safety stock level and S denotes the maximum inventory level. The policy operates such that when the inventory drops to the safety stock level s or below, an order is triggered to replenish the inventory to the maximum level S. In practice, this replenishment can take place in a continuous manner (order whenever inventory level drops to s or below), or in a periodic manner (review inventory at the start of each period and order if level is s or below). In this study, we employ a periodic policy that will run over a finite time horizon. Numerical values for s, S, the period length, the time horizon length and the initial inventory level will be discussed in the following subsection.

The (s, S) inventory system model used in this chapter also involves a number of randomly varying quantities (in contrast with the fixed quantities introduced in the preceding paragraph). To begin with, it is helpful to distinguish between the orders (which replenish the inventory level) and the demands (which deplete the inventory level) that arrive in the system. Neither the size of an order, nor the time at which it is created, are random, as they are determined by the (s, S) policy; however the time of an order's arrival to the system is random. This is due to the model's inclusion of randomly varying order lead times. Demands, on the other hand, are of randomly determined size, and arrive to the system at randomly determined time points. Probability distributions to characterise each of these random quantities will be discussed in the following subsection.

Finally, the (s, S) inventory system model also features a number of costs. Order costs are based on the order size, which is in turn based on the inventory level at the start of each period. Each order incurs a set-up cost and a variable cost per item. We also note that holding and shortage costs are included in the model, applying as the inventory level is positive or negative respectively. We can see that these quantities are not randomly determined, but rather are calculated based on the inventory level through simulated time. As with previously introduced model quantities, numerical values for these costs will be discussed in the following subsection.

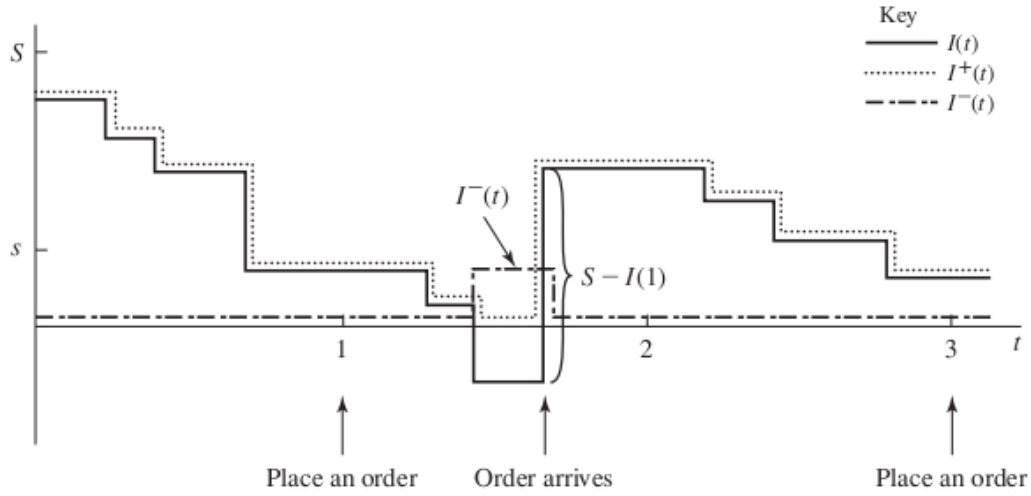

Figure 3.1: Terminology and inventory levels for periodic (s, S) policy, adapted from Law [154].

To illustrate the interaction between orders, demands and inventory level, we will now briefly discuss the operation of the (s, S) inventory policy, making use of the visualisation provided in Figure 3.1 (adapted from a similar figure found in Chapter 1 of Law [154]).

Let $I(t)$ indicate the inventory level at time point t. Here, we can see that reviews take place at the start of each period, that is, at time points $t = 1, 2, 3, \ldots$. If a review takes place, and the inventory level is found to be s or below, then an order of $S - I(t)$ units will be placed; otherwise no order is placed. For example, in Figure 3.1, at time point $t = 1$, an order of size $S - I(1)$ is placed. We can also see from Figure 3.1 that this order arrives with a random lead time before time point $t = 2$.

Each time a demand arrives, it will be satisfied immediately from the existing inventory provided the current level is sufficient. For example, we can observe the filling of three demands in Figure 3.1 before time point $t = 1$. If the inventory is insufficient, then the demand will be backlogged and satisfied later as soon as sufficient inventory is available. Such a situation results in a negative inventory level, as can be seen in Figure 3.1 just before time point $t = 2$. We further note that arriving orders are used first to eliminate any backlog (should this exist), and then to replenish the inventory level (should any remain).

Finally, we note that, as indicated in Figure 3.1, we can define the physical inventory level $I^+(t)$ (which is equal to $I(t)$ when $I(t) > 0$ and 0 otherwise) and backlog level $I^-(t)$ (which is equal to $-I(t)$ when $I(t) < 0$ and 0 otherwise).

For more information, we refer the interested reader to Axsäter [8] for a detailed discussion of (s, S) policy and its wide use in practice. This conceptual overview provides the structural context needed to interpret the model's behaviour under the different (s, S) settings systematically examined in the current numerical study.

3.1.2 Specifying Model Input Quantities

In the previous subsection, a range of model input quantities integral to the operation of the system were introduced and discussed. In this subsection, we will now discuss the specification of appropriate numerical values for these quantities.

To begin with, it is helpful to differentiate between the model input quantities to be varied during the numerical study, and those to be kept constant. The only model quantities to be varied during our experimentation are s, the safety stock level, and S, the maximum stock level. All other model quantities will kept constant during experimentation. By way of clarification for random quantities, being kept constant means that the probability distribution generating the random quantity will be kept constant during experimentation, rather than the quantity itself remaining constant. An experimental design describing the manner in which s and S are varied in the numerical study will be provided in Section 3.4. As such, the remainder of this subsection will focus on the specification of model input quantities kept constant in the numerical study.

To make the discussion in the last subsection more transparent, it is useful to list the model input quantities that now require specification. This information is captured in Table 3.1. The type of the model quantity in question, that is, whether it is deterministic or random, is also included in the table.

Before we present the necessary numerical specifications for the model input quantities identified, we note that the majority of these specifications will be kept iden-

Model Input Quantity	Type of Quantity
Period length	Deterministic
Time horizon length	Deterministic
Initial inventory level	Deterministic
Order arrival lead time	Random
Demand size	Random
Demand inter-arrival time	Random
Order set-up cost	Deterministic
Order variable cost	Deterministic
Holding cost	Deterministic
Shortage cost	Deterministic

Table 3.1: Model input quantities to be kept constant during experimentation.

tical to those presented in pages 48–50 of Law [154]. Where changes have been adopted, these will be highlighted, and the rationale for change discussed. As such, throughout our numerical experimentation with the (s, S) inventory model, model input quantity specifications will be as shown in Table 3.2.

Model Input Quantity	Specification
Period length	1 month
Time horizon length	12 months*
Initial inventory level	60 units
Order arrival lead times	Uniform ($\alpha = 0.5, \beta = 1$) months
Demand sizes	Truncated Poisson ($\mu = 2.69$) units*
Demand inter-arrival time	Exponential ($\mu = 0.1$) months
Order set-up cost	\$32 per order
Order variable cost	\$3 per unit
Holding cost	\$1 per unit per month
Shortage cost	\$5 per unit per month

Table 3.2: Specification of constant model input quantities.

In Table 3.2, deterministic model input quantities are assigned fixed numerical values, whilst random model input quantities are assigned fixed probability distributions. For example, in our model, orders involve a set-up cost of \$32, whilst demand inter-arrival times follow an exponential distribution, with a mean of 0.1 months.

Asterisks in Table 3.2 signify a deviation from the model input quantity specifications used by Law [154]. As may be seen, two changes were made to Law's specifications, relating to the demand size distribution and to the time horizon (or simulation run-length, in Law's terminology).

To motivate the changes adopted, however, it is useful to first recall the purpose of Law's model, comparing this to our current setting. Law's model was used to demonstrate the use of DES in choosing between different possible system configurations. His model input quantity specifications led to particularly low variability in model output quantities of interest, thus enabling a very straightforward choice between the different configurations. On the other hand, in our current numerical study, we wish to test the relative performance of two different estimation strategies. As such, the artificially low variability Law's settings create is unhelpful, and the changes made facilitate a more useful comparison of performance between EB and frequentist approaches. We now discuss each of these changes in turn.

Law opted to specify demand size using a simple, discrete probability distribution (with mean demand size 2.5 units) as described in Equation (3.1):

$$D = \begin{cases} 1 & \text{with} \quad \text{p=1/6} \\ 2 & \text{with} \quad \text{p=1/3} \\ 3 & \text{with} \quad \text{p=1/3} \\ 4 & \text{with} \quad \text{p=1/6} \end{cases} \tag{3.1}$$

In our experimentation, we instead specify demand size using a truncated approximation of a Poisson distribution (with mean demand size 2.69 units) as described in Equation (3.2):

$$D = \begin{cases} 1 & \text{with} \quad \text{p=0.22461} \\ 2 & \text{with} \quad \text{p=0.28075} \\ 3 & \text{with} \quad \text{p=0.23396} \\ 4 & \text{with} \quad \text{p=0.14623} \\ 5 & \text{with} \quad \text{p=0.07311} \\ 6 & \text{with} \quad \text{p=0.03046} \\ 7 & \text{with} \quad \text{p=0.01088} \end{cases} \tag{3.2}$$

This change was made for several reasons. Firstly, whilst full-scale realism is not of prime concern in our current numerical study, Law's original distribution was deemed unhelpfully simplistic and not well-suited to our current numerical study. Classically, a Poisson distribution is used to model demand size in (s, S) inventory models, and so by changing to Equation (3.2), our approach was more consistent with the literature. Relative to a Poisson distribution, Law's approach significantly reduces both the variability and the skewness of this model input quantity. As such, whilst our model is classified as a "toy" model, with this change it can also be considered a practically relevant one. Additionally, in opting for a truncated Poisson distribution, over a standard Poisson distribution, we were able to take advantage of greatly increased computational efficiency. This ensured the practical feasibility of the extensive simulation necessary for this numerical study.

Finally, preliminary testing was carried out to identify an appropriate mean value for our truncated Poisson distribution, a value compatible with the rest of the model input quantity specifications. More specifically, this testing enabled us to select a mean value ($\mu = 2.69$) that avoided edge-case behaviours and produced a level of variability suitable for a meaningful analysis.

The second change to Law's approach related to the time horizon, or simulation run-length. In Law's text, a time horizon of 120 months was used, yet a time horizon of 12 months was used in this study. Firstly, as discussed above, the data obtained using Law's time horizon was of particularly low variability. Whilst this was well suited to Law's original context (facilitating straightforward selection be-

tween configurations), it was less helpful in our current study (assessing relative performance in estimation). Adopting a time horizon of just 12 months yielded more practically useful data, and furthermore ensured the practical feasibility of the extensive simulation necessary for the study by reducing the computational cost involved.

At this point, we note that all preceding discussion on the specification of model quantities has focused solely on model input quantities. In the next subsection, we will proceed to discuss model output quantities and related questions of interest, thereby completing the description of the model elements and setting the scene for the later discussion of the (s,S)-based experimental design.

3.1.3 Model Output Quantities and Questions of Interest

In this subsection, we aim to discuss the context around the use of the DES (s, S) inventory model. We will begin by discussing the changes we are likely to observe in model behaviour as a result of changes made to the model input quantities (that is, changes in the values of s and S). Through this discussion, we will highlight the model output quantities of interest and the questions that are likely to be of key concern to a decision maker using the model. The purpose of this discussion is to provide the reader with a better appreciation of the technical details of the (s, S)-based experimental design outlined later in Section 3.4.

To begin, we make some simple observations about the model behaviour that is likely to be observed upon varying the model input quantities s and S. Looking at the safety stock level s, we note that smaller values of s, on average, lead to less accumulation of stock (or lower inventory levels), and thus likely lead to more frequent inventory shortages. Such a situation will likely incur lower holding costs, but greater backlog costs. On the other hand, larger values of s, on average, lead to greater accumulation of stock (or higher inventory levels), and thus likely lead to less frequent inventory shortages. This situation will conversely incur higher holding costs, but lower backlog costs.

Looking now at the maximum stock level S, we note that setting S very close to the safety stock level s will mean that stock orders will tend to be smaller and, all

things being equal, more frequent. Such an arrangement will likely incur higher order costs due to the repeated application of the set-up order cost. These costs, however, must be balanced against the increased holding costs that may be incurred from setting a higher value of S. Naturally, the inverse applies in setting S further away from s. An analyst using the (s, S) inventory model in practice would need to be aware of these cost related trade-offs in order to make useful recommendations regarding the real system.

The balancing of costs has emerged as a key theme of the foregoing discussion, and indeed the aim of a decision maker in such a context would be to strike the right balance between these cost related trade-offs. As such, frequently the key output quantity of interest to the decision maker is the *total operating cost*. The total operating cost incorporates the holding and shortage costs associated with the inventory level over the time period considered, as well as the set-up and variable costs associated with all orders created. Whilst situations do exist where other model output quantities are important (for example, the total number of backlogged demands in an organisation keenly focused on customer service), total operating cost is oftentimes the most important consideration. In fact, broadly speaking, the key objective in the use of such (s, S) inventory models is to determine the values of s and S that minimise the overall cost of operating the inventory system. In light of the importance of total operating cost, we have selected this as our model output quantity of interest.

Making this more precise, we note that for each pair of (s, S) values, this model scenario implicitly defines a distinct output distribution associated with the model output quantity total operating cost. In our upcoming numerical investigations, we will generate large quantities of data from a number of such (s, S) pairings as defined by our experimental design (to be discussed in Section 3.4). This data will be used to calculate a long-run estimate of the sample mean, which serves as a proxy for the true distribution mean (the true total operating cost), providing us with our estimation target. This data will also be batched into samples, and used to construct sets of EB and frequentist estimates of the true distribution mean in order to assess the relative performance of the two approaches.

In this subsection we discussed the model output quantity of interest, total op-

erating cost, and laid out, in general terms, how data generated will be used to compare our different estimation strategies. In the next section, we introduce and provide derivations of the EB and frequentist procedures to be compared in our current numerical study.

3.2 Double Shrinkage EB Point Estimator

The purpose of this section is to present and discuss the EB estimator used both in this and in later chapters of the thesis. The estimator we employ is the "double shrinkage" point estimator published by Zhao [260]. We outline the rationale for its selection and its suitability for use in the context of DES model experimentation.

We begin by discussing the "double shrinkage" nature of the estimator. As noted in Chapter 1 of the thesis, in the EB literature, the term population refers to what is, in our DES setting, a scenario. Double shrinkage means that the estimation of both means and variances involves a shrinkage component. Each population variance is estimated as a weighted average of two quantities: an estimate based on data from that population alone and an estimate based on data pooled across populations. The population-specific estimate is therefore 'shrunk' toward the pooled estimate. These variance estimates then feed into the subsequent mean estimation, which follows the same structure: each population mean is a weighted average of an individual estimate and an overall pooled estimate, hence the term "double shrinkage." This approach allows us to borrow strength on both mean and variance, yielding more efficient inference. The double shrinkage approach of Zhao [260], together with additional works by Hwang et al. [126] in interval estimation, and Hwang and Liu [125] in hypothesis testing, constitutes a step forward in the literature from previous EB estimation approaches. Prior to these works, EB estimation approaches featured single shrinkage, or shrinkage of means or variances, only. The interested reader is referred to these individual references for a more detailed review of the literature in this area.

The modelling assumptions used here offer a practical and well-motivated starting point for DES model experimentation, in particular, given the behaviour of the aggregated model output quantities of interest. Whilst DES output distributions are not guaranteed to be normal, aggregated output quantities, such as the total

operating cost over a fixed horizon, often display near-normal behaviour, owing to their construction from many small random contributions [16]. This makes a normal-based framework a practical choice for an initial investigation.[1] Allowing for heteroscedastic variances is also appropriate, as different scenarios frequently exhibit different levels of variability in practice [144].

Looking at the hierarchical structure of the modelling framework, the estimator adopts a normal–lognormal model, rather than the more standard normal–inverse-gamma model, based on a classical conjugate pairing. The lognormal model allows the utilisation of standard results available on normal models, and so enables a closed form estimator, a significant advantage in practice over the numerical methods required for normal-inverse-gamma formulations.

Finally, we note that aside from its theoretical advantages, rigorous empirical testing has been conducted in support of this estimator [125]. This extensive programme of testing features both simulated and area standard "spike-in" data sets, demonstrating improvements in performance over means only and variance only shrinkage estimation approaches. Having outlined the motivation for and broad characteristics of the estimator, we now introduce its modelling assumptions, its derivation and properties, and the estimation of its hyperparameters in the upcoming subsections.

3.2.1 Modelling Assumptions

We will now introduce the modelling assumptions underlying the double shrinkage (DS) EB point estimator presented by Zhao [260]. Here, we note that whilst the details and derivations of Subsections 3.2.1 to 3.2.3 follow that of Zhao, in a few places additional clarification has been provided. This is to ensure clarity for our intended audience of DES practitioners and researchers, as opposed to Zhao's intended audience of statistical researchers.

Letting X_i denote the sample mean from the i^{th} population, then for $i = 1, 2, \ldots, p$,

[1] Although the present study adopts a normal-based formulation, subsequent chapters (notably the case study in Chapter 6) examine settings where the output distributions are non-normal, including substantially skewed data. The robustness of the contributions of the thesis to such deviations is therefore assessed later in the thesis.

we make the following assumptions:

$$X_i|\theta_i, \sigma_i^2 \overset{iid}{\sim} N(\theta_i, \sigma_i^2), \tag{3.3}$$

$$\theta_i \overset{iid}{\sim} N(\mu, \tau^2), \tag{3.4}$$

$$\log \sigma_i^2 \overset{iid}{\sim} N(\mu_\upsilon, \tau_\upsilon^2), \tag{3.5}$$

$$\log(S_i^2/\sigma_i^2)|\sigma_i^2 \overset{iid}{\sim} N(m, \sigma_{ch}^2), \tag{3.6}$$

where S_i^2 is the sample estimate of the variance of X_i, and m and σ_{ch}^2 are:

$$m = E[\log(\chi_d^2)] = \psi(\frac{d}{2}) - \log(\frac{d}{2}),$$
$$\sigma_{ch}^2 = Var[\log(\chi_d^2)] = \psi'(\frac{d}{2}),$$

with ψ the digamma function, and d denoting the degrees of freedom (assumed equal across $i = 1, 2, \ldots, p$).

For the moment, we take hyperparameters $\mu, \tau^2, \mu_\upsilon$ and τ_υ^2 as known, however in Subsection 3.2.3, appropriate EB estimates will be derived for these quantities.

3.2.2 Derivation and Properties

Having outlined the modelling assumptions, we now present the derivation of our EB point estimator following Zhao [260]. As mentioned, we temporarily assume known hyperparameters.

Given observation X_i from our sampling distribution (3.3), and our prior distribution (3.4) on θ_i, then assuming for the moment that σ_i^2 is known, a standard application of Bayes rule yields:

$$\theta_i|X_i, \sigma_i^2 \ \sim \ N(M_i X_i + (1 - M_i)\mu, M_i\sigma_i^2),$$

for the posterior distribution of θ_i, where M_i is given by $M_i = \tau^2/(\tau^2 + \sigma_i^2)$ [46].

Typically, we would then use the posterior mean:

$$\hat{\theta}_i = M_i X_i + (1 - M_i)\mu, \tag{3.7}$$

as a point estimator of the population mean θ_i.

However, the true population variances σ_i^2 (for $i = 1, 2, \ldots, p$) are unknown. Consistent with Zhao [260], a lognormal prior (3.5) has been adopted for σ_i^2, with additional assumption (3.6) that S_i^2/σ_i^2 is lognormally distributed. In (3.6), parameters m and σ_{ch}^2 have been selected to coincide with those of χ_d^2/d, a standard distributional assumption regarding quantity S_i^2/σ_i^2 [147, Proposition 2.11].

Taking (3.6) and utilising a standard logarithmic identity results in:

$$\log S_i^2 | \sigma_i^2 \ \sim \ N(m + \log \sigma_i^2, \sigma_{ch}^2), \tag{3.8}$$

and, combining (3.5) and (3.8) using Bayes rule yields:

$$\log \sigma_i^2 | S_i^2 \ \sim \ N(M_\upsilon(\log S_i^2 - m) + (1 - M_\upsilon)\mu_\upsilon, M_\upsilon \sigma_{ch}^2), \tag{3.9}$$

$$\text{with } M_\upsilon \ = \ \tau_\upsilon^2/(\tau_\upsilon^2 + \sigma_{ch}^2). \tag{3.10}$$

As with Zhao [260], the variance σ_i^2 (for $i = 1, 2, \ldots, p$) is estimated by taking the exponential of the mean of (3.9), yielding:

$$\hat{\sigma}_i^2 = \exp(M_\upsilon(\log S_i^2 - m) + (1 - M_\upsilon)\mu_\upsilon). \tag{3.11}$$

Thus, assuming known hyperparameters $\mu, \tau^2, \mu_\upsilon$ and τ_υ^2, we have estimator $\hat{\theta}_i$ for the i^{th} population mean θ_i as follows:

$$\hat{\theta}_i \ = \ \hat{M}_i X_i + (1 - \hat{M}_i)\mu, \tag{3.12}$$

where $\hat{M}_i = \tau^2/(\tau^2 + \hat{\sigma}_i^2)$, and $\hat{\sigma}_i^2$ is as given by equation (3.11) above.

3.2.3 Estimation of Hyperparameters

It now remains to estimate the hyperparameters $\mu, \tau^2, \mu_\upsilon$ and τ_υ^2 from data pooled across our populations. Here, we note that only by estimating both μ and τ^2

(the mean hyperparameters) and μ_v and τ_v^2 (the variance hyperparameters) with pooled data do we obtain the desired DS EB point estimator. As with Zhao [260], we adopt the following estimation strategies.

Beginning with μ_v, a simple rearrangement of assumption (3.6) yields:

$$\log S_i^2 - m \,|\sigma_i^2 \;\sim\; N(\log \sigma_i^2, \sigma_{ch}^2),$$

which, in combination with (3.5) and the law of total expectation, yields:

$$\mu_v = E(\log \sigma_i^2) = E(E(\log S_i^2 - m \,|\sigma_i^2)) = E(\log S_i^2 - m).$$

Thus, the mean μ_v may be estimated by taking the observed, across population (or pooled) average of quantity $\log S_i^2 - m$, giving:

$$\hat{\mu}_v = \frac{1}{p}\sum_i (\log S_i^2 - m). \tag{3.13}$$

To estimate $\hat{\tau}_v^2$, we make use of the identity $E((\log S_i^2 - m)^2) = \sigma_{ch}^2 + \tau_v^2 + \mu_v^2$ and estimator (3.13). Similar to above, we take the across population average of quantity $(\log S_i^2 - m)^2$ as an estimate for $E((\log S_i^2 - m)^2)$. Substitution and rearrangement yields:

$$\hat{\tau}_v^2 = \left(\frac{1}{p}\sum_i (\log S_i^2 - m)^2 - \sigma_{ch}^2 - \hat{\mu}_v^2\right)_+, \tag{3.14}$$

where the positive part is taken to ensure non-negativity.

Before proceeding, we note that the estimation of μ_v and τ_v^2 immediately yields EB estimators for both the weight M_v in (3.10):

$$\hat{M}_v \;=\; \frac{\hat{\tau}_v^2}{\hat{\tau}_v^2 + \sigma_{ch}^2}, \tag{3.15}$$

and the population variance $\hat{\sigma}_i^2$ in (3.11):

$$\hat{\sigma}_{EB,i}^2 \;=\; \exp\left(\hat{M}_v(\log S_i^2 - m) + (1 - \hat{M}_v)\hat{\mu}_v\right). \tag{3.16}$$

Looking to define EB estimators for the hyperparameters of the prior distribution on θ_i, as with Zhao [260], a weighted average is used as an estimator for μ:

$$\hat{\mu} = \frac{\sum_i (X_i / \hat{\sigma}^2_{EB,i})}{\sum_i (1/\hat{\sigma}^2_{EB,i})}. \tag{3.17}$$

In (3.17), each observation X_i is weighted by the inverse of the corresponding population variance estimate, as given by (3.16).

Finally, to estimate τ^2, the approach proceeds as follows. Firstly, we note that:

$$E(X_i - \mu)^2 \mid \sigma_i^2 = \sigma_i^2 + \tau^2,$$

which follows from the general properties of the normal distribution.

Rearrangement of this identity suggests the following estimator for τ^2:

$$\hat{\tau}^2 = \left(\frac{\sum_i \left((X_i - \hat{\mu})^2 - \sigma_i^2 \right)}{p} \right)_+, \tag{3.18}$$

where, again, the positive part has been taken to ensure non-negativity.

However, $\hat{\sigma}^2_{EB,i}$ is not an unbiased estimator of $\hat{\sigma}_i^2$. Noting the lognormal distributional assumption (3.6), we recognise that:

$$E(S_i/\sigma_i^2) \mid \sigma_i^2 = e^{m + \sigma^2_{ch}/2}.$$

Some re-arrangement of this identity yields:

$$E(S_i e^{-m - \sigma^2_{ch}/2}) \mid \sigma_i^2 = \sigma_i^2,$$

indicating an unbiased estimator for σ_i^2.

As such, as with Zhao [260], we adjust (3.18) to provide an unbiased estimator of τ^2:

$$\hat{\tau}^2 = \left(\frac{\sum_i \left((X_i - \hat{\mu})^2 - S_i^2 exp(-m - \sigma^2_{ch}/2) \right)}{p} \right)_+. \tag{3.19}$$

Thus, the DS EB point estimator of θ_i is given as:

$$\hat{\theta}_{EB,i} = \hat{M}_{EB,i} X_i + (1 - \hat{M}_{EB,i})\hat{\mu}, \tag{3.20}$$

where $\hat{M}_{EB,i} = \hat{\tau}^2/(\hat{\tau}^2 + \hat{\sigma}^2_{EB,i})$, and $\hat{\mu}$, $\hat{\tau}^2$ and $\hat{\sigma}^2_{EB,i}$ are as given in (3.17), (3.19) and (3.16) above, respectively.

3.2.4 Comparative Frequentist Point Estimator

Having defined our EB estimator, in this subsection, we will now introduce the frequentist estimator used as a benchmark, facilitating our subsequent comparison between EB and frequentist approaches.

As such, we provide a formal definition of the sample mean $\bar{x}$ for given sample of data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ as follows (see, e.g. Kreyszig [149] for further details):

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n). \tag{3.21}$$

The standard sample mean was chosen as a benchmark estimator for several reasons. Firstly, it is a commonly used, intuitively simple, and easy to calculate estimator of a population mean parameter. Indeed, it may be considered the "natural" point estimator for the population mean [108]. In addition, it is computationally efficient and exhibits good statistical properties [108]. In particular, by the weak law of large numbers, the sample mean is a consistent estimator of the population mean, and by the Lehmann-Scheffe theorem, it is the minimum variance unbiased estimator (MVUE). Finally, we note its widespread acceptance within the DES literature as the standard or benchmark estimator for the population mean [1, 69, 108].

Having now introduced both the EB and frequentist estimation procedures, we next turn to discuss the framework by which we will undertake our comparison. As such, we begin by providing a brief review of statistical decision theory and a discussion of our error measurement approach in the next subsection, and then provide a technical description of the experimental design for the numerical study in the following subsection.

3.3 Decision Theory and Error Measures

In this section, we provide a brief overview of statistical decision theory, followed by a discussion on the error measures used in both and later chapters. This provides the reader with a framework within which to place and compare the different error measurement approaches used across the thesis in our comparative evaluations of EB and frequentist approaches. The framework outlined below broadly follows that presented on page 4 of Young and Smith [255]. The section concludes with a discussion relating the components of the framework to the evaluation undertaken in the current chapter.

In general, the formal formulation of a statistical decision problem involves several components, as follows:

1. A *parameter space* Θ, typically a subset of $\mathbb{R}^d$ for a given $d \geq 1$, giving a d-dimensional vector of unknown parameters. This set of vectors represents the set of possible unknown states of nature.

2. A *sample space* $\mathcal{X}$ containing possible data values x. Typically our data is of the form $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, where we have n different observations.

3. A *family of probability distributions* on the sample space $\mathcal{X}$, indexed by the parameter $\theta \in \Theta$. For our purposes, this will consist of a parametric family of distributions $f(x; \theta)$.

4. An *action space* $\mathcal{A}$ containing all possible actions (or decisions) available in the given situation.

5. A *loss function* L linking the action taken to the true state of nature. That is, if an action $a \in \mathcal{A}$ is taken and the true state of nature is $\theta \in \Theta$, then the loss function has a value $L(\theta, a)$.

6. A set $\mathcal{D}$ of *decision rules* containing elements d mapping each point x in the sample space $\mathcal{X}$ to an associated action $d(x)$ in the action space $\mathcal{A}$.

For further details regarding these terms, as well as for a comprehensive technical treatment of statistical decision theory, we refer the interested reader to Berger [24]. We now discuss this framework as it relates to the current research, first points

regarding loss functions and secondly points concerning decision rules.

Loss functions are concerned with error measurement. Essentially, they return a value that measures the impact of taking action a, based on decision rule d, when the true state of nature is θ. In our context, this may be thought of as the error incurred in approximating parameter θ with estimate $\hat{\theta}$ based on an estimation procedure d. Different loss functions emphasise different aspects of the statistical efficiency of the estimation approach in use.

In this research, there are two aspects of critical importance: measuring the overall error and measuring the maximum error. As illustrated by the baseball example of Efron and Morris [81] discussed in Chapter 1, an EB procedure may substantially reduce the overall error in estimation across a collection of related populations, provided that the underlying structure supports information sharing. However, the example also demonstrated that a reduction in overall error may coincide with an increase in the maximum error, as occurred for a notable outlier in that setting. For EB methods to be useful in DES model experimentation, where practitioners must balance efficiency gains against robustness at a scenario level, it is important to explicitly include both error measurements in our evaluations.

With this in mind, and supposing that we are estimating parameter θ with estimator $\hat{\theta}$ across k different populations (or model scenarios), we define the following error measures to evaluate and compare our different estimation approaches. To look at overall performance, we calculate the *total squared error* incurred across all k populations or model scenarios. Thus, we define:

$$Err_{overall} = \sum_{i=1}^{k} \left(\hat{\theta}_i - \theta_i\right)^2. \tag{3.22}$$

To examine at the *maximum possible error* incurred across all k populations or model scenarios, we define:

$$Err_{max} = \max_i |\hat{\theta}_i - \theta_i|. \tag{3.23}$$

Equations (3.22) and (3.23) will be used as the basis of error measurement strategies in this chapter, and in later chapters as well.

In considering decision rules, in the preceding subsections we introduced two point estimation procedures, namely the DS EB point estimator of Subsection 3.2.1, and the standard frequentist sample mean of Subsection 3.2.4. These decision rules seek to estimate the true, but unknown, parameter θ. As discussed in Subsection 3.1.3, currently parameter θ is defined as the total operating cost of the inventory system in question. As such, in the numerical current study, we will compare two different decision rules, based on EB and frequentism respectively, using the two different loss functions defined in Equations (3.22) and (3.23) respectively.

Having now introduced and discussed all requisite preliminaries, in the next section we proceed to discuss the experimental design for the current numerical study. This will be followed by the numerical results and their interpretation in Section 3.5, and the drawing together of concluding thoughts in Section 3.6.

3.4 Experimental Design for Numerical Study

In this section, we outline the experimental design of the numerical study that constitutes the core contribution of this chapter. This will involve the presentation and discussion of both the strategic aspects (such as, the choice of (s, S) combinations) and the tactical aspects (such as, the number of replications conducted) of the experimental design. In addition, we describe the data generated by this experimental design and how it is subsequently used in the comparative evaluation of EB and frequentist estimation approaches.

3.4.1 Strategic Aspects

We begin by discussing the strategic aspects of the study's experimental design. Firstly, we recall from Subsection 3.1.2 that the only model quantities treated as factors in the study are s, the stock re-order point, and S, the maximum stock level, with all other quantities held fixed. It is worth noting that in Law's experimentation with the (s, S) model, only nine inventory policies (that is, nine (s, S) combinations) are examined [154]. A more extensive range of inventory policies are examined in this chapter's numerical study. The reasoning behind this follows that set out in Subsection 3.1.2, namely the difference in the purpose of the exper-

imentation. Law's study is designed to illustrate the use of simulation in a setting with a relatively small number of clearly distinguishable (s, S) options, where the aim is simply to compare these alternatives and choose among them. By contrast, the present study uses the model as a testbed for a methodological evaluation, requiring a much larger and more variable set of (s, S) policies to support a rigorous comparison of the two estimation approaches. Details of the inventory policies, or (s, S) combinations, included in the numerical study are as presented in Table 3.3.

The table lists 30 different configurations, each a grid of (s, S) design points, or model scenarios, to be examined during the numerical study. Each configuration is defined by its s range, its S range and its step-size, with the additional condition that (s, S) design points not satisfying $s \leq S - 1$ are excluded due to infeasibility. The number of design points contained in each configuration is also provided in the table. Taking Configuration 8 (C8) as an example, we can see in Figure 3.2 how an s range of [40,50], an S range of [40,60], and a step-size of 2 give rise to 45 admissible (s, S) design points.

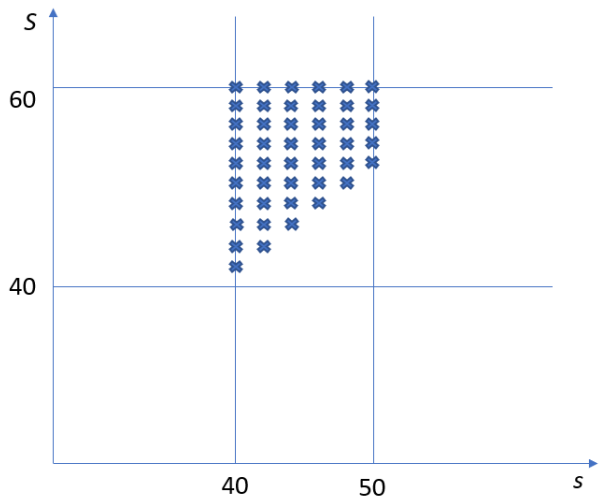

Figure 3.2: A visual representation of the grid of (s, S) points included in Configuration 8 (C8) of the numerical study.

The idea behind the overall experimental design represented in Table 3.3 is that each configuration may be thought of as its own practitioner-level experimental design, reflecting the kind of multi-scenario study a DES analyst might construct

Configuration	s range	S range	Step-size	# Grid points
C1	[20,60]	[40,100]	1	2270
C2	[20,60]	[40,100]	2	585
C3	[20,60]	[40,100]	5	102
C4	[20,60]	[40,100]	10	29
C5	[20,60]	[40,100]	20	9
C6	[20,60]	[40,100]	30	6
C7	[40,50]	[40,60]	1	165
C8	[40,50]	[40,60]	2	45
C9	[40,50]	[40,60]	5	9
C10	[40,50]	[40,60]	10	3
C11	[40,50]	[60,80]	1	231
C12	[40,50]	[60,80]	2	66
C13	[40,50]	[60,80]	5	15
C14	[40,50]	[60,80]	10	6
C15	[60,80]	[70,90]	1	375
C16	[60,80]	[70,90]	2	100
C17	[60,80]	[70,90]	5	19
C18	[60,80]	[70,90]	10	6
C19	[20,30]	[80,100]	1	231
C20	[20,30]	[80,100]	2	66
C21	[20,30]	[80,100]	5	15
C22	[20,30]	[80,100]	10	6
C23	[80,90]	[80,100]	1	165
C24	[80,90]	[80,100]	2	45
C25	[80,90]	[80,100]	5	9
C26	[80,90]	[80,100]	10	3
C27	[40,50] & [60,70]	[60,80]	10	9
C28	[30,40] & [80,90]	[40,60] & [80,100]	10	14
C29	[20,30] & [80,90]	[80,100]	10	9
C30	[40,50] & [70,80]	[80,100]	10	11

Table 3.3: Complete list of experimental configurations, each a grid of (s, S) values.

when exploring a policy space.[2] Within each configuration, these scenarios may

[2]The term "configuration" is adopted to avoid using "experimental design" in two different senses within this section: the overall numerical design of the study and the practitioner-level arrangements of model scenarios used as inputs to EB and frequentist estimation.

either be pooled to support EB estimation of the parameter of interest, or may be analysed separately to support frequentist estimation.

Each configuration listed in the table is also a subset of Configuration 1 (C1), the largest of the configurations. C1 serves as the master grid for the study: it contains every admissible (s, S) design point considered, and the data for all other configurations are obtained by extracting the relevant entries from C1. It is generated using $s \in [20, 60]$, $S \in [40, 100]$, and a step-size of 1, with a total of $2,270$ design points (or model scenarios). The horizontal lines in Table 3.3 demark groups of configurations with the same s and S ranges, but with step-size, and so grid density, decreasing from top to bottom.

Naturally, there is an almost limitless range of possible configurations that might be added to the study's experimental design. Rather than pursue exhaustive testing, which would add considerable computational burden without yielding commensurate insight, the study includes a set of configurations chosen to span substantively different experimental conditions. These configurations vary systematically in both the area of the (s, S) plane covered and the number of design points included, leading to substantial difference in grid density. Some comprise one contiguous area of the (s, S) plane, e.g. C1–C26, whilst others comprise two disjoint areas, e.g. C27–C30. This structured variation in the included configurations results in a rigorous testbed for evaluating and comparing the statistical efficiency of the EB and frequentist estimation approaches.

3.4.2 Tactical Aspects and Data Generation

After outlining the strategic aspects of the experimental design, we now turn to the practical details governing how the study is executed and how its data are generated. This includes the tactical choices made in running the simulation model, and the procedures used to construct the datasets for subsequent analysis.

We begin with the tactical aspects. Similar to both the model input quantity specifications and the strategic aspects of the experimental design, broadly speaking, decisions regarding the tactical aspects follow those outlined in Chapter 1 of Law [154]. Where there are deviations from Law's approach, these are highlighted

and discussed. The relevant points are as follows:

Initial condition:
The (s, S) model is run with initial condition $I(0) = 60$ (consistent with Law's approach, as discussed in Subsection 3.1.2).

- *Justification*: The value of the initial condition was determined by Law [154] via numerical testing and is retained here for comparability.[3]

Run-length:
The (s, S) model is run for a period of 12 months (differing from Law's 120 months horizon [154], as discussed in Subsection 3.1.2).

- *Justification*: A long (120 months) run-length produces an extremely accurate long-run total operating cost estimate, but it also artificially suppresses the within-policy variability. This suits Law's aim to illustrate policy comparison but distorts the comparative evaluation conducted in this study. Preliminary testing identified 12 months as preserving realistic variability while still providing a stable long-run estimate.

Number of replications:
For each configuration, $10,000$ independent replications are executed (with the number of replications left unspecified and to be explored by the reader in Law [154]).

- *Justification*: This value was chosen based on preliminary numerical testing, ensuring that long-run estimates of the total operating cost are accurate enough to serve as benchmarks, and to allow the creation of sufficiently many subsamples (at sizes 5, 10, and 20) for reliable comparative evaluation.

Having discussed the relevant strategic and tactical aspects of the experimental design, we now turn to discuss the final component: the data generated under these settings and the subsequent analysis required to evaluate and compare the two estimation procedures under consideration.

[3] As discussed in Subsection 3.1.2, the initial condition is fixed at $I(0) = 60$ for all (s, S) combinations. For policies with $S < 60$, e.g. C8, the system therefore begins in an overstocked state relative to the policy. While this would be atypical in a classical analytical treatment of the inventory model, it presents no difficulty for DES: the simulation simply starts from a transient condition and the system state adjusts accordingly during the run.

In terms of the data yielded, for C1 (the largest configuration), we obtain a 2,270 by 10,000 data matrix, where each entry represents the total operating cost for the given (s, S) policy and replication. C1 functions as the master grid for the numerical study, in the sense that it contains every (s, S) policy examined across all configurations. The remaining configurations (C2–C30) also have associated data matrices, each constructed via the extraction of the relevant entries of the aforementioned C1 matrix, thereby avoiding the need to re-simulate overlapping policies and ensuring full comparability between configurations. From Table 3.3, we see that the number of (s, S) policies varies across configurations, and so the dimensions of the data matrices obtained from experimentation also vary accordingly (e.g. C8 yields a 45 by 10,000 matrix).

In terms of how this data will be analysed, the data matrices facilitate the calculation of pairs of EB and frequentist estimates of the output quantity of interest, the true mean total operating cost, for each (s, S) policy included in each of the experimental configurations. As noted earlier, the tactical choices were made to ensure sufficiently accurate long-run estimates while enabling extensive subsampling. We evaluate the estimation procedures using practically feasible sample sizes of 5, 10 and 20. The 10,000 replications allow the construction of 500 non-overlapping subsamples of each size, which provides a sufficiently large number of independent estimate pairs to support stable comparisons of relative performance. For each pair of estimates, the error measures as given in Equations 3.22 and 3.23 are computed, and their ratio taken to measure relative performance, and the resulting ratios averaged to mitigate stochastic variability.

Finally, we discuss relevant computational aspects of the experimentation within this study, with this information provided so as to ensure the reproducibility of results. Firstly, all (s, S) inventory model simulation is conducted using the C code (modified as in Subsection 3.1.2) and SIMLIB library as provided in Law [154]. Secondly, all data analysis to calculate the EB and frequentist estimates of the total operating cost, and the assessment of their relative performance using the ratio of Equations 3.22 and 3.23, is implemented in MATLAB (2014a). Example MATLAB codes for this chapter may be found in Appendix A. Finally, we note that all computational work is implemented on a standard desktop PC (Intel Core-i5, 2.7GHz, 8GB RAM). With these experimental and computational details

established, we now turn to the numerical results and their interpretation.

3.5 Numerical Study Results and Interpretation

Having discussed all relevant aspects of the current study's experimental design, in this subsection, we present and discuss the numerical results obtained upon its implementation. In the following, final subsection of this chapter, we relate these findings to the on-going discussion of the relevance of EB to DES model experimentation.

The numerical results of the study are presented across three tables: Table 3.4 details the results for a sample size of 5; Table 3.5, the results for sample size 10; and Table 3.6, the results for sample size 20. It may be seen that their format is almost identical to the format of Table 3.3, save for the additional (final) two columns, MSSE and MMAE. These columns provide the error ratios used to assess the relative statistical efficiency of the EB and frequentist approaches. Whilst a broad overview of the error measurement approach used in the study was provided in the preceding subsections, it is helpful to restate their precise definitions here, as these form the basis of the comparisons reported in Tables 3.4–3.6. We therefore present the exact formulation of the error measurement ratios, before proceeding to discuss the results obtained upon their application.

Looking first at the mean sum of squared error (MSSE) ratio, we make use of Equation (3.22). For each configuration, we calculate $Err_{overall}$ for each replication by summing over the k different (s, S) design points, or model scenarios, that comprise that configuration. This calculation is carried out for both the DS EB point estimator of Equation (3.20), and the frequentist sample mean of Equation (3.21).

For each configuration, we therefore have:

$$Err_{overall,EB,j} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i,j} - \theta_i\right)^2,$$

and:

$$Err_{overall,freq,j} = \sum_{i=1}^{k} \left(\hat{\theta}_{freq,i,j} - \theta_i\right)^2,$$

for replication j. Averaging these quantities over our 500 replications yields:

$$MSSE_{EB} = \frac{1}{500}\sum_{j=1}^{500}\sum_{i=1}^{k}\left(\hat{\theta}_{EB,i,j} - \theta_i\right)^2,$$

and:

$$MSSE_{freq} = \frac{1}{500}\sum_{j=1}^{500}\sum_{i=1}^{k}\left(\hat{\theta}_{freq,i,j} - \theta_i\right)^2.$$

Therefore, for each configuration, the MSSE ratio is obtained as follows:

$$MSSE = MSSE_{EB}/MSSE_{freq}. \tag{3.24}$$

The MSSE ratio captures the relative performance of the two estimation approaches with respect to the average sum of squared error: a value less than 1 indicates the EB approach outperforms the frequentist approach, whilst a value greater than 1 indicates that the reverse is true.

A similar process is involved in the calculation of the mean maximum absolute error (MMAE) ratio. For each configuration, Equation (3.23) is used to calculate an error measure for each replication and estimator, that is:

$$Err_{max,EB,j} = \max_i |\hat{\theta}_{EB,i,j} - \theta_i|,$$

and:

$$Err_{max,freq,j} = \max_i |\hat{\theta}_{freq,i,j} - \theta_i|,$$

for replication j. Averaging these quantities over our 500 replications yields:

$$MMAE_{EB} = \frac{1}{500}\sum_{j=1}^{500}\max_i \left|\hat{\theta}_{EB,i,j} - \theta_i\right|,$$

and:

$$MMAE_{freq} = \frac{1}{500}\sum_{j=1}^{500}\max_i \left|\hat{\theta}_{freq,i,j} - \theta_i\right|.$$

Thus, for each configuration, the MMAE error ratio is given as follows:

$$MMAE = MMAE_{EB}/MMAE_{freq}. \tag{3.25}$$

Values of the MMAE ratio may be interpreted in a similar manner to values of the MSSE ratio. That is, values less than 1 indicate a superior EB performance, whilst values greater than 1 indicate an inferior EB perfomance, both relative to frequentist performance.

Having discussed the error measurement underlying the MSSE and MMAE results, we now turn to their interpretation. The layouts of Tables 3.4, 3.5 and 3.6 follow the same structure as Table 3.3, with the s and S ranges, step-size, and number of (s, S) design points listed alongside the corresponding error ratios. Taking Configuration 8 by way of example, we see its ranges s and S, [40,50] and [40,60], step-size of 2, and 45 design points reported in each table, alongside its error ratios of MSSE 0.5438 and MMAE 0.5605 for sample size 5, MSSE 0.7270 and MMAE 0.7322 for sample size 10, and MSSE 0.7396 and MMAE 0.7468 for sample size 20. So, for this configuration, EB outperforms the frequentist approach, across all sample sizes (5, 10 and 20) and both error measure (MSSE and MMAE).

This superiority of EB is not, however, observed consistently throughout the results. Configuration 5, for example, shows a mixed and largely neutral performance at sample size 5 (MSSE 0.9946, MMAE 1.0213), before displaying EB superiority on both measures at sample size 10. Conversely, Configuration 28 exhibits frequentist superiority across sample sizes; for instance, at sample size 10, the ratios are MSSE 1.0466 and MMAE 1.0565. As such, we can appreciate that the overall picture is varied, complicating the interpretation of the results.

Configuration	s range	S range	Step-size	Error ratios (EB/ST) MSSE	MMAE
C1 (2270)	[20,60]	[40,100]	1	0.7013	0.7434
C2 (585)	[20,60]	[40,100]	2	0.7094	0.7565
C3 (102)	[20,60]	[40,100]	5	0.7289	0.7688
C4 (29)	[20,60]	[40,100]	10	0.7857	0.8136
C5 (9)	[20,60]	[40,100]	20	0.9946	1.0213
C6 (6)	[20,60]	[40,100]	30	1.1260	1.0858
C7 (165)	[40,50]	[40,60]	1	0.4969	0.5116
C8 (45)	[40,50]	[40,60]	2	0.5438	0.5605
C9 (9)	[40,50]	[40,60]	5	0.6845	0.6946
C10 (3)	[40,50]	[40,60]	10	0.8952	0.9489
C11 (231)	[40,50]	[60,80]	1	0.5890	0.5936
C12 (66)	[40,50]	[60,80]	2	0.6335	0.6167
C13 (15)	[40,50]	[60,80]	5	0.7109	0.7348
C14 (6)	[40,50]	[60,80]	10	0.8662	0.9067
C15 (375)	[60,80]	[70,90]	1	0.6443	0.6506
C16 (100)	[60,80]	[70,90]	2	0.7442	0.7984
C17 (19)	[60,80]	[70,90]	5	0.9073	0.9437
C18 (6)	[60,80]	[70,90]	10	1.0851	1.0211
C19 (231)	[20,30]	[80,100]	1	0.4256	0.5031
C20 (66)	[20,30]	[80,100]	2	0.5394	0.5221
C21 (15)	[20,30]	[80,100]	5	0.6718	0.6583
C22 (6)	[20,30]	[80,100]	10	0.8077	0.8359
C23 (165)	[80,90]	[80,100]	1	0.7665	0.8012
C24 (45)	[80,90]	[80,100]	2	0.7707	0.8214
C25 (9)	[80,90]	[80,100]	5	0.8043	0.8645
C26 (3)	[80,90]	[80,100]	10	1.1258	1.0995
C27 (9)	[40,50] & [60,70]	[60,80]	10	1.0906	1.0876
C28 (14)	[30,40] & [80,90]	[40,60] & [80,100]	10	1.0793	1.0846
C29 (9)	[20,30] & [80,90]	[80,100]	10	1.1788	1.1904
C30 (11)	[40,50] & [70,80]	[80,100]	10	1.0347	1.0431

Table 3.4: Results of error ratios for configurations with sample size 5.

Configuration	s range	S range	Step-size	Error ratios (EB/ST) MSSE	MMAE
C1 (2270)	[20,60]	[40,100]	1	0.8290	0.8512
C2 (585)	[20,60]	[40,100]	2	0.8344	0.8688
C3 (102)	[20,60]	[40,100]	5	0.8484	0.8991
C4 (29)	[20,60]	[40,100]	10	0.8709	0.9183
C5 (9)	[20,60]	[40,100]	20	0.9566	0.9710
C6 (6)	[20,60]	[40,100]	30	1.0608	1.0470
C7 (165)	[40,50]	[40,60]	1	0.5742	0.5983
C8 (45)	[40,50]	[40,60]	2	0.7270	0.7322
C9 (9)	[40,50]	[40,60]	5	0.8359	0.8489
C10 (3)	[40,50]	[40,60]	10	0.9050	0.9410
C11 (231)	[40,50]	[60,80]	1	0.4930	0.5189
C12 (66)	[40,50]	[60,80]	2	0.5373	0.5515
C13 (15)	[40,50]	[60,80]	5	0.6560	0.6786
C14 (6)	[40,50]	[60,80]	10	0.9518	1.0030
C15 (375)	[60,80]	[70,90]	1	0.6244	0.6379
C16 (100)	[60,80]	[70,90]	2	0.8751	0.9128
C17 (19)	[60,80]	[70,90]	5	0.9490	0.9694
C18 (6)	[60,80]	[70,90]	10	1.0972	1.0554
C19 (231)	[20,30]	[80,100]	1	0.5763	0.6459
C20 (66)	[20,30]	[80,100]	2	0.6544	0.6340
C21 (15)	[20,30]	[80,100]	5	0.7658	0.7594
C22 (6)	[20,30]	[80,100]	10	0.8445	0.8298
C23 (165)	[80,90]	[80,100]	1	0.8548	0.8979
C24 (45)	[80,90]	[80,100]	2	0.8940	0.9296
C25 (9)	[80,90]	[80,100]	5	0.9373	0.9777
C26 (3)	[80,90]	[80,100]	10	1.0951	1.0741
C27 (9)	[40,50] & [60,70]	[60,80]	10	1.0685	1.0572
C28 (14)	[30,40] & [80,90]	[40,60] & [80,100]	10	1.0466	1.0565
C29 (9)	[20,30] & [80,90]	[80,100]	10	1.1441	1.1578
C30 (11)	[40,50] & [70,80]	[80,100]	10	1.0184	1.0238

Table 3.5: Results of error ratios for configurations with sample size 10.

Configuration	s range	S range	Step-size	Error ratios (EB/ST) MSSE	MMAE
C1 (2270)	[20,60]	[40,100]	1	0.9155	0.9368
C2 (585)	[20,60]	[40,100]	2	0.9190	0.9444
C3 (102)	[20,60]	[40,100]	5	0.9273	0.9797
C4 (29)	[20,60]	[40,100]	10	0.9402	0.9617
C5 (9)	[20,60]	[40,100]	20	0.9817	0.9937
C6 (6)	[20,60]	[40,100]	30	1.0122	0.9903
C7 (165)	[40,50]	[40,60]	1	0.6582	0.6687
C8 (45)	[40,50]	[40,60]	2	0.7396	0.7468
C9 (9)	[40,50]	[40,60]	5	0.8002	0.8152
C10 (3)	[40,50]	[40,60]	10	0.9558	0.9872
C11 (231)	[40,50]	[60,80]	1	0.7201	0.7384
C12 (66)	[40,50]	[60,80]	2	0.7654	0.7771
C13 (15)	[40,50]	[60,80]	5	0.8185	0.8230
C14 (6)	[40,50]	[60,80]	10	0.9067	0.9427
C15 (375)	[60,80]	[70,90]	1	0.7864	0.8138
C16 (100)	[60,80]	[70,90]	2	0.8670	0.9055
C17 (19)	[60,80]	[70,90]	5	0.9632	0.9848
C18 (6)	[60,80]	[70,90]	10	0.9986	0.9673
C19 (231)	[20,30]	[80,100]	1	0.6023	0.6485
C20 (66)	[20,30]	[80,100]	2	0.6935	0.6846
C21 (15)	[20,30]	[80,100]	5	0.7850	0.7711
C22 (6)	[20,30]	[80,100]	10	0.8969	0.8770
C23 (165)	[80,90]	[80,100]	1	0.9221	0.9508
C24 (45)	[80,90]	[80,100]	2	0.9241	0.9436
C25 (9)	[80,90]	[80,100]	5	0.9446	0.9890
C26 (3)	[80,90]	[80,100]	10	1.0663	1.0521
C27 (9)	[40,50] & [60,70]	[60,80]	10	1.0502	1.0466
C28 (14)	[30,40] & [80,90]	[40,60] & [80,100]	10	1.0222	1.0176
C29 (9)	[20,30] & [80,90]	[80,100]	10	1.1133	1.1259
C30 (11)	[40,50] & [70,80]	[80,100]	10	0.9997	1.0048

Table 3.6: Results of error ratios for configurations with sample size 20.

Speaking broadly, a number of general comments can be made concerning the results. To begin, many configurations yield one or both error ratios below 1, demonstrating that gains in statistical efficiency are indeed possible through the application of EB to DES experimentation. Others, however, yield neutral, mixed, or inferior performance relative to the frequentist approach. Classifying error ratios of close to 1 (i.e. between 0.99 and 1.01) as neutral, and those below or above as positive or negative respectively, enables us to get a clearer overall picture. On this basis, and regardless of sample size, close to 70% of configurations show improved performance for EB on both measures. Looking at the range of observed ratios, the differences can be substantial, from MSSE 0.4256 and MMAE 0.5031 for Configuration 19 with sample size 5, to MSSE 1.1133 and MMAE 1.1259 for Configuration 29 with sample size 20.

A consistent pattern emerges when comparing results across sample sizes. Taking Configuration 19 again as an example, we see the error ratios are MSSE 0.4256 and MMAE 0.5031 at sample size 5, MSSE 0.5763 and MMAE 0.6459 at sample size 10, and MSSE 0.6023 and MMAE 0.6485 at sample size 20. That is, in this case, the smaller the sample size, the more clearly observed the gains in statistical efficiency with EB. Looking across the results, a similar trend is clearly visible, with error ratios decreasing relatively consistently with decreasing sample size. This behaviour is consistent with the convergence of EB and frequentist estimates as sample size increases, as noted in Subsection 2.3.2.2, and is intuitively reasonable given as the smaller the sample size, the larger the potential impact of the additional strength leveraged through the pooling of data.

It is also clear that the structural characteristics of a configuration have a clear effect on the relative performance. Comparing groups of configurations separated by horizontal lines, we observe that EB performs increasingly poorly as configurations become sparser: that is, when step-sizes are large and the number of scenarios is small. This is to be expected for two reasons. Firstly, assuming a continuous response surface, scenarios spaced further apart will be less similar and so may introduce a greater bias when pooled in an EB analysis. Secondly, all else being equal, a smaller number of model scenarios from which to pool data decreases the relative information available and leads to less pronounced EB gains. Indeed, a poor relative performance for EB may be clearly observed in Configurations 6, 18

and 26, each with a low number of model scenarios and a high step-size. A similar effect is evident for non-contiguous configurations C27 to C30, where we observe a poor relative performance for EB, with both error ratios above 1, regardless of sample size. This observation also makes intuitive sense, with non-contiguous areas of the design space introducing unhelpful bias to the analysis when pooled together.

Finally, we examine the relationships between the MSSE and MMAE ratios themselves. Here, regardless of sample size, we see that the value of the MSSE ratio tends to be lower than the value of the MMAE ratio. Considering the nature of these error ratios, with the MSSE measuring the sum of squared error, and MMAE measuring the maximum absolute error, this pattern is unsurprising. We might naturally expect an error ratio taking into account the error across all model scenarios to be more moderate than one taking into account only the maximum error from the most outlying model scenario. This effect has been noted in the EB literature, such as for example by Efron [76]. However, in our results, exceptions occur, such as for example in Configurations 6 and 26 (across sample sizes), indicating that the relationship between error measures is related in a non-trivial manner to the underlying response surface, and is not entirely predictable a priori.

3.6 Conclusions

Whilst the preceding discussion established the potential for gains with the application of EB to DES model experimentation (Research Objective R2(a)), it also demonstrated the possible risks for DES practitioners (Research Objective R2(b)) and highlighted a need for further practitioner support and guidance (Research Objective R2(c)).

The classification of results as positive, neutral or negative in terms of relative EB performance revealed that close to 70% of configurations exhibited better EB performance, regardless of sample size. This, however, also meant that a significant proportion of configuration and sample size pairings resulted in a neutral or poor relative performance for EB. As discussed, the numerical values of the results ranged from around 0.43 to 1.13, clearly highlighting the potential upside and downside of the application of EB to DES model experimentation in terms of

statistical efficiency.

In addition to this overall pattern, several consistent structural features emerged from the numerical study. First, there was a clear relationship between sample size and relative efficiency: EB tended to provide the greatest gains when sample sizes were small, with its advantage diminishing as the amount of data increased. This aligns with theory and reflects the greater influence of information pooling when data within each scenario are relatively few. Secondly, the design structure of the experimental configuration played a significant role. Configurations featuring many, closely spaced model scenarios generally yielded stronger EB gains. Conversely, sparse configurations, with large step-sizes and few scenarios, tended to produce weaker results, and in some cases a frequentist approach proved superior. The poorest relative performance occurred for non-contiguous configurations, where EB pooling introduces substantial bias by drawing together observations from only distantly related model scenarios. These structural dependencies highlight that EB performance is not arbitrary but rather is clearly impacted by characteristics of the experimental environment.

A further observation concerned the behaviour of the two error measures. Whilst MSSE tended to produce lower ratios than MMAE, reflecting its averaging nature, this was not universal. There were notable cases in which MMAE and MSSE behaved differently from expectation, and these divergences indicate that the relationship between the error measures depends on the underlying response surface and cannot always be anticipated from theory alone.

Collectively, these findings represent the first systematic evidence that EB can deliver substantial efficiency gains in DES experimentation, but only under particular structural conditions. The numerical study therefore suggests a coherent set of principles governing EB performance in DES: its dependence on sample size, its sensitivity to structural similarity across scenarios, and the divergent behaviour of different error metrics.

The empirical results and their partial alignment with theoretical expectations highlight the need for clear practitioner guidance. It would be helpful to identify, in advance, the configurations and contexts in which EB is likely to deliver

improvements in statistical efficiency, as well as those in which it may not. The variation observed even within a single DES model illustrates that no simple rule of thumb will suffice; rather, systematic methods are required to predict when EB will be beneficial. Moreover, there remains a need to develop robust EB approaches that manage and contain the maximum error incurred across contexts. Finally, the divergences observed between certain theoretical expectations and empirical outcomes highlight the importance of rigorous empirical testing of any proposed methodological contributions. These themes motivate the work undertaken in Chapters 4 and 5.

Chapter 4

A Priori Prediction of EB Advantage

In this chapter, we turn our attention to the design and development of tools and techniques facilitating the practical application of EB to DES model experimentation. Whilst the numerical study of Chapter 3 demonstrated that EB can offer material gains in statistical efficiency relative to a frequentist approach, it also illustrated that these gains were not guaranteed. Several potential obstacles were exposed, highlighting a clear need for practitioner guidance on these fronts.

Thus, whilst Chapter 3 provides proof of concept, it also establishes the notion that some degree of operationalisation is first required before EB can be used reliably in DES experimentation. This thesis makes two explicit contributions towards this goal: the first is presented in the current chapter, and the second in the following chapter.

Chapter 3 highlighted two key issues for practitioners considering EB. The first concerns whether EB yields a reduction in the overall sum of squared error across scenarios (as measured by the MSSE ratio). The second concerns whether such gains can be realised without inducing a large maximum absolute error for any individual scenario (as measured by the MMAE ratio). This chapter addresses the first of these issues by developing a method for determining, in advance, whether a given DES experimental data set is likely to benefit from EB.

The work corresponds primarily to Research Objective R3, which seeks to provide principled guidance on EB suitability for DES practitioners. Theoretical results identifying when EB will offer advantage are limited in the broader literature, not because the question lacks importance, but because of its inherent difficulty [76]. Determining whether a collection of data populations forms a sufficiently coherent family to justify EB pooling remains a challenge dominated by subjective judgement [76]. The chapter also contributes to Research Objective R5 through the development of a Monte Carlo training and evaluation environment used in the construction and assessment of the method.

To address this gap, the chapter develops and evaluates a data-driven decision support tool that predicts when EB is likely to yield a reduction in the sum of squared error. The tool is constructed using a large, systematically varied training set and models how structural features of experimental data relate to the relative performance of EB and frequentist estimators. Its purpose is to provide DES practitioners with practical, empirically grounded guidance on whether EB is likely to be advantageous for their experimental data.

The remainder of the chapter proceeds as follows. Section 4.1 introduces the conceptual foundations of predictive suitability assessment. Section 4.2 sets out the empirical investigations underpinning the development of the decision support tool. Section 4.3 outlines its construction and training, while Section 4.4 evaluates its performance on an independent test set. Section 4.5 concludes with a discussion of the main findings and their implications for the chapters that follow.

4.1 Motivation and Conceptual Foundations

To begin the work of this chapter, we start by discussing the data characteristics likely to prove instructive concerning whether or not an EB analysis will offer more efficient inference as compared with a frequentist analysis.

Broadly speaking, the data analytic situations in which the EB approach enjoys the greatest relative advantage are those in which data from a large number of relatively "similar" populations[1] are pooled together, whilst simultaneously, a sig-

[1]What is meant by "similar" populations? Population homogeneity can be defined in formal

nificant degree of uncertainty exists about these individual populations. This holds whether the uncertainty is aleatory (i.e. from large population variances), or epistemic (i.e. from small sample sizes), in nature.

This insight follows from the very nature of the EB approach, and can be appreciated through consideration of the following decomposition of the mean squared error (MSE):

$$\text{MSE} = \text{variance} + (\text{bias})^2 \tag{4.1}$$

as presented in [46]. As noted in Carlin and Louis [46], the MSE is a useful and widely applied approach to statistical error measurement. Here, the mean squared error has been expressed as the sum of the variance and squared bias of the estimator in question. Frequentist approaches tend to focus on estimators with zero bias, in particular, favouring the use of minimum variance unbiased estimators (MVUEs). EB methods, and shrinkage methods in general, however, aim to increase overall efficiency by reducing variance (pooling effectively increases available data), at the cost of an increase in bias (pooling may introduce irrelevant data). As highlighted by Carlin and Louis [46], the essential point is that a balance must be struck between a reduction in variance and an increase in bias, in favour of the overall reduction in error. It seems clear that such a balance must depend critically upon the inherent structure and composition of the data set in question.

The bias–variance decomposition of Equation (4.1) helps us better understand the previous characterisation of an "ideal" data analytic situation for the application of EB. With few differences between a large number of pooled populations, there is low potential for the introduction of unhelpful bias, and a large effective increase in available data. With a high degree of uncertainty about each population, there is significant scope to reduce variance and so reduce the overall error in estimation. In such cases, it is straightforward to see how we can gain more reliable inference about the individual populations by pooling the data across all populations collectively. This leveraging effect is termed "borrowing strength" and constitutes a core advantage of the EB approach [45].

---

terms, but here it is sufficient to interpret it intuitively, taking it as similarity of moments and distribution shape between populations. Population homogeneity will be explicitly measured and taken account of in estimation in Chapter 5.

Whilst these observations help us understand the conditions under which EB should succeed in principle, determining when such conditions actually hold in practice is far from straightforward. As discussed previously in the thesis, results concerning when and where an EB approach is likely to offer advantage are lacking in the literature.[2] Such omissions are not because of a lack of importance, but because of the substantial technical difficulties involved in obtaining them [76]. Indeed, this issue is deemed to be a key obstacle, preventing more widespread use of EB methods in practice more generally, not only in DES settings. Due to the complexities involved, the question of whether or not a given set of data populations constitutes a family, that is, whether they bear a significant enough degree of relevance to one another to be considered together, is one which is still dominated by subjective judgement in practice [76].

In light of the great difficulty involved in obtaining a theoretical answer, and the potential value of such a result, we instead adopt a heuristic approach. Our aim is the development of theoretically motivated, empirically calibrated guidelines for the practitioner interested in applying EB in DES experimentation.

To begin to create such guidance, it is necessary to explore these how these ideas translate empirically. We begin with two small-scale empirical investigations: the first examining the roles of population count and sample size, and the second quantifying population similarity and uncertainty. Together, these studies motivate the development of a decision support tool for assessing EB suitability, the core contribution of this chapter, which is constructed in Section 4.3 and evaluated in Section 4.4.

4.2 Exploratory Empirical Studies

Having established the key data characteristics that appear to govern whether EB may outperform a frequentist approach, we now turn to empirically examine these ideas. The purpose of this section is diagnostic rather than confirmatory: to explore, in controlled numerical settings, how the components identified in Section 4.1 manifest in finite-sample performance across controlled Monte Carlo settings.

[2] Personal communication with Bradley Efron.

The first study investigates the roles of population count and sample size in isolation, while the second examines the influence of between population similarity and within population uncertainty, operationalised through an ANOVA-based decomposition. Together, these investigations provide the empirical basis required for the decision support tool developed in Section 4.3.

4.2.1 Empirical Investigation 1: Effects of Population Count and Sample Size

This first empirical investigation examines how population count and sample size influence the relative performance of EB. We begin by outlining the conceptual motivation before describing the experimental design, presenting the results, and offering an interpretation of the patterns observed.

4.2.1.1 Conceptual Motivation

At the beginning of Section 4.1, the following characterisation of an 'ideal' EB data analytic situation was provided:

> Broadly speaking, the data analytic situations in which the EB approach enjoys the greatest relative advantage are those in which data from a large number of relatively "similar" populations are pooled together, whilst simultaneously, a significant degree of uncertainty exists about these individual populations. This holds whether the uncertainty is aleatory (i.e. from large population variances), or epistemic (i.e. from small sample sizes), in nature.

In this subsection, we focus on exploring the consistency of the EB advantage with varying number of populations and size of samples, through both discussion and numerical experimentation.

In considering the impact of differing numbers of pooled populations, it seems reasonable to assume the impact to be highly dependent on the "similarity" (or lack thereof) of the pooled populations. Returning to the variance-bias decomposition of Equation (4.1), EB offers a reduction in the overall MSE through trading

reduced variance for increased bias. When populations are sufficiently similar, increasing the number of pooled populations yields a substantial effective increase in relevant data with little risk of introducing bias. When populations are dissimilar, however, pooling introduces increasing amounts of irrelevant information, leading to substantial bias. Thus, while a larger number of populations can in principle strengthen the EB advantage, the extent of this advantage will depend critically on the underlying population similarity. For this reason, our experimental design deliberately includes settings with both similar and dissimilar populations.

In considering the impact of differing sample sizes, we note that larger samples generally lead to more accurate inference for both EB and frequentist estimators. Indeed, as sample size grows, the two approaches tend to converge, reducing any potential EB advantage. In such cases, the additional complexity of EB may not be warranted. Conversely, for smaller sample sizes, differences in statistical efficiency become more apparent. Pooling data across similar populations increases the effective amount of information available, making the EB approach particularly attractive when individual sample sizes are small.

Having set out these informal hypotheses, the following subsection presents the experimental design through which they are examined empirically.

4.2.1.2 Experimental Design

To examine these expectations, a small numerical study was conducted. The purpose of the study is to assess the relative performance of the DS EB point estimator and standard sample mean (as presented in Chapter 3), as the number of pooled populations and sample size are varied. Consistent with the justification provided in Chapter 1, Monte Carlo simulation was adopted to generate data in a computationally efficient and model-agnostic manner, enabling a clear examination of the effects of k and n on relative performance.

The study comprises four sub-designs, summarised in Table 4.1. Each sub-design spans the same ranges for k (number of populations) and n (sample size), with $k \in [5,100]$ and $n \in [3,50]$, evenly partitioned on a logarithmic scale so that every (k, n) combination forms part of a full factorial design. For each combination,

n observations are sampled from each of k normal populations, with the population means equally spaced across the interval shown in Column 4 of the table, and population standard deviations equally spaced across the interval in Column 5.

Sub-design	k	n	Mean Range	SD Range
1	[5,100]	[3,50]	[0,5]	[1,3]
2	[5,100]	[3,50]	[0,10]	2
3	[5,100]	[3,50]	[0,15]	[1,2]
4	[5,100]	[3,50]	[0,20]	1

Table 4.1: Experimental design for the numerical investigation of k and n impact on the relative EB performance.

The sub-designs differ systematically in their population structures. Moving from Sub-design 1 to Sub-design 4, the mean interval widens, implying decreasing similarity between populations. At the same time, the standard deviation interval becomes narrower and lower in magnitude, implying decreasing underlying population uncertainty. Taken together, these differences create a spectrum of experimental conditions ranging from those highly favourable to EB (Sub-design 1) to those in which EB would be expected to perform poorly (Sub-design 4). This progression allows us to evaluate the informal hypotheses developed in the preceding subsection under increasingly challenging conditions.

All experimentation in this numerical study (and throughout the remainder of this chapter) is implemented in MATLAB (2014a), with sample code provided in Appendix B.

Before discussing the results, we briefly outline the approach to error measurement used in the study. As noted earlier, our primary interest is whether EB yields a reduction in the total squared error across populations. Accordingly, we use Equation (3.22) to compute:

$$EB_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2, \quad Freq_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{Freq,i} - \theta_i\right)^2,$$

where $\hat{\theta}_{EB,i}$ is the EB estimate of the i^{th} population mean, $\hat{\theta}_{Freq,i}$ is the frequentist

estimate of the i^{th} population mean, and θ_i is the true i^{th} population mean. This process is repeated 100 times for each (k, n) combination, and the results averaged to remove the issue of stochastic variability.

The relative performance of the two approaches is then summarised by the ratio:

$$EB_{error}/Freq_{error}\,, \tag{4.2}$$

with a value less than 1 indicating a superior result for EB.

4.2.1.3 Results

The results in terms of error ratio (4.2) for each sub-design are presented in the four scatter plots Figure 4.1 to Figure 4.4. In each of these figures, the blue circles pinpoint combinations of k and n where the EB estimator outperforms the frequentist estimator, whilst the red triangles denote combinations of k and n where the reverse is true, that is where the frequentist estimator outperforms the EB estimator.

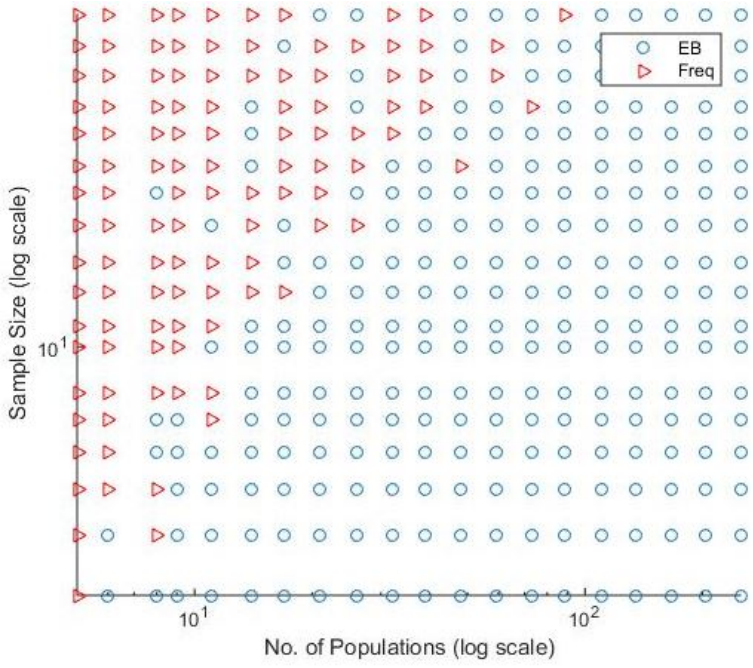

Figure 4.1: Sub-design 1, with mean in range [0,5], and standard deviation in range [1,3].

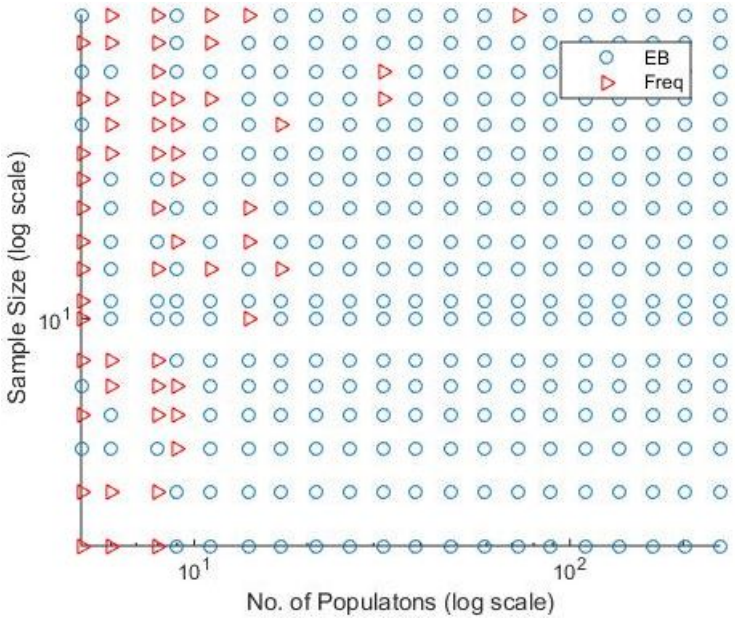

Figure 4.2: Sub-design 2, with mean in range [0,10], and standard deviation 2.

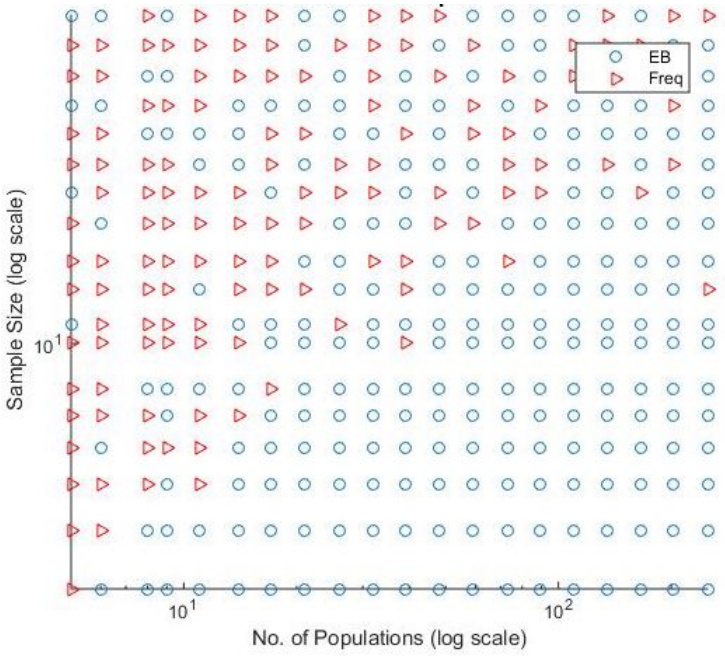

Figure 4.3: Sub-design 3, with mean in range [0,15], and standard deviation in range [1,2].

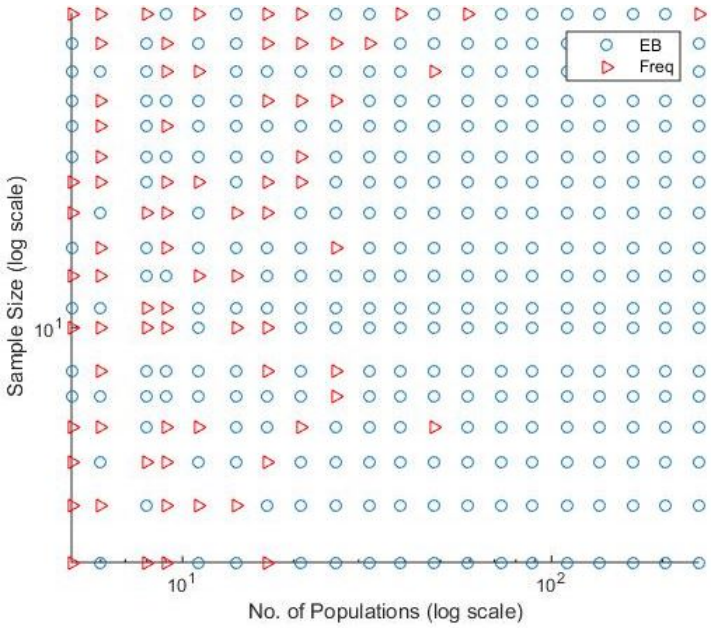

Figure 4.4: Sub-design 4, with mean in range [0,20], and standard deviation 1.

4.2.1.4 Interpretations and Implications

Several observations are revealed through examination of the four scatter plots. Firstly, across all of the figures, it seems clear that the blue points tend to be more prominent towards the lower right corners. This supports our proposition that the EB approach is favoured in data analytic situations featuring many samples of smaller size. The occurrence of red points, corresponding to frequentist advantage, appears to be most concentrated towards the upper left corners. However, even in these regions, the results are somewhat mixed, with many blue points scattered throughout. This observation is consistent with expectation, given the convergence of the approaches for larger size samples.

Secondly, in comparing the figures, we can see variation in the strength of this general pattern across sub-designs. The pattern is noticeably clearer for Sub-designs 1 and 2 than for Sub-designs 3 and 4, indicating a more consistent EB advantage under the experimental conditions represented by the first two sub-designs (that is, narrower mean ranges and larger underlying population variances, and thus

higher population similarity and uncertainty). By contrast, Sub-designs 3 and 4, with increasingly wide mean intervals and more constrained standard deviations, present conditions under which EB would be expected to perform less favourably. These results demonstrate a more consistent EB performance in the experimental conditions of Sub-designs 1 and 2, as opposed to those of Sub-designs 3 and 4. Furthermore, none of the scatter plots display a conclusively clear delineation of the boundary between the two sets of points.

Taken together, these findings suggest that k and n are indeed informative for predicting relative EB performance, and they should feature in the development of any decision support tool to guide practical use. At the same time, their explanatory power is clearly limited: they do not on their own yield a reliable decision rule. This motivates the need to consider additional data characteristics, most notably population similarity and uncertainty, which form the focus of the next empirical investigation.

4.2.2 Empirical Investigation 2: Quantifying Population Similarity and Uncertainty Using the RV Statistic

This second empirical investigation focuses on quantifying between population similarity and within population uncertainty through the proposed predictive statistic (later formalised as the RV statistic), and assessing its usefulness as a predictor of EB advantage. Following the structure adopted in the preceding empirical investigation, we begin with the conceptual foundations, outline the experimental design, then present the results and their interpretation.

4.2.2.1 Conceptual Motivation

Before proceeding, we return to our characterisation of an 'ideal' EB data analytic situation (introduced in Section 4.1):

> Broadly speaking, the data analytic situations in which the EB approach enjoys the greatest relative advantage are those in which data from a large number of relatively "similar" populations are pooled together, whilst simultaneously, a significant degree of uncertainty exists about these individual populations. This holds whether the uncer-

tainty is aleatory (i.e. from large population variances), or epistemic (i.e. from small sample sizes), in nature.

In the preceding subsection, the number of populations and the sample size were examined as drivers of EB advantage. Those results suggested that while k and n are informative, they are insufficient on their own to provide a reliable decision rule. In this subsection, we therefore turn to the remaining components of the ideal EB setting: between population similarity and within population uncertainty, and investigate how these characteristics influence relative EB performance.

A useful framework for formalising these ideas is provided by the classical analysis of variance (ANOVA) decomposition [164]. In this framework, the total variation present in the data is partitioned into two components: (i) the sum of squared deviations within populations (SSE), and (ii) the sum of squared deviations between population means (SSA). Written explicitly:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}\right)^2}_{SST}=\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}_{SSE}+\underbrace{\sum_{i=1}^{k}n\left(\bar{x}_i-\bar{x}\right)^2}_{SSA}, \tag{4.3}$$

where SST captures total variability, SSE quantifies variability within each population, and SSA reflects structural differences between populations.

The ANOVA decomposition is traditionally used to test for differences between population means, under the assumptions of normally distributed populations with equal variances and independent random samples [164]. In its conventional use, the central question is whether the between population component (SSA) is sufficiently large, relative to the within population component (SSE), to reject the null hypothesis of equal means.

In the present context, we do not use ANOVA as a hypothesis testing device. Rather, the decomposition provides a convenient way to quantify two key data characteristics of interest. The component SSA reflects structural differences between population means, and therefore serves as an indicator of population similarity: lower SSA corresponds to more similar populations. The component SSE captures variability within populations, and may be viewed as a data-based mea-

sure of aleatory uncertainty, with a higher SSE reflecting higher within population variation. All else equal, EB is expected to perform well when SSA is small and SSE is large.

These insights motivate the idea that the ratio SSE/SSA may serve as a useful quantitative indicator of the potential for EB to deliver improved estimation. Building on this insight, and informed by the findings of the previous empirical investigation, we propose an adapted version of this ratio, termed the Ratio of Variation (RV) statistic, defined as:

$$RV = \left(\frac{k}{n}\right)\frac{SSE}{SSA} = \left(\frac{k}{n}\right)\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_i\right)^2}{\sum_{i=1}^{k} n\left(\bar{x}_i-\bar{x}\right)^2}. \tag{4.4}$$

The normalising factor k/n reflects the empirical findings from the preceding subsection, where both quantities were shown to influence relative EB performance. We note that relative EB performance will be evaluated using the same error ratio as in the preceding subsection; a reminder of its definition is provided in the next subsection.

Taken together, these considerations lead to a clear set of informal expectations:

- EB-favourable data sets, characterised by low across population variation (low SSA) and high within population variation (high SSE), should yield large RV values, and error ratios below 1.
- Frequentist-favourable data sets, characterised by high across population variation (high SSA) and low within population variation (low SSE), should yield small RV values, and error ratios above 1.

Having set out these expectations, the following subsection presents the experimental design through which they are explored empirically.

4.2.2.2 Experimental Design

To assess whether the RV statistic (4.4) is predictive of the relative performance of the EB approach, a numerical study was conducted. The EB and frequentist estimators are exactly those defined in Chapter 3 and used in the preceding empirical

investigation. As before, Monte Carlo simulation is adopted, allowing controlled exploration of how between population similarity and within population uncertainty interact with EB performance.

The experimental design is summarised in Table 4.2. It comprises four distinct sub-designs, each defined by different fixed values of k (number of populations) and n (sample size). For example, Sub-design 2 sets $k = 50$ and $n = 5$. Each sub-design also specifies a set of allowable mean intervals and standard deviation intervals; within each interval, multiple partitioning schemes are used. Thus, for a given mean interval, the number of mean levels corresponds to the number of equally spaced partition levels within that interval; the same interpretation applies to the SD intervals. All allowable combinations of mean interval, mean-level count, SD interval, and SD-level count form part of the full factorial design for that sub-design.

Sub-design	1	2
k	100	50
n	3	5
μ Intervals	$[0,5],[0,10],[0,15],$ $[0,20],[0,25]$	$[0,5],[0,10],[0,15],$ $[0,20],[0,25]$
Number of μ Levels	$2, 50, 100$	$2, 25, 50$
σ Intervals	$1,[1,\sqrt{2}],$ $[1,2],[1,3]$	$1,[1,\sqrt{2}],$ $[1,2],[1,3]$
Number of σ Levels	$2, 50, 100$	$2, 25, 50$
Sub-design	**3**	**4**
k	10	$3, 4$
n	20	5
μ Intervals	$[0,5],[0,10],[0,15],$ $[0,20],[0,25]$	$[0,75],[0,90],[0,100],$ $[0,150],[0,200]$
Number of μ Levels	$2, 5, 10$	$2, 4$
σ Intervals	$1,[1,\sqrt{2}],$ $[1,2],[1,3]$	$[1,10],[1,15]$ $[1,25]$
Number of σ Levels	$2, 5, 10$	$2, 4$

Table 4.2: Experimental design for the numerical investigation of the usefulness of the RV statistic in predicting EB advantage.

The different sub-designs included in the study's experimental design are purpose-

fully varied, developed to provide a range of differing experimental settings in which to explore the predictive power of the RV statistic. The settings range from those highly favourable for EB, to those highly unfavourable for EB, as follows:

- Sub-design 1 ($k = 100$ and $n = 3$): almost perfectly suited to EB, with a very high number of populations, a very low sample size, and closely spaced population means and standard deviations (high population similarity).
- Sub-design 2 ($k = 50$ and $n = 5$): well suited to EB, with a moderately high number of populations, a moderately low sample size, and relatively closely spaced population means and standard deviations (relatively high population similarity)
- Sub-design 3 ($k = 10$ and $n = 20$): broadly neutral, with a moderate number of populations, a relatively large sample size and moderate population similarity, indicating likely convergence in performance between the two approaches
- Sub-design 4 ($k = 3, 4$ and $n = 10$): very poorly suited to EB, with low numbers of populations, and very low population similarity (both population means and standard deviations widely spaced)

The error measurement approach is identical to that employed in Subsection 4.2.1. We again calculate the total squared error across populations using Equation (3.22):

$$EB_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2, \quad Freq_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{Freq,i} - \theta_i\right)^2,$$

where $\hat{\theta}_{EB,i}$ is the EB estimate of the i^{th} population mean, $\hat{\theta}_{Freq,i}$ is the frequentist estimate of the i^{th} population mean, and θ_i is the true i^{th} population mean. This process is repeated 100 times and the results averaged to remove the issue of stochastic variability.

Relative performance is again summarised using the error ratio:

$$EB_{error}/Freq_{error}\,, \tag{4.5}$$

with values less than 1 indicating a superior performance by EB.

4.2.2.3 Results

The study results are presented in Figures 4.5 to 4.8, corresponding to Sub-designs 1 to 4, respectively. Each scatter plot displays, for every experimental parameter combination (that is, for each combination of k, n, mean interval, number of mean levels, SD interval, and number of SD levels), the observed mean value of the error ratio (4.5) plotted against the observed mean value of the RV statistic (4.4).

Several general patterns emerge from these results, most notably the contrast between Sub-designs 1–2 and Sub-designs 3–4. These patterns are examined in detail in the following subsection.

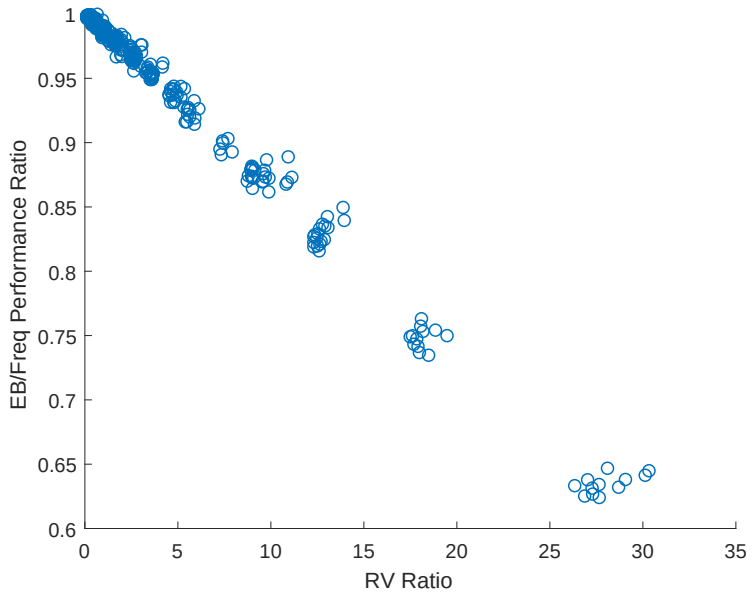

Figure 4.5: Sub-design 1 - RV values.

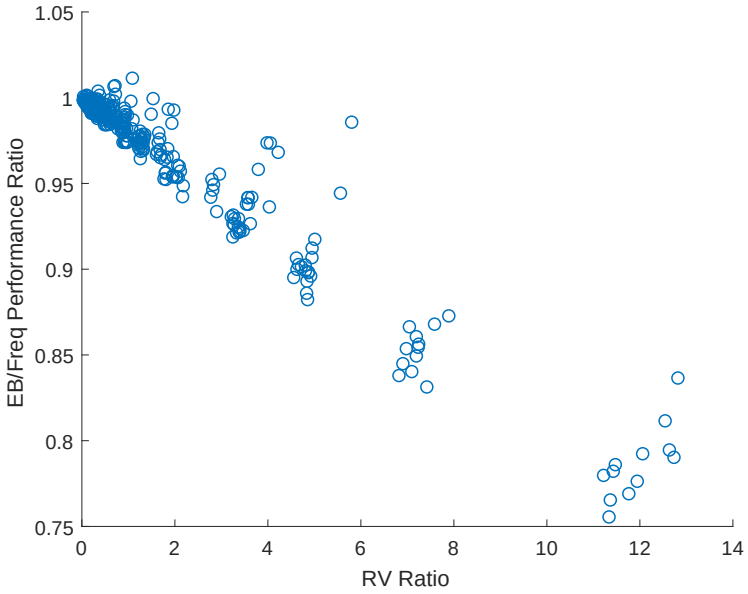

Figure 4.6: Sub-design 2 - RV values.

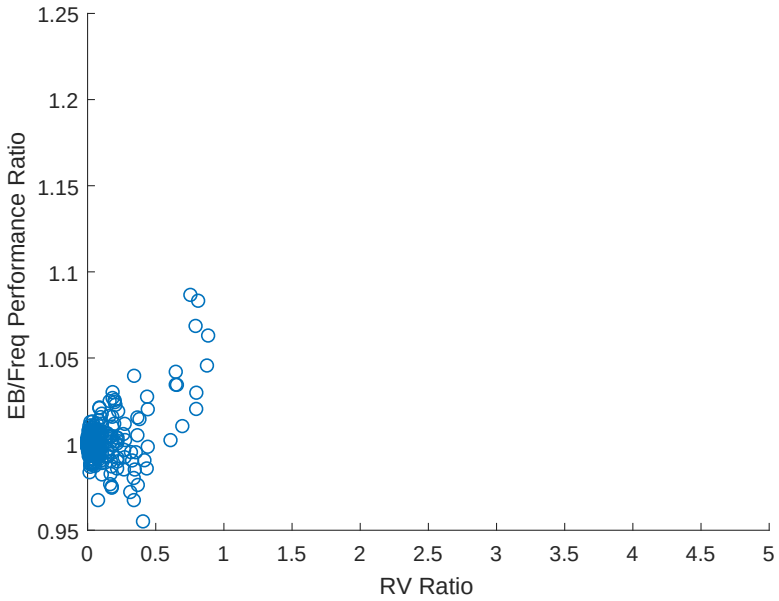

Figure 4.7: Sub-design 3 - RV values.

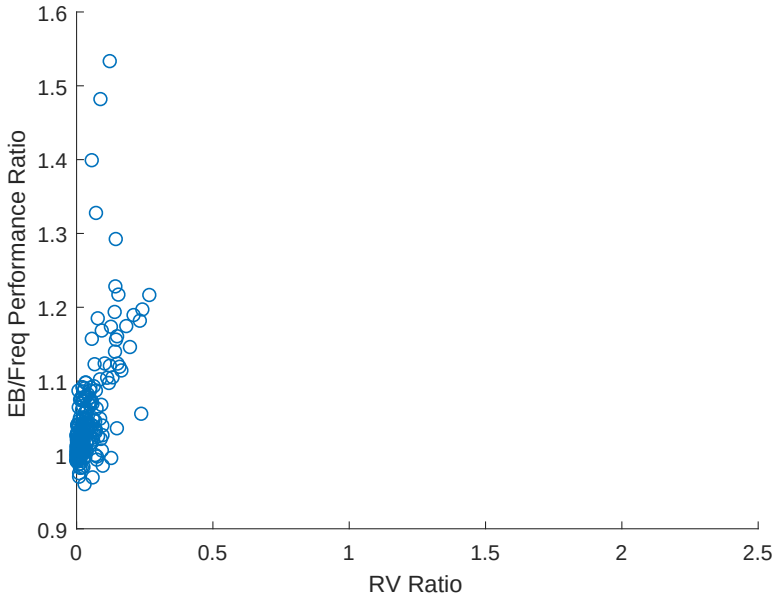

Figure 4.8: Sub-design 4 - RV values.

4.2.2.4 Interpretations and Implications

For the combinations of k (number of population) and n (sample size) that are most likely to favour the EB approach, namely Sub-designs 1 and 2, we observe a clear linear relationship between the RV statistic and the relative performance measure, the error ratio, as shown in Figures 4.5 and 4.6. This relationship is especially pronounced in Sub-design 1, whose experimental conditions strongly favour EB according to the earlier conceptual discussion. In these settings, the strength and consistency of the linear trend suggest that the RV statistic could, in principle, enable reliable prediction of the relative advantage of EB.

Turning to Sub-designs 3 and 4, in Figures 4.7 and 4.8 respectively, we see the expected erosion of EB's advantage. In these cases the RV values are noticeably lower, and in Sub-design 4, where population similarity is especially weak, the most pronounced frequentist superiority occurs at RV values well below 0.5.

Overall, Figures 4.5 to 4.8 indicate that the RV statistic offers a compelling basis for predicting when EB is likely to yield performance gains. Large RV values are

generally associated with EB superiority and display a strong linear association with the observed error ratios. As the RV value falls below 1, the methods become increasingly similar in performance, and at sufficiently small RV values, frequentist superiorirty emerges.

In the next subsection, we reflect on the findings from both empirical investigations and consider their implications for understanding when EB is most likely to offer a performance advantage.

4.2.3 Synthesis of Empirical Findings

Combined, the two empirical investigations of the current subsection reveal a coherent picture of the data characteristics that govern when EB is likely to offer a performance advantage. They also confirm many of the expectations outlined in Section 4.1.

The first study confirmed that, in line with the conceptual discussion of Section 4.1, EB tends to perform well when many populations are pooled and individual sample sizes are small, particularly when population means are relatively similar and underlying variances are relatively large. At the same time, the scattered pattern of results and the absence of a clear demarcation between EB- and frequentist-superior regions indicated that population count and sample size, while informative, are not sufficient on their own to yield a reliable decision rule.

The second investigation extended this analysis by explicitly quantifying between population similarity and within population uncertainty. Using the ANOVA decomposition, these characteristics were captured through the RV statistic, which combines SSE, SSA, and the empirically motivated normalising factor k/n. The resulting numerical study showed a strong and interpretable relationship between the RV statistic and the error ratio, particularly in settings that are a priori favourable to EB, and demonstrated the systematic erosion of EB superiority as RV decreases. As such, the findings of the study also confirmed our expectations concerning population similarity and uncertainty as outlined in Section 4.1

These findings suggest that a decision support tool based on the RV statistic and

possibly other data characteristics could provide practically useful guidance for DES practitioners as to the suitability of EB. The development and formalisation of such a heuristic is the focus of the next section.

4.3 Developing the EB Suitability Decision Heuristic

The smaller-scale numerical studies of the preceding subsections revealed that the RV statistic possesses promising predictive capabilities with respect to the relative statistical efficiency of the EB approach. The purpose of the current section is therefore to investigate the predictive power of the RV statistic more rigorously, and to exploit the relationship between RV and relative EB performance to construct a practically useful decision support tool. Such a tool would accomplish the overarching goal of the present chapter: to provide theoretically motivated and empirically calibrated guidelines concerning the suitability (or lack thereof) of the EB approach, given a particular DES data analytic context.

To achieve this, a more systematic and comprehensive approach is required than that afforded by the smaller exploratory studies. Those investigations were deliberately lightweight, designed to probe underlying mechanisms and to identify components most closely associated with EB advantage. However, they do not by themselves provide a sufficiently robust foundation for a practical decision support tool. The present study therefore expands the scope of investigation in two important respects. First, it employs a far more expansive and detailed experimental testbed, enabling the behaviour of the RV statistic to be examined across a dense and diverse range of population structures and uncertainty conditions. Second, it introduces a formal modelling step, the use of logistic regression, to quantify how the RV statistic and related data characteristics combine to determine whether EB or frequentist estimation is likely to perform better. This formal modelling step therefore transforms the observed data patterns into a predictive, classification-based decision rule, providing consistent and objective guidance for practical use. We refer to this resulting classification rule as the EB Suitability Decision Heuristic (hereafter, the EB suitability heuristic).

In the remainder of this section, the experimental testbed is first outlined in Subsection 4.3.1, followed by a presentation and discussion of preliminary results in Subsection 4.3.2. The study then proceeds, in Subsections 4.3.3 and 4.3.4, to apply a formal classification method, logistic regression, to construct and formalise the EB suitability heuristic.

4.3.1 Construction of the Experimental Testbed

Before presenting the design of the experimental testbed, it is helpful to briefly recall the requirements established at the outset of this section: the testbed must be broader, more varied, and more systematic than those used in the earlier exploratory investigations. The new testbed therefore expands the scope along three key dimensions: it spans a wider range of (k, n) combinations, incorporates more diverse patterns of population similarity and uncertainty, and employs a substantially finer grid of design points. This richer diversity is essential for building the logistic-regression model introduced later, which requires sufficient variation in both independent and dependent variables to estimate stable coefficients and, importantly, to support reliable probability statements about when EB will outperform a traditional frequentist analysis [121].

Turning to the structure of the testbed itself, each experimental data set is designed to reflect the structural form of typical DES experimental setting: it contains data drawn from multiple populations, where each population corresponds conceptually to a DES model scenario, and where each sampled observation represents an individual output that, in a DES setting, would arise from a single run of that model scenario.[3]

Before specifying the formal parameters and their values used in the experimental testbed, it is helpful to outline the structural considerations that guide its construction. The following list of factors provides an intuitive way to think about the kinds of variation a DES-like experimental environment should be able to express:

[3]Here, we note that the comparison between DES and the MC-based experimental testbed relates to structural aspects, and not necessarily distributional aspects. As discussed elsewhere in the thesis, it is recognised that DES models involve complexities not captured by Monte Carlo simulation of normal populations, e.g. feedback loops, threshold behaviour, non-normal and heavy-tailed distributions. A fuller discussion of these differences is provided in both Chapter 6 and Chapter 7.

- A range of different numbers of populations
- A range of different sample sizes
- Different degrees of variation in the population means, in terms of both interval spanned and step-size involved
- Different degrees of variation in the population standard deviations, in terms of both interval spanned and step-size involved
- The relationship between means and standard deviations, i.e. varying together or separately

In order to illustrate how altering these characteristics can produce a wide variety of different experimental data sets, we present and discuss some examples in Figures 4.9 to 4.12. Here we note that the first two examples, in Figures 4.9 and 4.10 depict experimental data sets with four populations, whilst the second two examples in Figures 4.11 and 4.12 depict data sets with twenty populations. Together they demonstrate how differing patterns of mean structure, variability, and mean—variance relationships give rise to markedly different experimental environments, highlighting the breadth built into the full experimental testbed.

First, Figure 4.9, as discussed, provides an example incorporating four normal populations. The population means are advanced in uniform steps between 0 and 10 (i.e. 0, 3.333, 6.667, 10). The standard deviations of the populations range between $\sqrt{2}$ and 1 in two levels, i.e. the first two populations have a standard deviation of $\sqrt{2}$, and the latter two a standard deviation of 1. Here, we note that whilst the mean is increasing, the standard deviation is decreasing, that is they vary separately.

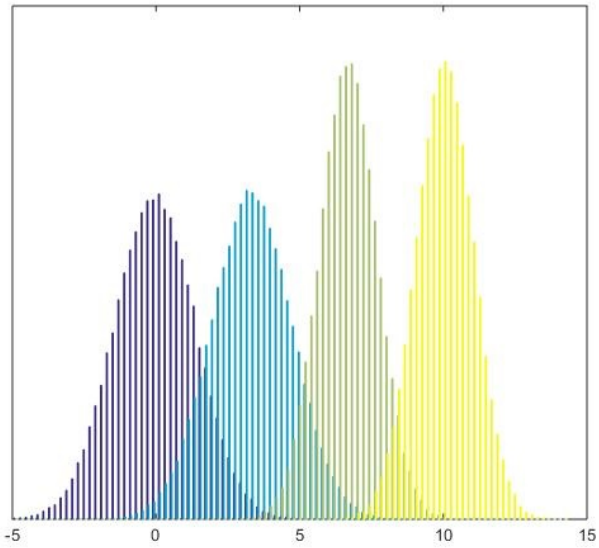

Figure 4.9: Example with 4 populations, mean increasing uniformly between 0 and 10, standard deviation increasing in two levels $\sqrt{2}$ and 1.

Next, in Figure 4.10, we present an example again featuring four normal populations, with their means identical to the previous example (that is, increasing in uniform steps between 0 and 10) but with standard deviations ranging from 1 to $\sqrt{2}$ in four uniform steps (i.e. 1, 1.138, 1.276, $\sqrt{2}$). Here, we note that contrary to the previous example, in this case, whilst the mean is increasing, the standard deviation is also increasing, that is they vary together.

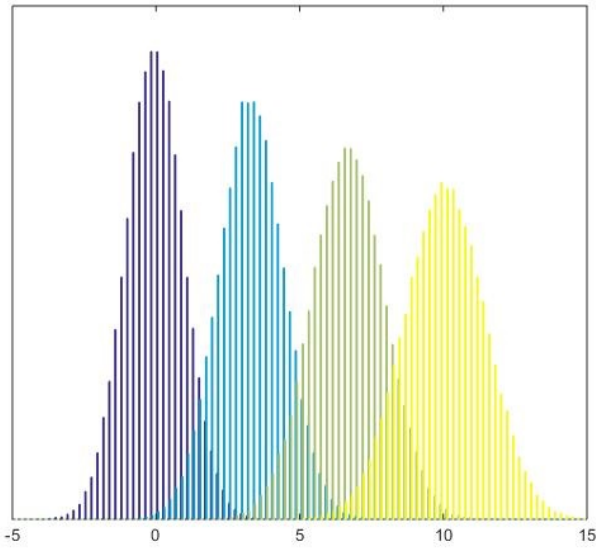

Figure 4.10: Example with 4 populations, mean increasing uniformly between 0 and 10, standard deviation decreasing in uniform steps $\sqrt{2}$ to 1.

In the final two examples, we examine data analytic situations with a far greater number of populations, twenty in each case. In Figure 4.11, the twenty population means advance in uniform steps through the interval [0,15]. The population standard deviations increase from 1 to 2 in ten uniform steps. As such, in this case the means and standard deviations are varying together. Finally, in Figure 4.12, we can see that the twenty population means are each either 0 or 10, with an equal number of populations distributed at each level. The population standard deviations decrease from 1 to 2, but this time in twenty uniform steps. In this case, the means and standard deviations are varying separately.

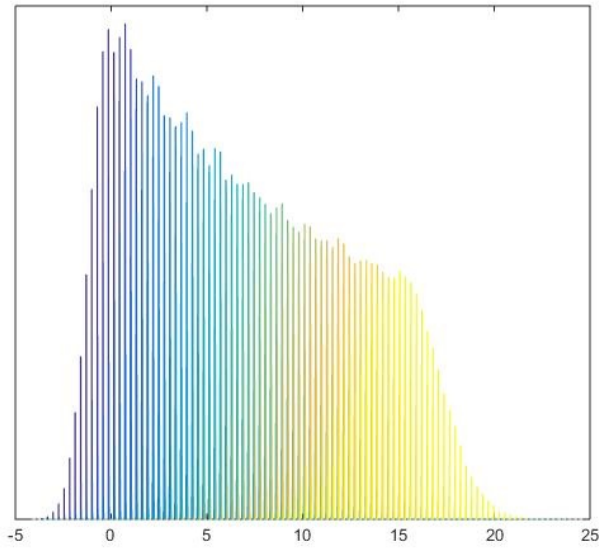

Figure 4.11: Example with 20 populations, mean increasing uniformly between 0 and 15, standard deviation increasing in uniform steps 1 to 2.

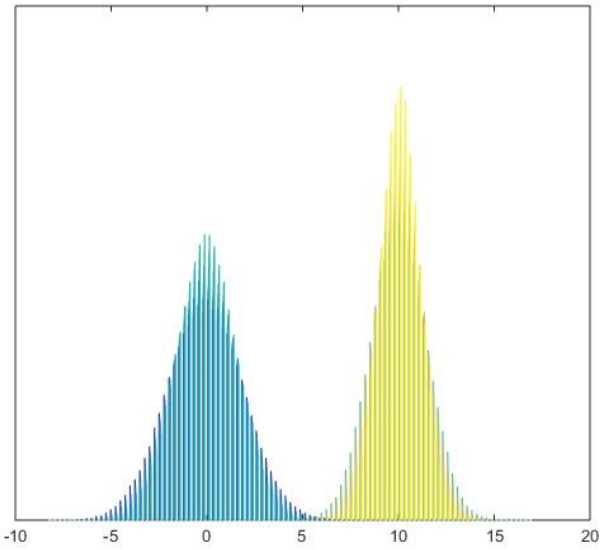

Figure 4.12: Example with 20 populations, mean in two levels 0 and 10, standard deviation decreasing in uniform steps 1 to 2.

The construction of the experimental testbed is of critical importance in ensuring that the results obtained are of value in real application. Without a diverse and comprehensive range of experimental data sets upon which to build and test our EB suitability heuristic, there can be little confidence in the results obtained. Even from the examination of this relatively small selection of visualisations, the diversity amongst the different data sets that can be obtained through varying these different characteristics is clear. Together, these characteristics capture the principal dimensions along which EB performance may vary, and they directly motivate the formal parameterisation adopted for the testbed.

As such, the experimental testbed will be constructed based on the parameters and values outlined in Table 4.3 below.

Population Count k	$5, 10, 20, 50, 200$
Sample Size n	$3, 5, 10, 20$
Range of μ Values	$[0, 5], [0, 10], [0, 15], [0, 20]$
Number of μ Levels	$2, k/2, k$
Range of σ Values	$1, [1, \sqrt{2}], [1, 2]$
Number of σ Levels	$2, k/2, k$
μ and σ Variation	increasing, decreasing

Table 4.3: Experimental design for the creation of the experimental testbed.

4.3.2 Study Data and Preliminary Observations

Having described and discussed the construction of the experimental testbed, in this section we discuss the generation of the study data and present some preliminary insights yielded by it. The data generated will be used to build the classification model presented in the following subsections, and so allow us to develop the EB suitability heuristic to guide the practitioner seeking to apply EB to DES experimentation.

As discussed in the previous subsection, the experimental testbed is constructed based on the parameters and their levels as outlined in Table 4.3. Taking all combinations of the parameter levels in Table 4.3 yields a total of $4,320$ *experimental*

parameter combinations. For each experimental parameter combination excluding sample size, a *large-scale experimental data set* comprising $2,000$ observations for each population is generated. This large-scale data set defines the full underlying data environment for that experimental parameter combination, and from it we construct the *practitioner-level experimental data sets* used in the analysis. Specifically, each large-scale data set permits the construction of 100 independent replications of the practitioner-level data sets for the largest sample size ($n = 20$), with practitioner-level data sets for smaller sample sizes obtained by extraction from this larger set. This structure provides, for each practitioner-level data set, the quantities required for evaluating EB performance, namely, pairs of EB and frequentist estimates, and the associated error ratio and RV value. We now describe explicitly how these quantities are computed.

As in our previous numerical studies, the different practitioner-level data sets were used to create pairs of EB and frequentist estimates of the population means for each experimental parameter combination. Again, we are interested in the sum of squared error across populations, and so again we make use of Equation (3.22) to calculate the following quantities:

$$EB_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2, \quad Freq_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{Freq,i} - \theta_i\right)^2.$$

These computations are performed for all 100 practitioner-level data sets, and the resulting values are averaged to mitigate stochastic variability. Again, we use ratio:

$$EB_{error}/Freq_{error}\,, \tag{4.6}$$

to evaluate the relative performance of the approaches, with a value less than 1 indicating a superior result for EB.

As a final step in the generation of this study's data, the mean value of the RV statistic, using Equation (4.4), was calculated over the same 100 practitioner-level data sets. This enables us to generate a scatter plot containing the observed pairings of the mean RV values against the mean error ratios, given by Equation (4.6), for each experimental parameter combination in the experimental testbed, as presented in Figure 4.13.

In Figure 4.13, each blue circle represents a pairing of the mean RV value and the mean error ratio associated with a given experimental parameter combination. The RV values are plotted on the vertical axis and error ratios plotted on the horizontal axis. By the construction of the error ratio of Equation (4.6), values less than 1 suggest EB-superior combinations, whilst those greater than 1 suggest frequentist-superior combinations. When the plot is viewed with interpretation in mind, a clear pattern emerges: EB-superior combinations, for which the error ratio < 1, are typically associated with large RV values, while frequentist-superior combinations, for which the error ratio > 1, tend to exhibit RV values close to zero.

Additionally, the spectra of lines extending upward and to the left from the horizontal value 1 mirror the approximately linear relationship between the RV statistic and the error ratio observed in the earlier exploratory studies (e.g. Figure 4.5).

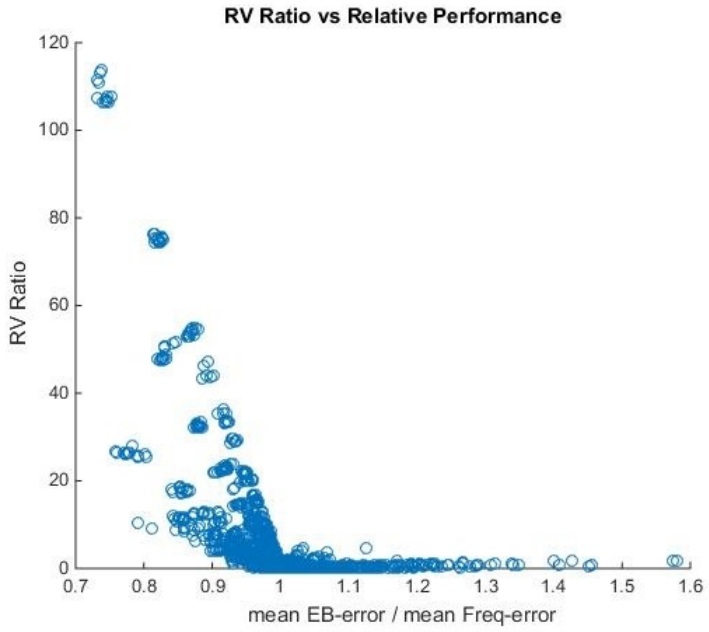

Figure 4.13: Scatter plot of mean RV statistic vs. mean error ratio.

The scatter plot of Figure 4.13 suggests the possibility of a simple threshold-based heuristic for choosing between EB and frequentist approaches in the analysis of

DES experimental data. One might imagine selecting an RV cut-off such that EB is recommended whenever the observed RV exceeds this threshold, and the frequentist estimator otherwise. However, any fixed threshold will inevitably misclassify some experimental data sets. Misclassifications may occur for combinations with a low RV value but an error ratio less than 1 (EB-superior yet misclassified as frequentist-superior), or for combinations with a high RV value but an error ratio greater than 1 (frequentist-superior yet misclassified as EB-superior). These misclassifications correspond directly to type I and type II errors in standard statistical hypothesis testing terminology.

While misclassification cannot be avoided entirely, these limitations motivate the adoption of a systematic approach that allows the balance between type I and type II errors to be managed in a principled way. As such, in the next subsection we introduce a formal classification model that operationalises this balance and forms the basis of the EB suitability heuristic.

4.3.3 Modelling Framework for Classification

Logistic regression is a generalised linear modelling approach used in data analytic situations featuring a binary dependent variable [121]. The model predicts the probabilities associated with the binary dependent variable taking either of the two possible values, based in turn on the values taken by a set of relevant independent variables. Whilst the particular values of the dependent variable vary by context (e.g. diseased/healthy, spam/not spam), they are frequently re-coded as 0/1. Frequently, the goal of logistic regression modelling is the classification of new items into one of the two possible categories, based on their predicted probabilities [121].

In the present study, logistic regression provides a natural way to develop a rule for determining whether a given DES experimental data set is EB-suitable or EB-unsuitable. The binary dependent variable to be modelled corresponds directly to this classification, and the relevant independent variables arise from structural characteristics of the practitioner-level data sets (e.g. RV value, number of populations, sample size) contained within the testbed. Logistic regression is well suited to this task because it produces probability statements rather than hard classifications, allowing the cut-off used to determine EB suitability to be adjusted in line

with the study's type I and type II error considerations.

Multiple similarities exist between linear and logistic regression. For example, the method of maximum likelihood is commonly used in both cases to fit the model coefficients. However, in the case of logistic regression, owing to the modelling of probability as opposed to a continuous dependent variable, the range of the model must be constrained to lie strictly between 0 and 1. This requires the use of the so-called logit transformation, where it is the log of the odds ratio (as opposed to the dependent variable itself) that is modelled by a function of the independent variables. As such, logistic regression involves the fitting of a hypothesis function, based on the logistic function, as follows:

$$h(x|\hat{\beta}) = \frac{1}{1 + e^{-g(x|\hat{\beta})}}, \tag{4.7}$$

where $g(x|\hat{\beta})$ is a function, usually a polynomial, of the independent variables x, dependent on the model coefficients $\hat{\beta}$ [77].

Two distinct data sets are required in a logistic classification framework: a training set, used to estimate the model coefficients, and a test set, used to assess predictive performance. In the present study, the training data are derived from the experimental testbed introduced in Subsection 4.3.1. Each experimental parameter combination's associated set of practitioner-level data sets is used to compute the summary statistics needed for model fitting (e.g. mean RV statistic and mean error ratio), and these processed values form the basis of the *modelling observations* used to train the logistic model. The separate test set, constructed with a similar structure but introduced in Subsection 4.4, is then used to evaluate out-of-sample accuracy.

In Subsection 4.3.4, as is standard in logistic regression modelling, the coefficients $\hat{\beta}$ will be estimated using the maximum likelihood method so that the hypothesis function $h(x|\hat{\beta})$ of Equation (4.7) achieves a good predictive fit on the training data. The resulting model will then be evaluated on the test set in Subsection 4.4, using standard measures such as overall accuracy, sensitivity, and specificity.

Before turning to the construction of the logistic model for the present study, it is

helpful to recall the behaviour of a simple logistic regression model. Figure 4.14 illustrates such a model with just one independent variable x. The blue dots scattered along the two horizontal lines (at 0 and 1) represent the labelled training set, each point representing either a "success" (a 1) or a "failure" (a 0). The s-shaped curve between the horizontal lines indicates the fitted logistic hypothesis function. This function may be interpreted as an approximation of the probability of a success for a given value of the independent variable. Predictive classifications on new items may be made by evaluating the hypothesis function at the points of interest, for the given values of x, and interpreting the output accordingly.

A simple and commonly used approach is to categorise any probability > 0.5 as a success, and any probability < 0.5 as a failure. While 0.5 is a natural default, any value between 0 and 1 may in principle be used. As discussed at the end of the previous subsection, selecting an appropriate cut-off requires a trade-off between type I and type II errors. That trade-off is central to the present study, and will be made explicit through the logistic modelling that follows.

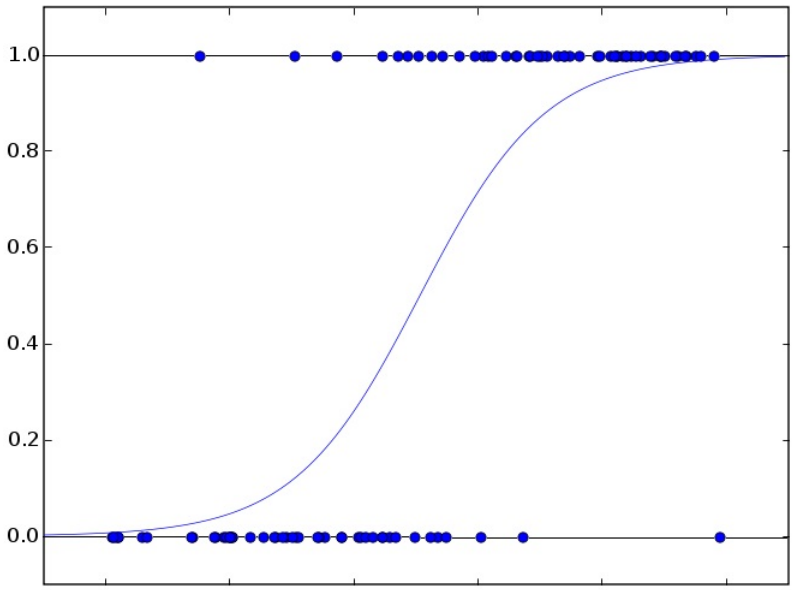

Figure 4.14: Typical logistic function fitted to data.

Having now introduced the classification method to be applied, we now turn to the construction of the logistic regression model used to determine EB suitability. This model forms the core of the EB suitability heuristic, whose development and subsequent evaluation are presented in the next two subsections.

4.3.4 Construction of the EB Suitability Heuristic

In this section, we apply the logistic regression framework introduced in Subsection 4.3.3 to the study data derived from the experimental testbed described in Subsection 4.3.2, in order to construct the decision heuristic that forms the core contribution of this chapter. Recall that for each of the $4,230$ experimental parameter combinations in the testbed, we have computed a mean RV value, the associated values of k and n, and a mean error ratio summarising EB versus frequentist performance. The aim is to translate the patterns observed in this testbed into a formal, data-driven predictive classification rule which can be used to determine whether a given DES experimental data set is EB-suitable or not.

The formulation of the EB suitability heuristic requires a number of components to be specified. First, a training parameter α must be defined to convert the continuous error ratio into a binary classification label. Second, a set of independent variables must be chosen (based on our preliminary empirical investigations) to capture the key structural characteristics of each DES experimental data set. Third, a classification cut-off γ must be introduced to translate predicted probabilities into EB-suitable or EB-unsuitable decisions. Finally, a systematic procedure must be adopted to choose the values of these tuning parameters (that is, α and γ) and to fit the logistic regression model. Each of these elements is described in turn below.

We begin by introducing the training parameter α which is used to construct the binary dependent variable required for logistic regression. As discussed in Subsection 4.3.2, for each experimental parameter combination in the testbed we have already computed a mean error ratio, averaged over 100 practitioner-level data sets. This quantity summarises the relative performance of the EB and frequentist estimators across that underlying experimental environment.

To convert this continuous error ratio into a label, we proceed as follows. For a given choice of α:

- if the mean error ratio is $< \alpha$, the corresponding experimental parameter combination is deemed EB-suitable, labelled a "success" and coded as 1;
- if the mean error ratio is $\geq \alpha$, the corresponding experimental parameter combination is deemed EB-unsuitable, labelled a "failure" and coded as 0.

In this way, α acts as a training parameter that encodes the level of performance advantage required for EB to be regarded as "suitable." Whilst this is straightforward, a key question remains regarding the selection of an appropriate value for α.

Earlier discussions in this chapter, and indeed in Chapter 3, have implicitly adopted $\alpha = 1$ as a natural benchmark. This choice has a clear interpretation: it precisely divides the space of mean error ratios into those combinations where EB is empirically superior (error ratio < 1) and those where the frequentist approach is superior (error ratio > 1). From a purely statistical perspective, $\alpha = 1$ therefore appears an obvious, and in some sense canonical, choice.

However, there are several reasons to consider α values below 1. While Zhao's DS EB estimator was deliberately chosen for its relative simplicity, it is nonetheless more complex than the standard sample mean. In practice, this increased complexity manifests as greater computational cost and a higher comprehension burden for DES practitioners. It is therefore reasonable to "penalise" EB slightly for this additional overhead, requiring it to demonstrate a more substantial gain in statistical efficiency before it is recommended.

In addition, there is a natural conservatism in many DES practice settings. Practitioners are often loyal to existing, well-understood methods, and may be reluctant to move away from familiar frequentist procedures. In such contexts, the consequences of incorrectly recommending EB (i.e. incorrectly labelling a frequentist-superior setting as EB-suitable, a type I error in this classification context) may be deemed more problematic than failing to recommend EB when it would have been marginally better (a type II error). Allowing α to take values below 1 therefore provides a way to encode this asymmetry in preferences: by setting a more stringent requirement for EB superiority, it becomes harder for a combination to

be labelled as EB-suitable, and any such recommendation is correspondingly more robust.

For these reasons, rather than fixing α a priori, we treat it as a tuning parameter whose optimal value is to be determined empirically. The next step is to specify the independent variables upon which the logistic model will be based.

The independent variables used in the logistic regression model must capture the principal data characteristics identified in Sections 4.1 and 4.2 as governing EB performance. These discussions and exploratory investigations highlighted the importance of:

- the RV statistic, summarising the proportions of within- and between-population variation, adjusted for the number of populations and sample size;
- the population count k;
- the sample size n.

To reflect these findings, and following the usual conventions of logistic regression, we define the following transformed variables:

$$x_0 = 1, \quad x_1 = log(\bar{RV}), \quad x_2 = log(k), \quad x_3 = log(n). \tag{4.8}$$

Here, $\bar{RV}$ denotes the mean value of the RV statistic of Equation (4.4) computed over 100 practitioner-level data sets for a given experimental parameter combination, while k and n are the population count and sample size, respectively.[4]

The inclusion of $x_0 = 1$ is standard practice in regression modelling, allowing for the estimation of an intercept term. The use of logarithmic transformations for $\bar{RV}$, k, and n is motivated by both conceptual and empirical considerations. In the mean estimation context, it is the relative differences in the number of populations and sample size that are most influential, rather than their absolute magnitudes, making a log scale more appropriate. Similarly, the $\bar{RV}$ values are often densely clustered near zero; transforming them via the logarithm spreads these values,

[4] $\bar{RV}$ is used only during model training as a stabilised estimate of the RV statistic. In a practical application of the EB suitability heuristic, practitioners should compute RV from their own data rather than using an averaged value.

providing better separation between experimental parameter combinations and, in turn, a clearer signal for the logistic model to use.

Taken together, the observed values of (x_0, x_1, x_2, x_3) for each experimental parameter combination, along with its label determined by α, form the modelling observations that comprise the training set and allow us to fit the logistic regression model.

Once a logistic regression model has been fitted, its output is a predicted probability that a given experimental parameter combination is EB-suitable. To turn these probabilities into a practical decision rule, we must introduce a classification cut-off, denoted by γ.

For any such combination, the fitted logistic model yields a predicted probability $h(x|\hat{\beta})$ (see Equation (4.7)). To convert this probability prediction into a binary decision, we adopt the following rule:

- if $h(x|\hat{\beta}) \geq \gamma$, classify the experimental data set as EB-suitable;
- if $h(x|\hat{\beta}) < \gamma$, classify the experimental data set as EB-unsuitable.

The value γ therefore plays an analogous role to α, but at a different stage of the process: α governs the creation of labels from the error ratio values, whereas γ governs the interpretation of the predicted probability values produced by the logistic model.

A natural default choice is $\gamma = 0.5$, which corresponds to classifying a data set as EB-suitable whenever the predicted probability of EB superiority exceeds one half. However, as implied through the discussion in Subsection 4.3.3, different choices of γ alter the balance between the two types of classification error:

- larger values of γ ($\gamma > 0.5$) make EB recommendations less frequent but more reliable;
- smaller values of γ ($\gamma < 0.5$) make EB recommendations more frequent but less reliable.

In the present context, for reasons similar to those discussed in relation to α, there are strong grounds for considering values of γ at or above 0.5. The additional complexity of EB, combined with practitioner conservatism, suggests that false positives (classifying a DES experimental data set as EB-suitable when EB is not clearly superior) are more costly than false negatives. Higher values of γ therefore offer an appealing way to encode this preference within the decision heuristic. As with α, we therefore treat γ as a tuning parameter whose value is to be determined empirically, rather than decided in advance.

To determine appropriate values for the tuning parameters α and γ, a systematic grid search was undertaken. The ranges and step-sizes used are summarised in Table 4.4 below.

Parameter	Interval	Step-size
α	0.6 to 1.0	0.01
γ	0.3 to 0.7	0.01

Table 4.4: Specification of tuning parameters α and γ.

Here, the interval for α (i.e. only values above 0.5) was selected based on the reasoning outlined earlier, namely the slight increase in cost (related to both computation and comprehension) of the EB approach, and the general preference for tried and tested approaches. For γ, a more balanced interval centred around 0.5 was selected to allow the tuning process to explore both more conservative and less conservative cut-offs, without pre-imposing a directional bias. Since the conservatism of the approach is already partially encoded through the α-range, additionally restricting γ to values above 0.5 was deemed unnecessarily restrictive. This interval therefore maintains flexibility while avoiding any strong skew in the model during tuning.

For each (α, γ) pair in the grid, a logistic regression model was fitted using the labels defined by α and the independent variables in Equation (4.8). The fitted model generated predicted probabilities of EB suitability for every experimental parameter combination, which were then converted into classifications using γ as the probability cut-off. These classifications were compared with the "true"

labels determined by the α-based classification of the error ratio, allowing sensitivity (correct identification of EB-suitable combinations) and specificity (correct identification of EB-unsuitable combinations) to be calculated. Each (α, γ) pair therefore yields a distinct classification rule with its own implied balance between sensitivity and specificity.

A convenient way to visualise the performance of the different (α, γ) combinations is through a Receiver Operating Characteristic (ROC) curve, as displayed in Figure 4.15. In this figure, each blue circle represents a particular pairing of α and γ. Its position in the ROC plane is determined by:

- the false positive rate (1 - specificity), plotted on the horizontal axis; and
- the true positive rate (sensitivity), plotted on the vertical axis.

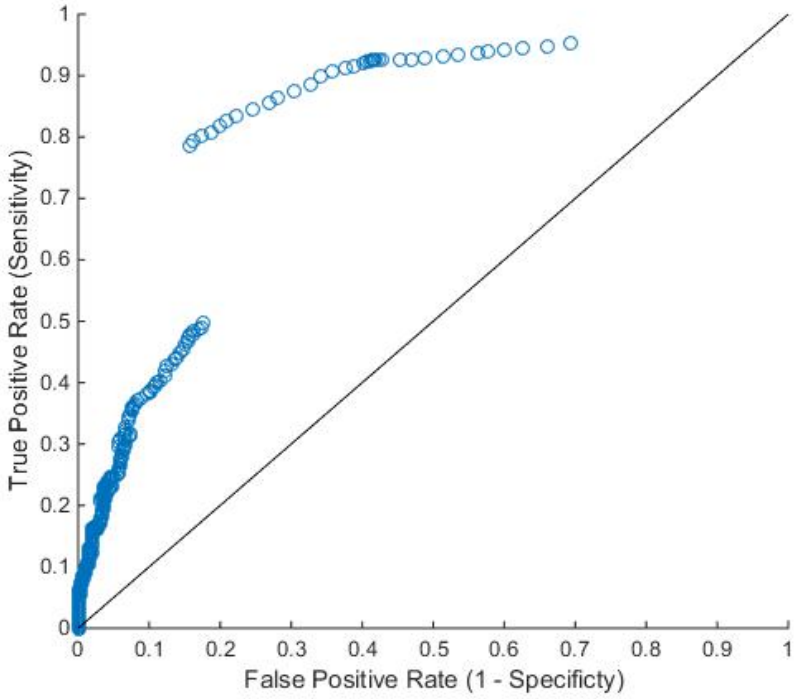

Figure 4.15: ROC curve for various α, γ choices.

The ROC curve thus makes explicit the trade-off between type I and type II errors inherent in any classification procedure. The upper left corner of the ROC plot

corresponds to an ideal classifier, with a true positive rate of 1 and a false positive rate of 0. Points lying closer to this corner represent classification rules that achieve a better overall balance between correctly identifying EB-suitable cases and avoiding incorrect EB recommendations. By contrast, points lying near the diagonal line correspond to classifiers little better than random guessing.

In the current context, the ROC representation allows us to compare all candidate combinations of α and γ in a single framework. A natural selection principle is to choose the (α, γ) pair whose ROC point lies closest to the upper left corner, thereby achieving, for the given data, the best available trade-off between sensitivity and specificity. Applying this approach to the present study led to the selection of $\alpha = 1$ and $\gamma = 0.69$ as the preferred tuning parameter combination. This pairing yields a classification rule that, within the constraints of the experimental testbed, offers a strong and balanced performance in terms of distinguishing EB-suitable from EB-unsuitable.

With α and γ fixed at their selected values, a final logistic regression model was fitted using the independent variables of Equation (4.8) and the corresponding EB-suitability labels. The resulting maximum-likelihood estimate of the coefficient vector was:

$$\hat{\beta} = (-3.3123, 0.2114, 1.6967, 0.2036), \tag{4.9}$$

where the coefficients apply respectively to (x_0, x_1, x_2, x_3) as defined in Equation (4.8). This vector fully characterises the EB suitability heuristic at the modelling level: for any DES experimental data set, the variables RV, k, and n may be computed, transformed, and supplied to the fitted logistic model to obtain a predicted probability of EB suitability. This probability is then converted to a classification using the cut-off $\gamma = 0.69$.

Within the original training environment defined by the experimental testbed, the fitted model achieved an overall accuracy rate of 0.8419, in the sense that this proportion of experimental parameter combinations were correctly classified as EB-suitable or EB-unsuitable according to the $\alpha = 1$ criterion. This represents a strong in-sample performance and provides initial evidence that the fitted logistic

model is able to capture the main structure linking the RV statistic and related data characteristics to relative EB performance.

However, as stressed throughout this chapter, in-sample performance alone is not sufficient to establish the practical usefulness of the EB suitability heuristic. There remains a clear need to assess how well the fitted logistic classification rule generalises to new, previously unseen data. The evaluation of this out-of-sample performance, using the separate test set noted in Subsection 4.3.3, forms the focus of the following subsection.

4.4 Evaluation of the EB Suitability Heuristic

With the EB suitability heuristic now formally constructed, it is now necessary to carry out a robust evaluation of its performance. This evaluation is essential for both statistical and practical reasons. Statistically, in fitting a logistic regression model to a given training set, there is an inherent risk of tuning the model too closely to the unique characteristics of the training set, or "overfitting" it, resulting in a model that does not generalise well to future data sets [112]. Practically, if the EB suitability heuristic is to offer meaningful support to DES practitioners, it must demonstrate stable behaviour across a broad range of data structures, including those that differ materially from the training set, the experimental testbed. A dedicated out-of-sample evaluation is therefore an essential part of establishing whether the tool generalises reliably beyond the conditions under which it was developed.

To conduct this evaluation, a new test set was created under an experimental design that differed in several respects from the training set, the experimental testbed of Section 4.3.1. Whilst the same experimental parameters were retained, namely population count, sample size, mean structure, standard deviation structure, and the joint mean-standard deviation variance pattern, the levels chosen for each were deliberately broadened to generate a more demanding and diverse assessment environment. In particular, wider mean and standard deviation intervals, altered spacing schemes, and larger sample sizes were included. The overall intention was to expose the EB suitability heuristic to significantly more challenging environments, thereby providing a stringent test of its ability to generalise.

The new test set for evaluating the classification rule was generated by means of the experimental design outlined in Table 4.5 below.

Whilst Table 4.5 and Table 4.3 (the experimental testbed design used to generate the training set) are similar in multiple ways, some notable differences are present. For example, the test set design incorporates more extreme mean and standard deviation ranges, e.g. up to $[0, 25]$ and $[1, 3]$ respectively, as compared with $[0, 20]$ and $[1, 2]$, respectively, in the experimental testbed. The levels of k and n were also adjusted to include values such as $k = 30$ and $n = 7$, not present in the earlier design. These adjustments ensure that the EB suitability heuristic is evaluated not merely through interpolations within familiar regions, but also by extrapolation into structurally distinct areas of the design space. In total, the experimental design of Table 4.5 represents a total of 648 different experimental parameter combinations.

Population Count k	$5, 30, 100$
Sample Size n	$3, 7, 25$
Range of μ Values	$[0, 1], [0, 10], [0, 25]$
Number of μ Levels	$2, k$
Range of σ Values	$1, [1, 1.5], [1, 3]$
Number of σ Levels	$2, k$
μ and σ Variation	increasing, decreasing

Table 4.5: Experimental design for creation of test set.

Computationally, the evaluation proceeded via the following steps (and in a similar manner to earlier empirical work conducted within the chapter):

1. For each experimental parameter combination specified in Table 4.5 with the largest sample size of 25, sufficient data was generated to create 100 practitioner-level data sets.

2. For all remaining experimental parameter combinations, 100 practitioner-level data sets were obtained by extraction from the large-scale data set generated for the corresponding $n = 25$ case.

3. For each practitioner-level data set, EB and frequentist estimates of the population means were computed.

4. For each practitioner-level data set, the total squared error across populations was calculated for both estimators using Equation (3.22). These values were then averaged across the 100 replications for each experimental parameter combination to obtain mean errors for each estimator.

5. Using the mean error ratio (4.6), each experimental parameter combination was labelled as EB-superior if the error ratio was less than 1, and as frequentist-superior if it was greater than 1.

6. For each practitioner-level data set, the fitted logistic regression model (with cut-off $\gamma = 0.69$ and coefficient vector $\hat{\beta}$) was applied to produce a predicted label of EB-suitable or EB-unsuitable.

7. The predicted classification from Step 6 was compared with the empirical performance label of its experimental parameter combination from Step 5, yielding a correct or incorrect classification for that practitioner-level data set.

8. Finally, these classification outcomes were first aggregated across the 100 practitioner-level data sets within each experimental parameter combination, and then across all 648 combinations, to produce overall measures of accuracy, sensitivity and specificity for the heuristic on the test set data.

Together, these steps mirror the logic used in the construction of the EB suitability heuristic itself, ensuring that the evaluation examines the tool under conditions directly comparable to, yet deliberately more demanding than, those of the training environment.

The performance of the heuristic on the test set is summarised in Table 4.6. The overall accuracy of 0.8032 is broadly consistent with the corresponding value of 0.8419 obtained on the training set (the experimental testbed), indicating no evident degradation in out-of-sample performance. Further, the sensitivity (0.8055) and specificity (0.7953) measures are closely aligned, indicating that the classification rule does not strongly favour either EB-superior or frequentist-superior

recommendations. Taken together, these results indicate that the heuristic generalises well to new experimental environments drawn from the broader design space.

Performance measure	Result
Overall accuracy	0.8032
Sensitivity	0.8055
Specificity	0.7953

Table 4.6: Summary of out-of-sample performance of the EB suitability heuristic on the test set.

Overall, the test set results indicate that the EB suitability heuristic retains its predictive performance beyond the confines of the experimental testbed, performing reliably across substantially altered distributional and structural settings. This work provides a sound basis for the broader discussion that follows and frames the chapter conclusion.

4.5 Conclusions

The purpose of the current chapter was to examine whether or not it was possible to predict, a priori, the EB suitability of a given DES experimental data set. As such, this chapter primarily addresses Research Objective R3, which concerns the provision of principled guidance on whether EB is likely to offer an advantage in DES experimentation contexts. In particular, the focus was on determining whether EB is likely to yield a reduction in the overall sum of squared error across scenarios. The chapter also contributes to Research Objective R5 through the development of a Monte Carlo training and evaluation environment used to construct and assess the proposed method.

To address this objective, the chapter developed a data-driven decision support tool, the EB Suitability Decision Heuristic, designed to predict whether EB is likely to outperform a frequentist approach in terms of overall squared error. The heuristic is built upon the ANOVA-based RV statistic, which quantifies the balance between within-population uncertainty and between-population similarity, and is formalised through a logistic regression classification model linking these struc-

tural features to relative estimator performance.

The numerical work of the chapter demonstrated that the EB suitability heuristic can predict EB advantage with a high degree of accuracy. Within the experimental testbed used for training, the fitted model achieved an overall classification accuracy of 0.8419. When evaluated on an independent and deliberately more challenging test set, the heuristic retained strong out-of-sample performance, achieving an overall accuracy of 0.8032, with closely aligned sensitivity and specificity. These results provide reassurance that the heuristic generalises beyond the specific conditions under which it was developed, at least within the Monte Carlo environments considered.

Beyond these numerical results, the chapter provides insight into how EB suitability may be assessed objectively in DES experimentation. By replacing subjective judgement about whether populations form a sufficiently coherent family with an empirically calibrated prediction rule, the heuristic represents a substantive step towards the operationalisation of EB in practice. Since the principal advantage of EB lies in its ability to reduce overall squared error through variance reduction, the heuristic offers a practical indication of whether an EB analysis is likely to be worthwhile in a given experimental context.

Taken together, these results demonstrate that Research Objectives R3(a) and R3(b) have been substantively addressed, with the EB suitability heuristic providing practical, empirically grounded guidance on when EB is likely to yield a reduction in overall squared error in a DES experimentation context. In addition, the evaluation of the tool within a controlled Monte Carlo environment establishes a systematic basis for its subsequent application in a genuine DES setting, representing an initial step towards Research Objective R3(c).

While the EB suitability heuristic represents a substantive step towards the operationalisation of EB for DES experimentation, it does not, in itself, resolve all of the challenges identified in Chapter 3. In particular, suitability prediction addresses only the question of when EB is likely to be advantageous in terms of overall squared error, and does not mitigate the risk of undesirable behaviour at the level of individual DES model scenarios. The development of robust EB esti-

mation approaches capable of controlling such behaviour, and thereby addressing Research Objective R4, forms the focus of Chapter 5. In addition, the methods developed in this chapter have been evaluated exclusively within controlled Monte Carlo environments. The extent to which the EB suitability heuristic translates to a genuine DES experimentation context therefore remains an open question. This is addressed in Chapter 6, where it is applied to an industrial-scale DES case study, further contributing to Research Objectives R3(c) and R5.

Chapter 5

A Robust Similarity-Weighted EB Estimation Procedure

In this chapter, we continue our efforts towards the design and development of tools and techniques facilitating the practical application of EB to DES model experimentation. The numerical study of Chapter 3 represented our first concrete application of EB to DES, revealing both the potential efficiency gains and the practical challenges associated with EB shrinkage-based inference.

As highlighted previously, two issues are central to determining the practical usefulness of EB estimators in DES experimentation. The first concerns whether EB yields a reduction in the overall sum of squared error across scenarios (as measured by MSSE), a question addressed in Chapter 4 through the development of the EB suitability heuristic. The second concerns whether such gains can be realised without introducing a large maximum absolute error for any individual scenario (as measured by MMAE). It is this second issue that motivates the present chapter.

The work of this chapter corresponds primarily to Research Objective R4, which seeks to understand and mitigate situations in which EB may exhibit undesirable behaviour at the level of individual model scenarios. Building on insights from Chapters 2 and 3, we consider the issue of scenario similarity, the difficulties posed by outlying or structurally atypical scenarios, and the extent to which these challenges can be addressed through methodological adaptations. The chapter also contributes to Research Objective R5 through the development of a Monte

Carlo training and evaluation environment required for its empirical investigations.

The difficulty of managing such outlier behaviour is well recognised in the EB literature, and remains a key concern where the similarity of populations cannot be guaranteed [76]. Given the practical importance of protecting individual scenarios in DES experimentation, particularly when decisions are sensitive to worst-case performance, there is clear value in developing an estimator that retains the efficiency of EB while offering improved robustness.

To this end, the chapter develops and evaluates an adapted EB estimator that incorporates a data-based similarity weighting mechanism designed to moderate shrinkage for potentially atypical scenarios. The aim is to retain the benefits of information sharing while reducing the risk of large individual errors. The adapted EB estimator is grounded in a detailed examination of scenario similarity, an exploration of data-based pooling, and the construction of a weighting scheme derived from Welch's t-test [242, 241].

The remainder of the chapter proceeds as follows. Section 5.1 outlines the motivation and conceptual background for addressing EB robustness to DES scenario heterogeneity. Section 5.2 examines population similarity and conducts a preliminary exploration of data-based pooling. Section 5.3 develops and tests the weighting mechanism that underpins the adapted estimation procedure. Section 5.4 formulates the adapted EB estimator, the core methodological contribution of the chapter. Section 5.5 presents a numerical evaluation of its performance under both standard and deliberately challenging experimental conditions. The chapter concludes in Section 5.6.

5.1 Motivation and Conceptual Foundations

A central theme of Chapter 4 was whether EB could achieve a reduction in overall error, through variance reduction, relative to a frequentist approach. The complementary issue, and the focus of this chapter, concerns the potential increase in bias introduced by shrinkage and the resulting possibility of large individual errors for particular populations.

To frame this discussion, we briefly return to the MSE decomposition [46] presented in Section 4.1:

$$\text{MSE} = \text{variance} + (\text{bias})^2. \tag{5.1}$$

This decomposition makes clear that reductions in variance may be offset by increases in bias, with the latter falling unevenly across populations. In particular, populations whose true means differ markedly from the pooled mean are most vulnerable to large errors.

A helpful illustration of the possible negative impact of bias in implementing an EB analysis comes from Efron [76]. In this experiment, data is simulated across ten normal distributions, with means as shown in Column 2 of Table 5.1. Two estimation approaches were applied to estimate the known means from the generated data, the maximum likelihood estimator (MLE) with results as in Column 3, and the James Stein (JS) estimator with results as in Column 4. The JS estimator [218] is an early EB estimator, its formulation discussed by Efron in [76].

i	μ_i	$MSE_i^{(MLE)}$	$MSE_i^{(JS)}$
1	-0.81	0.95	0.61
2	-0.39	1.04	0.62
3	-0.39	1.03	0.62
4	-0.08	0.99	0.58
5	0.69	1.06	0.67
6	0.71	0.98	0.63
7	1.28	0.95	0.71
8	1.32	1.04	0.77
9	1.89	1.00	0.88
10	4.00	1.08	2.04
	Total Err	10.12	8.13

Table 5.1: Table of simulation results illustrating the negative impact of bias from Efron [76].

Examining Column 2, we see the population means range from -0.81 to 1.89, with the exception of population 10, an outlier with a mean of 4, significantly different from the remainder. Comparing Columns 3 and 4, we see that for most of the populations the JS estimator results in a lower error measure than the MLE; this

is also evidenced by the total error measures in the final row, with a total error of 10.12 for the MLE versus 8.13 for JS. As such, we can clearly see a reduction in the overall squared error with the use of EB.

The key issue, however, is the inflated error associated with the use of the JS estimator ($MSE = 2.04$), as compared with the MLE ($MSE = 1.08$), for population 10.[1] This outcome makes intuitive sense. The JS estimator's shrinkage is helpful in estimating within groups of similar populations, with each sample mean shrunk towards a representative pooled mean, accelerating the convergence of the estimation process. However, it is unhelpful in estimating an outlying population, where we are shrinking the sample mean towards an unrepresentative pooled mean, and thus injecting unwelcome bias into, and slowing the convergence of, the estimation process. This possibility of large individual errors for outlying populations is a classic challenge associated with the application of EB in practice [76].

Given this is a standard challenge associated with practical EB estimation, the question remains regarding how it can reliably be navigated in the context of DES model experimentation. A number of different possible approaches present themselves. Classically, this issue was addressed through the pooling of relatively homogeneous sub-groups of populations, with subjective judgement used to guide pooling decisions [76]. In such an approach, only groups of populations deemed sufficiently "similar" are pooled together for analysis. Whilst this approach can lead to excellent results, it is also, however, tricky to implement in practice, as relevant and reliable information to inform subjective judgement is not always available. Beyond the use of subjective judgement, two technical approaches appear potentially useful: data-based pooling approaches, where data rather than judgement is used to guide pooling decisions, and robust estimation procedures, where the EB procedure is modified in some manner to limit the impact of bias. Both of these approaches will be explored in the remainder of this chapter. Such investigations help develop the overall theme of the thesis, the operationalisation of EB in DES model experimentation.

[1] Efron's experiment uses only the MSE to measure both the overall error and the individual errors across the different populations. In our studies, we have opted to use different error measures to more clearly and conveniently differentiate between the issues found in the practical application of EB, however, both approaches emphasise the same issue, that is, the possibility of inflated error in estimating outlying populations.

5.2 Preliminary Exploration: Population Similarity and Pooling

The discussion in Section 5.1 illustrated that shrinkage-induced bias can be highly problematic when dissimilar populations are pooled together. As such, it is useful to examine whether simple, data-based pooling decisions might offer a practical safeguard against this issue. This section conducts a preliminary exploration of population similarity in DES settings, using Welch's t-test as a quantitative measure of similarity and evaluating a basic pooling scheme informed by these results. As will be seen, this discrete pooling approach yields useful insights but also reveals important limitations, motivating adjusted estimation techniques developed in the remainder of the chapter.

5.2.1 Assessing Population Similarity Using Welch's t-test

In this section, we introduce and discuss Welch's t-test [242, 241] as an approach to the assessment of population similarity. A range of approaches exploring and assessing population similarity are available in the literature, from classical statistical testing [25] to sophisticated, machine learning clustering algorithms [156]. Here, we make use of Welch's test owing to its sufficiently strong empirical performance in practice, its constructive simplicity and transparency, and its widespread use and acceptance in the field of DES model experimentation [154]. Welch's test forms an important component of the methodology of this chapter, playing a role in the preliminary exploration of data-based pooling in Section 5.2.2, and appearing later as part of our broader investigation into robust EB estimation.

To begin, we briefly outline the problem at hand in appropriate mathematical notation. Let us suppose we have data from k different populations or DES model scenarios such that our data set comprises k samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ of sizes $n_1, n_2, \ldots, n_k$. An implicit assumption in the application of EB to such a data analytic situation is that the i^{th} and j^{th} populations are sufficiently similar that the data from one population is to some extent informative regarding the other population. Further, this assumption applies for all pairings of $i, j \in \{1, 2, \ldots, k\}$. In the case of DES model experimentation, k may be large and the underlying model scenarios diverse, so it is risky to take such an assumption for granted. It

is instead advisable to explicitly assess and take account of population similarity in the estimation process.

One method for undertaking the assessment of population similarity is Welch's t-test [242, 241]. Welch's test is a statistical method for testing the hypothesis that two data samples have been generated from underlying populations with equal means. It is an adaptation of the standard two-sample t-test for the difference in means between two normal populations. In particular, and in contrast to the standard test, Welch's test does not require the two populations to be of equal variance, nor does it require equal sample size, two useful generalisations in our current context. It is also useful to note that whilst we still require the assumption of normality, Welch's test is reasonably robust to deviations from normality, more so than the standard Student's t-test, for example. This observation is again pertinent given our current DES model experimentation context.

The testing procedure is as follows. Suppose we have two data samples $\mathbf{X}_1$ and $\mathbf{X}_2$, of size n_1 and n_2 respectively, and from population 1 and 2 respectively. We form the following test statistic t as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{5.2}$$

where $\bar{X}_i$ and s_i^2 are, respectively, the sample mean and sample variance of $\mathbf{X}_i$ for $i \in \{1, 2\}$. The degrees of freedom associated with the test statistic t may be approximated using the Welch–Satterthwaite approximation [206, 242], calculated as follows:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}. \tag{5.3}$$

Typically, the hypotheses adopted in undertaking Welch's test are:

$\mathbf{H_0}$: The means of population 1 and 2 are equal.
$\mathbf{H_1}$: The means of population 1 and 2 are different.

Under the null hypothesis H_0, the test statistic t approximately follows a t-distribution, with degrees of freedom equal to ν, rounded to the nearest integer.

From the test statistic t, we can calculate a p-value giving the probability (assuming the null hypothesis H_0 to be true) of obtaining a value of the test statistic as or more extreme than the one observed. Deviating slightly from the original definition of a p-value, we shall interpret $p_{i,j}$ as a relative measure of the likely similarity of the populations underlying samples $\mathbf{X}_i$ and $\mathbf{X}_j$. This makes intuitive sense for if $p_{i,j}$ is very low, this equates to a low likelihood of populations i and j actually having equal means, and vice versa. In this sense, a low p-value translates to low population similarity, at least in terms of their mean parameters.

Welch's test therefore provides a transparent and empirically reliable means of quantifying similarity between DES model scenarios. In the next subsection, we examine how this measure performs when used to guide a simple pooling rule within a preliminary EB analysis.

5.2.2 Preliminary Pooling Study Based on Welch's t-test

In this subsection, we summarise a preliminary study conducted by Blair et al. [29], with full details available in Appendix D of the thesis. This study examined population, or DES model scenario, similarity and its implications for the application of EB in DES model experimentation. In particular, it explored whether Welch's t-test could offer a practical basis for guiding scenario pooling decisions, by identifying which scenarios appeared sufficiently similar for their data to be combined. We briefly survey its methodology and conclusions below.

The DES test model used in the study was a MATLAB (2014a) implementation of an M/M/1 queuing model, selected for a number of reasons. Firstly, the theoretical results available for the M/M/1 model greatly facilitates the assessment of the relative performance of the estimation approaches. Secondly, the M/M/1 model has been used as a test model in various studies within the DES literature. Finally, the M/M/1 model execution time is modest, enabling sufficiently large-scale experimentation to be conducted. The estimation approaches compared were the DS EB estimator, proposed by Zhao [260] and presented in Equation (3.20), and the standard frequentist estimator, the sample mean, presented in Equation (3.21), both implemented in MATLAB (2014a). Further information on the specifics of the experimental design employed, both strategic and tactical, may be found in

Appendix D.

The statistical assessment using Welch's test enabled decisions regarding whether or not different model scenarios were sufficiently similar to justify pooling. The assessment was undertaken using a traditional hypothesis testing framework, with the null hypothesis describing the situation in which the underlying distributions have the same mean, and the alternative hypothesis describing the alternate situation in which they do not. If the value of the test statistic t led to a p-value below a pre-determined threshold, then the sample pairing was rejected for pooling. In this way, the p-value represented a measure of model scenario similarity, in the manner outlined at the end of Section 5.2.1. Appropriate pools for each model scenario were created by initially including all other model scenarios, and then discounting those deemed 'dissimilar' by means of the above process.

Whilst the use of Welch's test yielded an increased statistical performance for EB over the naive pooling of all DES model scenarios into a single family, a number of difficulties emerged. Firstly, the approach required the subjective determination of a threshold for the interpretation of p-values. In practice, this determination proved problematic, as the appropriate value for the threshold was found, through later experimentation, to be very sensitive to the choice of model and the problem context in question. Unlike the RV statistic used in Chapter 4, which was incorporated into a logistic regression model fitted across a large training set and thereby stabilised through aggregation and calibration, a Welch p-value is a lightweight, pairwise measure with no comparable opportunity for modelling or smoothing. Any threshold therefore operated as a hard, discontinuous decision rule.

Secondly, the discrete nature of the approach was also deemed unsound. The concept that one data sample be judged 'relevant' whilst another, potentially very similar, sample be judged 'irrelevant' solely because they fall on either side of a subjectively determined threshold seemed unsatisfactorily arbitrary in practice. Small changes in the data or in the model structure were often sufficient to reverse the decision, producing unstable and context-dependent pooling.

Taken together, these findings indicate that while Welch's test provides meaningful information about population similarity, a discrete pooling rule is too restrictive for

practical EB use in DES model experimentation. These insights therefore motivate an alternative approach in which the influence of each population is allowed to vary smoothly with its assessed similarity. The remainder of the chapter develops such an approach through a similarity-based weighting mechanism.

5.3 Developing a Similarity-Based Weighting Mechanism

In the following subsections, we develop a similarity-based weighting mechanism for robust EB estimation in DES settings. The aim is to modify the EB approach so that the degree of shrinkage reflects the similarity between populations, thereby limiting the negative impact of bias. As noted in Section 5.2, we draw on Welch's t-test to construct these weights. We begin our exploration by briefly reviewing the original EB estimation procedure, the DS EB estimator proposed by Zhao [260] and presented in Chapter 3.

As seen in Chapter 3, the DS EB estimator of the i^{th} population mean, $\hat{\theta}_{EB,i}$, is given by:

$$\hat{\theta}_{EB,i} = \hat{M}_{EB,i}\bar{X}_i + (1 - \hat{M}_{EB,i})\hat{\mu} \text{ for } i = 1, \ldots, k, \tag{5.4}$$

with shrinkage weight, $\hat{M}_{EB,i}$, given by:

$$\hat{M}_{EB,i} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2_{EB,i}}. \tag{5.5}$$

Equation (5.4) represents a weighted average of sample mean $\bar{X}_i$ and $\hat{\mu}$, an EB approximation of the 'grand mean' hyperparameter μ. The estimator $\hat{\mu}$ represents the shrinkage target of the estimation process, that is, the quantity towards which the individual sample means $\bar{X}_i$ are shrunk. It is constructed from the available data by means of the following weighted average formula:

$$\hat{\mu} = \sum_{j=1}^{k} \frac{\bar{X}_j/\hat{\sigma}^2_{EB,j}}{\sum_{l=1}^{k} 1/\hat{\sigma}^2_{EB,l}}. \tag{5.6}$$

Examining Equation (5.6), we observe that the estimator is based on data from

all populations. Further, the j^{th} sample mean, $\bar{X}_j$, is weighted only by the inverse of the EB estimate of the variance of the j^{th} population, $\hat{\sigma}^2_{EB,j}$. Intuitively, it does makes sense for the contribution of sample mean $\bar{X}_j$ to be weighted by the best available estimate of the uncertainty in sample data $\mathbf{X_j}$, however, it is also clear that no explicit account is taken of the degree of similarity between populations i and j in the estimation process. This point is apparent from jointly considering the formulation of the estimator as given in Equation (5.4), the shrinkage weights as given in Equation (5.5), and the grand mean estimator as given in Equation (5.6).[2]

The practical implications of this design were already evident in the numerical study of Chapter 3, where homogeneous groups were handled well but substantial individual errors arose when populations differed materially. Further, even in situations where most of the underlying populations are similar, it is unlikely that all possible sub-groups of populations are sufficiently similar to avoid large individual errors, as noted in Efron's illustration [76] presented at the beginning of the chapter. As such, it is clear that a naive pooling approach, in which all populations are pooled together, cannot always be justified. Integrating information on population similarity into the estimation process would therefore seem a sensible additional step in operationalising the approach for use in DES practice, and in avoiding the possibility of large individual errors arising from the pooling of dissimilar model scenarios.

In the following subsections, our focus is therefore on developing and examining the similarity-based weighting mechanism that underpins the robust EB estimator introduced later in the chapter. Subsection 5.3.1 sets out the construction of the weights, using Welch's t-test to quantify the relative similarity between populations. Subsection 5.3.2 then provides a visual examination of the resulting weight patterns across a range of representative data sets, allowing us to assess whether the mechanism behaves as expected in practice. These subsections together establish the foundation on which the robust EB estimation approach of Section 5.4 is built.

[2]We refer the interested reader to Chapter 3 for a comprehensive presentation of the development of the DS EB estimator [260].

5.3.1 Construction of Similarity-Based Weights

In this subsection, we develop the weighting mechanism underlying the robust EB estimation approach of Section 5.4. The weighting mechanism is based on normalised p-values resulting from the pairwise application of Welch's t-test (as introduced in Section 5.2.1). In the following subsection, we then use visualisations to explore the behaviour of the weighting mechanism in practice.

To begin, we briefly review the mathematical formulation of our data analytic situation, as outlined in Section 5.2.1. Suppose we have data from k different populations, so that our data set is comprised of k samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ of sizes $n_1, n_2, \ldots, n_k$. As previously discussed, an implicit assumption made in applying EB to such a setting is that the i^{th} and j^{th} populations are sufficiently similar that the data from one population is informative regarding the other, and that this holds for all pairings of $i, j \in \{1, 2, \ldots, k\}$. We questioned the validity of this assumption in DES model experimentation and intend to explicitly take account of population similarity to ensure robust EB estimation approaches for DES practice. We do so through a weighting mechanism derived from Welch's test.

We proceed by temporarily fixing our attention solely on the i^{th} data sample $\mathbf{X}_i$, undertaking pairwise comparisons with each of the remaining samples in turn. Let $p_{i,j}$ denote the p-value arising from the application of Welch's test to the comparison of the i^{th} sample mean $\bar{X}_i$ with the j^{th} sample mean $\bar{X}_j$ for $j = 1, 2, \ldots, k$. Intuitively, and as per previous discussion in Section 5.2.1, the values $p_{i,j}$ provide a relative measure of the similarity between the i^{th} and j^{th} populations.

To normalise the $p_{i,j}$ values so that they sum to k, and thus place them on a common scale for later use, we divide each by their mean, as follows:

$$w_{i,j} = \frac{k p_{i,j}}{\sum_{\ell=1}^{k} p_{i,\ell}}, \quad \text{for } j = 1, 2, \ldots, k. \tag{5.7}$$

Although this process has been outlined for the data from population i, it should be repeated for each of the data samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$. In doing so, we obtain a set of standardised weights $\{w_{i,j}\}_{i,j=1}^{k}$ providing a measure of similarity between each pairing of data samples, and by extension their respective populations.

Whilst the proposed weighting mechanism makes intuitive sense, it is important to confirm that it operates as expected in practice. To provide such confirmation, in the next subsection we visually explore a number of data sets and their associated weight sets as generated via Equation (5.7). This exploration will provide insight into how data set composition affects the pattern of weights yielded by the mechanism, and reassurance that it places weight primarily on relevant model scenarios, thereby limiting the impact of bias in practice.

5.3.2 Behaviour of Similarity-Based Weights

In this subsection, we explore the weighting mechanism proposed in the preceding subsection through visualising its behaviour on a number of contrasting examples. Here, we select a range of illustrative data sets generated via the Monte Carlo simulation approach, with rationale as outlined in Chapter 1 and as used in Chapter 4. For each, we display the associated weight sets derived from our proposed mechanism. The contrasting data sets allow us to observe how different data structures influence the resulting weights, and to understand how the mechanism appears to determine the relevance of populations in the pooled estimation process.

The names given to the data sets provide an initial sense of their structure:

- Clustered data set
- Continuously varying data set
- Data set with an outlier

Each data set comprises samples of size 10 ($n = 10$), drawn from 50 normal populations ($k = 50$). In all cases, the means range from 0 to 10, and the standard deviations from 1 to 2. Whilst the standard deviations vary uniformly from 1 to 2, the manner in which the means vary from 0 to 10 differs, leading to highly distinct data sets. Further, we will consider both cases in which the means and the standard deviations vary together and separately. As in previous chapters, all experimentation is implemented and carried out in MATLAB (2014a), with the sample code provided in Appendix C.

Clustered Data Set

To begin, we first outline the structure of the clustered data set as follows:

- Twenty-five populations are centred at 0 and twenty-five at 10, giving two clusters of population means.
- Population standard deviations increase uniformly over $[1, 2]$ as means increase.

The resulting populations are displayed in Figure 5.1. As expected, it is clear that the 50 populations do indeed form two clusters. It is also straightforward to appreciate that data from the first cluster (centred at 0) is unlikely to be relevant in the estimation of the second cluster (centred at 10), and vice versa. Ideally, we would like the generated weights to reflect this point.

The weights obtained from the proposed weighting mechanism are displayed in Figure 5.2. As anticipated, weights for population pairs within the same cluster are strongly positive and relatively uniform, whereas weights for pairs across clusters are close to zero. In this case, the behaviour of the proposed weighting mechanism is well-aligned with the relevance structure inherent in the data.

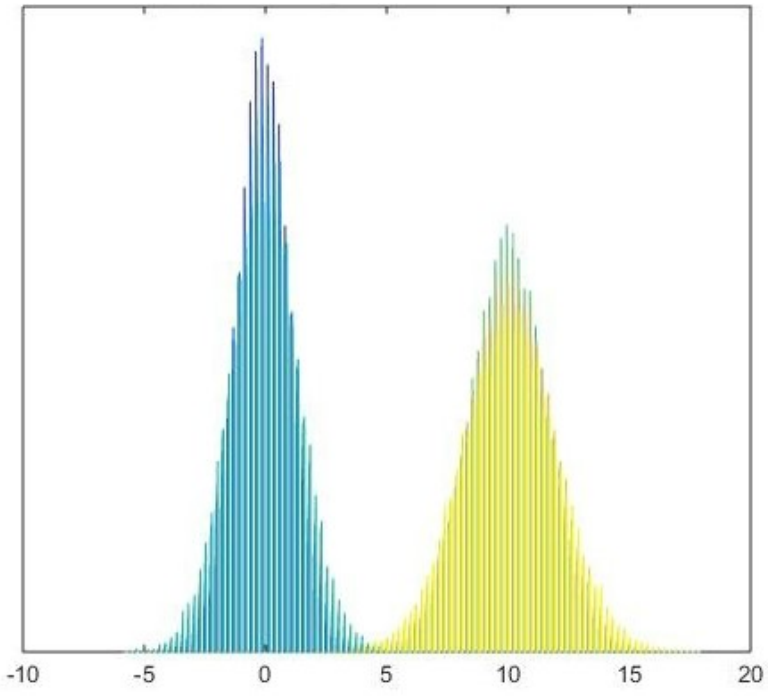

Figure 5.1: Clustered data set, population means distributed across two levels.

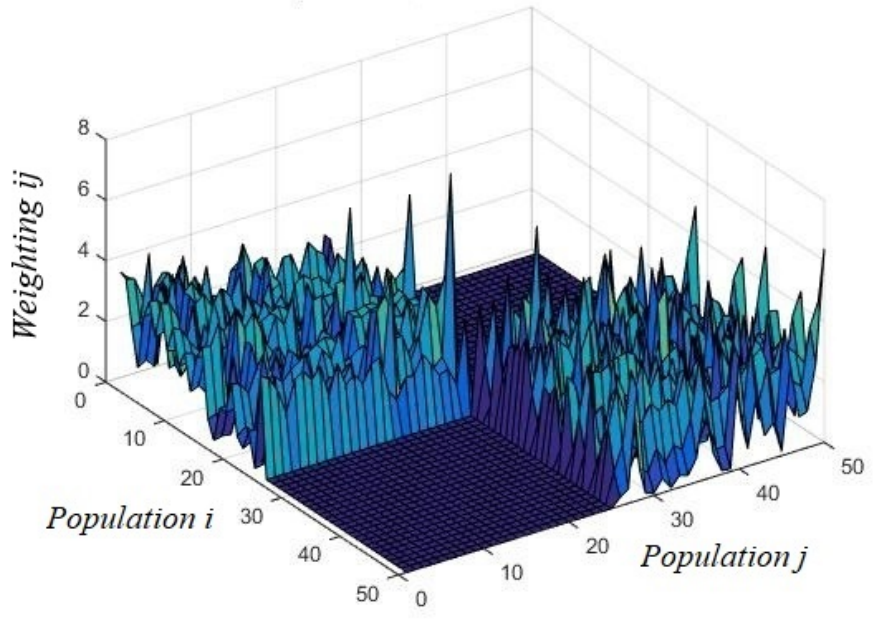

Figure 5.2: Weights derived for clustered data set

Continuously Varying Data Set

Next, we outline the structure of the continuously varying data set as follows:

- Population means increase uniformly over the mean interval $[0, 10]$.
- Population standard deviations decrease uniformly over $[1, 2]$ as means increase.

The resulting collection of populations is illustrated in Figure 5.3. Here, rather than forming clusters, the population means are evenly distributed across the mean interval. In this case, we would expect populations closely adjacent to one another to be most informative in terms of mean inference.

The weight structure is illustrated in Figure 5.4. We note immediately that the pattern of weights exhibited is markedly different from that observed for the clustered data set, as expected given the differences in the construction of the two

data sets.

Examining Figure 5.4 and taking an individual population i, the highest weights for population i tend to occur in those populations immediately adjacent to it. This leads to the distinctive diagonal ridge observed in Figure 5.4, somewhat analogous to the diagonal of ones in a correlation matrix. This is unsurprising, as these are the populations we would expect to be most similar to population i. Again, the weights obtained in application are well-aligned with our desired outcomes for the continuously varying data set.

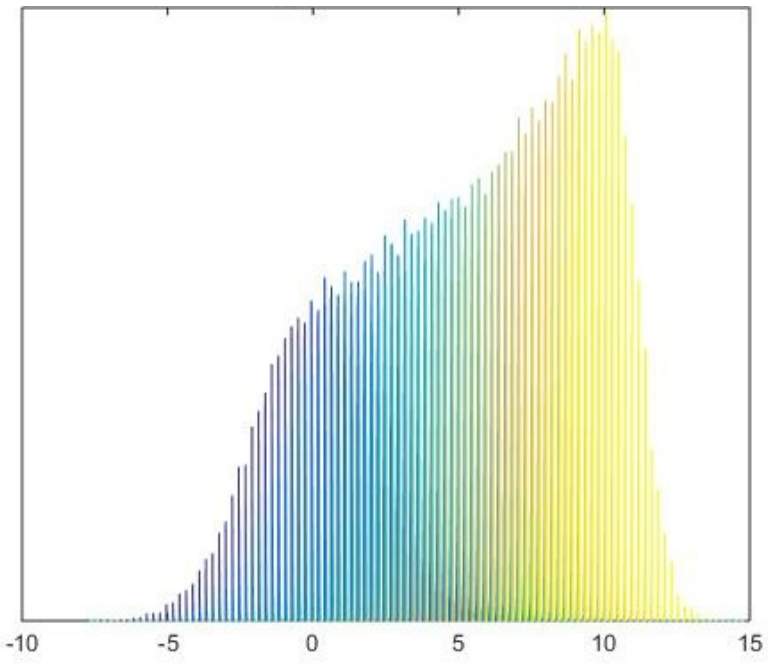

Figure 5.3: Continuously varying data set, population means uniformly spaced across interval.

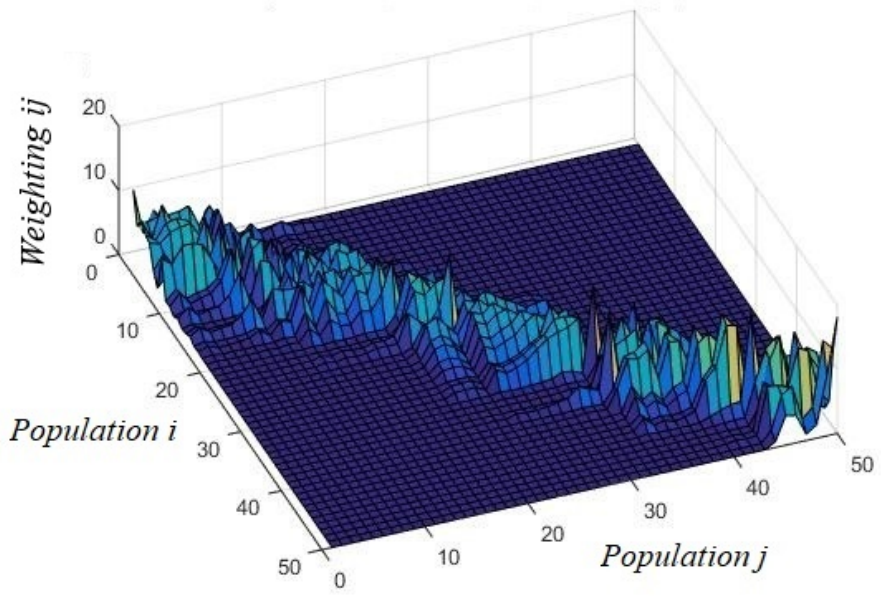

Figure 5.4: Weights derived for continuously varying data set

Data Set with an Outlier

Finally, we outline the structure of the data set with an outlier as follows:

- Forty-nine populations are centred at 0 and one is centred at 10.
- The populations with mean 0 have standard deviation 1, whilst the population with mean 10 has standard deviation 2.

The resulting populations appear in Figure 5.5. Here, the data are homogeneous with the exception of a single outlying population, markedly different from the first two data sets. Clearly, we would expect data from the first 49 populations to be informative in inferring about the lower mean endpoint, and only data from the outlying population to be informative about the higher mean endpoint. Therefore, for this structure, we would expect weights among the first 49 populations to be broadly similar, and weights involving the outlier to be close to zero unless the pair includes only the outlier itself.

The weights constructed via the weighting mechanism are shown in Figure 5.6. As expected, weights among the first 49 populations form a relatively flat surface, while the outlier population places most of its weight on itself, producing the distinctive spike. Again, the behaviour exhibited is consistent with the relevance structure inherent in the data, and with the intended operation of the weighting mechanism.

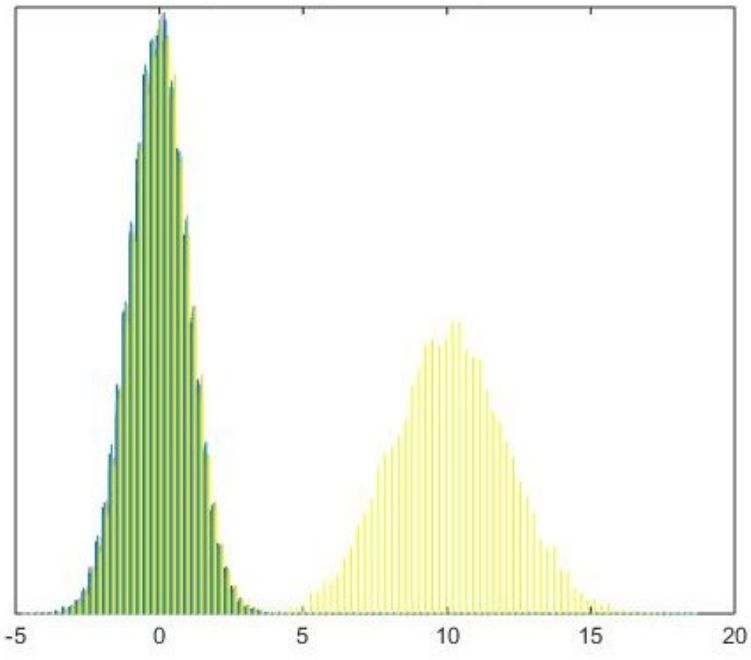

Figure 5.5: Data set with an outlying population, uneven distribution of population means.

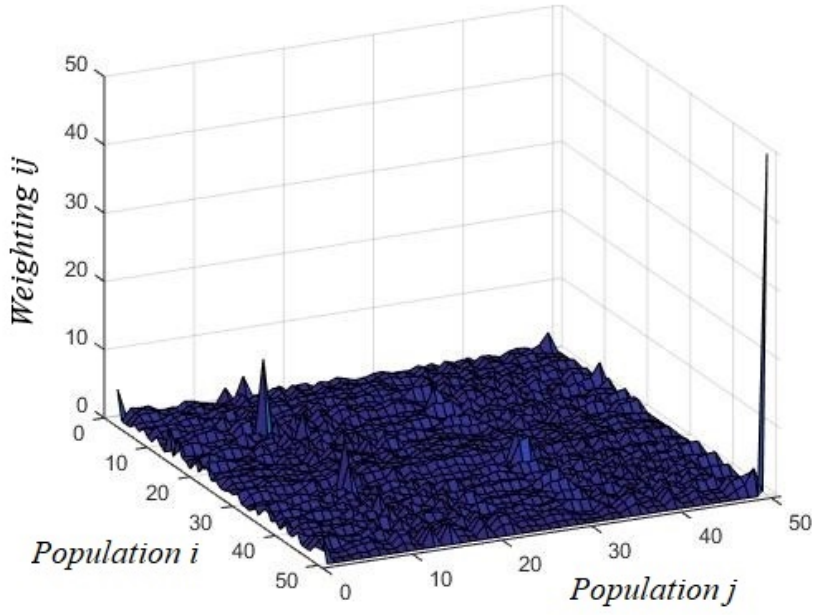

Figure 5.6: Weights derived for data set with an outlying population

These visualisations confirm that the weighting mechanism behaves as intended across a range of data structures, assigning weight to relevant populations and suppressing weight from dissimilar ones. In the following section, we integrate this weighting mechanism into a robust EB estimation procedure tailored to DES model experimentation.

5.4 Adapting the DS EB Estimator Using Similarity-Based Weights

Having developed and tested the weighting mechanism, we now incorporate it into an adapted version of Zhao's DS EB estimator [260]. The goal is to obtain a more robust estimator for DES settings by ensuring that shrinkage is driven primarily by data from populations judged to be similar. To motivate the construction, we first compare the individual weighted shrinkage target derived from the weighting mechanism with the common shrinkage target $\hat{\mu}$ used in the original DS EB

procedure. These comparisons, carried out for the illustrative data sets of Subsection 5.3.2, provide the foundation for the revised hyperparameter estimation of Subsection 5.4.2, and the formulation of the adapted DS EB estimator in Subsection 5.4.3.

Let us again consider Zhao's [260] DS EB estimator $\hat{\theta}_{EB,i}$. In this estimation procedure, a *common grand mean estimator* $\hat{\mu}$ is used as a shrinkage target in the estimation of the mean of each of our k populations. Going forward, however, in our adapted estimation procedure, we will make use of the weights constructed in the previous subsection to create an *individual weighted grand mean estimator* $\hat{\mu}_i$ to be used in place of $\hat{\mu}$ in the estimation of each of the k populations.

We proceed by fixing $i \in \{1, 2, \ldots, k\}$. As indicated above, we denote the individual weighted estimator for the i^{th} population by $\hat{\mu}_i$. In forming this estimator, we weight the contribution from the j^{th} data sample according to the likely similarity of the populations underlying the i^{th} and j^{th} samples. This is accomplished through the multiplication of each sample mean $\bar{X}_j$ by the corresponding weight $w_{i,j}$, giving rise to the following estimation formula:

$$\hat{\mu}_i = \sum_{j=1}^{k} \frac{w_{i,j}\bar{X}_j/\hat{\sigma}^2_{EB,j}}{\sum_{l=1}^{k} 1/\hat{\sigma}^2_{EB,l}} \qquad \text{for} \quad i = 1, 2, \ldots, k. \tag{5.8}$$

Here, we note that the formula in Equation (5.8) is identical to the original common estimator $\hat{\mu}$, as given by Equation (5.6), aside from the multiplication by the normalised weights $w_{i,j}$ in the numerator, as defined in Equation (5.7).

Using this weighting mechanism, we would naturally expect each of the individual weighted estimators $\hat{\mu}_i$ to be closer to the true population means μ_i than the original common estimator $\hat{\mu}$, and thus to provide a more efficient shrinkage target for the observed sample means $\bar{X}_i$. In the next subsection, we will examine this expectation empirically.

5.4.1 Behaviour of the Adapted Shrinkage Targets

To assess whether the adapted shrinkage targets behave as intended, we compare the common estimator $\hat{\mu}$ with the set of individual weighted estimators $\{\hat{\mu}_i\}$ across

the three illustrative data sets defined in Subsection 5.3.2. For each data set, both estimators are plotted against the known true population means μ_i, allowing us to observe how closely the individual weighted estimators track the underlying structure and whether they offer the anticipated improvement.

First briefly we recall only the names of the illustrative data sets:

- Clustered data set
- Continuously varying data set
- Data set with an outlier

Full structural details were provided in Subsection 5.3.2; here we focus solely on comparing the original common estimator $\hat{\mu}$ with the individual weighted estimators $\hat{\mu}_i$. Figures 5.7 to 5.9 below present these comparisons.

Comparison of Shrinkage Target for Clustered Data Set

For the clustered setting (with two groups, one of mean 0 and one of mean 10), the comparison between $\hat{\mu}$ and $\hat{\mu}_i$ appears in Figure 5.7.

In Figure 5.7, we illustrate the values taken by the original common estimator $\hat{\mu}$ and the individual weighted estimators $\hat{\mu}_i$ against the known, true population means μ_i. The common estimator $\hat{\mu}$ is depicted in red, the individual weighted estimators $\hat{\mu}_i$ in blue, whilst the true population values μ_i are depicted in black.

Examining Figure 5.7, we can clearly observe the superiority of the individual weighted estimators $\hat{\mu}_i$ over the common estimator $\hat{\mu}$. The common estimator $\hat{\mu}$ overestimates the true population values μ_i at the lower end of the interval, yet underestimates them at the upper end of the interval. The individual weighted estimators $\hat{\mu}_i$, on the other hand, closely follow the true population values μ_i across the mean range.

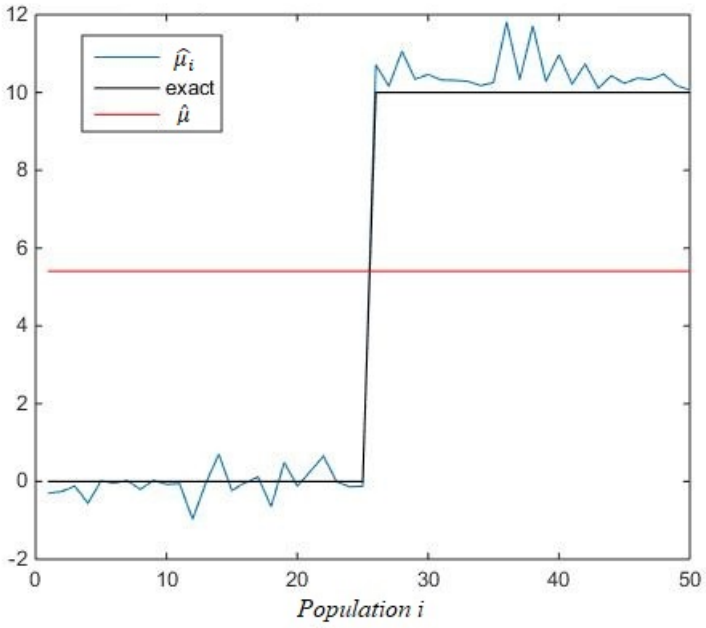

Figure 5.7: Common and individual weighted estimators of μ_i on clustered data set.

Comparison of Shrinkage Target for Continuously Varying Data Set

For the continuously varying setting, where population means change smoothly across the interval 0 to 10, Figure 5.8 displays the corresponding comparison.

The values taken by the common estimator $\hat{\mu}$ and the individual weighted estimators $\hat{\mu}_i$ are illustrated in Figure 5.8, alongside the known, true population means μ_i. Again, the common estimator $\hat{\mu}$ is depicted in red, the individual weighted estimators $\hat{\mu}_i$ in blue, and the true population values μ_i in black.

Once again, the individual weighted estimators $\hat{\mu}_i$ are seen to be superior, providing a far closer fit to the true population values μ_i than the common estimator $\hat{\mu}$. Here, $\hat{\mu}$ first overestimates, then underestimates, the true values μ_i. The individual weighted estimators $\hat{\mu}_i$, however, closely track the true values, without systematic bias.

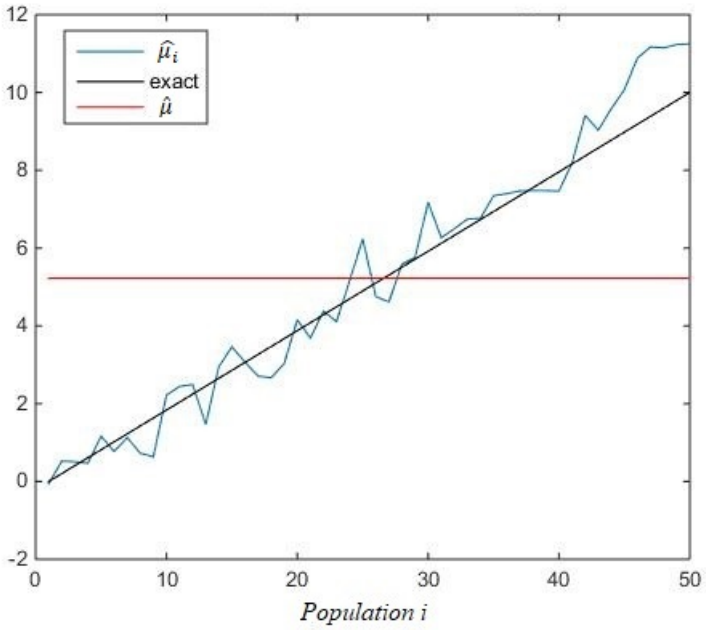

Figure 5.8: Common and individual weighted estimators of μ_i on continuously varying data set.

Comparison of Shrinkage Target for Data Set with an Outlier

For the setting with a single outlying population (with 49 populations with mean 0 and a single population with mean 10), Figures 5.9 to 5.10 present the comparison.

The common and individual weighted estimators for this data set are shown in Figure 5.9, alongside the true population values. Once again the common estimator $\hat{\mu}$ is in red, the individual weighted estimators $\hat{\mu}_i$ in blue, whilst the true population values μ_i are in black.

Examining Figure 5.9, we separate the horizontal population axis into two sections: the first 49 populations (mean 0) and the single outlying population (mean 10). Looking at the first section, we see that the presence of the outlier visibly skews the common estimator $\hat{\mu}$ upward, causing it to overestimate the true values. The individual weighted estimators $\hat{\mu}_i$, however, remain close to the true values, oscillating around 0.

Turning to the second section of the population axis with the aid of magnifications in Figure 5.10, we see the contrasting behaviour at the boundary between the non-outlying populations and the outlier. The left-hand panel zooms in on population 49, immediately preceding the outlier. Here, the blue line of the individual weighted estimators $\hat{\mu}_i$ is visible, relatively close to the true values in black: the red line of the common estimator $\hat{\mu}$, however, is not even visible. The right-hand magnification then focuses on population 50 itself. Again, the red line of the common estimator $\hat{\mu}$ is absent, whereas the individual weighted estimators $\hat{\mu}_i$ in blue track the outlying value far more closely.

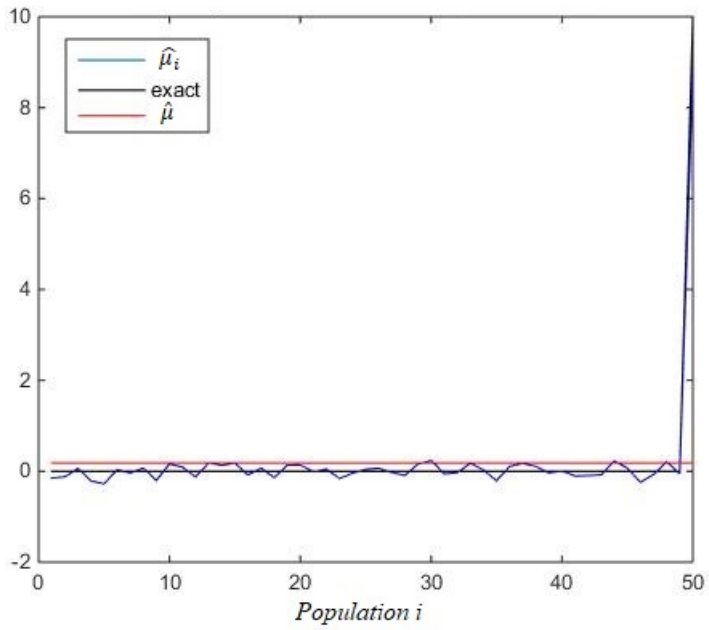

Figure 5.9: Common and individual weighted estimators of μ_i on data set with an outlying population.

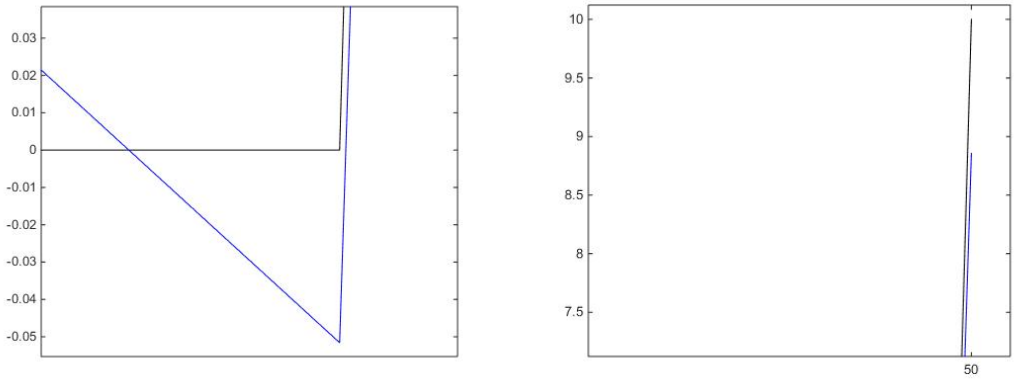

Figure 5.10: Magnifications from Figure 5.9

The results for the individual weighted grand mean estimator $\hat{\mu}_i$ illustrated in Figures 5.7, 5.8, 5.9 and 5.10 appear promising. Across all three settings, the individual weighted estimator $\hat{\mu}_i$ provides a much better fit concerning the intricacies of the true underlying population means μ_i, and therefore represents a markedly improved shrinkage target.

The overall statistical performance of the adapted DS EB estimator, built upon the individual weighted estimator $\hat{\mu}_i$, will be rigorously evaluated in the following section. Before doing so, however, the next subsection reviews the estimation of the other hyperparameters and, where necessary, provides the adaptations necessary to ensure their compatibility with $\hat{\mu}_i$.

5.4.2 Adapted Hyperparameter Estimation

In the preceding subsection, the individual weighted grand mean estimator $\hat{\mu}_i$ was introduced and shown to behave more responsively as a shrinkage target across a range of illustrative data sets. We now turn to the other components of the original DS EB estimation procedure and examine whether the introduction of $\hat{\mu}_i$ requires any adjustments to the estimation of the remaining hyperparameters.

In the DS EB estimator proposed by Zhao [260] and detailed in Chapter 3, multiple hyperparameters require estimation, namely $\mu, \tau^2, \mu_\upsilon$ and τ_υ^2.

Briefly recalling their formulation, for μ_υ and τ_υ^2 we have:

$$\hat{\mu}_\upsilon = \frac{1}{p}\sum_i (\log S_i^2 - m), \tag{5.9}$$

and:

$$\hat{\tau}_\upsilon^2 = \left(\frac{1}{p}\sum_i (\log S_i^2 - m)^2 - \sigma_{ch}^2 - \hat{\mu}_\upsilon^2\right)_+, \tag{5.10}$$

respectively. For μ and τ^2 we have:

$$\hat{\mu} = \frac{\sum_i (X_i/\hat{\sigma}_{EB,i}^2)}{\sum_i (1/\hat{\sigma}_{EB,i}^2)}, \tag{5.11}$$

and:

$$\hat{\tau}^2 = \left(\frac{\sum_i \left((X_i - \hat{\mu})^2 - S_i^2 exp(-m - \sigma_{ch}^2/2)\right)}{p}\right)_+, \tag{5.12}$$

respectively.

Examining Equations (5.9) to (5.12), we first note that Zhao's common grand mean estimator $\hat{\mu}$, as given by Equation (5.11), has now been replaced by our individual weighted grand mean estimator $\hat{\mu}_i$, as given by Equation (5.8). This is our only modification to Zhao's original approach at this stage, and it directly determines which additional hyperparameter estimators require attention. In particular, only those expressions that explicitly depend on $\hat{\mu}$ require updating.

It is straightforward to verify that the estimators for μ_υ and τ_υ^2 do not involve $\hat{\mu}$, and thus remain unchanged. However, the estimator for τ^2 does depend on $\hat{\mu}$, and so does require adaptation to ensure compatibility with our individual weighted grand mean estimator $\hat{\mu}_i$.

An appropriate alternative estimator for τ^2 follows from substituting the individual values of $\hat{\mu}_i$ for the common value $\hat{\mu}$, leading to our adapted τ_w^2 estimator as follows:

$$\hat{\tau}_w^2 = \left(\frac{1}{k}\sum_{j=1}^{k}\left(\left(\bar{X}_j - \hat{\mu}_j\right)^2 - S_j^2 \exp(-m_j - \sigma_{ch,j}^2/2)\right)\right)_+. \tag{5.13}$$

Note that, given the construction of $\hat{\mu}_j$, the value $\bar{X}_j$ necessarily receives the largest

weight. It follows that $\hat{\mu}_j$ must be at least as close to $\bar{X}_j$ as $\hat{\mu}$. As such, we would expect the typical magnitude of $\left(\bar{X}_j - \hat{\mu}_j\right)^2$ to be less than that of $\left(\bar{X}_j - \hat{\mu}\right)^2$, and hence $\hat{\tau}_w^2$ will typically be smaller than $\hat{\tau}^2$. This reduction carries implications for the overall degree of shrinkage applied by the adapted DS EB estimator, and we return to this point immediately after presenting its full formulation in the next subsection.

5.4.3 Final Formulation of the Adapted Similarity-Weighted DS EB Estimator

Bringing together the adapted shrinkage target $\hat{\mu}_i$, the revised variance component $\hat{\tau}_w^2$, and the unchanged hyperparameters μ_υ and τ_υ^2, we are now able to present the full formulation of the adapted similarity-weighted DS EB estimator (hereafter, the adapted DS EB estimator).

The adapted DS EB point estimator, denoted by $\hat{\theta}_{wEB,i}$ is given by the formula:

$$\hat{\theta}_{wEB,i} = \hat{M}_{wEB,i}\bar{X}_i + (1 - \hat{M}_{wEB,i})\hat{\mu}_i, \tag{5.14}$$

where the adapted shrinkage weight $\hat{M}_{wEB,i}$ is given by:

$$\hat{M}_{wEB,i} = \frac{\hat{\tau}_w^2}{\hat{\tau}_w^2 + \hat{\sigma}_{EB,i}^2}, \tag{5.15}$$

and where the EB estimate of the population variance, $\hat{\sigma}_{EB,i}^2$, is given by:

$$\hat{\sigma}_{EB,i}^2 \;=\; \exp\left(\hat{M}_\upsilon(\log S_i^2 - m) + (1 - \hat{M}_\upsilon)\hat{\mu}_\upsilon\right). \tag{5.16}$$

Our individual weighted grand mean estimator $\hat{\mu}_i$ is given by:

$$\hat{\mu}_i = \sum_{j=1}^{k} \frac{w_{i,j}\bar{X}_j/\hat{\sigma}_{EB,j}^2}{\sum_{l=1}^{k} 1/\hat{\sigma}_{EB,l}^2} \qquad \text{for} \;\; i = 1,\, 2,\, \ldots,\, k, \tag{5.17}$$

and our adapted τ_w^2 estimator is given by:

$$\hat{\tau}_w^2 = \left(\frac{1}{k}\sum_{j=1}^{k}\left(\left(\bar{X}_j - \hat{\mu}_j\right)^2 - S_j^2\exp(-m_j - \sigma_{ch,j}^2/2)\right)\right)_+. \tag{5.18}$$

Finally, hyperparameters μ_v and τ_v^2 are, as in the original approach, as follows:

$$\hat{\mu}_v = \frac{1}{p}\sum_i (\log S_i^2 - m),$$

and:

$$\hat{\tau}_v^2 = \left(\frac{1}{p}\sum_i (\log S_i^2 - m)^2 - \sigma_{ch}^2 - \hat{\mu}_v^2\right)_+ .$$

Returning to our earlier observation concerning $\hat{\tau}_w^2$ and $\hat{\tau}^2$, namely that $\hat{\tau}_w^2$ will typically be less than $\hat{\tau}^2$, it follows that the adapted shrinkage weight $\hat{M}_{wEB,i}$ will tend to be less than the original shrinkage weight $\hat{M}_{EB,i}$. Comparing Zhao's original DS EB estimator (as given in Equation (5.4)) with our adapted DS EB estimator (as given in Equation (5.14)), the reduction in $\hat{M}_{wEB,i}$ indicates a stronger pull towards the individual weighted shrinkage target $\hat{\mu}_i$, and this stronger pull is beneficial precisely because the shrinkage target $\hat{\mu}_i$ tends to be more accurate.

Taken together, this implies that the adapted DS EB estimator $\hat{\theta}_{wEB,i}$ should, in principle, offer improved performance over Zhao's original DS EB estimator $\hat{\theta}_{EB,i}$. The next section evaluates this expectation numerically over a large and diverse collection of data sets.

5.5 Numerical Study of Relative Performance

The purpose of this section is to numerically compare the adapted DS EB estimator, developed in the preceding subsections and defined in Equation (5.14), with the original DS EB estimator proposed by Zhao [260]. Whilst Zhao's estimator was shown in Chapter 3 to be of value within a DES experimentation context, the discussion and empirical investigations of this chapter suggest that the adapted version may be better suited to the demands of DES practice. In particular, a key challenge for standard EB approaches is the potential for large individual errors in the presence of outlying populations. The weighting mechanism and adapted shrinkage target introduced in this chapter were designed specifically to limit bias introduced into estimation by such outliers, and thereby reduce maximum error behaviour. As noted in Chapter 2, DES model experimentation can feature complex and sometimes highly irregular response surfaces, making robustness to such

effects an important practical requirement.

In this evaluation we therefore compare the original DS EB point estimator, $\hat{\theta}_{EB,i}$, with the adapted DS EB estimator $\hat{\theta}_{wEB,i}$. No comparison with the frequentist estimator is required here: proof of concept for the potential advantage of EB over the sample mean (Equation (3.21)) has already been demonstrated in Chapter 3. The present task is instead to assess how EB may be operationalised to best effect within DES experimentation, and thus we focus solely on the relative performance of the two DS EB estimators.

Following the approach adopted in Chapter 4, the evaluation is conducted using Monte Carlo simulated data sets, with the rationale for this methodology outlined in Chapter 1. A carefully constructed experimental design is required to ensure that a wide variety of data-generating conditions are included, including both those representing standard DES experimental environments and those representing more challenging, highly heterogeneous configurations. This is essential for assessing not only average performance but also robustness in the types of settings where the adapted DS EB estimator is expected to provide benefit.

The remainder of this section proceeds as follows. Subsection 5.5.1 outlines the experimental design used to generate the Monte Carlo simulated data sets. Subsection 5.5.2 sets out the error measurement framework, based on the MSSE and MMAE measures defined in Chapter 3. Subsection 5.5.3 presents the results for the main experimental design, and Subsection 5.5.4 reports the results of an extended study incorporating deliberately more heterogeneous data sets. Together, these analyses provide a comprehensive evaluation of the adapted DS EB estimator's performance.

5.5.1 Experimental Test Set Design

Now we turn to the construction of the experimental test set used to evaluate the original and adapted DS EB estimators. As in Chapter 4, we base our evaluation on Monte Carlo simulated data sets comprised of normal populations, with the rationale for this approach discussed in Chapter 1. The purpose here is not to attempt to reproduce the full complexity of DES model data, but rather to create

a large and systematically varied collection of data sets that captures some of the structural challenges that arise in DES experimentation. More specifically, we aim to introduce substantial population heterogeneity through purposefully varying the mean and variance structures of the included populations. The impact of outlying populations is, however, examined separately in an extension study discussed later in Subsection 5.5.4.

The experimental design's structure broadly follows that introduced in Subsection 4.3.1 for the construction of the experimental testbed. Each large-scale experimental data set consists of observations drawn from multiple normal populations, with each population playing the role of a DES model scenario in which the estimator must perform well. The emphasis here, as in Chapter 4, is on building a broad and controlled range of structures along the key dimensions that are expected to influence EB performance, namely:

- the number of populations, k;
- the sample size, n;
- the range and structure of the population means;
- the range and structure of the population standard deviations;
- whether the means and standard deviations vary together or separately.

Using these factors, the experimental test set is generated using the experimental parameter combinations as outlined in Table 5.2. Taken together, this factorial design results in a total of $1,296$ distinct experimental parameter combinations.

Population Count k	$10, 30, 100, 200$
Sample Size n	$3, 5, 10$
Range of μ Values	$[0, 5], [0, 10], [0, 20]$
Number of μ Levels	$2, k/2, k$
Range of σ Values	$1, [1, \sqrt{2}], [1, 2]$
Number of σ Levels	$2, k$
μ and σ Variation	increasing, decreasing

Table 5.2: Experimental design for the creation of the experimental test set.

As is evident from the experimental parameter combinations in Table 5.2, this design supports a broad range of mean—variance structures and degrees of heterogeneity across populations. For example, consider parameter combination (a): $k = 200$, $n = 3$, a mean range $[0, 5]$, with only 2 mean levels, and a common standard deviation of 1. This parameter combination yields many closely aligned populations with very limited within-population information, a structure that is highly conducive to EB inference. By contrast, parameter combination (b): $k = 10$, $n = 10$, a mean range $[0, 20]$ over $k = 10$ levels, and standard deviation range of $[1, 2]$ also over $k = 10$ levels. This yields a collection of populations that differ markedly in both mean and variance, providing an example of an EB-unfavourable setting in which pooling has the potential to introduce substantial bias. Taken together, these examples illustrate the spectrum of heterogeneity captured by the design, ranging from strongly EB-favourable to clearly EB-unfavourable settings.

As in Chapter 4, each experimental parameter combination excluding sample size corresponds to a large-scale experimental data set. Each such large-scale data set is generated with sufficient size so as to permit the extraction of 100 practitioner-level data sets, each with sample size up to the maximum n specified in Table 5.2. These practitioner-level data sets form the basis for the replication and averaging used in the error measurement approach described in the following subsection.

Having now discussed the experimental design and resultant data sets, we next outline the error measurement approach used to assess estimator performance over these data sets.

5.5.2 Error Measurement Approach

In evaluating the performance of the original and adapted DS EB estimators, we again make use of the error measures introduced in Chapter 3. Since the adapted DS EB estimator is designed to address the challenge posed by heterogeneous and, in particular, outlying populations, both the mean sum of squared error (MSSE) and the mean maximum absolute error (MMAE) should be employed. Re-introducing the MMAE will provide a more comprehensive picture of the errors incurred, allowing us to better examine how the approaches perform in estimating outlying populations, as opposed to examining only how they perform

in aggregate across all populations. As noted in its development, the adapted DS EB estimator's individual weighted shrinkage target more closely tracks the true population means than the original DS EB estimator's common shrinkage target, and so any improvements should be reflected across both error measures.

As discussed above, performance is evaluated using 100 practitioner-level data sets extracted from each large-scale experimental data set. Since the true means θ_i are known by construction, these replications allow direct evaluation of the MSSE and MMAE for each estimator, and the required averaging to mitigate stochastic variability, following the same general rationale as in Chapters 3 and 4. However, unlike in those chapters, once these averages have been obtained for every parameter combination, the resulting ratios are then averaged across the full experimental design to provide a reliable overall measure of the relative performance of the two estimation approaches.

To compute the relative mean sum of squared error (MSSE), we calculate the quantities:

$$wEB_{SSE} = \sum_{i=1}^{k} \left(\hat{\theta}_{wEB,i} - \theta_i\right)^2 \text{ and } EB_{SSE} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2, \tag{5.19}$$

based on Equation (3.22) given in Chapter 3.

These quantities are calculated over 100 replications, and the mean taken. Taking the ratio of the resulting mean values provides a relative measure of the observed accuracy of the two methods. Therefore, we define the relative mean sum of squared error (MSSE) ratio as follows:

$$wEB/EB_{SSE} = \text{mean}wEB_{SSE}/\text{mean}EB_{SSE}. \tag{5.20}$$

This process is carried out for each experimental parameter combination in our experimental design.

Turning attention to the relative mean maximum absolute error (MMAE), we

likewise calculate the quantities:

$$wEB_{max} = \max_i \left|\hat{\theta}_{wEB,i} - \theta_i\right| \text{ and } EB_{max} = \max_i \left|\hat{\theta}_{EB,i} - \theta_i\right|, \tag{5.21}$$

based on Equation (3.23) given in Chapter 3.

Again, these quantities are calculated over 100 replications, and the mean taken. We take the ratio of these resulting mean values to obtain a relative measure of the observed accuracy of the two methods over the populations in question. Therefore, we define the relative mean maximum absolute error (MMAE) ratio as follows:

$$wEB/EB_{max} = \text{mean}wEB_{max}/\text{mean}EB_{max}. \tag{5.22}$$

Once again, this process is carried out for each experimental parameter combination in our experimental design.

At this point, each experimental parameter combination in the experimental design now has two error measures associated with it, an MSSE ratio and a MMAE ratio, each comparing the adapted DS EB estimator to Zhao's original DS EB estimator over the experimental conditions represented by that parameter combination. To obtain a stable, design-wide assessment of the relative performance of these estimators, these error ratios are then averaged across all parameter combinations.

Having now discussed both the experimental design and error measurement strategy for the study, in the next subsection, we present and discuss the results obtained upon their implementation.

5.5.3 Main Study Results

The main numerical results, obtained from the experimental design outlined in Subsection 5.5.1 and evaluated using the error measurement approach of Subsection 5.5.2, are summarised below. These results reflect performance averaged across all experimental parameter combinations and practitioner-level data sets in the experimental test set, thereby providing a stable overall assessment of the relative behaviour of the adapted and original DS EB estimators.

Table 5.3 presents the resulting error ratios. As may be seen, a mean value of 0.8910 was observed for wEB/EB_{SSE}, the relative MSSE error measure, whilst a mean value of 0.9044 was observed for wEB/EB_{max}, the relative MMAE error measure. Both values indicate improved performance for the adapted DS EBestimator, with reductions of approximately 10% relative to the original DS EB estimator.

Error measure	Result
wEB/EB_{SSE}	0.8910
wEB/EB_{max}	0.9044

Table 5.3: Results of numerical evaluation of adapted versus original DS EB point estimators on the experimental design.

These results may be interpreted in light of the heterogeneity present in the study's experimental design. As discussed in Subsection 5.5.1, the data sets generated from Table 5.2 span a broad range of structures, from strongly EB-favourable settings, such as many populations with limited within-population information clustered in a narrow region, to more EB-unfavourable settings in which populations are widely dispersed across their respective mean and variance ranges. Importantly, however, none of these settings produce the kind of extreme population separation considered in the next subsection.

Against this backdrop, the adapted DS EB estimator performs well for two main reasons. First, the individual weighted shrinkage target introduced in Section 5.4 provides a more appropriate point of attraction than the original common shrinkage target. It improves alignment with the underlying mean structure, and reduces systematic bias when populations differ in structured but non-extreme ways. Second, the weighting mechanism limits borrowing of strength when a population's mean departs substantially from the dominant pattern across the data set. Taken together, these features help stabilise performance across the range of DES-relevant conditions represented in the study's experimental design.

The similarity of the MSSE and MMAE improvements is also informative. In heterogeneous settings these two measures typically diverge, with EB shrinkage

often improving average accuracy (MSSE) while leaving the worst individual errors (MMAE) largely unchanged or inflated. Their near equal improvement here therefore indicates that the adapted DS EB estimator is improving overall accuracy without inducing the disproportionate errors sometimes produced by shrinkage under population heterogeneity. This balanced behaviour is a central design objective of the adapted DS EB estimation approach and underscores its practical value in DES experimentation settings with mixed levels of population similarity.

However, the experimental parameter combinations in Table 5.2 will not, by definition, produce data sets exhibiting the more pronounced population separation seen earlier in Figure 5.5. Whether such sharply outlying populations arise frequently or only occasionally in DES practice is difficult to ascertain, but they do represent an important class of estimation problems, particularly given the motivating concerns of this chapter. The main study therefore cannot speak directly to the adapted DS EB estimator's performance in these types of highly heterogeneous data analytic situations. Thus, in the next subsection, we explore an extension to the main study that better enables us to comment on the relative performance of the estimators in settings featuring outlier populations.

5.5.4 Extension Study: Design and Results

A central motivation for this chapter has been the development of an estimator that remains reliable even when confronted with markedly heterogeneous populations, including settings exhibiting clearly outlying populations. To examine the adapted DS EB estimator under more demanding conditions than those represented in the main study, an extension study was conducted in which each experimental parameter combination deliberately includes an extreme population mean. This subsection first outlines the necessary modifications to the main study's experimental design, then briefly recaps the error measurement strategy used, before presenting and discussing the implications of the extension study's results.

To introduce controlled outliers while preserving comparability with the main study, the data-generation process associated with each experimental parameter combination expressed in Table 5.2 was modified as follows. A value $p \sim \mathrm{U}(10, 50)$ was first generated. The mean range was then constrained such that $k-1$ popula-

tions were centred uniformly within the interval spanning 0 to $p\%$ of the original upper mean limit, while the final population was centred at the original upper mean limit. This construction ensures that, in general, all but one of the populations will be centred no more than halfway along the original mean range, whilst the last population will be centred at the upper end of the original mean range, thereby creating a controlled outlier in each large-scale experimental data set.

Providing a concrete illustration helps clarify the construction process. If the baseline mean range is $[0, 20]$ with $k = 30$, and $p = 25$ is generated, then 29 of the populations are centred across $[0, 5]$, while the remaining population is centred at 20. Taken together, these modifications introduce a qualitatively different population structure from that of the main study. The consolidation of $k - 1$ populations into a relatively narrow portion of the mean range creates settings that are, in principle, favourable to EB pooling, since most populations now share broadly similar locations. At the same time, the deliberately displaced final population creates a situation where shrinkage is unlikely to provide genuine improvement and the main concern becomes avoiding pull towards an inappropriate target. This combination of an EB-favourable bulk structure and a structurally challenging outlier frames the interpretation of the results.

The error measurement strategy mirrors that set out in Subsection 5.5.2. For each modified experimental parameter combination excluding sample size, a large-scale experimental data set was generated. This was sufficiently large so as to permit the extraction of 100 practitioner-level data sets of sample size n (with n as specified by the experimental parameter combination). The adapted and original DS EB estimates were calculated for each practitioner-level data set, and the use of the true mean values permitted the calculation of both MSSE and MMAE error measures. These were averaged over the 100 practitioner-level data sets, and then averaged across all experimental parameter combinations.

Applying the modified experimental design and error measurement strategy as outlined above leads to the error ratio results as shown in Table 5.4.

The adapted DS EB estimator again shows improved performance, with a mean value of 0.7529 observed for wEB/EB_{SSE}, the relative MSSE error measure, and

Error measure	Result
wEB/EB_{SSE}	0.7529
wEB/EB_{max}	0.9085

Table 5.4: Results of numerical evaluation of adapted versus original DS EB point estimators on the extension study experimental design.

a mean value of 0.9085 observed for wEB/EB_{max}, the relative MMAE error measure. The magnitude and pattern of these improvements reflect the dual structure of the modified experimental data sets. The substantial reduction in MSSE, approximately 25%, arises because the bulk of the populations now form a compact and highly EB-favourable cluster. The individual weighted shrinkage target allows estimation within these clusters to borrow strength more effectively than under the original DS EB estimator, thereby reducing systematic error across the majority of the data set.

In contrast, the improvement in MMAE is more modest, around 10%. This is to be expected: MMAE is governed by the single most difficult population, which in this design is almost always the intentionally displaced outlier. For such a population, shrinkage generally cannot produce genuine improvement, since borrowing strength from highly dissimilar populations only introduces bias, and the best achievable outcome is therefore to avoid exacerbating the error. The adapted DS EB estimator tends to limit shrinkage for this population, behaving similarly to a lightly pooled or even nearly unpooled estimator. This restraint reduces the bias that the original DS EB estimator would otherwise incur and explains the modest but consistent improvement in MMAE.

Taken together, the extension study's results illustrate a key qualitative property of the adapted DS EB estimator: its ability to pool information selectively. When population similarity is strong, the adapted DS EB estimator takes advantage of this structure; when similarity is weak, it suppresses shrinkage to avoid an unhelpful pull towards an inappropriate target. This selective behaviour underpins the gains observed in both studies and provides evidence of robustness in the types of heterogeneous conditions that motivated the developments of this chapter.

5.6 Conclusions

The purpose of the current chapter was to explore the idea of scenario similarity, and examine how EB estimation might be adapted to account for heterogeneous or outlying DES experimental conditions. As such, this chapter primarily addresses Research Objective R4, which concerns the mitigation of undesirable EB behaviour at the level of individual DES model scenarios. In particular, it examined whether the variance reduction offered by EB can be realised without inducing large maximum absolute errors in heterogeneous or outlying settings. In doing so, the chapter also contributes to Research Objective R5 through the construction of a Monte Carlo training and evaluation environment used to assess robustness under controlled population heterogeneity.

To address this objective, the chapter developed an adapted DS EB estimation procedure designed to account explicitly for scenario similarity. Building on the DS EB estimator of Zhao [260], the proposed method replaces the common shrinkage target with an individual similarity-weighted target constructed using pairwise measures of scenario similarity derived from Welch's t-test.

The numerical work of the chapter revealed a strong performance for the adapted DS EB estimator, with an improvement of around 10% on both error measures on the general experimental test set, and an improvement of up to 25% on the extension study's modified test set which integrated deliberate outliers. These results provide reassurance that the adapted DS EB estimator is robust to highly heterogeneous data, at least in a Monte Carlo simulation environment.

Beyond the numerical improvements observed, the results provide insight into how EB estimation can be stabilised in heterogeneous DES settings. The similarity-weighted formulation enables selective borrowing of strength: when populations exhibit strong structural similarity, the estimator pools information effectively, while in the presence of dissimilar or outlying scenarios, shrinkage is moderated to limit the introduction of bias. This behaviour is reflected in the concurrent improvements observed in both MSSE and MMAE, indicating that gains in average accuracy are achieved without disproportionately inflating worst case errors. These findings reinforce the importance of explicitly accounting for scenario simi-

larity when operationalising EB in DES experimentation.

Taken together, these results demonstrate that Research Objectives R4(a) and R4(b) have been substantively addressed, with the adapted DS EB estimator offering improved robustness in the context of DES scenario heterogeneity while retaining the efficiency gains of EB. In addition, the evaluation of the estimator within controlled Monte Carlo environments establishes a systematic basis for its subsequent application in a genuine DES setting, representing an initial contribution towards Research Objective R4(c) and further advancing Research Objective R5.

While the results of this chapter demonstrate that the adapted DS EB estimator represents a substantive step towards the operationalisation of EB for DES experimentation, the methods developed in Chapters 4 and 5 have thus far been evaluated only within controlled Monte Carlo environments based on normal populations. The extent to which these tools translate to a genuine DES experimentation context therefore remains to be established. In the next chapter, both the EB suitability heuristic and the adapted DS EB estimator are applied to an industrial-scale DES case study, completing Research Objectives R3(c) and R4(c) through assessment of their practical applicability in a real modelling setting.

Chapter 6

Case Study: Offshore Wind Farm Installation Logistics Model

In this chapter, we apply the tools and techniques developed thus far in the thesis to a genuine DES case study. The case study in question concerns the installation phase of an offshore wind farm (OWF) and its inclusion at this point in the thesis serves several key purposes. While the numerical testing in previous chapters demonstrated the statistical efficiency of the methods, the current chapter allows us to assess their performance within a real-world DES experimentation context. The work of the chapter therefore corresponds primarily to Research Objective R3(c) and R4(c), which seek to evaluate the practical applicability and performance of the EB suitability heuristic and the adapted DS EB estimator in applied DES settings.

Although a rich literature exists on methods for the design and analysis of DES experiments, recent research by Hoad et al. [117] revealed this does not always translate to effective and informed practice. Several reasons are posited by the authors for this disparity, including limited technical knowledge among practitioners, and a perceived lack of need of advanced analytical methods. This highlights the importance of careful dissemination, and the value of approaches that are accessible, intuitive and demonstrably useful in practice.

This practical applicability may be measured along a number of dimensions. First, accessibility, meaning the ease with practitioners can acquire the knowledge re-

quired to use the methods. Secondly, robustness, the extent to which performance is affected by deviations from modelling assumptions. Third, usability, referring to whether the methods remain workable for problems of realistic size and scale. Finally, practical value, that is, the extent to which improvements in statistical efficiency translate into outcomes that matter for decision making. Some of these dimensions can be partially evaluated a priori. For example, the closed form of the adapted EB estimator developed in Chapter 5, removes the need for numerical methods, supporting accessibility and usability. However, a holistic assessment of real-world applicability requires testing in practice, motivating the case study undertaken in this chapter.

Whilst all of the above dimensions are important, robustness and practical value warrant particular emphasis. The need for methods that remain dependable under real-world complexity is widely recognised in both management science [247] and DES [38]. Much of the work carried out so far has focused on normally distributed data, with Zhao's original DS EB estimator [260] itself built around a normality assumption. For the methods developed here to be applicable in general DES practice, their behaviour under departures from normality must be examined.

Practical value is equally important. Although reducing in the sum of squared error across model scenarios is desirable, the discovery of an improved installation schedule, or the identification of the same schedule within a reduced timeframe or computational budget, is typically of greater significance in applied settings.

We also note that the case study used originates from a Strathclyde Technology and Innovation Centre (TIC) Low Carbon Power and Energy (LCPE) grant funded by industrial partners SSE Renewables, Scottish Power Renewables and Technip Offshore Wind Limited. As such, the tools developed in the project are the intellectual property of the industrial partners. This necessitates care in the presentation of results due to confidentiality restrictions, and limits the extent to which the work carried out can be described in detail. In the remainder of the chapter, specific remarks will be made to indicate such restrictions.

The remainder of the chapter is organised as follows. Sections 6.1 and 6.2 introduce useful information about the OWF context, whilst Section 6.3 provides an overview

of the work and impact of the original project. Section 6.4 discusses the use of the case study DES model to trial the methods developed in the thesis, including the experimental design used, the results obtained and their implications. The chapter concludes in Section 6.5.

6.1 Problem Context: OWF Trends

In this section, we provide context for the current case study, discussing a simple classification of wind farm location, and both the practical and political drivers shaping decisions regarding wind farm location. The reason for our interest in wind farm location is its significant impact in terms of the uncertainty involved in OWF installation and operation, giving some sense of the complexities involved in the case study's simulation modelling efforts.

Wind farms, previously considered the gold standard in green energy generation, have recently been subject to a variety of environmental and social criticisms [60]. Despite such criticisms, there remains a clear need for sustainable energy. As a consequence, this has led to wind farm installations being located further and further offshore, in an effort not only to mitigate the aforementioned criticisms, but also to optimise energy production. For the purposes of discussion, any wind farms on land are referred to as onshore wind farms, any wind farms close to shore (i.e. less than 5 miles) are referred to as coastal wind farms, and those further from shore are referred to as offshore wind farms.

The practical drivers for the move further offshore in wind farm installation are summarised in Table 6.1, based on discussions outlined by Barlow et al. [13], Clegg [60], the UK Department of Trade and Industry (DTI) [177], and Snyder and Kaiser [217]. Perhaps the most critical driver is the improved weather conditions found further from shore, with stronger and more continuous wind enabling greater and more reliable energy generation [13]. This driver, combined with the reduced visual impact, a current lack of regulations concerning turbine blade size, the reduced vessel traffic and the greater availability of area, make further offshore locations more and more attractive for wind farm installation [177]. As might be expected however, these advantages do not come without cost, and the more severe weather conditions that facilitate greater energy generation also make the

Driver	Impact
More suitable weather conditions	(onshore, coastal) → offshore
Lack of turbine blade size regulations	onshore → (coastal, offshore)
Reduced visual impact on environment	onshore → coastal → offshore
Reduced vessel traffic	coastal → offshore
Greater availability of area	(onshore, coastal) → offshore

Table 6.1: Practical drivers involved in the move further offshore in windfarm installations.

installation and operation of the wind farm more difficult, often leading to significant increases in build duration and cost [217].

The main political driver behind the push for further offshore in wind farm location is RenewableUK's vision as illustrated in The Crown Estate map [84] shown in Figure 6.1. This map indicates Round 1 and Round 2 OWFs (i.e. those already built, and currently being built, respectively) in colour, whereas Round 3 OWF sites (i.e. those agreed / in planning) are indicated by the hatched zones. The push for further offshore discussed in this section is clearly discernable in this visual representation.

6.2 Technical Overview: OWF Lifecycle

With a clear picture of the trend for increasingly offshore locations, in this section, we present an overview of the various stages involved in the lifecycle of an OWF project, featuring relevant engineering information. This information, coupled with the information contained in the previous subsection, helps build a more comprehensive picture of the practical context and modelling challenges of the current case study.

In Figure 6.2, we can see the lifecycle of an OWF decomposed into four broad stages: Design, Build, Operate and Decommission. In this figure, taken from

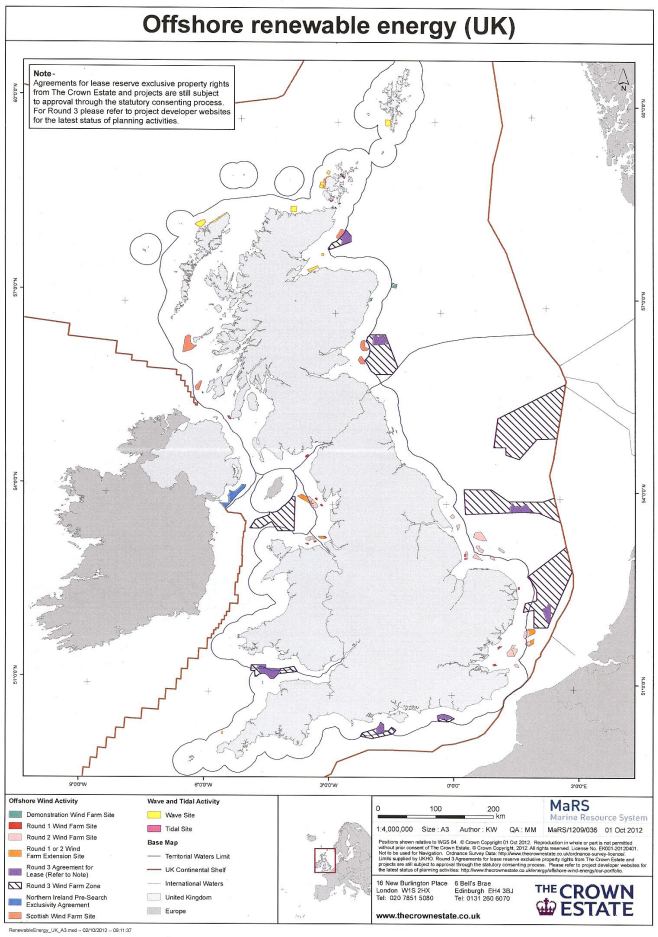

Figure 6.1: RenewableUK's vision for OWFs with hatched area indicating Round 3 zones.

Scottish Enterprise [83], the providers (individuals and/or companies) responsible for each stage are noted in bold at the head of each column. Underneath this, the various resources and processes involved in each corresponding stage are listed.

It is helpful to position the work of the current case study within the scope of this broad framework. As discussed, the current case study concerns the logistics of OWF installation, placing it within the logistics-related elements of the Build stage. The first five elements shown in Figure 6.2 may therefore be excluded as they relate to supply rather than logistics. The work here is instead concerned with processes such as turbine and foundation installation, and with resources such as support vessels and ports. It is worthwhile noting that whilst the case study's

Figure 6.2: Overview of the lifecycle of an OWF, as illustrated in Scottish Enterprise report [83].

decision-support tools address the problems and challenges associated with the installation logistics, chronologically they are in fact applied during the Design stage.

Looking now to provide a clearer picture of the technical specifics involved in installation, we next list the key assets assembled together in the construction of an OWF:

1. Turbines:
   - Each turbine consists of several components: blades, switch gear, transformer and tower.
   - OWFs typically involve tens, if not hundreds, of turbines.
2. Foundations: with structure dependent on the specific ground conditions.

3. Substations: support maintenance activities and provide control over power generation.

4. Interarray cables: connect turbines to the substation.

5. Export cables: transfer power generated to the shore.

In order to provide a low-level understanding of how such assets are combined in the installation of an OWF, the diagram as provided in Figure 6.3 represents the simplest possible schematic, with a single turbine connected to the shore. This schematic provides a minimal illustration of how the key assets combine, offering a baseline against which the more complex full-farm layout in the next figure may be contrasted.

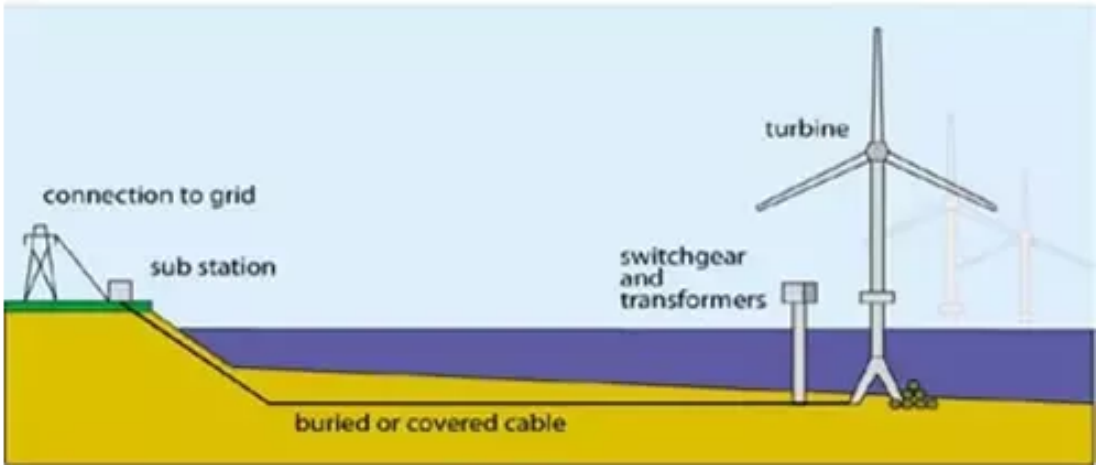

Figure 6.3: A very simple OWF layout, as illustrated in UK DTI report [177].

To illustrate a full wind farm layout, the London Array is currently the largest OWF in the world, with 630MW generation capacity, 175 turbines and two substations, located 20 km away from shore [196]. Figure 6.4 gives an approximate picture of the layout of its turbines, substations and connecting cables. Here, we note that whilst wind farm layout is an important decision problem, it is not within the scope of installation logistics, and as such is considered an input.

During the installation process, a range of different vessels are used to install the assets discussed above. Representative vessel types are shown in Figure 6.5. Each type of vessel has its own specific capabilities with regards to installation

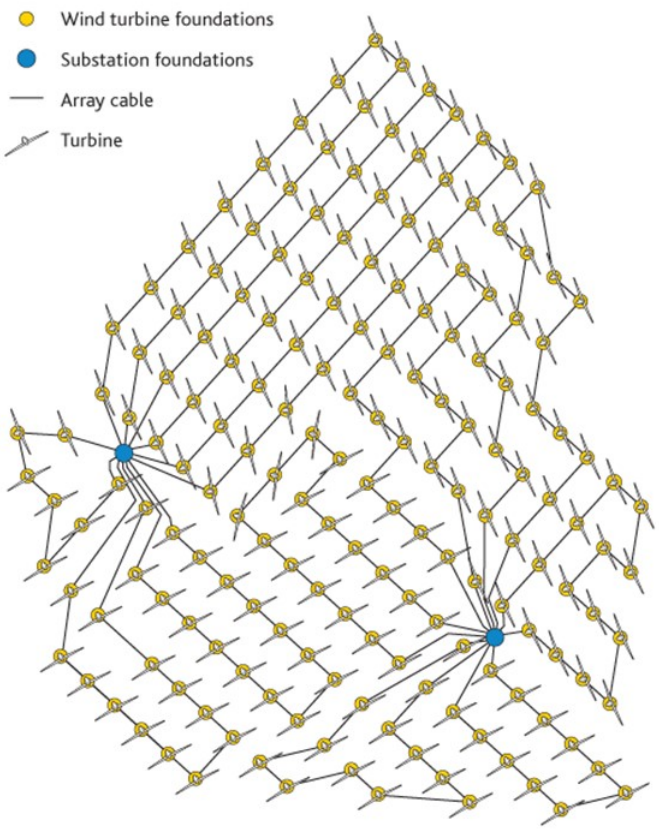

Figure 6.4: Layout of the London Array OWF, as illustrated in Sanchez [196].

operations, as well as its own limitations, in particular with regards to weather conditions, i.e. the maximum permissible wave height and wind speed for safe operation [14]. These constraints complicate the installation process and must be accounted for appropriately in any decision-support tools involved in the planning stage.

To conclude this section, we note that readers interested in a more comprehensive overview of the technical specifics of OWF installation are referred to works by The Crown Estate [84], UK DTI [177] and Barlow et al. [13].

6.3 Case Study Work and Results

Having outlined the broader context of OWF installation logistics, this section turns to the original case study. We review the case study context, the nature

Figure 6.5: Some vessels used in the OWF installation (clockwise from top-left corner): jack-up vessel, heavy lift crane, cable laying vessel, and heavy lift semi-submersible.

of the tools developed and their interrelationships, their practical application to generate insight for the industrial partners, and the impact and recognition of this work. A clear understanding of the original case study is essential to appreciate the modelling and experimentation challenges encountered by the team, and it provides a useful foundation for interpreting the experimentation conducted later in this chapter in relation to the tools developed in the thesis.

To begin, we note the case study was funded by the Strathclyde Technology Innovation Centre (TIC) Low Carbon Power and Energy Grant (£220,221, Grant reference: TIC/LCPE/FI-03), and was carried out by an academic team involving Dr. Kerem Akartunali and Dr. Matthew Revie from the Department of Management Science, and Prof. Sandy Day and Dr. Evangelos Boulougouris from the Department of Naval Architecture, Ocean and Marine Engineering as investigators, and Dr. Euan Barlow and Dr. Diclehan Tezcaner Ozturk as postdoctoral research associates. The project was developed through consultation with industry partners, SSE Renewables, Scottish Power Renewables and Technip Offshore Wind Ltd., and culminated in a model that was iteratively refined and validated through the collaborative process.[1]

[1]Each industrial partner provided support from at least two subject matter experts, with

The model developed encompasses two components, one optimisation-based and the other DES-based, with both implemented in MATLAB, as per the request of the industrial partners given the engineering context of the case study. Each of the components can be used independently, however to provide greatest insight on the problem at hand they are best used jointly, as feedback channels exist between them to faciliate synergistic interaction, as detailed in Barlow et al. [13].

The optimisation component aims to find a schedule of operations that is robust to changes in model inputs and/or external factors such weather conditions or vessel availability. The assignment of fleets of vessels to various tasks is first accomplished based on a greedy algorithm, followed by rolling horizon schedule optimisation, allowing planners to re-optimise the schedule to incorporate changes in weather conditions [13]. Finally, the model employs a robust optimisation framework to ensure the final schedule can accommodate deviations in task durations while still satisfying essential problem constraints such as precedence of operations and deadlines.

The simulation component aims to explore the impact of a range of logistical decisions on key performance measures associated with the OWF installation. The model development broadly followed the facilitated process outlined by Robinson et al. [191], with a series of workshops held to collaboratively decide on model-scope and the level of modelling detail [12]. Modelling developments were proposed by the industrial stakeholders, and integrated or not based on team consideration of their impact on model accuracy and fidelity, and model coding and run time [12].

The simulation model requires an order of operations as an input, whether a default schedule typically adopted in such installations, or an optimised schedule obtained by means of the optimisation sub-model. The decision parameters that can be explored are: the number and operational capabilities of vessels employed in the installation of each asset type, the scheduling of the vessels involved in each asset type, the subset of ports used as operational bases, and as mentioned above, the proposed installation schedule for each of the stages involved [14].

---

titles such as Head of Installation and Logistics and Head of Offshore Marine and Construction Engineering [12], demonstrating significant industrial involvement.

To provide an overview of the activities included in the simulation model's operation, two flowcharts from technical report [12] are included in Figure 6.6. The first flowchart (a) presents the flow of operations involved in the installation of a single OWF asset, whilst the second flowchart (b) provides a high-level overview of the relationships between the different OWF asset installation processes.

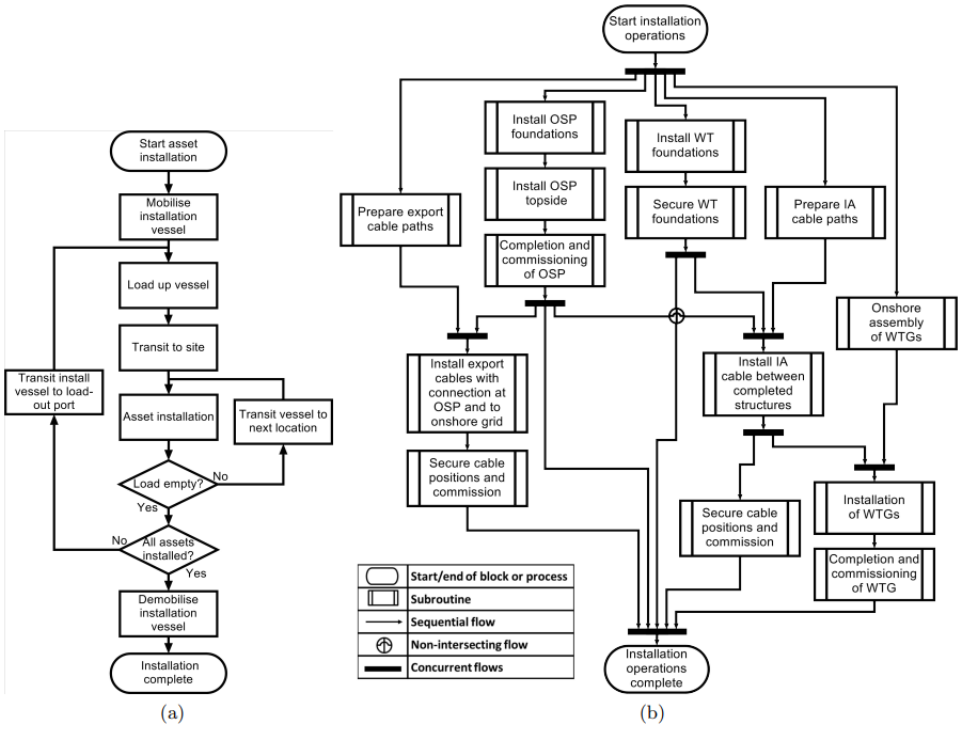

Figure 6.6: Flowcharts illustrating the logical flow of the simulated activities from Barlow et al. [12].

The simulation model also features a synthetic weather time-series model to generate typical weather patterns likely to be encountered in practice [12]. Important weather variables modelled in the model include wave height and wind speed. This allows the impact of weather conditions to be evaluated in relation to scheduling choices, and for seasonal variations to inform the start date for potential OWF installation projects.

Finally, as might be expected from the nature of the thesis, we note the work carried out later in this chapter relates only to the simulation model, and not the optimisation model. As such, only the experimentation originally carried out with

the simulation model, and not the optimisation model, will now be discussed.

Examining the original simulation experimentation, a few important aspects are noted. Eighteen simulation scenarios were explored, with a total of 1,000 replications collected for each of these scenarios. Data was collected on two key performance measures: mean project cost and duration. The generation of this volume of data required around 3 hours per simulation scenario, resulting in a total experimentation time of around 1.5 working weeks. This information is summarised in Table 6.2 below.

Original experimentation	
Number of simulation scenarios	18
Number of replications	1,000
Total time for experimentation	1.5 working weeks
Simulation performance measures	Mean cost and duration

Table 6.2: Summary of the experimentation conducted in the original case study.

Finally, we discuss the results and practical impact of the work carried out. The model developed in the course of this project has already been used in the logistical planning of the Beatrice OWF Installation Project, a 600MW OWF located off the North-East coast of the UK, scheduled for installation over 2017–2019. SSE Renewables, one of the industrial partners, estimate that the use of the model will deliver a saving of tens of millions of pounds from the cost of a typical OWF installation [13]. The work was also shortlisted for the recent EURO Excellence in Practice Award 2016.

6.4 Using the Case Study Simulation Model

Having introduced and discussed the original case study, its simulation model, and the experimentation conducted with it, this section turns to how the simulation model will be used to apply and evaluate the methods developed in the thesis. This work forms the core contribution of the chapter, and is an important component of the broader evaluation of the tools and techniques developed in this thesis.

We begin with some general comments on the preparatory work and early experimentation undertaken to better understand the simulation model and make necessary adaptations. We next outline, in more technical terms, the experimental design and error measurement strategy for the numerical studies of this chapter: first for evaluating the EB suitability heuristic of Chapter 4, and second for evaluating the adapted DS EB estimator of Chapter 5. This is followed by a discussion of the results and their implications. Finally, we discuss this chapter's contribution to the work of the thesis in the concluding section.

6.4.1 Preparatory Work and General Considerations

This subsection sets out the preparatory work and general experimental considerations underlying the numerical studies of this chapter. It describes the adaptations made to the simulation model, the preliminary experimentation undertaken to better understand its behaviour, and considerations regarding how the original experimentation broadly compares with the experimentation undertaken here. These foundations support the more detailed studies presented later in the chapter.

As noted in Barlow [12], the original case study's simulation model included an Excel-based interface, enabling the industrial partners to explore and execute it without accessing its MATLAB-based core. However, because this interface was configured specifically for their needs, it was not initially well-suited to our experimental requirements. Accordingly, several adaptations to the simulation model were required.

Adaptations to the model were required along two dimensions: first, to increase the granularity of the outputs, and second, to expand the scale of experimentation possible. Regarding the first, the model's Excel interface originally provided only visualisations and summary statistics of the performance measures of interest. While this facilitated client use and comprehension, it did not provide access to the raw data needed for experimental evaluation, and so adaptation was required. Regarding the second, some simple modifications were made to the simulation code to allow greater flexibility in the scenarios examined and the number of replications executed. Overall, these adaptations offered the flexibility required to undertake the experimentation conducted later in this chapter.

After making the necessary adaptations to the simulation model, exploratory experimentation and analysis were undertaken to increase understanding around its behaviour. This involved examining the convergence and observed distributions of the relevant performance measures, and gaining a clearer view of the model's observed response surface. Each of these topics are discussed in turn.

The first stage of experimentation focused on convergence. Specifically, we examined the degree of convergence achieved in the performance measures of interest (project duration and cost) during the original experimentation. The aim was to confirm whether $1,000$ replications provided sufficient accuracy, for our purposes, in estimating the true mean performance measures.

Given the project context, with its large budget, significant real-world investment and potential cost savings, and the substantial amount of experimentation conducted ($1,000$ replications per scenario), it was expected that a high degree of convergence was both important to and likely to have been achieved by the original research team. Discussion with the original researchers, together with preliminary experimentation, confirmed this to be true.[2]

To verify this formally, exploratory experiments were conducted to assess the inherent variability of the simulation model's response data for both project cost and duration across all eighteen original simulation scenarios. The five scenario-measure pairings with the highest sample variances were identified, and additional experimentation was then carried out on these pairings to assess the extent to which $1,000$ replications achieved sufficient convergence. Negligible gains were observed beyond $1,000$ replications, and as such, the long-run mean estimates based on this replication level were retained as our approximations to the true underlying mean response values.

Next, preliminary experimentation was carried out to explore the nature of the simulation model's response data. This revealed that the response data tended not to be normally (or even approximately normally) distributed, but were instead

[2]Personal communication with Dr Kerem Akartunali and Dr Euan Barlow.

positively skewed. Discussion with the original research team[3] confirmed the predominance of positively skewed response data throughout their experimentation with the simulation model. This creates a stringent evaluation environment for the methods developed in the thesis, all of which assume normally distributed data. Although the assumption of normality cannot be removed, applying the methods in such a context offers a demanding test of their robustness. Should they perform well despite this departure from assumptions, such performance would provide meaningful reassurance regarding their potential use in broader simulation practice.

Preliminary experimentation also revealed the presence of thresholds in the model's response surface. This observation was likewise confirmed by the original research team,[4] who noted, in particular, that relatively small changes in the specifics of the installation scenario could result in very large cost savings. This further contributes to the creation of a challenging test environment, particularly for the adapted DS EB estimator developed in Chapter 5. As discussed throughout the thesis, EB methods rely on similarity between populations, and so can to perform poorly in situations featuring substantial population heterogeneity. The adapted DS EB estimator was specifically developed to offer greater robustness to such population heterogeneity through its weighting mechanism based on the observed similarity between populations. Accordingly, the OWF simulation model provides a particularly informative test case for evaluating the adapted DS EB estimator, alongside its frequentist and original DS EB counterparts.

Having discussed the simulation model adaptations and preliminary explorations, we now turn to the broad requirements for the experimental design of the numerical studies of this chapter. Specifically, we comment on how the scale of the original experimentation compares to the scale of experimentation required for the evaluations conducted here. More technical aspects of the experimental design are deferred until the following subsections.

In the original experimentation, just eighteen simulation scenarios were deemed of interest, and a substantial $1,000$ replications were conducted of this experimental

[3]Personal communication with Dr Kerem Akartunali and Dr Euan Barlow.
[4]Personal communication with Dr Kerem Akartunali and Dr Euan Barlow.

design. In general terms this represent a relatively small number of scenarios and a very large sample size, and is unlikely to be representative of typical simulation practice.

This arrangement is understandable in context. Extensive industry collaboration provided expert knowledge for narrowing the scenario set, while the long timescales and high financial outlay associated with OWF installation justified a large computational budget. However, such resources are not usually available to support simulation experimentation more broadly. Practitioners are often unable to restrict the set of interesting scenarios to such a manageable number, and they rarely have access to such extensive computational budgets due to cost and/or time constraints.

Moreover, the eighteen scenarios correspond to a single experimental configuration (in the terminology of Section 3.4). To adequately assess the methods developed in the thesis, the experimental design for this chapter must therefore span a range of configurations. Ideally, these should vary both in the number of scenarios and in the sample sizes employed, so that the experimental conditions better reflect the diversity encountered in general simulation practice. Experimental design decisions taken to ensure this representativeness are discussed in the following subsection.

6.4.2 Experimental Design

In this section, we outline the experimental design of the numerical study that forms the main contribution of this chapter. This includes the presentation and discussion of both the strategic aspects (such as, the number of scenarios) and the tactical aspects (such as, the size of samples) of the experimental design. The purpose of the experimentation is to explore the extent to which the methods developed in the thesis can support more effective and efficient DES experimentation. To this end, a range of experimental configurations are examined, featuring different numbers of scenarios and different sample sizes, to gain an understanding of how these methods may perform in real simulation practice.

To create an experimental design appropriate for the numerical evaluation at hand, the approach adopted mirrors that used in the numerical study presented in Chap-

ter 3. A large-scale "master" data set was generated, from which different experimental configurations were created as smaller subsets. As such, we first discuss the creation of the master data set, then its use in forming a range of different experimental configurations intending to represent general DES practice.

As highlighted previously, the 18 simulation scenarios explored in the original experimentation are insufficient for our experimental purposes. Therefore, an additional 82 simulation scenarios were identified and added to the original 18 simulation scenarios, yielding a total of 100 simulation scenarios. These additional simulation scenarios were determined through sensible interpolation and extrapolation of the original design space (with experimental factors listed in Section 6.3), informed by discussion with the original research team.[5] Also as noted in the previous subsection, preliminary work revealed $1,000$ replications as sufficient to guarantee convergence of the performance measures of interest, specifying the second dimension of the master data set. A large-scale data set comprising 100 simulation scenarios, each executed over $1,000$ replications was generated to form the master data set, as summarised in Table 6.3:

Master data set	
Number of scenarios k	100
Sample size n	1,000

Table 6.3: Structure of the master data set for the evaluation of the methods developed in the thesis.

Before proceeding, we briefly discuss the computational effort required to generate a master data set of this scale. As noted in the previous subsection, the original experimentation with 18 scenarios required 1.5 working weeks of computer time, owing to the size of the model and the large number of replications. Extending the experimental design to incorporate 100 simulation scenarios leads to an approximate five-fold increase in computational time. To avoid practical difficulties, the required experimentation was executed in parallel on multiple machines (100 desktop computers) using overnight access to a university computer laboratory. The computation was carried out on standard desktop PCs (Intel Core i5, 2.7GHz,

[5]Personal communication with Dr Kerem Akartunali and Dr Euan Barlow.

4GB RAM) and using MATLAB (2014a).

Decisions on the experimental configurations to be extracted from the master data set were guided largely by the experimentation conducted earlier in the thesis, informed in particular by discussions in Sections 3.4, 3.5, 4.2, 4.3.1, 4.4, 5.5.1 and 5.5.4. The configurations were selected so as to provide a diverse range of experimental environments, as representative as possible of real DES practice. Details of the configurations extracted from the master data set are presented in Tables 6.4 (showing the factors and levels) and 6.5 (listing the configurations comprising the experimental design).

Factor	Levels
Number of scenarios k	5, 10, 20, 50, 100
Sample size n	3, 5, 10, 20

Table 6.4: Factors and levels characterising the experimental design.

Although strict confidentiality restrictions apply to the original case study, it is helpful to clarify, in general terms, how subsets of model scenarios were selected. Any subset of size k was chosen so as to be spaced as uniformly as possible throughout the design space, subject to constraints such as the exclusion of practically impossible scenarios. For example, in Configuration 9, with $k = 20$ simulated scenarios, the selected scenarios would be distributed as evenly as feasible across the design space.

Having discussed the experimental design and data-generating process for the numerical studies of the chapter, the following subsections consider how this experimental data has been used to evaluate the methods developed in the thesis.

6.4.3 Evaluation of Methods

The purpose of this section is to apply the methods developed in the thesis to experimentation with the case study simulation model, thereby enabling an evaluation of their applicability practice. The section is divided into two subsections. In the first, we examine the EB suitability heuristic of Chapter 4, applying it across a

Experimental Configurations	Number of Populations	Sample Size
Configuration 1	5	3
Configuration 2	5	5
Configuration 3	5	10
Configuration 4	5	20
Configuration 5	10	3
Configuration 6	10	5
Configuration 7	10	10
Configuration 8	10	20
Configuration 9	20	3
Configuration 10	20	5
Configuration 11	20	10
Configuration 12	20	20
Configuration 13	50	3
Configuration 14	50	5
Configuration 15	50	10
Configuration 16	50	20
Configuration 17	100	3
Configuration 18	100	5
Configuration 19	100	10
Configuration 20	100	20

Table 6.5: Range of experimental configurations included in master data set.

range of experimental configurations to evaluate its performance in predicting the EB suitability of a given data analytic situation. In the second, we compare the adapted DS EB estimator of Chapter 5, the original DS EB estimator [260] and the frequentist sample mean in terms of their ability to estimate the mean performance measures, project cost and duration, over a range of different experimental configurations.

6.4.3.1 Evaluation of EB Suitability Heuristic

In this subsection, we broaden our evaluation of the EB suitability heuristic developed in Chapter 4 by applying it to the experimental data generated using the OWF installation simulation model. The purpose of this application is to assess how accurately the tool classifies configurations as EB-suitable or not, using the error-ratio labels as the empirical benchmark in a practical DES context.

To establish these benchmark labels, EB and frequentist estimates of the mean

performance measures were obtained for each scenario in each experimental configuration. EB estimates were generated using Zhou's DS estimator [260] (as given in Equation (3.20)), and frequentist estimates using the standard sample mean (as given in Equation (3.21)). Using the long-run sample means based on $1,000$ replications, the corresponding estimation errors were calculated as:

$$EB_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2 \quad \text{and} \quad Freq_{error} = \sum_{i=1}^{k} \left(\hat{\theta}_{Freq,i} - \theta_i\right)^2,$$

where $\hat{\theta}_{EB,i}$ and $\hat{\theta}_{Freq,i}$ denote the EB and frequentist estimates for scenario i, and θ_i the corresponding long-run mean.

A maximum practitioner-level sample size of 20 and $1,000$ replications per configuration yielded 50 replications of these estimate pair calculations. Averaging the error values across the 50 replications mitigated stochastic variability, and the relative empirical performance was then summarised using the ratio:

$$EB/Freq_{error} = (\text{mean}EB_{error})/(\text{mean}Freq_{error}),$$

Here, values below one indicated EB-superior performance, and values above one indicated frequentist-superior performance. These labels serve as the benchmark against which the EB suitability heuristic's classifications are evaluated.

To apply the EB suitability heuristic the RV statistic was calculated for each configuration and each replication using:

$$RV = \left(\frac{k}{n}\right)\frac{SSE}{SSA} = \left(\frac{k}{n}\right)\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2}{\sum_{i=1}^{k} n\left(\bar{x}_i - \bar{x}\right)^2}.$$

The RV values, together with k and n values, provide the predictor variable values required by the tool. Overall, this produces 50 predictions for each configuration, allowing accuracy, sensitivity and specificity to be calculated over 50 independent replications per configuration.

For convenience, we also recall that the EB suitability heuristic is a logistic regression classifier constructed in Chapter 4 using the transformed variables $log(RV)$,

$log(k)$, and $log(n)$. The estimated coefficient vector:

$$\hat{\beta} = (-3.3123, 0.2114, 1.6967, 0.2036),$$

together with the decision threshold $\gamma = 0.69$, determines whether a configuration is classified as EB-suitable or not.

Applying the EB suitability heuristic to the OWF simulation model data yielded an overall accuracy rate of 0.7622, a sensitivity of 0.7815, and a specificity of 0.7453. Table 6.6 summarises these results and, for context, includes the corresponding accuracy, sensitivity and specificity obtained on the Monte Carlo simulated test set of Chapter 4.

Data source	Overall accuracy	Sensitivity	Specificity
Monte Carlo test set data	0.8032	0.8055	0.7953
OWF simulation model data	0.7622	0.7815	0.7453

Table 6.6: Performance of the EB suitability heuristic on OWF simulation data, with results from the Monte Carlo test set of Chapter 4 provided for comparison.

These findings, together with those obtained for the adapted DS EB estimator in the next subsection, are discussed in Subsection 6.4.4.

6.4.3.2 Evaluation of Adapted DS EB estimator

In this subsection, we broaden our evaluation of the adapted DS EB estimator developed in Chapter 5 by applying it, alongside the original DS EB estimator [260] and the frequentist sample mean, to the experimental data generated using the industrial OWF installation simulation model. The purpose of this application is to assess the relative performance of these approaches in estimating the mean performance measures of interest, namely project cost and duration, in a practical DES context.

Proceeding similarly to the previous subsection, the first step is to compute, for each estimator, the estimates of the mean performance measures for each scenario and experimental configuration. Using these estimates together with the long-run sample means based on $1,000$ replications, two forms of estimation error were cal-

culated: overall squared error and maximum absolute error.

We began by calculating the sum of squared errors for each estimator as follows:

$$wEB_{SSE} = \sum_{i=1}^{k} \left(\hat{\theta}_{wEB,i} - \theta_i\right)^2, \quad EB_{SSE} = \sum_{i=1}^{k} \left(\hat{\theta}_{EB,i} - \theta_i\right)^2$$
$$\text{and } \; Freq_{SSE} = \sum_{i=1}^{k} \left(\hat{\theta}_{Freq,i} - \theta_i\right)^2,$$

where $\hat{\theta}_{wEB,i}$ denotes the adapted DS EB estimator of Chapter 5, $\hat{\theta}_{EB,i}$ Zhao's original DS EB estimator [260], and $\hat{\theta}_{Freq,i}$ the frequentist sample mean.

As before, the mean of these error values, taken across 50 replications, was calculated to minimise the impact of stochastic variability, and the overall error ratios constructed, as follows:

$$wEB/EB_{SSE} = \text{mean}wEB_{SSE}/\text{mean}EB_{SSE} \text{ and}$$
$$wEB/Freq_{SSE} = \text{mean}wEB_{SSE}/\text{mean}Freq_{SSE}.$$

Following this, the maximum absolute errors for each of the estimators were calculated using the formulas:

$$wEB_{max} = \max_i \left|\hat{\theta}_{wEB,i} - \theta_i\right|, \quad EB_{max} = \max_i \left|\hat{\theta}_{EB,i} - \theta_i\right|$$
$$\text{and } \; Freq_{max} = \max_i \left|\hat{\theta}_{Freq,i} - \theta_i\right|.$$

Here, wEB_{max} denotes the maximum absolute error for the adapted DS EB estimator, EB_{max} for Zhao's DS EB estimator [260], and $Freq_{max}$ for the frequentist sample mean.

Again, taking the means of these values over 50 replications, the relative maximum error ratios are as follows:

$$wEB/EB_{max} = \text{mean}wEB_{max}/\text{mean}EB_{max} \text{ and}$$
$$wEB/Freq_{max} = \text{mean}wEB_{max}/\text{mean}Freq_{max}.$$

This approach yields four error ratios per experimental configuration: wEB/EB_{SSE},

$wEB/Freq_{SSE}$, wEB/EB_{max} and $wEB/Freq_{max}$. These compare the adapted DS EB estimator to both the original DS EB estimator and the frequentist sample mean, firstly in terms of the overall squared error and secondly in terms of the maximum absolute error.

The results, averaged across experimental configurations, are presented in Table 6.7, together with the corresponding Monte Carlo–based results from Chapter 5 for ease of comparison:

	wEB/EB_{SSE}	$wEB/Freq_{SSE}$	wEB/EB_{max}	$wEB/Freq_{max}$
Monte Carlo: standard	0.8910	NA	0.9044	NA
Monte Carlo: with outlier	0.7529	NA	0.9085	NA
OWF simulation data	0.8253	0.7811	0.9161	0.9248

Table 6.7: Relative performance of the adapted DS EB estimator, original DS EB estimator, and frequentist sample mean under the OWF simulation model, with Monte Carlo results from Chapter 5 included for comparison.

These results are considered further in the following subsection.

6.4.4 Implications of Results

The broad aim of this chapter was to apply the methods developed in the thesis to experimentation with a real DES model, enabling evaluation of their applicability for real DES practice. While the Monte Carlo based evaluations of previous chapters demonstrate the statistical efficiency of the methods, the evaluation of their real-world applicability requires assessment on a genuine simulation model. We therefore reflect upon the results obtained using the OWF installation simulation model [13].

The results for the EB suitability heuristic of Chapter 4 show only a modest reduction in performance when moving from Monte Carlo simulated data to the industrial OWF model. The overall accuracy decreased from 0.8032 to 0.7622, with sensitivity falling from 0.8055 to 0.7815 and specificity from 0.7953 to 0.7453. Given the transition from controlled Monte Carlo environments to a large, structurally complex and positively skewed DES model, a far more substantial degradation in performance could reasonably have been expected. The observed consistency

across all measures therefore represents a strong outcome, suggesting that the EB suitability heuristic transfers effectively to a practical DES context.

Turning to the estimation procedures, the comparisons between the adapted DS EB estimator of Chapter 5 and Zhao's original DS EB estimator [260] show a similar pattern. The wEB/EB_{SSE} ratio for the OWF data lies between the corresponding values for the standard Monte Carlo data and the Monte Carlo with an outlier data. This is intuitively reasonable: the OWF data represent a level of heterogeneity that lies between these two designed extremes. The wEB/EB_{max} ratios show very little separation at all, with the three values clustering around 90%, the OWF value being only marginally higher.

In comparing the adapted DS EB estimator with the standard frequentist sample mean, its first such comparison in the thesis, we observe a $wEB/Freq_{SSE}$ of 0.7811 and a $wEB/Freq_{max}$ of 0.9248. As may be anticipated, the improvement is more pronounced in terms of overall squared error than for the maximum error, given that the latter depends on a single scenario rather than an averaged performance. Nevertheless, both values indicate a favourable comparative performance for the adapted estimator.

It is also worth noting that the OWF model exhibits two characteristics that, in principle, pose challenges for the methods developed in this thesis. First, the performance measures are positively skewed, whereas both the EB suitability heuristic and the adapted DS EB estimator were built around approximate normality assumptions. Second, the model's threshold-driven behaviour introduces marked heterogeneity across scenarios, a feature that is particularly demanding for EB-based shrinkage estimators. Despite these departures from the ideal modelling assumptions, all methods performed well in this application. While it is not unusual for well-behaved statistical procedures to show some resilience to such deviations, the degree to which both the EB suitability heuristic and the adapted DS EB estimator maintained performance is nonetheless notable, and adds weight to the case for their robustness in practical DES settings.

The empirical findings also feed directly into the broader dimensions of real-world applicability introduced earlier in the chapter.

In terms of accessibility, the methods require no subjective pooling decisions and their closed-form expressions make them straightforward to implement. Although the mathematical derivations may lie beyond the immediate reach of some practitioners, the underlying intuition is accessible, and once embedded in software, the methods could be applied without specialist statistical expertise.

With respect to robustness, the OWF model presents several departures from idealised EB modelling assumptions, including skewness, irregular response structures and threshold effects. The strong empirical performance of the methods under these conditions provides clear evidence of robustness, reinforcing the conclusions drawn from the earlier Monte Carlo-based studies.

Regarding usability, the successful application of the methods to a large, industrial-scale DES model indicates that they are indeed feasible for problems of realistic size and complexity. This offers reassurance that the methods can be operationalised in practical settings rather than only within controlled experimental studies.

Finally, in terms of practical value, confidentiality constraints prevent explicit reporting of improved OWF schedules. Nevertheless, the increased statistical efficiency demonstrated here suggests practical benefits: enhanced search capacity, reduced computational burden, and potentially better informed experimentation decisions. These advantages indicate that the methods have tangible potential value for real DES practice.

These implications lead directly to the chapter's concluding remarks.

6.5 Conclusions

The EB suitability heuristic maintained a high level of classification accuracy, with only a modest decrease relative to its performance on the independent test set of Chapter 4. The adapted DS EB estimator demonstrated strong comparative performance against both Zhao's DS estimator and the frequentist sample mean, particularly in terms of overall squared error.

Taken together, these findings provide a sound evidential basis for the broader claims made throughout the thesis: that appropriately adapted EB methods and EB-informed decision support can enable the practitioner to achieve meaningful gains in statistical efficiency within DES experimentation, and that these gains can be realised in practice, beyond the controlled conditions of artificial numerical studies.

The implications for real-world accessibility, robustness, usability and practical value further reinforce the potential of the methods developed, and motivate the concluding reflections and future research directions presented in Chapter 7.

The purpose of the current chapter was to assess the real-world applicability of the tools developed in Chapters 4 and 5 by applying them to an industrial-scale DES case study concerning offshore wind farm (OWF) installation. As such, this chapter addresses Research Objectives R3(c) and R4(c), which concern the evaluation of the methods developed in R3 and R4 in a genuine DES experimentation context. Through providing an additional, complementary evaluation environment, it also contributes to Research Objective R5.

To this end, both the EB suitability heuristic and the adapted DS EB estimator were applied to the OWF installation simulation model. This case study departs in several important respects from the controlled Monte Carlo environments used in earlier chapters, exhibiting non-normality, skewness, irregular response structures and threshold effects. Despite these departures from underlying modelling assumptions, the behaviour of both methods was broadly consistent with expectations formed from the earlier numerical investigations.

The EB suitability heuristic maintained a high level of classification accuracy, with only a modest reduction relative to its out-of-sample performance observed in Chapter 4. In parallel, the adapted DS EB estimator demonstrated strong comparative performance against both Zhao's original DS EB estimator and the frequentist sample mean, particularly in terms of overall squared error. These findings provide evidence that the methodological advantages identified in earlier chapters are not confined to idealised experimental settings.

Taken together, the results of this chapter demonstrate that the tools developed in this thesis can be operationalised within a realistic DES experimentation environment. The successful application of both the EB decision support and the adapted EB estimation to a large, industrial-scale model provides a critical validation step, illustrating that gains in statistical efficiency can be realised in practice, beyond the confines of artificial numerical studies.

The implications for accessibility, robustness, usability and practical value further reinforce the contribution of the proposed methods, and motivate the broader reflections and future research directions presented in Chapter 7.

Chapter 7

Conclusion

This chapter draws together the findings of the thesis and reflects on their collective implications. The overarching research aim and related research objectives are revisited, and the extent to which they have been addressed through the preceding chapters is considered. The chapter then situates the contributions of the work in relation to the existing literature and DES experimentation practice, before acknowledging key limitations and outlining directions for future research.

7.1 Overview and Research Aims

This section revisits the motivation underpinning the research, and the aims and objectives defined at the start of the thesis. The purpose here is not to restate these elements verbatim, but rather to re-establish the rationale for investigating the use of EB within DES experimentation, and to recall the specific objectives that structured the programme of work. Doing so provides a reference point for the synthesis and reflection that follow in the later sections of the chapter.

7.1.1 Concluding Overview

DES is a well-established methodology for analysing complex real-world systems, but its effective use in practice is often constrained by computational concerns. As both simulation models and the questions posed of them grow in scale and complexity, practitioners are frequently required to draw inference from limited numbers of simulation runs. While computational resources continue to improve, they do not scale commensurately with this growth in experimental scope, due in

part to the curse of dimensionality. Conventional approaches to the analysis of simulation data typically treat such scenarios independently, even where strong structural similarity exists across large collections of model scenarios, potentially leading to inefficient use of available data.

EB methods offer a principled statistical framework for borrowing strength across related estimation problems and for trading variance against bias through shrinkage. Their successful application in fields such as genomics and brain imaging, where inference is conducted across many parallel estimation problems, each with limited data, implies a natural conceptual alignment with DES experimentation. Despite this apparent compatibility, EB methods have received little explicit attention in the simulation literature, and practical guidance on their use in DES remains essentially non-existent.

Motivated by this gap, the research undertaken in this thesis sought to investigate the utility of EB methods within the context of DES model experimentation. Rather than advocating their indiscriminate application, the work was concerned with establishing whether material gains in statistical efficiency are achievable in practice, identifying the conditions under which such gains arise, and developing tools and procedures to support its operationalisation in practice.

7.1.2 Research Aims and Objectives

This subsection revisits the research aims and objectives defined at the beginning of the thesis. Its purpose is to provide a concise reference point for the synthesis of and reflections upon the findings presented in the remainder of the chapter.

The overarching aim of this research was to investigate the utility of adopting an EB approach to DES model experimentation, considering both potential gains in statistical efficiency and the extent to which such methods can be operationalised for use in practice. To address this aim in a structured manner, a set of research objectives was defined to guide the work undertaken in the thesis.

The first objective concerned positioning the research within the existing literature, through a review of DES experimentation approaches and EB inference method-

ology, with the aim of identifying a promising and tractable overlap between the two fields. Building on this foundation, the second objective involved a preliminary computational study to provide proof of concept for the application of EB in DES experimentation, while also identifying key challenges, including data heterogeneity and bias, and highlighting opportunities for practical operationalisation.

The remaining objectives focused on methodological development and implementation. The third objective addressed the design of a decision support tool to guide practitioners on the suitability of EB in a given DES experimentation context. The fourth objective investigated the development of adapted EB procedures intended to mitigate the challenges identified earlier and to ensure robustness in practice. A final objective concerned the supporting mechanics required to deliver this programme of work, including the selection of suitable DES models and simulation environments, and the development of appropriate error measurement strategies for comparative evaluation.

Having reviewed the fundamental research objectives of the project, we next discuss how they have been fulfilled through the content of the thesis.

7.2 Synthesis of Findings Relative to Research Objectives

This section synthesises the main findings of the thesis relative to the research objectives, drawing together results from across the preceding chapters.

In relation to Research Objective R1, the literature review established a clear asymmetry between the maturity of EB methodology and its uptake within DES model experimentation. On the DES side, the review highlighted the wide range of experimentation goals and approaches, as well as their uneven development across Barton's taxonomy [16], with early-stage, understanding-oriented experimentation remaining comparatively underdeveloped in methodological terms. On the EB side, a mature body of methodology was identified. Despite this, the review confirmed that EB has gained little traction in DES experimentation, where the very limited number of existing applications tend to be narrow in scope, and

insufficiently elaborated. Taken together, these findings suggested a non-trivial but largely unexplored overlap between the challenges of DES experimentation, such as parallel scenarios and limited computational budgets, and the types of problems for which EB methods have been developed. This work thereby directly motivated the empirical investigation pursued in Chapter 3.

In relation to R2, the numerical study presented in Chapter 3 provides clear proof of concept for the application of EB to DES model experimentation. Across a wide range of experimental configurations, EB was shown to deliver improvements in estimation performance relative to frequentist approaches, confirming that the intuitive appeal of EB in this context is supported empirically. However, the study also demonstrated that EB does not uniformly outperform the frequentist approach, with neutral or inferior performance observed in a non-negligible subset of settings. This variability complicates naive interpretation and highlights the potential risks associated with indiscriminate pooling.

Closer examination of the results revealed a clear set of principles governing EB performance in DES experimentation. Gains were most pronounced when sample sizes were small, consistent with the increased influence of sharing information under limited data. Structural features of the experimental design were also important: dense collections of closely related scenarios tended to favour EB, while sparse or non-contiguous configurations introduced bias and worsened performance. Taken together, these findings show that EB can offer substantial benefits in DES experimentation, but only in specific settings, thereby motivating the development of the EB suitability heuristic and adapted DS EB estimator in subsequent chapters.

Research Objective R3 concerned the development of principled decision support to guide practitioners on the a priori suitability of EB methods within a given DES experimentation context. This objective arose directly from the proof of concept findings of Chapter 3, which demonstrated that while EB can deliver substantial gains in statistical efficiency, such gains are highly context-dependent and cannot be assumed a priori. The observed variability in EB performance across experimental structures highlighted the need for systematic guidance to inform whether or not EB should be used in a given setting.

To address this need, Chapter 4 developed the EB suitability heuristic based on the observable characteristics of experimental data sets, rather than subjective judgement. The tool links measurements of the structural features of the experimental environment to expected relative performance, using the RV statistic as a summary linking between scenario similarity to within scenario uncertainty. Embedded within a formal classification model, this approach provides an explicit and reproducible mechanism for anticipating whether EB is likely to outperform the frequentist benchmark in terms of overall squared error.

Research Objective R4 addressed the development of EB estimation procedures that remain reliable in the presence of heterogeneous and structurally challenging DES experimentation contexts. This objective was motivated by deeper inspection of the numerical results underpinning R2 and R3, which showed that although EB can reduce overall error through variance reduction, it may also introduce substantial bias and inflated individual errors when similarity assumptions are weak or violated. For EB to be operationally viable in DES practice, it must therefore offer not only average efficiency gains, but also protection against poor performance at the level of individual scenarios.

Chapter 5 addressed this challenge by developing the adapted DS EB estimator designed to remain reliable when used in heterogeneous DES experimentation contexts. Building on Zhao's [260] original framework, the adapted DS EB estimator introduces an individual weighted shrinkage target through a similarity-sensitive weighting mechanism, enabling shrinkage to be moderated when populations deviate from the dominant structure. These modifications are explicitly aimed at preserving the variance-reduction benefits of EB, whilst limiting bias and controlling large individual errors in unfavourable settings.

Taken together, the findings of Research Objectives R2 to R4 establish that the application of EB methods to DES experimentation is a conditional and risk-aware undertaking, rather than a universally beneficial alternative to frequentist estimation. The numerical investigations demonstrated that EB performance depends critically on both the structural characteristics of the observed experimental data and the behaviour of the estimator under population heterogeneity. This naturally

partitions the problem into two distinct but complementary challenges: determining whether EB is suitable at the analysis stage, given the observed structure of the experimental data, and ensuring that EB estimation remains reliable when similarity assumptions are imperfect. Research Objective R3 addressed the former by providing data-driven guidance on EB suitability prior to estimation, whilst Research Objective R4 addresses the latter through the development of estimation procedures that mitigate adverse behaviour when EB is applied. Research Objective R5 then underpins this programme by establishing a coherent experimental and evaluative framework within which these complementary contributions can be developed and interpreted.

Research Objective R5 concerned the identification and development of the methodological foundations required to support the programme of work undertaken in pursuit of Research Objectives R1 to R4. Rather than introducing a distinct substantive contribution, this objective focused on establishing the experimental, computational, and evaluative infrastructure necessary to ensure that the numerical and applied investigations conducted throughout the thesis were coherent and comparable.

Three complementary experimental environments were employed for this purpose. The proof of concept investigations of Chapter 3 were based on a well-established "toy" DES model, enabling controlled exploration of EB behaviour while retaining a recognisable DES structure. For the development and evaluation of the EB suitability heuristic in Chapter 4, and the adapted DS EB estimator in Chapter 5, large-scale Monte Carlo data sets, collectively functioning as experimental testbeds, were constructed. These testbeds were designed to reflect key structural features of DES experimentation, such as parallel scenarios, limited sampling, and scenario heterogeneity, while permitting systematic variation across design dimensions. Finally, Chapter 6 employed an industrial-scale DES case study, providing a realistic applied environment in which the practical implications of the methods could be examined.

A consistent error measurement framework was adopted across Chapters 3 to 6 to support meaningful comparison between the inferential approaches. Performance was assessed using both overall and individual criteria, specifically the sum

of squared error across scenarios, and the maximum absolute error incurred on any individual scenario. This dual perspective reflects the central bias–variance trade-off inherent in EB methods and ensured that comparisons between frequentist, EB, and adapted EB approaches were aligned with concerns relevant to DES experimentation practice.

Viewed in retrospect, this experimental and evaluative framework reflects the fundamental tension underlying the application of EB methods in DES experimentation. The programme of work developed in this thesis can be understood as managing the balance between variance reduction through information pooling and the risk of bias arising from imperfect similarity. The contributions of R3 and R4 address these complementary aspects directly, while R5 provides the coherent foundation within which such trade-offs can be explored and assessed in a manner relevant to both methodological development and practical DES experimentation.

7.3 Relationship to Literature and Practice

This section situates the contributions of the thesis within the existing literature and DES experimentation practice. The discussion focuses on how the thesis extends current understanding at the intersection of EB methodology and DES experimentation, and on the practical implications of these contributions for simulation analysts. Rather than advancing new theoretical results in EB methodology, the emphasis is on clarifying the role, scope, and limitations of EB methods when applied in DES experimentation contexts.

Implications for the Literature

From a literature perspective, the thesis makes a contribution at the intersection of three largely disconnected bodies of work: EB methodology, DES experimentation, and methodological guidance for statistical inference in simulation analysis.

First, with respect to the use of EB in DES experimentation, the thesis helps to fill a clear gap identified in the literature review. Existing DES studies make little systematic use of EB methods, and where such methods do appear, they are typically applied in a narrow or ad hoc fashion, without sustained attention to

when it is is appropriate or how its risks should be managed. By providing both a structured proof of concept and subsequent methodological developments, the thesis establishes EB as a viable, but conditional, tool for DES experimentation analysis, particularly in early-stage, understanding-oriented experimentation contexts where data are limited and many related scenarios are of interest.

Second, the work contributes to the broader EB literature by examining EB behaviour in a setting that differs in important ways from many classical and modern application domains. DES experimentation is characterised by small scenario-specific sample sizes, parallel scenario structures, and heterogeneous response behaviour reflecting complex system dynamics. By examining EB development and evaluation in these conditions, the thesis extends existing EB perspectives beyond traditional multigroup problems and high-throughput scientific settings, highlighting how similarity, heterogeneity, and robustness considerations play out in simulation-based environments.

A key conceptual contribution in this regard is the explicit separation between EB suitability and EB estimation. The thesis shows that assessing whether EB should be applied at all (R3) is analytically distinct from determining how EB estimation should be conducted, if used (R4). This distinction is not always made explicit in the EB literature, where methodological development often proceeds under implicit assumptions of suitability. By formalising this separation, the thesis reframes EB application as involving two complementary concerns: first assessing whether pooling is likely to be beneficial given the observed data structure, and then ensuring that estimation procedures remain reliable when similarity assumptions are imperfect.

Taken together, these contributions provide not only a clearer conceptual framing of EB within DES experimentation, but also a foundation for translating that framing into practical guidance for simulation analysts.

Implications for Practice

In practical terms, the thesis demonstrates how EB methods can be incorporated into DES experimentation practice in a systematic and usable way. The practical

implications of this work concern when EB decisions are made, the burden placed on the analyst, how risk is managed when assumptions are imperfect, and how EB methods integrate with existing DES analysis practice.

A first key practical implication concerns when the decision to apply EB is made. The thesis shows that EB need not be treated as a commitment at the experimental design stage, rather it can be considered as a post-experimental analytical option, informed by the observed structure of the experimental data obtained. The EB suitability heuristic enables this assessment using quantities directly computable from practitioner-level data, allowing analysts to evaluate whether pooling is likely to be beneficial after data have been collected. This lowers the barrier to EB use by removing the need to anticipate its appropriateness in advance, and supports exploratory DES studies in which data limitations and structural uncertainties are common.

Beyond the timing of the EB decision, a second practical implication concerns the reduction in subjective judgement required of the analyst. In DES settings, the uptake of EB methods has been hindered by the need to intuit scenario similarity or justify pooling decisions on largely informal grounds. The EB suitability heuristic developed in this thesis directly addresses this issue by basing such decisions on observable features of the experimental data. As a result, practitioners are not required to defend ad hoc similarity assessments when deciding whether EB should be applied, reducing the cognitive and interpretive burden placed on the analyst.

This reduction in judgement burden is reinforced by a corresponding lowering of technical and computational barriers to use. By relying on closed-form estimators, the proposed approach avoids the need for numerical optimisation or iterative fitting procedures, along with the associated risks of instability or failure. In practical terms, this allows EB estimation to be carried out using simple and transparent calculations, rather than requiring bespoke programming, repeated adjustment, or specialist statistical intervention.

Alongside these considerations, the thesis addresses the explicit management of risk associated with EB estimation. Practitioners are often less concerned with average error reductions than with the possibility of large individual errors arising

from inappropriate shrinkage. The adapted DS EB estimator developed in the thesis directly addresses this concern by moderating shrinkage in the presence of weak similarity or outlying scenarios. This leads to more conservative and predictable behaviour when EB assumptions are imperfect. From a practical perspective, this makes EB outcomes easier to trust and defend, particularly in decision-making contexts where credibility and accountability are important.

Taken together, these features support compatibility with existing DES analysis practice. The proposed methods require only existing simulation experimental data and minimal user input, and can be layered onto established analysis procedures without altering experimental design or core modelling choices. This creates scope for automation and software integration, allowing EB suitability assessment and robust EB estimation to be embedded alongside existing experimental analysis tools, rather than treated as separate, specialist steps.

Overall, the implications for practice reinforce the central message of the thesis: EB methods are neither universally superior nor inherently risky in DES experimentation, but instead require informed, context-sensitive application. By providing tools that support both suitability assessment and robust implementation, the thesis helps bridge the gap between the statistical potential of EB and its real-world usability in DES practice.

7.4 Limitations and Scope

This thesis has investigated the application of EB methods to DES model experimentation, with a particular focus on suitability assessment, robust estimation, and practical applicability. As with any methodological study, the scope and conclusions of the work are shaped by a number of modelling and design choices. The principal limitations are outlined below.

A first and substantive limitation concerns the distributional assumptions underpinning the EB framework employed. The estimation procedures developed and analysed in the thesis are grounded in parametric EB formulations, with normality assumptions playing an important role in both theoretical development and controlled numerical evaluation. This choice provides analytical transparency and

enables systematic investigation, but it constrains the formal validity of the methods under strong departures from normality. While the industrial case study in Chapter 6 demonstrated encouraging robustness under skewness and threshold effects, further investigation across a wider range of distributional regimes remains warranted.

A second limitation relates to the generality of the empirical evidence base. The thesis adopts a staged experimental strategy, progressing from a toy DES model, through Monte Carlo development environments based on normal populations, to a single large-scale industrial case study exhibiting non-normal, skewed, and threshold-driven behaviour. This progression supports controlled exploration of heuristic and estimator behaviour during development and initial evaluation, followed by stress-testing in a realistic applied setting. However, it also entails a relatively direct transition from idealised development environments to a complex industrial application, without the inclusion of an intermediate toy model. While the industrial case study provides valuable evidence of robustness under such conditions, a more gradual escalation in experimental complexity could offer additional insight into how specific departures from modelling assumptions affect EB performance. Accordingly, the findings should be interpreted as indicative rather than universally generalisable across all DES contexts

Third, the thesis focuses exclusively on univariate point estimation. This deliberate restriction allows clear exposition of the bias—variance trade-offs inherent in EB methods and facilitates meaningful comparison with frequentist estimators. However, DES experimentation often involves multivariate outputs, dependence structures, or interest in extreme rather than average behaviour alone. The extent to which the proposed EB suitability heuristic and adapted DS EB estimator extend naturally to such settings has not been examined.

In addition to these primary scope limitations, the thesis makes use of several methodological approximations and pragmatic choices that warrant acknowledgement. Performance evaluation throughout the numerical studies relies on comparative error measures rather than full sampling-distribution characterisation. ANOVA and Welch's t-test are employed heuristically as diagnostic tools rather than as strict inferential models, reflecting their practical role in guiding explo-

ration rather than formal hypothesis testing. Finally, confidentiality constraints associated with the industrial case study necessarily limited the level of experimental detail that could be reported, although these restrictions do not affect the substantive conclusions drawn regarding feasibility and robustness.

Taken together, these limitations define the scope within which the thesis' conclusions should be interpreted and point to directions for future work.

7.5 Directions for Future Research

This section outlines several avenues for future research that arise naturally from the findings of the thesis. These directions are intended to extend and deepen the contributions made, and reflect the broader potential of EB methods within and beyond DES experimentation.

A first direction concerns the extension of the proposed framework to a wider range of DES experimentation contexts. The numerical and applied studies in this thesis focused on settings that allowed controlled examination of EB behaviour under structured forms of heterogeneity. Future work could involve systematic consideration of more complex output characteristics commonly encountered in DES practice, including stronger skewness, heavy-tailed or multimodal responses, and more intricate forms of scenario dependence. In addition, while the present work focused on point estimation in early-stage, understanding-oriented experimentation, EB methods could also be explored in relation to other DES objectives, such as screening, ranking and selection, or robustness analysis. Investigating the interaction between EB and such objectives would help to clarify the broader role of EB within the DES experimentation lifecycle.

A second promising avenue relates to sequential and adaptive experimentation. The methods developed in this thesis treat EB as a post-experiment analytical tool applied once simulation experimentation data have been generated. In many DES studies, however, data are accumulated iteratively as computational budgets allow or experimental priorities evolve. Extending the proposed framework to exploit sequential updating, while retaining an EB perspective, could enable more adaptive experimentation strategies and further efficiency gains. The relation-

ship between EB estimation and fully Bayesian sequential updating is non-trivial, but the alignment between EB principles and incremental data collection in DES makes this a particularly attractive direction for further investigation.

Further opportunities lie in the refinement and integration of the tools developed in this thesis. The EB suitability heuristic and adapted DS EB estimator were deliberately designed to be simple, interpretable, and deterministic. Future work could explore alternative predictive statistics, similarity measures, or robustness mechanisms, as well as more systematic evaluation of their behaviour under departures from modelling assumptions. In addition, these tools could be more tightly integrated to provide unified guidance on both the suitability and implementation of EB within a given experimentation context. Finally, the closed-form nature of the methods and their limited reliance on subjective inputs also make them well suited to automation within DES analysis environments, potentially allowing EB assessment and estimation to be embedded alongside existing analysis routines.

A longer term conceptual extension of the work concerns the principled reuse of simulation data across related experimentation studies. While storing and reusing simulation data is not new in itself, EB provides a formal mechanism for borrowing strength from historical or auxiliary scenarios while explicitly controlling the influence of dissimilar data. Exploring how the proposed suitability assessment and robust estimation ideas could support such cross-experiment reuse represents a natural, but as yet unexplored, extension of the framework.

Finally, although this thesis is motivated by DES experimentation, the underlying setting considered, that is, parallel populations with limited sample sizes and varying degrees of similarity, also arises in other data analytic contexts, such as high-throughput scientific applications. Exploring the applicability of the proposed EB suitability heuristic and the adapted DS EB estimator beyond DES may therefore offer further insight into the generality of the framework and its potential impact across a broader range of data analytic settings.

Bibliography

[1] T. Aktaran-Kalayci, C. Alexopoulos, D. Goldsman, and J. R. Wilson. Jackknifed variance estimators for simulation output analysis. In *Winter Simulation Conference (WSC) Proceedings*, pages 459–471, 2015.

[2] A. Alan and B. Pritsker. Model evolution: a rotary index table case history. In *Winter Simulation Conference (WSC) Proceedings*, pages 703–707, 1986.

[3] J. Albert. Revisiting efron and morris's baseball study. Exploring Baseball Data with R [Blog], 2016. Last accessed on 24 June 2017.

[4] C. Alexopoulos. A comprehensive review of methods for simulation output analysis. In *Winter Simulation Conference (WSC) Proceedings*, pages 168–178, 2006.

[5] P. M. Allen, S. Maguire, and B. McKelvey. *The Sage Handbook of Complexity and Management*. Sage, 2011.

[6] S. Amaran, N. V. Sahinidis, B. Sharda, and S. J. Bury. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research*, 240(1):351–380, 2016.

[7] S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

[8] S. Axsäter. *Inventory Control*. Springer, 2007.

[9] O. Balci and R. G. Sargent. A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM*, 24(11):190–197, 1981.

[10] O. Balci and R. G. Sargent. Validation of simulation models via simultaneous confidence intervals. *American Journal of Mathematical and Management Science*, 4(3):375–406, 1984.

[11] J. Banks, J. Carson, B. L. Nelson, and D. Nicol. *Discrete-event system simulation*. Prentice-Hall, 2010.

[12] E. Barlow, D. T. Ozturk, M. Revie, K. Akartunalı, A. H. Day, and E. Boulougouris. On using simulation to model the installation process logistics for an offshore wind farm. Working paper, University of Strathclyde, 2017.

[13] E. Barlow, D. T. Ozturk, M. Revie, K. Akartunalı, A. H. Day, and E. Boulougouris. A mixed-method optimisation and simulation framework for supporting logistical decisions during offshore wind farm installations. *European Journal of Operational Research*, 264(3):894–906, 2018.

[14] E. Barlow, D. T. Ozturk, M. Revie, A. H. Day, E. Boulougouris, and K. Akartunalı. Exploring the impact of innovative developments to the installation process for an offshore wind farm. *Ocean Engineering*, 109:623–634, 2015.

[15] R. R. Barton. Simulation optimization using metamodels. In *Winter Simulation Conference (WSC) Proceedings*, pages 230–238, 2009.

[16] R. R. Barton. Designing simulation experiments. In *Winter Simulation Conference (WSC) Proceedings*, pages 342–353, 2013.

[17] R. R. Barton and M. Meckesheimer. Metamodel-based simulation optimization. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, chapter 18, pages 535–574. Elsevier, 2006.

[18] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763.

[19] R. E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25(1):16–39, 1954.

[20] R. E. Bechhofer, T. J. Santner, and D. M. Goldsman. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. John Wiley & Sons, 1995.

[21] R. E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[22] Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):405–416, 2010.

[23] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a new and powerful approach to multiple comparisons. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995.

[24] J. O. Berger. *Statistical decision theory: foundations, concepts, and methods*. Springer, 2013.

[25] R. L. Berger and G. Casella. *Statistical Inference*. Duxbury, 2008.

[26] B. Bettonvil. *Detection of Important Factors by Sequential Bifurcation*. PhD thesis, 1990.

[27] B. Bettonvil and J. Kleijnen. Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research*, 96(1):180–194, 1996.

[28] G. Bharathy and B. Silverman. Validating agent based social system models. In *Winter Simulation Conference (WSC) Proceedings*, pages 441–453, 2010.

[29] S. Blair, T. Bedford, and J. Quigley. Empirical Bayes methods for discrete event simulation performance measure estimation. In *3rd Student Conference on Operational Research*, OpenAccess Series in Informatics (OASIcs), pages 21–30.

[30] J. Borwein and D. Bailey. *Mathematics by Experiment: Plausible Reasoning in the 21st Century*. CRC Press, 2008.

[31] G. E. P. Box. The exploration and exploitation of response surfaces: Some general considerations and examples. *Biometrics*, 10(1):16–60, 1954.

[32] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces*. Wiley, 1987.

[33] G. E. P. Box and N. R. Draper. *Response surfaces, mixtures, and ridge analyses*. Wiley, 2007.

[34] G. E. P. Box, W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. Wiley, 2005.

[35] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 2015.

[36] G. E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 13(1):1–45, 1950.

[37] S. Brailsford, L. Churilov, and B. Dangerfield. *Discrete-Event Simulation and System Dynamics for Management Decision Making*. Wiley, 2014.

[38] S. Brailsford, E. Silverman, S. Rossiter, J. Bijak, R. Shaw, J. Viana, J. Noble, S. Efstathiou, and A. Vlachantoni. Complex systems modeling for supply and demand in health and social care. In *Winter Simulation Conference (WSC) Proceedings*, pages 1125–1136, 2011.

[39] R. J. Brooks. Some thoughts on conceptual modelling: performance, complexity and simplification. In *Proceedings of the 3rd UK Simulation Workshop*, 2006.

[40] R. J. Brooks and A. M. Tobias. Choosing the best model: level of detail, complexity, and model performance. *Mathematical and Computer Modelling*, 24(1):1–14, 1996.

[41] L.D. Brown and E. Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 37(4):1685–1704, 2009.

[42] T. E. Burk and A. R. Ek. Application of empirical Bayes / James-Stein procedures to simultaneous estimation problems in forest inventory. *Forest Science*, 28(4):753–771, 1982.

[43] J. Cares. The use of agent-based models in military concept development. In *Winter Simulation Conference (WSC) Proceedings*, pages 935–939, 2002.

[44] B. P. Carlin and A. E. Gelfand. A sample reuse method for accurate parametric empirical Bayes confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(1):189–200, 1991.

[45] B. P. Carlin and T. A. Louis. Empirical bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289, 2000.

[46] B. P. Carlin and T. A. Louis. *Bayesian methods for data analysis*. CRC Press, 2009.

[47] J. Carson. Model verification and validation. In *Winter Simulation Conference (WSC) Proceedings*, pages 52–58, 2002.

[48] Y. Carson and A. Maria. Simulation optimization: methods and applications. In *Winter Simulation Conference (WSC) Proceedings*, pages 118–126, 1997.

[49] P. Checkland. *Systems Thinking, System Practice*. Wiley, 1981.

[50] L. Chen. Curse of dimensionality. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 545–546. Springer, 2009.

[51] P. Chen. On selecting the best of k systems: an expository survey of subset-selection multinomial procedures. In *Winter Simulation Conference (WSC) Proceedings*, pages 440–444, 1988.

[52] R. Cheng. Searching for important factors: Sequential bifurcation under uncertainty. In *Winter Simulation Conference (WSC) Proceedings*, pages 275–280, 1997.

[53] S. E. Chick. Selecting the best system: a decision-theoretic approach. *Winter Simulation Conference (WSC) Proceedings*, pages 326–333, 1997.

[54] S. E. Chick and K. Inoue. New procedures to select the best simulated system using common random numbers. *Management Science*, 47(8):1133–1149, 2001.

[55] S. E. Chick and K. Inoue. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49(5):732–743, 2001.

[56] L. Chwif, M. R. P. Barretto, and R. J. Paul. On simulation model complexity. In *Winter Simulation Conference (WSC) Proceedings*, pages 449–455, 2000.

[57] L. Chwif, P. S. Muniz, and L. M. Shimada. A prescriptive technique for v&v of simulation models when no real-life data are available. In *Winter Simulation Conference (WSC) Proceedings*, pages 911–918, 2006.

[58] L. Chwif, R. J. Paul, and M. R. P. Barretto. Discrete event simulation model reduction: a causal approach. *Simulation Modelling Practice and Theory*, 14(7):930–944, 2006.

[59] T. Cioppa, T. Lucas, and S. M. Sanchez. Military applications of agent-based simulations. In *Winter Simulation Conference (WSC) Proceedings*, pages 171–180, 2004.

[60] B. Clegg. Optimising the wind. *Impact Magazine*, 3(1):32–35, 2017.

[61] R. Conway. Some tactical problems in digital simulation. *Management Science*, 10(1):47–61, 1963.

[62] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

[63] M. A. Crane and D. L. Iglehart. Simulating stable stochastic systems: I. general multiserver queues. *Journal of the ACM (JACM)*, 21(1):103–113, 1974.

[64] M. A. Crane and D. L. Iglehart. Simulating stable stochastic systems: II. markov chains. *Journal of the ACM (JACM)*, 21(1):114–123, 1974.

[65] M. A. Crane and D. L. Iglehart. Simulating stable stochastic systems: III. regenerative processes and discrete-event simulations. *Operations Research*, 23(1):33–45, 1975.

[66] H. Damerdji. Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research*, 19(2):494–512, 1994.

[67] R. Davies, P. Roderick, and J. Raftery. The evaluation of disease prevention and treatment using simulation models. *European Journal of Operational Research*, 150(1):53–66, 2003.

[68] B. Dengiz and O. Belgin. Paintshop production line optimization using response surface methodology. In *Winter Simulation Conference (WSC) Proceedings*, pages 1667–1672, 2007.

[69] K. D. Dingeç, C. Alexopoulos, D. Goldsman, J. R. Wilson, W. Chiu, and T. Aktaran-Kalayci. On the mean-squared error of variance estimators for computer simulations. In *Winter Simulation Conference (WSC) Proceedings*, pages 549–555, 2011.

[70] D. Dudenhoeffer, M. Permann, and M. Manic. CIMS: a framework for infrastructure interdependency modeling and analysis. In *Winter Simulation Conference (WSC) Proceedings*, pages 478–485, 2006.

[71] E. J. Dudewicz and S. R. Dalal. Allocation of observations in ranking and selection with unequal variances. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 471–474. Academic Press, 1971.

[72] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[73] B. Edmonds and S. Moss. From kiss to kids – an 'anti-simplistic' modelling approach. In Logan B. Davidsson, P. and K. Takadama, editors, *MABS 2004: Multi-Agent and Multi-Agent-Based Simulation*, volume 3145 of *Lecture Notes in Computer Science*, pages 130–144. Springer, 2005.

[74] Efron. *Journal of the American Statistical Association*, 102(477):93–103, 2007.

[75] B. Efron. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

[76] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

[77] B. Efron and T. Hastie. *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*. Cambridge University Press, 2016.

[78] B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes estimators – Part ii: The empiricial Bayes case. *Journal of the American Statistical Society*, 67(377):130–139, 1972.

[79] B. Efron and C. Morris. Stein's estimation rule and it competitors – an empiricial Bayes approach. *Journal of the American Statistical Society*, 68(341):117–130, 1973.

[80] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[81] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[82] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Society*, 96(456):1151–1160, 2001.

[83] Scottish Enterprise. Offshore wind factsheet. Scottish Enterprise [Factsheet], 2017. Last accessed on 19 June 2017.

[84] The Crown Estate. A guide to an offshore wind farm. The Crown Estate [Report], 2017. Last accessed on 19 June 2017.

[85] P. J. Farrell, B. Macgibbon, and T. J. Tomberlin. Bootstrap adjustments for empirical Bayes interval estimates of small-area proportions. *Canadian Journal of Statistics*, 25(1):75–89, 1997.

[86] J. Fisher and T. Henzinger. Executable biology. In *Winter Simulation Conference (WSC) Proceedings*, pages 1675–1682, 2006.

[87] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368, 1922.

[88] R. A. Fisher. *Statistical Methods For Research Workers*. Cosmo Publications, 1925.

[89] G. S. Fishman. *Spectral Methods in Econometrics*. Harvard University Press, 1969.

[90] G. S. Fishman. Estimating sample size in computing simulation experiments. *Management Science*, 18(1):21–38, 1971.

[91] G. S. Fishman. Statistical analysis for queueing simulations. *Management Science*, 20(3):363–369, 1973.

[92] G. S. Fishman. Estimation in multiserver queuing simulations. *Operations Research*, 22(1):72–78, 1974.

[93] G. S. Fishman. *Principles of Discrete Event Simulation*. Wiley, 1978.

[94] G. S. Fishman and D. Gross. Concepts and methods in discrete event digital simulation. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(5):390–390, 1976.

[95] M. C. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215, 2002.

[96] M. C. Fu and K. J. Healy. Simulation optimization of (s,S) inventory systems. In *Winter Simulation Conference (WSC) Proceedings*, pages 506–514, 1992.

[97] D. P. Gaver and D. H. Worledge. Contemporary statistical procedures (parametric empirical Bayes) and nuclear plant event rates. Technical Report EPRI-NP–3912-SR-Vol3, Electric Power Research Institute, Washington DC (USA), 1984.

[98] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 1995.

[99] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

[100] Paul Glasserman. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, 1991.

[101] Fred Glover. Tabu search — part i. *ORSA Journal on Computing*, 1(3):190–206, 1989.

[102] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[103] D. Goldsman. A multinomial ranking and selection procedure: simulation and applications. In *Winter Simulation Conference (WSC) Proceedings*, page 258–264, 1984.

[104] D. Goldsman. On selecting the best of k systems: an expository survey of indifference-zone multinomial procedures. In *Winter Simulation Conference (WSC) Proceedings*, page 106–112, 1984.

[105] D. Goldsman. Ranking and selection procedures using standardized time series. In *Winter Simulation Conference (WSC) Proceedings*, pages 120–124, 1985.

[106] D. Goldsman, S.-H. Kim, W. S. Marshall, and B. L. Nelson. Ranking and selection for steady-state simulation: procedures and perspectives. *INFORMS Journal on Computing*, 14(1):2–19, 2002.

[107] D. Goldsman, R. Nance, and J. Wilson. A brief history of simulation revisited. In *Winter Simulation Conference (WSC) Proceedings*, pages 567–574, 2010.

[108] D. Goldsman and B. Schmeiser. Computational efficiency of batching methods. In *Winter Simulation Conference (WSC) Proceedings*, pages 202–207, 1997.

[109] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. Wiley, 1998.

[110] S. S. Gupta. *On a Decision Rule for a Problem in Ranking Means*. PhD thesis, 1956.

[111] S. S. Gupta and T. J. Santner. On selection and ranking procedures – a restricted subset selection rule. In *Proceedings of the 39th Session of the International Statistical Institute*, pages 478–486, 1973.

[112] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[113] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85, 1995.

[114] J. R. Hill, A.S. Heger, and B. V. Koen. Application of Stein and related parametric empirical Bayes estimators to the nuclear plant reliability data system. Technical Report NUREG/CR–3637, Texas University, Austin (USA); Idaho National Engineering Laboratory, Idaho Falls (USA), 1984.

[115] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments: Advanced Experimental Design*. Wiley, 2005.

[116] Yu-Chi Ho and Xi-Ren Cao. Perturbation analysis of discrete event dynamic systems. *IEEE Transactions on Automatic Control*, 32(7):563–572, 1987.

[117] K. Hoad, T. Monks, and F. O'Brien. The use of search experimentation in discrete-event simulation practice. *Journal of the Operational Research Society*, 66(7):1155–1168, 2015.

[118] K. Hoad, S. Robinson, and R. Davies. Automating warm-up length estimation. *The Journal of the Operational Research Society*, 61(9):1389–1403, 2010.

[119] K. Hoad, S. Robinson, and R. Davies. Classification of discrete event simulation models and output: creating a sufficient model set. In *Operational Research Society Simulation Workshop 2010 (SW10)*, pages 137–143, 2010.

[120] J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[121] D. Hosmer, S. Lemeshow, and R. Sturdivant. *Applied Logistic Regression*. Wiley, 2013.

[122] M.-H. Hsieh, D. L. Iglehart, and P. W. Glynn. Empirical performance of bias-reducing estimators for regenerative steady-state simulations. *ACM Transactions on Modeling and Computer Simulation*, 14(4):325–343, 2004.

[123] J. C. Hsu. Constrained simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics*, 12(3):1136–1144, 1984.

[124] J. C. Hsu and B. L. Nelson. Optimization over a finite number of system designs with one-stage sampling and multiple comparisons with the best. In *Winter Simulation Conference (WSC) Proceedings*, pages 451–457, 1988.

[125] G. Hwang and P. Liu. Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, 9(1):1544–1615, 2010.

[126] G. Hwang, J. Qiu, and Z. Zhao. Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):265–285, 2009.

[127] INFORMS. What is operations research? INFORMS [Webpage], 2017. Last accessed on 19 June 2017.

[128] G. Innis and E. Rexstad. Simulation model simplification techniques. *SIMULATION*, 41(1):7–15, 1983.

[129] K. Inoue and S. E. Chick. Comparison of Bayesian and frequentist assessments of uncertainty for selecting the best system. In *Winter Simulation Conference (WSC) Proceedings*, pages 727–734, 1998.

[130] M. D. L. A. Irizarry, J. Wilson, and J. Trevino. A flexible simulation tool for manufacturing-cell design, II: response surface analysis and case study. *IIE Transactions*, 33(10):827–836, 2001.

[131] R. Jin, W. Chen, and T.W. Simpson. Comparative studies of metamodeling techniques under multiple modeling criteria. *Structural and Multidisciplinary Optimization*, 23(1):1–13, 2001.

[132] I.M. Johnstone and B.W. Silverman. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

[133] I.M. Johnstone and B.W. Silverman. Empirical bayes selection of wavelet thresholds. *The Annals of Statistics*, 33(4):1700–1752, 2005.

[134] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[135] J. Joyce. Bayes' theorem. In E.N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. 2016.

[136] Y. Jun and S. H. Ng. An entropy based sequential calibration approach for stochastic computer models. In *Winter Simulation Conference (WSC) Proceedings*, pages 589–600, 2013.

[137] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[138] J. P. Kearney, R. L. Sedlmeyer, W. B. Thompson, M. A. Gray, and M. A. Adler. Software complexity measurement. *Communications of the ACM*, 29(11):1044–1050, 1986.

[139] W. D. Kelton and A. M. Law. A new approach for dealing with the startup problem in discrete event simulation. *Naval Research Logistics Quarterly*, 30(4):641–658, 1983.

[140] S.-H. Kim and B. L. Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation*, 11(3):251–273, 2001.

[141] S.-H. Kim and B. L. Nelson. Selecting the best system. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, chapter 17, pages 501–534. Elsevier, 2006.

[142] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[143] J. Kleijnen. White noise assumptions revisited: regression metamodels and experimental designs in practice. In *Winter Simulation Conference (WSC) Proceedings*, pages 107–117, 2006.

[144] J. Kleijnen. *Design and Analysis of Simulation Experiments*. Springer, 2015.

[145] J. Kleijnen, S. M. Sanchez, T. Lucas, and T. Cioppa. State-of-the-art review: a user's guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing*, 17(3):263–289, 2005.

[146] J. Kleijnen and R. G. Sargent. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120(1):14–29, 2000.

[147] K. Knight. *Mathematical Statistics*. CRC Press, 1999.

[148] L. W. Koenig and A. M. Law. A procedure for selecting a subset of size m containing the l best of k independent normal populations, with applications to simulation. *Communications in Statistics - Simulation and Computation*, 14(3):719–734, 1985.

[149] E. Kreyszig. *Advanced Engineering Mathematics*. Wiley, 1999.

[150] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[151] N. M. Laird and T. A. Louis. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399):739–750, 1987.

[152] M. Landry, J.-L. Malouin, and M. Oral. Model validation in operations research. *European Journal of Operational Research*, 14(3):207–220, 1983.

[153] A. M. Law. Statistical analysis of simulation output data: the practical state of the art. In *Winter Simulation Conference (WSC) Proceedings*, pages 77–83, 2007.

[154] A. M. Law. *Simulation modeling and analysis*. McGraw-Hill, 2015.

[155] A. M. Law, J. S. Carson, J. G. Fox, S. K. Halladin, K. J. Musselman, and O. M. Ulgen. A forum on crucial issues in the simulation of manufacturing systems. In *Winter Simulation Conference (WSC) Proceedings*, pages 916–922, 1993.

[156] P. Linton, W. Melodia, A. Lazar, D. Agarwal, L. Bianchi, D. Ghoshal, G. Pastorello, L. Ramakrishnan, and K. Wu. Understanding data similarity in large-scale scientific datasets. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4525–4531, 2019.

[157] T. Lucas, S. M. Sanchez, F. Martinez, L. R. Sickinger, and J. W. Roginski. Defense and homeland security applications of multiagent simulations. In *Winter Simulation Conference (WSC) Proceedings*, pages 138–149, 2007.

[158] J. Luo and L. J. Hong. Large-scale ranking and selection using cloud computing. In *Winter Simulation Conference (WSC) Proceedings*, pages 4051–4061, 2011.

[159] H. Madsen and D. Rosbjerg. Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. *Water Resources Research*, 33(4):771–781, 1997.

[160] J. S. Maritz. *Empirical Bayes Methods*. Methuen, 1970.

[161] H. F. Martz, P. H. Kvam, and L .R. Abramson. Empirical Bayes estimation of the reliability of nuclear-power-plant emergency diesel generators. *Technometrics*, 38(1):11–24, 1996.

[162] H.F. Martz and R.A. Waller. *Bayesian Reliability Analysis*. Wiley, 1982.

[163] M. S. Meketon and B. Schmeiser. Overlapping batch means: something for nothing? In *Winter Simulation Conference (WSC) Proceedings*, pages 226–230, 1984.

[164] D. Montgomery. *Design and Analysis of Experiments*. Wiley, 2006.

[165] C. N. Morris. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

[166] M. K. Nakayama. Multiple comparisons with the best using common random numbers for steady-state simulations. *Journal of Statistical Planning and Inference*, 85(1-2):37–48, 2000.

[167] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[168] B. L. Nelson. Robust multiple comparisons under common random numbers. *ACM Transactions on Modeling and Computer Simulation*, 3(3):225–243, 1993.

[169] B. L. Nelson and J. C. Hsu. Control-variate models of common random numbers for multiple comparisons with the best. *Management Science*, 39(8):989–1001, 1993.

[170] C. M. Nicol, O. Balci, R. M. Fujimoto, P. A. Fishwick, P. L'Ecuyer, and R. Smith. Strategic directions in simulation research. In *Winter Simulation Conference (WSC) Proceedings*, pages 1509–1520, 1999.

[171] J. Noguera and E. Watson. Response surface analysis of a multi-product batch processing facility using a simulation metamodel. *International Journal of Production Economics*, 102(2):333–343, 2006.

[172] D. H. Ockerman and D. Goldsman. The impact of transients on simulation variance estimators. In *Winter Simulation Conference (WSC) Proceedings*, pages 234–239, 1997.

[173] Editors of Collins English Dictionary. Ockam's razor. Collins English Dictionary [Online], 2017. Last accessed on 19 June 2017.

[174] Editors of Encyclopaedia Britannica. Computational complexity. Encyclopaedia Britannica [Online], 2017. Last accessed on 19 June 2017.

[175] Editors of Oxford English Dictionary. Complexity. Oxford English Dictionary [Online], 2017. Last accessed on 19 June 2017.

[176] Editors of The Britannica Dictionary. Experimentation. The Britannica Dictionary [Online], 2017. Last accessed on 19 June 2017.

[177] UK Department of Trade and Industry. Future offshore: A strategic framework for the offshore wind industry. UK Department of Trade and Industry [Report], 2017. Last accessed on 19 June 2017.

[178] T. Ören. The many facets of simulation through a collection of about 100 definitions. *SCS M&S Magazine*, 2(2):82–92, 2011.

[179] M. C. Overstreet and R. E. Nance. A specification language to assist in analysis of discrete event simulation models. *Communications of the ACM*, 28(2):190–201, 1985.

[180] K. Pearson. Method of moments and method of maximum likelihood. *Biometrika*, 28(1-2):35–39, 1936.

[181] C. D. Pegden, R. P. Sadowski, and R. E. Shannon. *Introduction to Simulation Using SIMAN*. McGraw-Hill, 1995.

[182] B. Persaud and C. Lyon. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. *Accident Analysis & Prevention*, 39(3):546–555, 2007.

[183] M. Pidd. *Tools for Thinking: Modelling in Management Science*. Wiley, 2003.

[184] M. Pidd. *Computer simulation in management science*. Wiley, 2004.

[185] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, 1973.

[186] Y. Rinott. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics - Theory and Methods*, 7(8):799–811, 1978.

[187] H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 157–163. University of California Press, 1956.

[188] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[189] S. Robinson. Conceptual modelling for simulation Part i: definition and requirements. *Journal of the Operational Research Society*, 59(3):278–290, 2008.

[190] S. Robinson. *Simulation: The Practice of Model Development and Use*. Wiley, 2014.

[191] S. Robinson, C. Worthington, N. Burgess, and Z. J. Radnor. Facilitated modelling with discrete-event simulation: Reality or myth? *European Journal of Operational Research*, 234(1):231–240, 2014.

[192] J. D. Salt. Simulation should be easy and fun! In *Winter Simulation Conference (WSC) Proceedings*, pages 1–5, 1993.

[193] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. Wiley, 2008.

[194] A. Samvedi and V. Jain. Studying the impact of various inventory policies on a supply chain with intermittent supply disruptions. In *Winter Simulation Conference (WSC) Proceedings*, pages 1641–1649, 2011.

[195] P. Sanchez. Fundamentals of simulation modeling. In *Winter Simulation Conference (WSC) Proceedings*, pages 54–62, 2007.

[196] P. Sanchez. London array: Where the amazing happens. Escuela de Organizacion Industrial [Blog], 2014. Last accessed on 19 June 2017.

[197] S. M. Sanchez. Robust design: seeking the best of all possible worlds. In *Winter Simulation Conference (WSC) Proceedings*, pages 69–76, 2000.

[198] S. M. Sanchez. Data farming: methods for the present, opportunities for the future. *ACM Transactions on Modeling and Computer Simulation*, 30(4):1–30, 2020.

[199] S. M. Sanchez and T. Lucas. Exploring the world of agent-based simulations: simple models, complex analyses. In *Winter Simulation Conference (WSC) Proceedings*, pages 116–126, 2002.

[200] S. M. Sanchez and H. Wan. Better than a petaflop: the power of efficient experimental design. In *Winter Simulation Conference (WSC) Proceedings*, pages 1441–1455, 2011.

[201] S. M. Sanchez, H. Wan, and T. Lucas. Two-phase screening procedure for simulation experiments. *ACM Transactions on Modeling and Computer Simulation*, 19(2):1–24, 2009.

[202] T. J. Santner. A restricted subset selection approach to ranking and selection problems. *The Annals of Statistics*, 3(2):334–349, 1975.

[203] R. G. Sargent. Verification and validation of simulation models. In *Winter Simulation Conference (WSC) Proceedings*, pages 183–198, 2011.

[204] R. G. Sargent. An introductory tutorial on verification and validation of simulation models. In *Winter Simulation Conference (WSC) Proceedings*, pages 1729–1740, 2015.

[205] R. G. Sargent, D. Goldsman, and T. Yaacoub. Use of the interval statistical procedure for simulation model validation. In *Winter Simulation Conference (WSC) Proceedings*, pages 60–72, 2015.

[206] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, 1946.

[207] H. Sauro, C. Harel, M. Kwiatkowska, C. Shaffer, A. Uhrmacher, M. Hucka, P. Mendes, L. Strömback, and J. Tyson. Challenges for modeling and simulation methods in systems biology. In *Winter Simulation Conference (WSC) Proceedings*, pages 1720–1730, 2006.

[208] S. Schlesinger, R. E. Crosbie, R. E. Gagné, G. S. Innis, C. S. Lalwani, J. Loch, R. J. Sylvester, R. D. Wright, N. Kheir, and D. Bartos. Terminology for model credibility. *SIMULATION*, 32(3):103–104, 1979.

[209] L. Schruben. Confidence interval estimation using standardized time series. *Operations Research*, 31(6):1090–1108, 1983.

[210] L. Schruben. Simulation modeling with event graphs. *Communications of the ACM*, 26(11):957–963, 1983.

[211] H.-P. Schwefel. *Evolution and Optimum Seeking*. Wiley, 1995.

[212] R. Shannon. *Systems Simulation – The Art and Science*. Prentice-Hall, 1975.

[213] R. Shannon. Introduction to simulation. In *Winter Simulation Conference (WSC) Proceedings*, pages 65–73, 1992.

[214] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[215] T. Smith, L. Marshall, and A. Sharma. Predicting hydrologic response through a hierarchical catchment knowledgebase: a Bayes empirical Bayes approach. *Water Resources Research*, 50(2):1189–1204, 2014.

[216] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, 2004.

[217] B. Snyder and M. J. Kaiser. Ecological and economic cost-benefit analysis of offshore wind energy. *Renewable Energy*, 34(6):1567–1578, 2009.

[218] C. M. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 197–206. University of California Press, 1956.

[219] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

[220] J.D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[221] D. W. Sullivan and J. R. Wilson. Restricted subset selection procedures for simulation. *Operations Research*, 37(1):52–71, 1989.

[222] W. Sun and T.T. T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

[223] J. J. Swain. Simulation software survey – simulation: a better reality? *ORMS–Today*, 40(5), 2013.

[224] J. Swisher, S. Jacobson, and E. Yücesan. Discrete-event simulation optimization using ranking, selection, and multiple comparison procedures: A survey. *ACM Transactions on Modeling and Computer Simulation*, 13(2):134–154, 2003.

[225] C. Szabo and Y. M. Teo. An analysis of the cost of validating semantic composability. *Journal of Simulation*, 6(3):152–163, 2012.

[226] G. Taguchi. *Introduction to Quality Engineering: Designing Quality into Products and Processes*. Quality Resources, 1986.

[227] E.-G. Talbi. A taxonomy of hybrid metaheuristics. *Journal of Heuristics*, 8(5):541–564, 2002.

[228] S. J. E. Taylor, P. Lendermann, R. J. Paul, S. W. Reichenthal, S. Strassburger, and S. J. Turner. Panel on future challenges in modeling methodology. In *Winter Simulation Conference (WSC) Proceedings*, pages 327–335, 2004.

[229] A. Tolk, N. R. Adam, E. Cayirci, S. Pickl, R. Shumaker, J. A. Sullivan, and W. F. Waite. Defense and security applications of modeling and simulation - grand challenges and current efforts. In *Winter Simulation Conference (WSC) Proceedings*, pages 2351–2365, 2012.

[230] L. Trocine and L. Malone. Finding important independent variables through screening designs: a comparison of methods. In *Winter Simulation Conference (WSC) Proceedings*, pages 749–754, 2000.

[231] J. W. Tukey. The problem of multiple comparisons. Unpublished, 1953.

[232] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[233] J. M. Ver Hoef. Parametric empirical Bayes methods for ecological applications. *Ecological Applications*, 6(4):1047–1055, 1996.

[234] W. Vinyard and T. Lucas. Exploring combat models for non-monotonicities and remedies. *Phalanx*, 35(1):1,36–38, 2002.

[235] J. C. Wallace. The control and transformation metric: toward the measurement of simulation model complexity. In *Winter Simulation Conference (WSC) Proceedings*, pages 597–603, 1987.

[236] H. Wan, B. Ankenman, and B. L. Nelson. Controlled sequential bifurcation: a new factor-screening method for discrete-event simulation. *Operations Research*, 54(4):743–755, 2006.

[237] H. Wan, B. Ankenman, and B. L. Nelson. Improving the efficiency of controlled sequential bifurcation for simulation factor screening. *INFORMS Journal on Computing*, 22(3):482–492, 2010.

[238] S. C. Ward. Arguments for constructively simple models. *Journal of the Operational Research Society*, 40(2):141–153, 2006.

[239] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

[240] R. Waupotitsch, S. Eidenbenz, J. Smith, and L. Kroc. Multi-scale integrated information and telecommunications system (MIITS): first results from a large-scale end-to-end network simulator. In *Winter Simulation Conference (WSC) Proceedings*, pages 2132–2139, 2006.

[241] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3-4):350–362, 1938.

[242] B. L. Welch. The generalisation of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[243] P. D. Welch. On the problem of the initial transient in steady-state simulation. Technical Report 37, IBM Watson Research Center, New York (USA), 1981.

[244] K. P. White. An effective truncation heuristic for bias reduction in simulation output. *SIMULATION*, 69(6):323–334, 1997.

[245] K. P. White, M. J. Cobb, and S. C. Spratt. A comparison of five steady-state truncation heuristics for simulation. In *Winter Simulation Conference (WSC) Proceedings*, pages 755–760, 2000.

[246] P. White and R. Ingalls. Introduction to simulation. In *Winter Simulation Conference (WSC) Proceedings*, pages 12–23, 2009.

[247] T. Williams. *Modelling Complex Projects*. Wiley, 2002.

[248] W. Winston. *Operations Research: Applications and Algorithms*. Thomson Brooks/Cole, 1998.

[249] R. Yaesoubi and S. Roberts. Important factors in screening for colorectal cancer. In *Winter Simulation Conference (WSC) Proceedings*, pages 1475–1482, 2007.

[250] R. Yaesoubi, S. Roberts, and R. Klein. A modification of Cheng's method: an alternative factor screening method for stochastic simulation models. In *Winter Simulation Conference (WSC) Proceedings*, pages 1034–1047, 2010.

[251] W.-N. Yang and B. L. Nelson. Optimization using common random numbers, control variates and multiple comparisons with the best. In *Winter Simulation Conference (WSC) Proceedings*, pages 444–449, 1989.

[252] W.-N. Yang and B. L. Nelson. Using common random numbers and control variates in multiple-comparison procedures. *Operations Research*, 39(4):583–591, 1991.

[253] H. Y. Yin and Z. N. Zhou. Simplification techniques of simulation models. In *Proceedings of Beijing International Conference on System Simulation and Scientific Computing*, pages 782–786. International Academic Publishers, 1989.

[254] J. Yin, K. M. Ng, and S. H. Ng. A Bayesian metamodeling approach for stochastic simulations. In *Winter Simulation Conference (WSC) Proceedings*, pages 1055–1066, 2010.

[255] G. A. Young and R. L. Smith. *Essentials of Statistical Inference*. Cambridge University Press, 2005.

[256] E. Yourdon. *Modern Structured Analysis*. Sage, 1988.

[257] M. Yuan and B. L. Nelson. Autoregressive-output-analysis methods revisited. *Annals of Operations Research*, 53:391–418, 1994.

[258] E. Yücesan and L. Schruben. Complexity of simulation models: a graph theoretic approach. *INFORMS Journal on Computing*, 10(1):94–106, 1998.

[259] B.P. Zeigler. *Theory of Modelling and Simulation*. Wiley, 1976.

[260] Z. Zhao. Double shrinkage empirical Bayesian estimation for unknown and unequal variances. *Statistics and its Interface*, 3(4):533–541, 2010.

[261] F. Zouaoui and J. R. Wilson. Accounting for parameter uncertainty in simulation input modeling. In *Winter Simulation Conference (WSC) Proceedings*, pages 354–363, 2001.

Appendix A

Example MATLAB Codes for Chapter 3

naiveEB.m

```
%EB point estimation for population mean based on NLN model
%of data. Pools all model configurations (naive pooling strategy).

%Inputs:
%b - sample size
%data - simulation model output data

%Outputs are:
%avg - standard estimates (for each config and sample, B by dim)
%thetahat - EB estimates (for each config and sample, B by dim)
%exact - long run config averages (for each config and sample,
%        B by dim)

function [thetahat,avg,exact] = naiveEBunmod(data,b)
%Calculating size of simulation data, rep by dim
[rep,dim]=size(data);

%Calculating number of samples, B
B=fix(rep/b);

%Calculating EXACT values, long run configuration averages
exact=mean(data);
```

```
exact=repmat(exact,B,1);

%Initialising avg and vara
avg=zeros(B,dim);
vara=zeros(B,dim);

%Calculating avg and vara
for i=1:B
    for j=1:dim
        avg(i,j)=mean(data(((i-1)*b+1):(i*b),j));
    end
end

for i=1:B
    for j=1:dim
        vara(i,j)=var(data(((i-1)*b+1):(i*b),j))/b;
    end
end

%Calculating fixed constants
d=b-1; m=psi(0,d/2)-log(d/2); sigch2=psi(1,d/2);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Estimating hyperparameters of variance prior%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
q1=log(vara)-m; q2=q1.^2;

%Initialising muhatv_c
muhatv_c=zeros(B,1);

%Calculating muhatv_c
for i=1:B
    muhatv_c(i)=sum(q1(i,:))/dim;
end

%Initialising tauhatv2_c
tauhatv2_c=zeros(B,1);

%Calculating tauhatv2_c
for i=1:B
    tauhatv2_c(i)=max(sum(q2(i,:))/dim-muhatv_c(i)^2-sigch2,0);
end
```

```
65
66 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
67 %Calculating EB estimates of variance%
68 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
69
70 %Initialising Mhatv_c
71 Mhatv_c=zeros(B,1);
72
73 %Calculating Mhatv_c
74 for i=1:B
75     Mhatv_c(i)=tauhatv2_c(i)/(tauhatv2_c(i)+sigch2);
76 end
77
78 %Initialising sighat2_EB
79 sighat2_EB=zeros(B,dim);
80
81 %Calculating sighat2_EB
82 for i=1:B
83     for j=1:dim
84         sighat2_EB(i,j)=exp(Mhatv_c(i)*q1(i,j)
85         +(1-Mhatv_c(i))*muhatv_c(i));
86     end
87 end
88
89 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
90 %Estimating hyperparameters of mean prior%
91 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
92 q3=avg./sighat2_EB; q4=1./sighat2_EB;
93
94 %Initialising muhat_c
95 muhat_c=zeros(B,1);
96
97 %Calculating muhat_c
98 for i=1:B
99     muhat_c(i)=sum(q3(i,:))/sum(q4(i,:));
100 end
101
102 %Initialising mat
103 mat=zeros(B,dim);
104
105 %Calculating mat
106 for i=1:B
```

```
107     for j=1:dim
108         mat(i,j)=(avg(i,j)-muhat_c(i))^2-vara(i,j)*exp(-m-sigch2/2);
109     end
110 end
111
112 %Initialising tauhat2_c
113 tauhat2_c=zeros(B,1);
114
115 %Calculating tauhat2_c
116 for i=1:B
117     tauhat2_c(i)=max(sum(mat(i,:))/dim,0);
118 end
119
120 %Initialising Mhat and thetahat
121 Mhat=zeros(B,dim);
122 thetahat=zeros(B,dim);
123
124 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
125 %Calculating EB estimates of mean%
126 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
127
128 for i=1:B
129     for j=1:dim
130         Mhat(i,j)=tauhat2_c(i)/(sighat2_EB(i,j)+tauhat2_c(i));
131     end
132 end
133
134 for j=1:dim
135     for i=1:B
136         thetahat(i,j)=Mhat(i,j)*avg(i,j)+(1-Mhat(i,j))*muhat_c(i);
137     end
138 end
139 end
```

performance.m

```
function [ratio_SSE,ratio_ABS,ratio_MAX_SE] = ...
    performance(thetahat,avg,exact)

%calculating B
[B,dim]=size(avg);

%squared errors
sq_err_eb=(thetahat-exact).^2; sq_err_st=(avg-exact).^2;
abs_err_eb=abs(thetahat-exact); abs_err_st=abs(avg-exact);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% SSE and MSSE %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Sum of squared errors across configurations (per sample)
%for eb and st. Sum across rows of sq_err_eb and sq_err_st
%(creates vector length B each).
SSE_ac_conf_eb=sum(sq_err_eb,2); SSE_ac_conf_st=sum(sq_err_st,2);

% Absolute error summation:
ABS_ac_conf_eb=max(abs_err_eb,[],2); ...
    ABS_ac_conf_st=max(abs_err_st,[],2);

% Max error identification:
MAX_SE_ac_conf_eb=max(sq_err_eb,[],2); ...
    MAX_SE_ac_conf_st=max(sq_err_st,[],2);

%Mean of SSE across configs (across samples) for eb and st:
%Take average of above vectors, single value for EB and ST
Mean_SSE_ac_conf_sample_eb=sum(SSE_ac_conf_eb)/B;
Mean_SSE_ac_conf_sample_st=sum(SSE_ac_conf_st)/B;

Mean_ABS_ac_conf_sample_eb=sum(ABS_ac_conf_eb)/B;
Mean_ABS_ac_conf_sample_st=sum(ABS_ac_conf_st)/B;

Mean_MAX_SE_ac_conf_sample_eb=sum(MAX_SE_ac_conf_eb)/B;
Mean_MAX_SE_ac_conf_sample_st=sum(MAX_SE_ac_conf_st)/B;

ratio_SSE=Mean_SSE_ac_conf_sample_eb/Mean_SSE_ac_conf_sample_st;
```

```
37 ratio_ABS=Mean_ABS_ac_conf_sample_eb/Mean_ABS_ac_conf_sample_st;
38 ratio_MAX_SE=Mean_MAX_SE_ac_conf_sample_eb/
39 Mean_MAX_SE_ac_conf_sample_st;
```

Appendix B

Example MATLAB Codes for Chapter 4

classific.m

```
%Fits logistic regression model based on the input variables of
%interest (sample size, number of populations and variation ratio)
%and dataset labels from the ratio of mean ss errors.
%The programme considers a range of classification thresholds
%when attaching labels to the datasets. Predictions are then made
%using these logistic regression models and a range of prediction
%thresholds.

function [trueclassification] = classific(aveSSEeb_aveSSEst,
no_of_datasets, Ratios, sizes,B)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Inputs:

%aveSSEeb_aveSSEst - ratio of mean sse errors (EB/Freq) for ...
    each dataset

%no_of_datasets - number of datasets in our collection

%Ratios -  the variation ratio for each of our datasets

```

```
%sizes -  array of dataset sizes

%B - number of batch replications

%Outputs:

%trueclassification - vector of EB or Freq prevalence (1-EB, ...
    0-Freq)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Creating vectors of thresholds for classifying datasets as
%EB or Freq and for predicting datasets as EB or Freq based on
%the output of our logistic regression.

Classification_threshold = 0.6:0.01:1;
n = length(Classification_threshold);
Prediction_threshold = 0.3:0.01:0.7;
m = length(Prediction_threshold);

%Preallocating storage arrays for quantities of interest:
sens = zeros(1,n*m);
spec = zeros(1,n*m);
CT = zeros(1,n*m);
PT = zeros(1,n*m);
accuracy = zeros(1,n*m);
Youden = zeros(1,n*m);
Bcoeff = zeros(n*m,4);

count = 1; %Initialising count variable

%Working through all combinations of the two thresholds
for ii = 1:n
    for jj = 1:m

        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
        %Classifying datasets as to whether EB or Freq is superior
        trueclassification = zeros(1,no_of_datasets);
        classification = zeros(1,no_of_datasets);


```

```
63          %True classification based of repeated sampling
64          for i = 1:no_of_datasets
65              if aveSSEeb_aveSSEst(1,i) < 1
66                  trueclassification(1,i) = 1;
67              else
68                  trueclassification(1,i) = 0;
69              end
70          end
71
72          %Classification based on selected threshold (ie not 1)
73          for i = 1:no_of_datasets
74              if aveSSEeb_aveSSEst(1,i) < ...
                   Classification_threshold(ii)
75                  classification(1,i) = 1;
76              else
77                  classification(1,i) = 0;
78              end
79          end
80          %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
81
82          %Extracting the no. of populations and sample sizes for
83          %our datasets
84          Pops = sizes(2,:)
85          Sample_size = sizes(1,:)/B
86
87
88          %Fitting the logistic regression model to
89          %the threshhold classified data
90          X = [log(Ratios).',log(Pops).',log(Sample_size).'];
91          b = glmfit(X,classification.','binomial');
92          Bcoeff(count,:) = b; %storing coefficients of
93                               %log reg model
94
95          z1 = b(1)+b(2)*log(Ratios)+b(3)*log(Pops)
96          +b(4)*log(Sample_size);
97          gamma = Prediction_threshold(jj)-0.5;
98          prediction = round((1./(1+exp(-z1)))-gamma);
99
100
101         %Calculating no. of datasets where EB prevails
102         no_of_EB = sum(trueclassification);
103
```

```
104         %Calculating no. of datasets where frequentist prevails
105         no_of_Freq = no_of_datasets - no_of_EB;
106
107         %Categorising each prediction as false/true positive or
108         %false/true negative. Done by computing two digit value
109         %trueclassification*10+prediction:
110         % 00 corresponds to a prediction of neg which is correct
111         % 01      "              "       of pos which is incorrect
112         % 10      "              "       of neg which is incorrect
113         % 11      "              "       of pos which is correct
114
115         %Initialising category count
116         true_freq = 0; false_freq = 0; true_EB = 0; false_EB = 0;
117
118         %Preallocating storage vector
119         score = zeros(1,no_of_datasets);
120
121         %Categorising each dataset prediction
122         for i = 1:no_of_datasets
123             score(1,i) = ...
                   trueclassification(1,i)*10+prediction(1,i);
124             if score(1,i) == 0
125                 true_freq = true_freq+1;
126             elseif score(1,i) == 1
127                 false_EB = false_EB+1;
128             elseif score(1,i) == 10
129                 false_freq = false_freq+1;
130             else
131                 true_EB = true_EB+1;
132             end
133         end
134
135         %Calculating and storing sensitivity and specificity
136         sens(1,count) = true_EB/no_of_EB;
137         spec(1,count) = true_freq/no_of_Freq;
138
139         %Storing the threshold configuration
140         CT(1,count) = Classification_threshold(ii);
141         PT(1,count) = Prediction_threshold(jj);
142
143         %Recording overall accuracy
144         accuracy(1,count) = (true_EB+true_freq)/no_of_datasets;
```

```
145
146         %Recording Youden statistic
147         Youden(1,count)= true_EB/(true_EB+false_freq)
148         +true_freq/(false_EB+true_freq)-1;
149
150         count = count+1;    %Advancing count variable
151
152     end
153 end
154
155 oneminspec = 1-spec;
156
157
158 %Determining the decision rule which provides the best overall
159 %accuracy, and the accuracy characteristics of this rule:
160 [Max_acc,I] = max(accuracy); Max_acc
161 classificationThreshold = CT(I)
162 predictionThreshold = PT(I)
163 oneminspec(I)
164 sens(I)
165 Bcoeff(I,:)
166
167 %Determining the decision rule that is furthest from the diagonal
168 %on the ROC curve, and the accuracy characteristics of this rule:
169 dist = abs(oneminspec-sens)./sqrt(2);
170 [Max_dist,J] = max(dist); Max_dist
171 accuracyAttained = accuracy(J)
172 classificationThreshold = CT(J)
173 predictionThreshold = PT(J)
174 oneminspec(J)
175 sens(J)
176 Bcoeff(J,:)
177
178 %Determining the decision rule that is closest to (0,1) on the ROC
179 %curve, and the accuracy characteristics of this rule:
180 dist2 = sqrt(oneminspec.^2+(1-sens).^2);
181 [Min_dist,K] = min(dist2); Min_dist
182 accuracyAttained = accuracy(K)
183 classificationThreshold = CT(K)
184 predictionThreshold = PT(K)
185 oneminspec(K)
186 sens(K)
```

```
187 Bcoeff(K,:)
188
189 %Plotting the ROC curve, graphically representing the strength of
190 %all considered prediction rules:
191 scatter(oneminspec,sens)
192
193 a = sum(sens+spec-1-Youden);
194
195 z = 1 %#ok<*NOPRT> suppressing warnings about ; at end of lines
```

freqbayes.m

```
%Calculates the proportions of data sets for which empirical Bayes
%outperforms frequentist and vice versa, produces a scatter plot
%indicating this classification against the number of populations
%and sample size within the data set.

function [propb, propf] = freqbayes(trueclassification,sizes,B)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Inputs:

%trueclassification - vector of 1's and 0's indicating whether
%                     EB or freq performed best

%sizes              - array of dataset sizes

%B                  - no. of batch replications

%Outputs:

%propb -  proportion of datasets in which EB outperformed

%propf - proportion of datasets in which freq outperformed

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Pops = sizes(2,:);                          %Extracting no. of pop.
                                            %from each dataset
Sample_size = sizes(1,:)/B;                 %Extracting sample size
                                            %from each dataset
total_number = length(trueclassification); %Extracting number of
                                            %datasets

%Finding the indices of the data sets which favour EB:
[~,J]=find(trueclassification);

%Calculating the number of data sets which favour EB:
bayes_number = length(J);

```

```
%Calculating the number of data sets which favour frequentist:
freq_number = total_number-bayes_number;

%calculating the proportion of data sets which favour
%EB and frequentist:
propb = bayes_number/total_number;
propf = freq_number/total_number;

%Preallocating vectors to store the number of populations
%and sample sizes in the data sets which favour EB:
pop = zeros(1,bayes_number);
ss = zeros(1,bayes_number);

%Determining the number of populations and samples sizes
%in the EB favourable data sets:
for i = 1:bayes_number
    pop(i) = Pops(J(i));
    ss(i) = Sample_size(J(i));
end

%Creating scatter plot indicating number of populations and
%sample sizes in EB data sets:
scatter(pop,ss)
hold

%Finding the indices of the data sets which favour frequentist:
altern = classification-ones(1,total_number);
[~,J]=find(altern);

%Preallocating vectors to store the number of populations
%and sample sizes in the datasets which favour frequentist:
pop2 = zeros(1,freq_number);
ss2 = zeros(1,freq_number);

%Determining the number of populations and samples sizes
%in the freq favourable data sets:
for i = 1:freq_number
    pop2(i) = Pops(J(i));
    ss2(i) = Sample_size(J(i));
end

%Creating scatter plot indicating number of populations and
```

```
82 %sample sizes in freq data sets (overlaying EB scatter plot):
83 scatter(pop2,ss2,'r>')
84
85 %Setting axis to a log scale:
86 set(gca,'xscale','log')
87 set(gca,'yscale','log')
88
89 hold
```

multinormdata.m

```
1  %Generates collection of data sets.
2
3  function [B,DATA,sizes,exact] = multinormdata()
4
5  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
6  %Inputting configuration parameters:
7
8  k = 1;
9
10 B = 10; %No. of replications/batches for each config.
11
12 K = [10,20,50];
13 Klen = length(K); %No. of pops in data sets
14
15 N = [3,5,10];
16 Nlen = length(N); %Sample sizes
17
18 UM = [5,10];
19 UMlen = length(UM); %Upper limits of mean range in data sets
20
21 USD = [1,sqrt(2)];
22 USDlen = length(USD); %Upper limits of std devs range in data sets
23
24 meanstep = round([2,3]);
25 msteplen  = length(meanstep); %No. of step values assumed
26                               %by means (assigned fully later)
27
28 sdstep = round([2,3]);
29 sdsteplen  =  length(sdstep); %No. of step values assumed
30                               %by std devs (assigned fully later)
31
32 am = 0; sdm = 1; %Lowest means and std devs in data sets
33
34
35 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
36 %Outputs
37
38 %B       - No. of replications/batches for each config.
39
```

```
40 %DATA    - 3D array of data sets
41
42 %sizes   - Matrix of data set sizes
43 %            1st row - sample size
44 %            2nd row - no. of pops
45
46 %exact   - Matrix of exact population means
47 %            for each data set
48
49 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
50 %Generating dataset
51
52 count = 0; %Initialisng count variable
53
54 %#ok<NOPRT> %Calculating the number of data sets in collection:
55 dataset_size = 2*Klen*Nlen*UMlen*USDlen*msteplen*sdsteplen
56
57 %Preallocating matrix to store data sets (oversized):
58 DATA = zeros(B*max(N),max(K),dataset_size);
59
60 %Preallocating matrix to store data set sizes for later use:
61 sizes = zeros(2,dataset_size);
62
63 %Preallocating matrix to store exact population means:
64 exact = zeros(dataset_size,max(K));
65
66 for i = 1:Klen
67     k = K(i);  %Setting the no. of populations within the data set
68
69     %Vector outlining the no. of step values assumed by means:
70     meanstep = round([2,k]); msteplen  = length(meanstep);
71
72     %Setting the no. of step values assumed by means:
73     sdstep = round([2,k]); sdsteplen  =  length(sdstep);
74
75     for j = 1:Nlen
76
77         %Setting the sample size for data set:
78         n = N(j);
79
80         %Calculating the total number of reps required:
81         tot_reps = n*B;
```

```
82
83          for i2 = 1:UMlen
84
85              %Setting the upper end of mean range:
86              aM = UM(i2);
87
88              for i3 = 1:msteplen
89
90                  %Setting the no. of distinct values for mean:
91                  aS = meanstep(i3);
92
93                  %Forming the mean values vector and calculating
94                  %the no. needed to allocate all pops:
95                  av = linspace(am,aM,aS); temp1 = ceil(k/aS);
96
97                  %Tiling the mean values vector to create pop means
98                  %vector (oversize):
99                  avg = repmat(av,1,temp1);
100
101                 %Cutting down mean value vector to size, removing
102                 %one each level until appropriately sized,
103                 %starting at LHS or RHS randomly:
104                 c1 =  round(rand); L1 = length(avg);
105                 avg = avg((c1+(1-c1)*(L1-k+1)):(c1*k+(1-c1)*L1));
106
107                 %Sorting to obtain mean value vector:
108                 avg = sort(avg);
109
110                 %Creating matrix of average values
111                 %for generating data set:
112                 Avg = repmat(avg,tot_reps,1);
113
114                 for j2 = 1:USDlen
115                     sdM = USD(j2);
116
117                     %Creating vector of std devs:
118                     for j3 = 1:sdsteplen
119                         sdS = sdstep(j3);
120                         sd = linspace(sdm,sdM,sdS);
121                         temp2 = ceil(k/sdS);
122                         sdev = repmat(sd,1,temp2);
123
```

```
124                         %Cutting down std dev vector:
125                         c2 =  round(rand); L2 = length(sdev);
126                         sdev =
127                         sdev((c2+(1-c2)*(L2-k+1)):(c2*k+(1-c2)*L2));
128                         sdev = sort(sdev);
129
130                         %Creating matrix of std devs:
131                         Sdev = repmat(sdev,tot_reps,1);
132                         %Creating reversed matrix of std devs:
133                         vedS = fliplr(Sdev);
134
135                         %Generating dataset with std dev ...
                               increasing with mean:
136                         data1 = Sdev.*randn(tot_reps,k)+Avg;
137                         %Generating dataset with std dev ...
                               increasing against mean:
138                         data2 = vedS.*randn(tot_reps,k)+Avg;
139
140                         %Calculating size of data set:
141                         [hpop,wpop] = size(data1);
142
143                         %Storing data set:
144                         DATA(1:hpop,1:wpop,2*count+1) = data1;
145                         DATA(1:hpop,1:wpop,2*count+2) = data2;
146
147                         exact(2*count+1,1:k) = avg;
148                         exact(2*count+2,1:k) = avg;
149
150                         %Storing data set sizes:
151                         sizes(1,2*count+1) = hpop; ...
                               sizes(2,2*count+1) = wpop;
152                         sizes(1,2*count+2) = hpop; ...
                               sizes(2,2*count+2) = wpop;
153
154                         %Advancing count variable:
155                         count = count+1;
156
157                         %If no. of datasets is <= 50,
158                         %then histograms generated.
159
160                         if dataset_size <51
161                             figure(2*count-1)
```

```
162                                 hist(data1(1:n,:),100)
163                                 figure(2*count)
164                                 hist(data2(1:n,:),100)
165                             end
166
167                         end
168                     end
169                 end
170             end
171         end
172 end
```

naiveEBunmod.m

```
%EB point estimation for population mean based on NLN model
%of data. Pools all model configurations (naive pooling
%strategy).

function [Thetahat,AVG] = ...
    naiveEBunmod(DATA,sizes,B,no_of_datasets)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Inputs:
%DATA - data sets
%sizes - matrix of data set sizes
%B - no. of batch replications
%no_of_datasets - the number of data sets in DATA

%Outputs are:
%AVG - standard estimates for each pop in each data set
%Thetahat - EB estimates for each pop in each data set

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Calculating the max no. of pops in any data set:
dim = max(sizes(2,:));

%Preallocating matrix for emp Bayes estimators:
Thetahat = zeros(B,dim,no_of_datasets);

%Preallocating matrix for standard estimator:
AVG = zeros(B,dim,no_of_datasets);

for i9 = 1:no_of_datasets

    %Calculating size of simulation data, h by w:
    h = sizes(1,i9);
    w = sizes(2,i9);

    %Recovering individual data sets:
    data = DATA(1:h,1:w,i9);
    sample_size = h/B;

```

```
    %Initialising avg and vara:
    avg = zeros(B,w);
    vara = zeros(B,w);

    %Calculating sample mean of each batch of each population:
    for i=1:B
        for j=1:w
            avg(i,j)=
            mean(data(((i-1)*sample_size+1):(i*sample_size),j));
        end
    end

    %Storing avg as frequentist estimates:
    AVG(1:B,1:w,i9) = avg;


    %Calculating sample variance of each batch of each population:
    for i=1:B
        for j=1:w
            vara(i,j)=
            var(data(((i-1)*sample_size+1):(i*sample_size),j))/
            sample_size;
        end
    end

    %Calculating fixed constants:
    d=sample_size-1; m=psi(0,d/2)-log(d/2); sigch2=psi(1,d/2);

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Estimating hyperparameters of variance prior%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    q1=log(vara)-m; q2=q1.^2;

    %Initialising muhatv_c:
    muhatv_c=zeros(B,1);

    %Calculating muhatv_c
    for i=1:B
        muhatv_c(i)=sum(q1(i,:))/w;
    end

    %Initialising tauhatv2_c:
```

```
81      tauhatv2_c=zeros(B,1);
82
83      %Calculating tauhatv2_c:
84      for i=1:B
85          tauhatv2_c(i)=max(sum(q2(i,:))/w-muhatv_c(i)^2-sigch2,0);
86      end
87
88      %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
89      %Calculating EB estimates of variance%
90      %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
91
92      %Initialising Mhatv_c:
93      Mhatv_c=zeros(B,1);
94
95      %Calculating Mhatv_c:
96      for i=1:B
97          Mhatv_c(i)=tauhatv2_c(i)/(tauhatv2_c(i)+sigch2);
98      end
99
100     %Initialising sighat2_EB:
101     sighat2_EB=zeros(B,w);
102
103     %Calculating sighat2_EB:
104     for i=1:B
105         for j=1:w
106             sighat2_EB(i,j)=exp(Mhatv_c(i)*q1(i,j)
107             +(1-Mhatv_c(i))*muhatv_c(i));
108         end
109     end
110
111     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
112     %Estimating hyperparameters of mean prior%
113     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
114     q3=avg./sighat2_EB; q4=1./sighat2_EB;
115
116     %Initialising muhat_c:
117     muhat_c=zeros(B,1);
118
119     %Calculating muhat_c:
120     for i=1:B
121         muhat_c(i)=sum(q3(i,:))/sum(q4(i,:));
122     end
```

```
123
124     %Initialising mat:
125     mat=zeros(B,w);
126
127     %Calculating mat:
128     for i=1:B
129         for j=1:w
130             mat(i,j)=(avg(i,j)-muhat_c(i))^2
131             -vara(i,j)*exp(-m-sigch2/2);
132         end
133     end
134
135     %Initialising tauhat2_c:
136     tauhat2_c = zeros(B,1);
137
138     %Calculating tauhat2_c:
139     for i=1:B
140         tauhat2_c(i)= max(sum(mat(i,:))/w,0);
141     end
142
143     %Initialising Mhat and thetahat:
144     Mhat=zeros(B,w);
145     thetahat=zeros(B,w);
146
147     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
148     %Calculating EB estimates of mean%
149     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
150
151     %Calculating the shrinkage weights for the EB estimate:
152     for i=1:B
153         for j=1:w
154             Mhat(i,j)=tauhat2_c(i)/(sighat2_EB(i,j)+tauhat2_c(i));
155
156         end
157     end
158
159     %Calculating the EB estimates:
160     for i=1:B
161         for j=1:w
162             thetahat(i,j)=Mhat(i,j)*avg(i,j)
163             +(1-Mhat(i,j))*muhat_c(i);
164         end
```

```
165       end
166
167       %Storing EB estimates for i9th data set:
168       Thetahat(1:B,1:w,i9) = thetahat;
169
170   end
171   end
```

testing.m

```
%Calculates the sensitivity, specificity and overall accuracy
%of a given logistic regression predictor over a collection of
%data sets, based on multiple replication testing of EB vs freq.

function [sens,spec,accuracy] = ...
    testing(prediction_threshold,b,aveSSEeb_aveSSEst, ...
    no_of_datasets, Ratios, sizes, B)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Inputs:

%prediction_threshold - value s.t. if hypothesis exceeds this,
%                       then we predict EB positive.

%b                    - coefficients of fitted predictor.

%aveSSEeb_aveSSEst    - ratio of mean SSE EB vs mean SSE freq.

%no_of_datasets       - number of data sets.

%Ratios               - variation ratios for each data set.

%sizes                - array of data set sizes.

%B                    - no. of batch replications.


%Outputs:

%sens     - the sensitivity of our predictor.

%spec     - the specificity of our predictor.

%accuracy - overall accuracy of our predictor,
%           percent of predictions correct.


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
38
39 %Calculating the no. of pops and sample size in each data set:
40 Pops = sizes(2,:);
41 Sample_size = sizes(1,:)/B;
42
43 %Classifying data sets regarding whether EB or freq is superior:
44
45 %Preallocating:
46 trueclassification = zeros(1,no_of_datasets);
47
48 %True classification based on repeated sampling:
49 for i = 1:no_of_datasets
50     if aveSSEeb_aveSSEst(1,i) < 1
51         trueclassification(1,i) = 1;
52     else
53         trueclassification(1,i) = 0;
54     end
55 end
56
57 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
58
59 z1 = b(1)+b(2)*log(Ratios)+b(3)*log(Pops)+b(4)*log(Sample_size);
60
61 gamma = prediction_threshold-0.5;
62 prediction = round((1./(1+exp(-z1)))-gamma);
63
64 no_of_EB = sum(trueclassification);
65 no_of_Freq = no_of_datasets - no_of_EB;
66
67
68 %Initialising counts of each category:
69 true_freq = 0; false_freq = 0; true_EB = 0; false_EB = 0;
70
71 %Preallocating vector to store:
72 score = zeros(1,no_of_datasets);
73
74 %Categorising each prediction as false/true positive,
75 %or false/true negative. Done by computing the two digit value
76 %trueclassification*10+prediction:
77
78 % 00 corresponds to a prediction of negative which is correct
79 % 01      "              "       of positive which is incorrect
```

```
% 10      "               "       of neagtive which is incorrect
% 11      "               "       of positive which is correct

for i = 1:no_of_datasets
    score(1,i) = trueclassification(1,i)*10+prediction(1,i);
    if score(1,i) == 0
        true_freq = true_freq+1;
    elseif score(1,i) == 1
        false_EB = false_EB+1;
    elseif score(1,i) == 10
        false_freq = false_freq+1;
    else
        true_EB = true_EB+1;
    end
end

%Calculating the sensitivity of the predictor:
sens = true_EB/no_of_EB;

%Calculating the specificity of the predictor:
spec = true_freq/no_of_Freq;

%Calculating the overall accuracy of the predictor:
accuracy = (true_EB+true_freq)/no_of_datasets;
```

variationratio.m

```
1  %Calculates ratio of within sample variation to
2  %across sample variation.
3
4  function [Ratios,ratios,no_of_datasets] = ...
       variationratio(DATA,sizes,B)
5
6  %Inputs:
7
8  % DATA     -  Collection of data sets.
9  % sizes    -  Matrix of sizes of each data set to allow extraction.
10 % B        -  Number of batches/replications.
11
12 %Outputs:
13
14 % no_of_datasets  - Number of data sets.
15 % ratios          - 'anova' ratios for each batch of
16 %                    each data set.
17 % Ratios          - 'anova' ratio for each data set
18 %                    averaged across batches.
19
20
21 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
22
23 %Calculating the number of datasets in collection.
24
25 %For multiple(>1) datasets:
26 %size returns 3-vector such that temp2 = 3,
27 %and number of datasets is given by
28 %the third element of temp.
29
30 %For single data set:
31 %size returns a 2-vector and temp2 = 2.
32
33 temp = size(DATA);
34 temp2 = size(temp);
35
36 if temp2(2) == 3
37     no_of_datasets = temp(3);
38 else
```

```
39      no_of_datasets = 1;
40  end
41
42  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
43
44  %Calculating the variation ratios for data sets:
45
46  %Preallocating vector to store ratios:
47  ratios = zeros(B,no_of_datasets);
48
49  %Stepping though each data set 1 by 1:
50  for i = 1:no_of_datasets
51
52      %Extracting size of the ith data set, h = height w = width:
53      h = sizes(1,i);
54      w = sizes(2,i);
55
56      %Extracting the ith data set:
57      data = DATA(1:h,1:w,i);
58
59      %Calculating the sample size from the ith data set:
60      sample_size = h/B;
61
62      for j = 1:B %Stepping through batches for ith data set:
63
64          %Extracting data relevant to jth batch:
65          working_data = data((j-1)*sample_size+1:j*sample_size,:);
66
67          %Calculate mean of each population in jth batch:
68          group_means = mean(working_data);
69
70          %Calculate overall mean of jth batch:
71          total_mean = mean(working_data(:));
72
73          %Calculate measure of variability across groups:
74          across_group_dev = sample_size*sum((group_means - ...
                repmat(total_mean,1,w)).^2)/w;
75
76          %Calculate measure of variability within groups:
77          within_group_dev = sum(sum((working_data - ...
                repmat(group_means,sample_size,1)).^2))/sample_size;
78
```

```
79          %Checking denominator non-zero, calculate variation ratio:
80          if across_group_dev ~= 0
81              %Take ratio of measures for each batch of data set:
82              ratios(j,i) = (within_group_dev)/(across_group_dev);
83          else
84              0     %#ok<NOPRT> If denominator of ratio is 0,
85                    %then outputs 0 as warning.
86          end
87      end
88  end
89
90  Ratios = mean(ratios); %Calculates mean ratio across batches ...
        for each data set.
```

Appendix C

Example MATLAB Codes for Chapter 5

EB.m

```
%EB point estimation for population mean based on NLN model
%of data. Using both a naive pooling strategy (pooling all data
%equally) and a weighted pooling strategy, whereby for
%each population estimate more significance is placed on samples
%which are similar.

function [Thetahat,Thetahatp,muhat_c,muhatp_c] = ...
    EB(DATA,sizes,B,no_of_datasets,Weights)

%Inputs:
%DATA - data sets.
%sizes - matrix of data set sizes.
%B - batch size (constant).
%no_of_datasets - number of data sets in DATA.

%Outputs are:
%AVG - standard estimates for each pop. in each data set.
%Thetahat - EB estimates for each pop. in each data set.

%Calculating the max no. of pops in any data set:
dim = max(sizes(2,:));

```

```
%Initialising matrix for emp Bayes estimators:
Thetahat = zeros(B,dim,no_of_datasets);
%Initialising matrix for emp Bayes estimators:
Thetahatp = zeros(B,dim,no_of_datasets);

for i9 = 1:no_of_datasets

    %Calculating size of simulation data, h by w:
    h = sizes(1,i9);
    w = sizes(2,i9);

    %Recovering individual data sets:
    data = DATA(1:h,1:w,i9);
    sample_size = h/B;

    %Initialising avg and vara:
    avg = zeros(B,w);
    vara = zeros(B,w);

    %Calculating batched sample means and variances:
    for i=1:B
        for j=1:w
            avg(i,j) = ...
               mean(data(((i-1)*sample_size+1):(i*sample_size),j));
            vara(i,j) = var(data(((i-1)*sample_size+1)
            :(i*sample_size),j))/sample_size;
        end
    end

    %Calculating fixed constants:
    d = sample_size-1; m = psi(0,d/2)-log(d/2); sigch2 = ...
       psi(1,d/2);

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Estimating hyperparameters of variance prior%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    q1 = log(vara)-m; q2 = q1.^2;

    %Initialising muhatv_c:
    muhatv_c = zeros(B,1);

    %Calculating muhatv_c:
```

```
    for i=1:B
        muhatv_c(i) = sum(q1(i,:))/w;
    end

    %Initialising tauhatv2_c:
    tauhatv2_c = zeros(B,1);

    %Calculating tauhatv2_c:
    for i=1:B
        tauhatv2_c(i) = ...
           max(sum(q2(i,:))/w-muhatv_c(i)^2-sigch2,0);
    end

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Calculating EB estimates of variance%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    %Calculating Mhatv_c:
    Mhatv_c = tauhatv2_c./(tauhatv2_c+sigch2);

    %Initialising sighat2_EB:
    sighat2_EB = zeros(B,w);

    %Calculating sighat2_EB:
    for i=1:B
        t1 = (1-Mhatv_c(i))*muhatv_c(i);
        for j=1:w
            sighat2_EB(i,j) = exp(Mhatv_c(i)*q1(i,j)+t1);
        end
    end

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Estimating hyperparameters of mean prior%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    q3 = avg./sighat2_EB; q4=1./sighat2_EB;

    %Initialising grand means:

    %Preallocating vector of grand means:
    muhat_c = zeros(B,1);
    %Preallocsting matrix of weighted grand means:
    muhatp_c = zeros(B,w);
```

```
103
104      %Calculating the grand mean muhat_c for each batch ...
             replication:
105      for i=1:B
106          muhat_c(i) = sum(q3(i,:))/sum(q4(i,:));
107      end
108
109     %Calculating the vector of weighted grand means muhatp_c
110     %for each batch replication:
111      for i=1:B
112          t2 = sum(q4(i,:));
113          for j = 1:w
114               muhatp_c(i,j) = ...
                     sum(transpose(squeeze(Weights(i9,j,1:w,i))).
115               *q3(i,1:w))/t2;
116          end
117      end
118
119      %Preallocating vectors and calculating quantities used in
120      %upcoming calculations of the estimators of tau^2:
121      mat=zeros(B,w);
122      matp=zeros(B,w);
123
124      for i = 1:B
125          mat(i,1:w) = ...
                 (avg(i,1:w)-muhat_c(i)).^2-vara(i,1:w)*exp(-m-sigch2/2);
126          matp(i,1:w) = (avg(i,1:w)-muhatp_c(i,1:w)).^2
127          -vara(i,1:w)*exp(-m-sigch2/2);
128      end
129
130      %Preallocating tauhat2_c and tauhat2p_c:
131      tauhat2_c = zeros(B,1);
132      tauhat2p_c = zeros(B,1);
133
134      %Calculating tauhat2_c:
135      for i=1:B
136          tauhat2_c(i)= max(sum(mat(i,:))/w,0);
137      end
138
139      %Calculating the weighted version of tauhat2_c:
140      for i = 1:B
141         tauhat2p_c(i)= max(sum(matp(i,:))/w,0);
```

```
    end

    %Preallocating shrinkage weights and EB estimates of mean:

    Mhat=zeros(B,w);
    Mhatp=zeros(B,w);
    thetahat=zeros(B,w);
    thetahatp=zeros(B,w);

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Calculating EB estimates of mean%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    %Calculating shrinkage weights both standard
    %and fully weighted:
    for i=1:B
        for j=1:w
            Mhat(i,j)=tauhat2_c(i)/(sighat2_EB(i,j)+tauhat2_c(i));
            Mhatp(i,j)=tauhat2p_c(i)/
            (sighat2_EB(i,j)+tauhat2p_c(i));
        end
    end

    %Calculating EB estimates of mean:
    for i=1:B
        for j=1:w
            thetahat(i,j)=Mhat(i,j)*avg(i,j)
            +(1-Mhat(i,j))*muhat_c(i);      %standard EB estimates
            thetahatp(i,j)=Mhatp(i,j)*avg(i,j)
            +(1-Mhatp(i,j))*muhatp_c(i,j);  %weighted EB estimate
        end
    end

    %Storing EB estimates of mean:
    Thetahat(1:B,1:w,i9) = thetahat;
    Thetahatp(1:B,1:w,i9) = thetahatp;

end
end
```

errors.m

```
%Calculates the average ratio of sum of squared error and
%ratio of maximum absolute error for weighted EB methods
%against the standard EB method across collection of data sets.

function [SSE,ME] = ...
    errors(Thetahat,Thetahatp,exact,no_of_datasets,B,sizes)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%Inputs:

%Thetahat, Thetahatp    - EB mean estimates
%exact                  - exact values of population means
%no_of_datasets         - no. of data sets in our collection
%B                      - no. of batch replications
%sizes                  - matrix of data set sizes

%Outputs:

%SSE - mean (across data sets) of the ratio of mean (across) batch
%of sum of squared error for fully weighted.
%ME - mean (across data sets) of the ratio of mean (across) batch
%of max absolute error for fully weighted.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Preallocating vectors to store error values:
sum_sq_err_ratio = zeros(1,no_of_datasets);
max_err_ratio = sum_sq_err_ratio;
E1 = zeros(1,B); E2 = E1; E3 = E1; E4 = E1;

%Stepping through data sets one by one:
for i = 1:no_of_datasets

    %Extracting the no. of populations in ith data set:
    w = sizes(2,i);

    for j = 1:B

        %Extracting the EB estimates and exact population means
```

```
39          %for data set/batch:
40          thetahat = Thetahat(j,1:w,i);
41          thetahatp = Thetahatp(j,1:w,i);
42          exactvalues = exact(i,1:w);
43
44          %Calculating the sum of squared error for each EB estimate
45          %for data set/batch:
46          sumsqerEB = sum((thetahat-exactvalues).^2);
47          sumsqerWEB = sum((thetahatp-exactvalues).^2);
48
49          %Calculating the max absolute error for each EB estimate
50          %for data set/batch:
51          maxerEB = max(abs(thetahat-exactvalues));
52          maxerWEB = max(abs(thetahatp-exactvalues));
53
54          %Storing the errors for the jth batch:
55          E1(j) = sumsqerEB;
56          E2(j) = sumsqerWEB;
57          E3(j) = maxerEB;
58          E4(j) = maxerWEB;
59      end
60
61      %Taking the ratio (weighted vs standard) of the means of
62      %each error across batch replications:
63
64      %Fully weighted vs standard:
65      sum_sq_err_ratio(i) = mean(E2)/mean(E1);
66      %Fully weighted vs standard:
67      max_err_ratio(i) = mean(E4)/mean(E3);
68  end
69
70  %Taking the mean of each error ratio across our data sets:
71  SSE = mean(sum_sq_err_ratio);    % sse error
72  ME = mean(max_err_ratio);        % max abs error
73
74  end
```

weights.m

```
1  %Calculates weights for the pooling of data in the EB
2  %calculations. Weights are based on p-values from Welch's t test
3  %for equality of population mean between two samples.
4  %The higher the p-value, the larger the weighting placed on
5  %the sample data in the estimation process.
6
7  function [no_of_datasets,Pvalues,Weights] = ...
       weights(B,DATA,sizes,l)
8
9  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
10
11 %Inputs:
12
13 %B     - no. of batch replications within our data sets
14
15 %DATA  - our collection of data sets
16
17 %sizes - matrix containing the sizes of each data set
18 %        to allow extraction
19
20 %l     - parameter to select a data set for which to produce
21 %        a weight heatmap
22
23 %Outputs:
24
25 %no_of_dataset - number of datasets in our collection
26
27 %P-values      - array of p-values from p-wise t tests between
28 %                samples in data sets
29
30 %Weights       - weightings for the EB estimate, based on p-values
31
32 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
33
34 %Computing the number of data sets in collection:
35
36 temp = size(DATA);
37 temp2 = size(temp);
38
```

```
39 if temp2(2) == 3
40     no_of_datasets = temp(3);
41 else
42     no_of_datasets = 1;
43 end
44
45
46 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
47
48 %Finding the largest no. of samples in a data set:
49 maxpops = max(sizes(2,:));
50
51 %Preallocating arrays to store p-values and weights:
52 Pvalues = zeros(no_of_datasets,maxpops,maxpops,B);
53 Weights = zeros(no_of_datasets,maxpops,maxpops,B);
54
55 %Stepping through the data sets one by one:
56 for K = 1:no_of_datasets
57
58     %Extracting the size of the Kth data set:
59     h = sizes(1,K);
60     w = sizes(2,K);
61
62     %Extracting the Kth data set:
63     data = DATA(1:h,1:w,K);
64
65     %Calculating the sample size for the Kth data set:
66     b = fix(h/B);
67
68     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
69
70     %Calculating batch means and variances:
71
72     %Initialising avg and vara matrices:
73     avg = zeros(B,w);
74     vara = zeros(B,w);
75
76     for i = 1:B
77         for j = 1:w
78             avg(i,j) = mean(data(((i-1)*b+1):(i*b),j));
79             vara(i,j) = var(data(((i-1)*b+1):(i*b),j));
80         end
```

```
    end

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    %Calculating numerator of t-statistic:

    Diff = zeros(w,w,B);

    for k = 1:B
        for i = 1:w-1
            for j = i+1:w
                Diff(i,j,k) = avg(k,i) - avg(k,j);
            end
        end
    end

    diff = zeros(w,w,B);

    for k = 1:B
        diff(:,:,k) = Diff(:,:,k)-transpose(Diff(:,:,k));
    end

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    %Calculating denominator of t-statistic:

    Ssterr_sq = vara./b;

    summation = zeros(w,w,B);

    %For each i,j summation, Ssterr from each:

    for k = 1:B
        for i = 1:w
            for j = 1:w
                summation(i,j,k) = Ssterr_sq(k,i)+Ssterr_sq(k,j);
            end
        end
    end

    %Take square root of summation, to obtain the standard error
    %for difference in two means:
```

```
123
124     div = sqrt(summation);
125
126     %Calculating t-stat, difference divided by standard error:
127
128     tstat=diff./div;
129
130     %Taking absolute value of t-stat:
131
132     tstatabs = abs(tstat);
133
134     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
135
136     %Calculating degrees of freedom:
137
138     %Numerator:
139     num=summation.^2;
140
141     %Denominator:
142     denom = Ssterr_sq.^2;
143     denomf = zeros(w,w,B);
144
145     for k = 1:B
146         for i = 1:w
147             for j = 1:w
148                 denomf(i,j,k)=denom(k,i)+denom(k,j);
149             end
150         end
151     end
152
153     denomfnl = (1/(b-1))*denomf;
154     df = num./denomfnl;
155
156     %Preallocating data set p-value array:
157     pvalues = zeros(w,w,B);
158
159     %Calculating the p-values for the Kth dataset:
160     for k = 1:B
161         for i = 1:w
162             for j = 1:w
163                 pvalues(i,j,k) = ...
                        (2*(1-tcdf(tstatabs(i,j,k),df(i,j,k))));
```

```
164             end
165         end
166     end
167
168     %Storing the p-values:
169     Pvalues(K,1:w,1:w,1:B) = pvalues;
170
171     %Computing the weights for the Kth dataset:
172     for batch = 1:B
173         for i = 1:w
174             for j = 1:w
175                 Weights(K,i,1:w,batch) =  w*Pvalues(K,i,1:w,batch)
176                 /sum(Pvalues(K,i,1:w,batch));
177             end
178         end
179     end
180
181     K  %#ok<NOPRT>  Counter to indicate progress.
182 end
183
184 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
185
186 %Heatmap displaying the weights of the l^th dataset:
187
188 W = sizes(2,l);
189 heat = zeros(W,W);
190
191 for i = 1:W
192     for j = 1:W
193         heat(i,j) = Weights(l,i,j,1);
194     end
195 end
196
197 imagesc(p)
```

Appendix D

Pooling Investigation - Chapter 5

Empirical Bayes Methods for Discrete Event Simulation Performance Measure Estimation

Shona Blair, Tim Bedford and John Quigley
Dept. of Management Science, University of Strathclyde, Glasgow, G1 1QE, UK,
{shona.blair, tim.bedford, j.quigley}@strath.ac.uk

Discrete event simulation (DES) is a widely-used operational research methodology facilitating the analysis of complex real-world systems. Although, generally speaking, simplicity is greatly desirable in DES modelling applications, in many cases the nature of the underlying system results in simulation models which are large in scale, complex, and expensive to run. As such, the careful design and analysis of simulation experiments is essential to ensure valid and efficient inference concerning DES model performance measures. It is envisaged that empirical Bayes (EB) methods, which enable data to be pooled across a set of populations to support inference of the parameters of a single population, may be of use within this context. Despite this potential, EB has so far been neglected within the DES literature. This paper presents a preliminary computational investigation into the efficacy of EB procedures in the estimation of DES performance measures. The results of this investigation, and their significance, are explored. Additionally, likely directions for future research are also addressed.

1998 ACM Subject Classification: I.6.8 Types of Simulation

Key words: Discrete Event Simulation, Analysis Methodology, Empirical Bayes.

1. Introduction and Motivation

Discrete event simulation (DES) is a powerful and flexible methodology, widely utilized in OR applications for the design, analysis and improvement of complex, dynamic and stochastic real-world systems. At its core, DES involves abstracting the fundamental structure of the system of interest and using this information to construct a computer model of the system. A process of experimentation is conducted with the computer model in order to gain insight into and understanding of the performance of the real system. One of the key advantages of discrete event simulation is its ability to incorporate a "realistic" level of system complexity into the analysis process, when compared with the more rigid assumptions of alternative modelling techniques; indeed, DES is frequently referred to as a "method of last resort" [154]. Whilst the benefits of simple models are well understood and widely disseminated (see, for example, [184, 189]), there are many instances when the nature of the underlying system being studied necessitates the use of simulation models which are large-scale, structurally complex, difficult to interpret and computationally expensive to run. As such, the careful design and analysis of simulation experiments is necessary to ensure valid and efficient inference concerning model performance [145].

Empirical Bayes (EB) procedures offer a structured and theoretically sound framework for the pooling of data obtained across a set of populations to support inference concerning the parameters of an individual population. This often enables more efficient inference in situations which feature a repeated structure, providing that sufficient "similarity" exists between component populations. (For a general EB reference, see [46]). It seems intuitively reasonable that such an approach may be of benefit in simulation model experimentation, owing to the underlying similarity between simulation model configurations. In light of the computational expense involved in executing simulation models (touched on above), such increased efficiency in estimation would likely prove highly advantageous in practice. Yet, in spite of this apparent potential, EB has so far been neglected within the simulation literature.

This paper presents the results of a preliminary computational investigation into the use of EB procedures in the estimation of DES model performance measures.

It begins, in Section 2, with the presentation and brief derivation of the EB procedures which are to be applied. Then, in Section 3, the DES model selected for testing is introduced, the reasons behind this choice outlined and certain theoretical results regarding performance measures of interest are presented. After this, in Section 4, the experimental design of the study is described, before the results are summarised in Section 5. The paper concludes, in Section 6, with some discussion of how this research area might be further explored in the future.

2. Introducing the Empirical Bayes Procedures

Empirical Bayes procedures feature a hierarchical model structure, identical to that of a traditional Bayesian analysis. As such, we typically have a situation in which model parameters are themselves represented by probability distributions, termed "prior" distributions. In a Bayesian analysis, the prior distribution would be subjectively determined, usually elicited from subject matter experts. However, in an empirical Bayes setting the data themselves determine the prior distribution. As mentioned in the previous section, EB methods are well-suited to applications featuring a large number of "similar" populations or processes. In this situation, the data obtained from each of the populations are pooled and used to provide inference on a general prior distribution. This general prior distribution is then combined with the individual samples from each of the populations using standard Bayesian updating to obtain a "posterior" distribution specific to each of the populations. (For a detailed overview of the above theory and terminology, please refer to [46].)

EB methods have a long history, with their roots in actuarial work on credibility theory, the first major publication by Robbins [187] in 1955 and a series of landmark papers in the 1970's by Efron and Morris [78, 79, 80] (see [160] for a more detailed account of their development). However, recent years have seen a huge upsurge in the volume of EB publications. This is due predominantly to scientific advances such as microarray technology, facilitating high-throughput biological screening and generating massive data-sets that demand a fresh approach to statistical analysis [76]. A frequent feature of such data sets is their large number of populations, contrasted with relatively few observations from each. Such structures, as mentioned before, are ideally suited to an empirical Bayes analysis.

Indeed, many successful applications have been published; a recent survey being [22]. Not surprisingly, this renewed interest has led to methodological and theoretical developments, as well as applications.

Here, however, we focus on some specific results which show particular promise in terms of their potential applicability within the context of DES model analysis. The empirical Bayes estimator shall employ is the "double-shrinkage" estimator presented in article [260]. This estimator assumes a normal/lognormal model of the data and its derivation is presented in the following subsections.

2.1. Assumptions

For $i = 1, 2, \ldots, p$, we assume:

$$X_i|\theta_i, \sigma_i^2 \overset{ind}{\sim} N(\theta_i, \sigma_i^2) \tag{D.1}$$

$$\theta_i \overset{iid}{\sim} N(\mu, \tau^2) \tag{D.2}$$

$$\log \sigma_i^2 \overset{iid}{\sim} N(\mu_\upsilon, \tau_\upsilon^2) \tag{D.3}$$

$$\log(S_i^2/\sigma_i^2) \overset{iid}{\sim} N(m, \sigma_{ch}^2) \tag{D.4}$$

where $m = E[\log(\chi_d^2)] = \psi(\frac{d}{2}) - \log(\frac{d}{2})$ and $\sigma_{ch}^2 = Var[\log(\chi_d^2)] = \psi'(\frac{d}{2})$, with d denoting the degrees of freedom. Initially, we suppose that hyperparameters $\mu, \tau^2, \mu_\upsilon$ and τ_υ^2 are known.

2.2. Derivation

Given a sample of n observations, x_i, from sampling distribution (D.1) for population i, the prior distribution (D.2) on θ_i, and, for the moment, the additional assumption that σ_i^2 is known, a standard application of Bayes rule yields:

$$\theta_i|x_i, \sigma_i^2 \sim N(M_i\bar{x}_i + (1 - M_i)\mu, M_i\sigma_i^2), \tag{D.5}$$

where $M_i = \tau^2/(\tau^2 + \sigma_i^2/n)$ and $\bar{x}_i = \frac{1}{n}\sum_j x_{ij}$, for the posterior distribution of θ_i.

In such a case, we would use the posterior mean:

$$\hat{\theta}_i = M_i\bar{x}_i + (1 - M_i)\mu, \tag{D.6}$$

as a point estimator of θ_i.

However, the true population variances σ_i^2 for $i = 1, 2, \ldots, p$ are unknown, and as with [260] we adopt a lognormal prior (D.3) for σ_i^2, with the additional assumption (D.4) that S_i^2/σ_i^2 is also lognormally distributed (with parameters selected to coincide with those of χ_d^2/d, the standard distributional assumption regarding the quantity S_i^2/σ_i^2).

From (D.4), it follows that:

$$\log S_i^2 | \log \sigma_i^2 \quad \sim \quad N(m + \log \sigma_i^2, \sigma_{ch}^2), \tag{D.7}$$

and combining (D.3) and (D.7) using Bayes rule yields:

$$\log \sigma_i^2 | \log S_i^2 \quad \sim \quad N(M_v(\log S_i^2 - m) + (1 - M_v)\mu_v, M_v\sigma_{ch}^2), \tag{D.8}$$

where $M_v = \tau_v^2/(\tau_v^2 + \sigma_{ch}^2)$.

As in [260], we estimate this quantity using:

$$\hat{\sigma}_i^2 = \exp(M_v(\log S_i^2 - m) + (1 - M_v)\mu_v). \tag{D.9}$$

Thus, assuming known hyperparameters, we have the estimator:

$$\hat{\theta}_i = M_i\bar{x} + (1 - M_i)\mu, \tag{D.10}$$

with $M_i = \tau^2/(\tau^2 + \hat{\sigma}_i^2/n)$, where $\hat{\sigma}_i^2$ is as given by equation (D.9) above.

2.3. Estimation of Hyperparameters

All that remains is the estimation of the hyperparameters μ, τ^2, μ_v and τ_v^2. As with [260], we adopt the following estimators.

For μ_v and τ_v^2, we have:

$$\hat{\mu}_{\upsilon} = \frac{1}{p}\sum_{i}(\log S_i^2 - m) \quad \text{and} \quad \hat{\tau}_{\upsilon}^2 = (\frac{1}{p}\sum_{i}(\log S_i^2 - m)^2 - \sigma_{ch}^2 - \hat{\mu}_{\upsilon}^2)_+,$$

from which we obtain:

$$\hat{M}_{\upsilon} = \frac{\hat{\tau}_{\upsilon}^2}{\hat{\tau}_{\upsilon}^2 + \sigma_{ch}^2} \quad \text{and} \quad \hat{\sigma}_{EB,i} = \exp(\hat{M}_{\upsilon}(\log S_i^2 - m) + (1 - \hat{M}_{\upsilon}).$$

To estimate μ and τ^2, we use:

$$\hat{\mu} = \frac{\sum_i(\bar{x}_i/\hat{\sigma}_{EB,i})}{\sum_i(1/\hat{\sigma}_{EB,i})} \quad \text{and} \quad \hat{\tau}^2 = (\frac{\sum_i(\bar{x}_i - \hat{\mu})^2}{p})_+.$$

Thus, the 'double-shrinkage' empirical Bayes point estimator is given as:

$$\hat{\theta}_{EB,i} = \hat{M}_{EB,i}\bar{x}_i + (1 - \hat{M}_{EB,i})\hat{\mu}, \tag{D.11}$$

where $M_{EB,i} = \hat{\tau}^2/(\hat{\tau}^2 + \hat{\sigma}_{EB,i}^2/n)$, and $\hat{\mu}$, $\hat{\tau}^2$ and $\hat{\sigma}_{EB,i}^2$ are as given above.

3. Introducing the DES Test Model

The purpose of this study is to evaluate the application of the EB methodology to the estimation of DES model performance measures. Having already introduced the EB procedures which are to be evaluated, this section aims to discuss an appropriate DES model upon which to test the EB procedures.

The model to be used is a computer-based implementation of an $M/M/1$ queuing model. This simple model consists of a single-server queuing system with exponentially distributed inter-arrival and service times and a first in - first out (FIFO) queuing discipline.

Simple "artificial" DES models such as this are frequently used in research for the evaluation of DES model analysis techniques [119]. These test models offer the key

advantage that theoretical values are available for many performance measures of interest, and this knowledge greatly facilitates the testing of output analysis methods. One criticism which can be leveled at this approach is that such models bear little resemblance to the majority of DES models encountered in practice (those being significantly more complex). In light of this point, it is helpful to highlight the exploratory nature of the study; it should be emphasized that much of the value of this investigation lies in the issues it raises and the directions for further research which surface. This discussion is taken up again in Section 6 after the results are presented.

For now, some relevant queuing model theory and results, extracted from [**?**] and used in later sections of the paper, are presented.

3.1. $M/M/1$ Theory

As discussed above, an appropriate choice of DES test model for this investigation appears to be an $M/M/1$ model with a first in - first out queuing discipline. In this case, there are only two additional model parameters which may be varied, the arrival rate, denoted by λ, and the service rate, denoted by μ. In order to simplify our analysis, we note that these parameters may be combined to give a single parameter, namely the traffic intensity, which uniquely specifies a particular $M/M/1$ configuration. The traffic intensity parameter, denoted by ρ, is given by $\rho = \frac{\lambda}{\mu}$. We note that in future discussions, the particular $M/M/1$ configuration will be specified solely by reference to the traffic intensity, ρ.

Our performance measure of interest will be the steady-state (or long-run) expected time in system, which we denote by W. The exact value of this quantity for any $M/M/1$ model is given as a simple function of the arrival (λ) and service rate (μ) parameters, $W = \frac{1}{\mu-\lambda}$. This enables us to calculate the steady-state expected time in system exactly for any given $M/M/1$ model configuration.

4. Experimental Design

The aim of this section is to provide an overview of the experimental design employed during the computational testing. Although the key points will be pre-

sented, a more comprehensive picture of the research design may be obtained from visiting the following web address: `http://personal.strath.ac.uk/shona.blair/research/SCOR2012/`

Initially, it is important to mention that in this experiment, the performance of the EB estimator, $\hat{\theta}_{EB,i}$, (as given by (D.11) in Section 2), is evaluated relative to that of the standard estimator of a population mean, the sample mean $\bar{X}_i$. The sample mean is most prevalent point estimator of DES performance measures [154]; it offers the advantage of being easily interpretable and constitutes a convenient standard by which other methods may be compared.

In order to estimate the steady-state time in system, the test model had to be executed and time in system data collected. The relevant details concerning this stage of the experimentation are as follows:

- **Traffic intensities**: a mesh with lower limit $\rho = 0.02$, upper limit $\rho = 0.9$ and step-size $\rho = 0.02$ was used in the experiment. This resulted in 45 model configurations and gave a nice, fine grid.
- **Number of replications**: $200,000$ independent replications were made at each traffic intensity configuration to ensure sufficient data was collected.
- **Warm-up period**: a warm-up of 500 customers was used, ensuring that the model was in steady-state prior to data collection.
- **Run-length**: a run-length of 600 customers was used. The relatively short run-length (in comparison to the warm-up period) ensured the experiment reflected possible real-life data collection constraints.

This scheme yields a $200,000$ by 45 data matrix, where each data value represents an average time in system based on the first 100 customers after the aforementioned warm-up period.

Having obtained the necessary $M/M/1$ time in system data, we now discuss how this data can be used to calculate both EB and standard estimates. In the standard setting, we simply collect a batch of data from each traffic intensity configuration and calculate the sample mean to make inference regarding the true population

mean. The only decision to be made concerns the size of the batch of data. In the empirical Bayes setting, however, data obtained from other model configurations can be pooled to support inference of the population mean of a given configuration. Thus immediate decisions need to be made concerning the size of the batches to be used and the pooling strategy to be employed.

To gain as much insight as possible, a range of batch sizes were explored, from 3 to 20. This enabled us to understand how the batch size affected the performance of the estimation technique. Additionally, in terms of the pooling strategy employed, a simple two sample t-test was conducted systematically for each pair of traffic intensity configurations to test for homogeneity. This approach was adopted as it avoids subjective decisions, based on theoretical knowledge of the system, biasing the results of the study. A range of significance levels, from $0.05, 0.1, 0.15, ..., 0.95$, were used to explore the relationship between pooling strategy and EB performance. This design resulted in 342 possible combinations of batch size and significance level, each of which was evaluated in the course of the computational testing. For each of these batch size / significance level combinations, the large volume of $M/M/1$ data collected permitted the calculation of $10,000$ pairs (both EB and standard) of estimates for each of the 45 traffic intensities. It is thus convenient to consider two $10,000$ by 45 matrices (one for each estimator) associated with each of the 342 batch size / significance level combinations.

To assess the performance of each estimator, the estimated values were subtracted from the true values, the errors squared, and the averages calculated over the 45 traffic intensities. This created, for each estimator, $10,000$ mean squared error (MSE) values. These were averaged, to remove the issue of stochastic variability, and the square root was taken to obtain the root mean squared error (rMSE) for each estimator. In order to gauge the relative efficacy, the ratio of the EB rMSE over the standard rMSE was taken as our critical statistical metric of interest. Finally, note that both the $M/M/1$ model and the EB and standard data analysis procedures were implemented in Matlab R2011a and run on a standard desktop PC (Intel Core-i5, 3GHz, 8GB RAM).

5. Summary and Discussion of Experimental Results

This section of the paper briefly presents and discusses the key results obtained from the aforementioned program of experimentation. As in the previous section, we begin by noting that a comprehensive set of numerical results and a detailed analysis may be found at the same web address. As mentioned in the previous section, the key statistical metric of interest is the ratio of the rMSE of the EB estimator to the rMSE of the standard estimator. The value of this quantity for each combination of batch size and significance level can found in Table D.1 of Appendix A, which has been illustrated in Figure D.1.

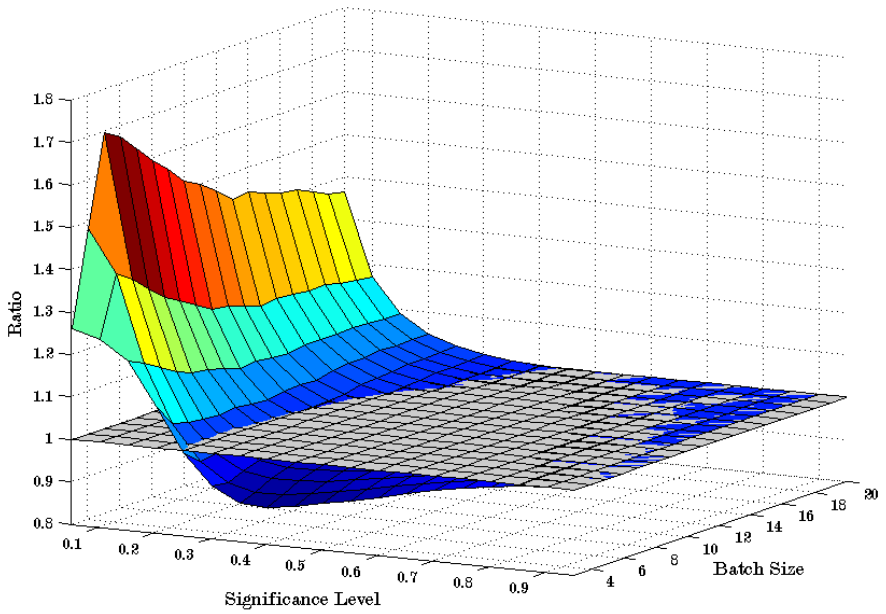

Figure D.1: Ratio of EB rMSE to standard rMSE: This depicts the relative performance of the estimators, and how this varies with batch size and significance level. The plane at ratio= 1 enables us to easily identify areas where the EB estimator outperforms the standard estimator.

From examination of Table D.1, the most favourable value of this ratio (0.8852) was obtained using a batch size of 3 and a significance level of 0.4, whilst the least favourable (1.6980) occurred for batch size 5 and a significance level of 0.05. Additionally, it appears that the optimal significance level, in terms of EB performance, is 0.4. As such, Figure D.2 illustrates the relative EB performance for this

significance level over varying batch size.

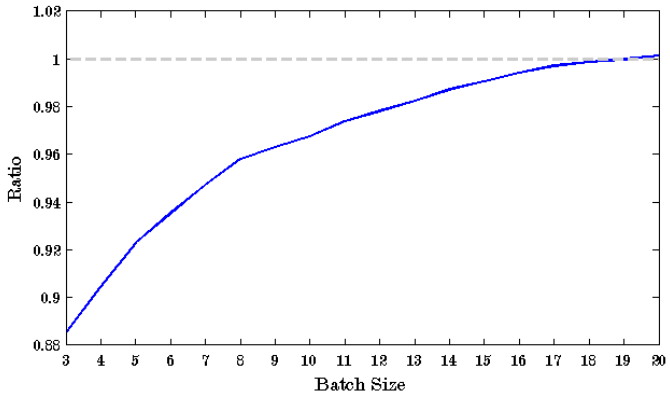

Figure D.2: rMSE ratio over varying batch size for optimal significance level.

6. Conclusion and Future Research

As may be seen from the results illustrated in the previous section, quantifiable benefit can be achieved from the application of EB procedures to DES performance measure estimation. This is a positive outcome to our pilot study which goes a significant way towards establishing the feasibility of the application of empirical Bayes to DES model performance measure estimation. However, it is also apparent that the batch size used and the pooling strategy adopted are critical to the realization of this benefit. In this study, the decision as to whether or not to pool model configurations was made on the outcome of a simple two sample t-test. Although there are many more options for statistical tests of homogeneity, little in the way of formal guidance exists specifically concerning empirical Bayes pooling strategies. More exploration should be done on this subject with the aim of providing statistical indicators to guide practitioners in how to apply this method.

DES models exhibit great variety, differing vastly in characteristics such as purpose, structure, scale, nature of input parameters and nature of output distributions. This complexity increases the challenges involved in attempting to apply empirical Bayes to model analysis and in attempting to provide detailed guidelines enabling practitioners to make use of this methodology. A comprehensive study evaluating the benefits of EB in relation to more “realistic” DES models forms part of our ongoing research.

Acknowledgements We would like to thank Stephen Chick for several fruitful discussions on this topic and for his involvement as external supervisor of S. Blair, and the Simul8 Corporation for their technical advice and general project support.

A. Detailed Results

Batch	Significance Level									
Size	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
3	1.2616	1.2425	1.1975	1.0992	**0.9934**	**0.9226**	**0.8887**	**0.8852**	**0.8939**	**0.9085**
4	1.4835	1.3842	1.1791	1.0449	**0.9616**	**0.9177**	**0.9026**	**0.9044**	**0.9134**	**0.9242**
5	1.6980	1.3588	1.1560	1.0339	**0.9666**	**0.9334**	**0.9215**	**0.9227**	**0.9290**	**0.9382**
6	1.6748	1.3238	1.1351	1.0287	**0.9711**	**0.9447**	**0.9346**	**0.9352**	**0.9393**	**0.9481**
7	1.6339	1.2898	1.1187	1.0255	**0.9782**	**0.9555**	**0.9471**	**0.9472**	**0.9506**	**0.9575**
8	1.5932	1.2686	1.1127	1.0302	**0.9877**	**0.9670**	**0.9585**	**0.9580**	**0.9602**	**0.9657**
9	1.5593	1.2473	1.1020	1.0265	**0.9891**	**0.9715**	**0.9635**	**0.9629**	**0.9651**	**0.9695**
10	1.5176	1.2244	1.0913	1.0243	**0.9907**	**0.9751**	**0.9682**	**0.9674**	**0.9697**	**0.9742**
11	1.4973	1.2189	1.0944	1.0311	**0.9990**	**0.9839**	**0.9763**	**0.9739**	**0.9755**	**0.9790**
12	1.4691	1.2055	1.0893	1.0323	1.0019	**0.9867**	**0.9801**	**0.9780**	**0.9790**	**0.9824**
13	1.4381	1.1922	1.0863	1.0312	1.0032	**0.9897**	**0.9842**	**0.9825**	**0.9826**	**0.9849**
14	1.4411	1.1983	1.0938	1.0391	1.0126	**0.9970**	**0.9900**	**0.9872**	**0.9870**	**0.9883**
15	1.4262	1.1947	1.0946	1.0444	1.0166	1.0018	**0.9939**	**0.9906**	**0.9904**	**0.9924**
16	1.4135	1.1900	1.0945	1.0439	1.0179	1.0042	**0.9976**	**0.9942**	**0.9932**	**0.9936**
17	1.4074	1.1910	1.0972	1.0492	1.0239	1.0092	1.0015	**0.9971**	**0.9961**	**0.9961**
18	1.3900	1.1816	1.0937	1.0476	1.0231	1.0094	1.0009	**0.9986**	**0.9965**	**0.9971**
19	1.3735	1.1769	1.0912	1.0484	1.0249	1.0119	1.0043	**0.9997**	**0.9979**	**0.9985**
20	1.3641	1.1718	1.0913	1.0496	1.0269	1.0131	1.0057	1.0013	**0.9992**	**0.9996**

Batch	Significance Level								
Size	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
3	**0.9216**	**0.9359**	**0.9509**	**0.9651**	**0.9774**	**0.9878**	**0.9954**	**0.9993**	1.0000
4	**0.9368**	**0.9499**	**0.9618**	**0.9735**	**0.9838**	**0.9921**	**0.9975**	**0.9999**	1.0000
5	**0.9490**	**0.9591**	**0.9698**	**0.9803**	**0.9886**	**0.9947**	**0.9989**	1.0000	1.0000
6	**0.9569**	**0.9669**	**0.9761**	**0.9849**	**0.9915**	**0.9966**	**0.9993**	1.0000	1.0000
7	**0.9657**	**0.9737**	**0.9812**	**0.9885**	**0.9945**	**0.9983**	**0.9997**	1.0000	1.0000
8	**0.9718**	**0.9788**	**0.9861**	**0.9922**	**0.9967**	**0.9990**	**0.9999**	1.0000	1.0000
9	**0.9754**	**0.9820**	**0.9886**	**0.9937**	**0.9974**	**0.9995**	**0.9999**	1.0000	1.0000
10	**0.9788**	**0.9845**	**0.9907**	**0.9953**	**0.9981**	**0.9994**	1.0000	1.0000	1.0000
11	**0.9831**	**0.9884**	**0.9925**	**0.9964**	**0.9990**	**0.9999**	1.0000	1.0000	1.0000
12	**0.9859**	**0.9906**	**0.9945**	**0.9975**	**0.9995**	**0.9999**	1.0000	1.0000	1.0000
13	**0.9878**	**0.9928**	**0.9961**	**0.9985**	**0.9997**	1.0000	1.0000	1.0000	1.0000
14	**0.9913**	**0.9946**	**0.9974**	**0.9991**	**0.9999**	**0.9999**	1.0000	1.0000	1.0000
15	**0.9939**	**0.9962**	**0.9983**	**0.9996**	**0.9999**	1.0001	1.0000	1.0000	1.0000
16	**0.9950**	**0.9972**	**0.9991**	**0.9999**	1.0000	1.0001	1.0000	1.0000	1.0000
17	**0.9976**	**0.9995**	1.0003	1.0002	1.0000	1.0000	1.0000	1.0000	1.0000
18	**0.9979**	**0.9991**	**0.9997**	1.0001	1.0000	1.0000	1.0000	1.0000	1.0000
19	1.0001	1.0006	1.0002	1.0000	1.0001	1.0000	1.0000	1.0000	1.0000
20	**0.9996**	1.0002	1.0002	1.0001	1.0000	1.0000	1.0000	1.0000	1.0000

Table D.1: rMSE ratios for 342 combinations: EB outperforming standard highlighted.