Ecotoxicological Risk Assessment: Developments in PNEC Estimation

HICKEY, GRAEME,LEE

Ecotoxicological Risk Assessment: Developments in PNEC Estimation

Graeme Lee Hickey

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

February 2010

Dedication

To my wife, Angelika, for all the love and faith you have given to me.

Ecotoxicological Risk Assessment: Developments in PNEC Estimation

Graeme Lee Hickey

Submitted for the degree of Doctor of Philosophy
February 2010

Abstract

Ecotoxicological risk assessment must be undertaken before a chemical can be deemed safe for application. The assessment is based on three components: hazard assessment, exposure assessment and risk characterisation. The latter is a combination of the former two. One standard approach is based on the deterministic comparison of exposure concentration estimates to the concentration of the toxicant below which adverse effects are unlikely to occur to the potentially exposed ecological assemblage. This concentration is known as the 'predicted no effect concentration' (PNEC).

At the level of hazard assessment we are concerned with, there is a requirement that procedures be straightforward and efficient, as well as being transparent. The PNEC is in general currently determined using either a fixed assessment factor applied to a summary statistic of observed laboratory derived toxicity data, or as a percentile of a distribution over the ecological community sensitivity. Often it is the situation that a hazard assessment will be based on substantially small samples of data.

In this thesis we evaluate proposals for determining a PNEC according to regulatory guidance and scientific literature. In particular, we explore these methods under the context of alternative probabilistic models. We also focus on the determination of conservative probabilistic estimators, which may be appropriate for this level of risk assessment. Additionally, we also discuss the detection of species non-exchangeability, a concept which is recognised by scientists and risk assessors, yet typically discounted in practice. A proposal on incorporating knowledge of a non-

exchangeable species for probabilistic estimators is discussed and evaluated. The final topic of research examines a generalised deterministic estimator proposed in a recent European Food Safety Agency report. In particular, we analyse the robustness and analytical properties of some cases of this estimator which (at least) maintains the expected level of protection currently attributed.

Proposals made within this thesis, many of which extend upon what is currently scientifically accepted, satisfy the requirements of being tractably straightforward to apply and are scientifically defensible. This will appeal to end users and increase the chances of gaining regulatory acceptance. All developments are fully illustrated with real-life examples.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematical Sciences, Durham University, UK. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all the author's original work unless referenced to the contrary in the text. Parts of Chapter 3 were collaborated on with Dr. Peter Craig (Department of Mathematical Sciences, Durham University, UK) and Dr. Andy Hart (The Food and Environment Research Agency, York, UK), leading to the publication Hickey et al. (2009). In addition, parts of Chapters 4 and 5 were collaborated on by Dr. Peter Craig, Dr. Andy Hart and Dr. Robert Luttik (The National Institute for Public Health and the Environment, Bilthoven, The Netherlands), leading to an article in revision to the Journal of the Royal Statistical Society: Series A.

Acknowledgements

First I would like to thank my supervisor, Dr. Peter Craig, for all the guidance he has offered me. In the many times I frequented his office, he would always make time for me and I would always leave more knowledgeable (and with more questions!).

I wish to also acknowledge the help of everyone in the statistics group within The Department of Mathematical Sciences at Durham University. In so little time, I have learnt so much. I especially thank Prof. Frank Coolen who made it possible for me to be back at Durham.

I am indebted to Dr. Andy Hart for his time and especially for the valuable trips he took me along on so that I may gain an appreciation of risk assessment in practice. These experiences and the discussions I had with him undoubtedly influenced the direction I went in. In particular I thank ECETOC for funding my trip to Oslo, Norway and CSL (now FERA) for funding my trip to Oxford.

I am grateful to The Dutch National Institute for Public Health and the Environment (RIVM) for granting me access to their data. I also extend my thanks to researchers at RIVM, especially Dr. Tom Aldenberg and Dr. Robert Luttik, for many stimulating discussions.

My thanks go to my mum and Andy for their belief in me. In particular my mum, who from an early age instilled in me a desire to learn and worked tirelessly to help make my ambitions a reality.

Finally, my greatest thanks is reserved for my wife, Angelika, for her love and support. Her smile, fantastic cooking and love of baby elephants has kept me going through even the toughest days of research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the motivation behind the need for greater research into probabilistic ecotoxicological risk assessment for the purposes of chemical safety assessment is discussed. In addition, an outline of the focus of subsequent chapters is provided, including a description of ancillary chapters.

1.1 Motivation

Chemical substances are widely embedded in modern society. They are used as: pesticides on our crops; hydraulic fluids in our cars; and detergents in our cleaners — these are just a few examples. It is the role of the governments to ensure that the chemicals presented for application by manufacturers are safe for use, whilst maintaining their use for advantage. Therefore, legislation exists across the broad chemical market to enforce that adequate risk assessments are performed. As an example, the new REACH regulation which was introduced on July 1st 2007 deals with the "Registration, Evaluation, Authorisation and Restriction of CHemical substances" within the European Union (EU) (EC, 2006). Different legislation governs other major categories of chemicals, for example pesticides are regulated within the EU under Council Directive 91/414/EEC (EC, 1991).

Under the REACH guidance, for chemicals manufactured or imported in quantities of greater than 10 metric tonnes per year it is required that a chemical safety report is produced. The report will contain a large volume of information, includ-

ing a chemical safety assessment which is comprised of many different aspects such as: human risk; exposure assessments; environmental impact assessment; etc. The latter component of the chemical safety assessment is referred to as *ecotoxicological* risk assessment from here on. The compiled report is then reviewed by a risk manager who will evaluate it, as well as other aspects such as socio-economic evidence; similar safety report requirements are required for pesticides also.

The risk assessment itself is a tiered approach, with the lowest tier being inexpensive, simple to implement and highly conservative. The level of conservatism is a reflection of the degree to which uncertainty is quantitatively described. As one climbs the tiers, expense and complexity increases while conservatism is lowered in exchange for more refined assessment. It is the failure to pass a preceding tier risk assessment, or the identification of unacceptable uncertainty levels, that triggers a higher one. At the very lowest tier, it may be possible to base the entire risk assessment on qualitative reasoning where evidence is sufficient. At the highest tier, risk assessment will be substantially detailed and focused on specific aspects of concern; subsequently being reviewed by scientific experts.

Our research focuses on a single safety issue of the ecotoxicological risk assessment, what the relevant REACH guidance document (ECHA, 2008a) entitles: 'characterisation of dose [concentration]-response for environment'. Essentially, this problem entails attempting to predict the concentration of a toxicant below which adverse effects are unlikely to happen to ecological communities — collections of different interacting biological species [populations]. In the literature, the sought-after value is called the 'predicted no effect concentration' (PNEC). In fact, EC (2003, p. 99) state, in the context of industrial substance risk assessment (of which REACH replaces):

> 'It is not intended to be a level below which the substance is considered safe. However, again, it is likely that an unacceptable effect will not occur.'

Therefore, we refrain from using the term 'safe concentration', as used by Emans et al. (1993) and others. Additionally, we acknowledge the existence of a 'true' ecological community level fixed point PNEC is sketchy. Although the aim is to

protect the structure of ecological communities, PNECs are derived based on protecting individuals; this is subsequently used as a proxy for protecting *all* ecological communities in *all* ecosystems — the wider interaction of biological and physical components of an environment.

The PNEC, in addition to the corresponding degree of uncertainty, falls under the umbrella of *hazard assessment* (ECHA, 2008b). The hazard assessment in some cases is sufficient for the purposes of risk assessment, e.g. when the substance is classed as non-dangerous. In other cases it is evaluated in conjunction with corresponding exposure assessments to evaluate safety; this process is known as *risk characterisation*. For all intents and purposes, the associated technical guidance documentation — which stipulates the accepted scientific methodology for application — for hazard assessment of general chemicals and pesticides is very similar for the scope of this research.

At the lowest tier of quantitative risk assessment, the PNEC is deterministically arrived at by little more than dividing the lowest measured tolerance of a sample of species from the potentially exposed assemblage by a large number with limited attributable meaning. This is based primarily on the *precautionary principle*, which Forbes and Calow (2002a) describe as:

> 'applying controls to chemicals in advance of scientific understanding if there is a presumption that harm will be caused.'

However, the premise of such extrapolation methodology is highly criticised as it introduces unquantifiable levels of conservatism. The inverse of the PNEC determination problem is the estimation of impact (sometimes erroneously referred to as 'risk') to an ecological community conditional upon an environmental exposure concentration. This is highly motivated for prioritisation of environmental clean-up operations with limited resources, and is important for higher tier risk assessments where one needs to characterise the overall risk by jointly treating all components as stochastic (ECHA, 2008b).

It is important to recognise that ecological communities which are intended to be protected are typically comprised of many species belonging to different taxonomic groups, i.e. the classification of biological species. There are multiple sources of

uncertainty in the hazard assessment component of a risk assessment; the variability across species tolerance to the toxicant is considered to be one of the most influential. Refinement of this aspect of uncertainty leads to an intermediate tier of quantitative risk assessment for determining the PNEC.

Underpinning the aforementioned tier of ecotoxicological risk assessment, as well as our research, is the fundamental assumption that there is a statistical model that adequately represents the interspecies variability in tolerance for a given toxicant. These models are known as *species sensitivity distributions* (SSDs), on which a great volume of research exists, as for example Posthuma et al. (2002a) and references therein. Application of SSDs at such tiers of risk assessments is restricted by the resource costs needed to satisfy current regulatory requirements; one such requirement is an adequate sample size n of data pertaining to the tolerance of non-target species, yet currently there is little gain for chemical manufacturers to implement this. In addition, there is still debate about many of the underlying assumptions that proponents of SSDs must subscribe to; the most prominent are described in Forbes and Calow (2002b). This has led to a degree of hesitation in their application by stakeholders and regulators.

We explore current methodology, both deterministic and probabilistic, for calculation of PNECs which account for this aspect of uncertainty in accordance with the lower and intermediate quantitative tiers of risk assessment. In particular, we substantially build on the developments of a recent European Food Safety Authority (EFSA) report (EFSA, 2005); however the remit of the former report differs from the scope of this research. Currently the handling and quantification of uncertainty within the regulatory setting of ecotoxicological risk assessment is lacking defensibility, thus underpinning the research presented here. In addition, a noticeable element of this research is the mathematical tractability of PNEC estimators. There is a definite requirement that risk assessments at the tiers we focus upon both pragmatically and transparently balance robustness, protection and tractability if they are to be adopted by current risk managers. This is contrary to the growing increase in complexity of statistical modelling found in other areas of risk assessment.

1.2 Outline of Thesis

We begin by exploring the background of key concepts in ecotoxicological risk assessment, focusing on the PNEC estimation problem in greater depth, during Chapter 2. In particular, we give emphasis to procedures which are currently recommended, or perceivably valid, in regulatory guidance documents subject to the lower and intermediate quantitative tiers of hazard assessment which we are concerned with. Additionally, we describe and critique recent developments within associated scientific literature which might be introduced to strengthen or replace the *status quo*. During this chapter, we construct necessary definitions, notation and model descriptions which are paramount throughout the entire research report.

The remaining four chapters are ordered so that they cover three distinct strands of improvement to the way current ecotoxicological risk assessment is understood and performed. Although these topics of research are considered independently in order to aide transparency in understanding to regulators, there is potential for overlap which may facilitate more refined hazard assessment. Here we briefly outline the key research topics and chapters.

In Chapter 3 we examine in detail the probabilistically determined PNEC estimators. Particular focus is given to the concept of conservative estimation by reconsidering the estimator derivation from a decision theoretic loss function perspective. Suggestions of possible improvements across the spectrum of quantitative tier hazard assessment are provided. A key conclusion is that there is a need for risk managers to articulate what conservatism is required for PNEC estimators beyond what is currently provided. A number of coincidences between estimators based on statistical inference yielded by different behavioural models and decision theory based estimators are found and discussed; including a relation between a versatile class of estimators which have become a well established practice and gained significant scientific acceptance (the Aldenberg and Jaworska 2000 estimators) and the utility of asymmetric log-linear estimation error.

One of the main contributions of this thesis is the concept of *species non-exchangeability*, as described in Chapter 4. The concept has been discussed by a number of authoritative authors (Forbes and Calow, 2002b; Dwyer et al., 2005), yet

the statistical implications — required in order to defensibly refine the uncertainty of the hazard assessment — remained unaddressed until the EFSA instigated research into the problem (EFSA, 2005). The idea is that if one identifies the species whose tolerance values have been measured, then this provides the assessor with additional information. Relevance of species non-exchangeability is enhanced by the regulatory use of standard dossier species; for example, in the context of plant protection products under Directive 91/414/EEC (EC, 1991) it is required that the rainbow trout (*Oncorhynchus mykiss*) — a species belonging to the salmonid family — is assessed with the toxicant. In addition, the sample size of additional fish species tested is typically very small, as low as one in some cases.

A number of different methods of detecting this property are discussed in this chapter for this standard test species. Whilst it is plausible that a statistical modeller would want to fit some sort of model where each species and chemical has an effect, thus to a degree circumventing the issue of non-exchangeability, such methods are unlikely to have sufficient potential for adoption within the current regulatory arena to be of practical use. In Chapter 5 we explore a probabilistic model proposal, which is effectively an adaptation of current scientifically accepted probabilistic ecotoxicological methodology, in order to take account of species non-exchangeability. Issues related to performance and uncertainty of the derived decision rules are subjected to extensive scrutiny and the results discussed in order to defend any contentious assumptions made.

The final topic, which is presented in Chapter 6, explores the strictly deterministic decision rules from a probabilistic perspective. The purposes of our findings are not to influence the precautionary factors to which a risk assessor appeals in order to construct an initial PNEC estimate, but rather the way one utilises the limited laboratory data to which the assessor applies the factors. The former requires a prerequisite expertise tailored to the assessment, whereas the latter can be fully reasoned for by means of probabilistic consideration. This naturally leads to a generalised class of 'decision rules', briefly introduced in EFSA (2005) but limited to the Gaussian context, that maintain straightforward application and level of protection. The latter is a difficult concept to define as it remains unspecified in associated tech-

nical guidance documentation. We extend the viewpoint, and show a high degree of robustness to alternative distributional models which we use to support the stance that going between the strictly deterministic risk assessments (usually with $n \leq 3$) and the probabilistic SSD derived PNECs (usually with $n \geq 10 - 15$), i.e. where sample sizes are deemed intermediate ($4 \leq n \leq 11$), can be handled with mutual benefits to the chemical manufacturer and risk manager.

We evaluate and summarise the research discussed above in Chapter 7. The conclusions drawn indicate that the methods proposed in this thesis have, in our opinion, potential for adoption within the regulatory arena. Notwithstanding this point, there are branches of this risk assessment field that require further research in order to validate aspects of our proposals and lead to more refined intermediate tiers of hazard assessment.

Six appendices are provided at the end of the thesis. Appendix A provides instruction on the Bayesian calculations used for Chapters 2 and 3, as well as an alternative derivation of a class of indirectly proposed PNEC estimators described in EFSA (2005). Appendix B is central to Chapters 2–5 by providing details regarding the hyper-parameter estimation required for probabilistically derived estimators derived under different behavioural and data models proposed here. In addition, we also give details of two hyper-parameter estimation procedures pertinent to models proposed in EFSA (2005) to which we compare ours. Appendix C sketches the derivation of decision rules we derive and discuss in Chapter 3 from the loss function perspective. In exploring the consequences of restricting decision rules which account for species non-exchangeability to be tractable, Appendix D supports the arguments made in Chapter 5. Appendix E displays equations relevant to the analytical analysis of the robustness of a generalised deterministic PNEC estimator to various distributional assumptions. Appendix F gives details on how to increase the precision of numerical quadrature for calculating a mean level of protection evaluated at a generalised decision rule which accounts for species non-exchangeability. Finally, a glossary of acronyms is provided at the end of the thesis.

Chapter 2

Background

When assessing the safety of a new or existing chemical substance to the environment, it is impractical and unethical to determine its impact on every species present in the potentially exposed ecological community. Moreover, it is legally implausible due to endangered species protection legislation existing in many industrialised nations. Therefore, one must appeal to risk assessment, i.e. assessing the potential — under uncertainty — for harm to be caused to the environment. Risk assessors should consider the aforementioned uncertainty, but the degree to which this is done varies according to the tier of assessment undertaken. The following sections discuss the relevant background of ecotoxicological risk assessment, with particular focus on the tier of assessment which uses PNECs as a decision making tool. We will also introduce some notation and definitions that will be used throughout this thesis.

2.1 The Risk Assessment Procedure

An ecotoxicological risk assessment (ERA) for a chemical substance which is to be used in considerable quantity or classed as dangerous can be reduced to consideration of, or rather characterised by, the risk characterisation ratio [1] (RCR), defined as:

$$\mathrm{RCR} = \frac{\mathrm{PEC}}{\mathrm{PNEC}},$$

[1] Alternatively, the toxicity exposure ratio (TER) is calculated (EFSA, 2005). This is equivalent to RCR = AF / TER where AF is a constant defined in Section 2.3.

whereby PNEC is the predicted no effect concentration — the concentration of the toxicant below which adverse effects are unlikely to occur; and the PEC is the predicted environmental concentration — the concentration of the toxicant that might be expected to be found in an ecosystem. The former falls under the umbrella of *hazard assessment*, whilst the latter falls under that of *exposure assessment*.

There are basically three ways of considering the RCR: (1) qualitatively; (2) deterministically; (3) probabilistically. These three perspectives effectively define the three tiers [levels] of uncertainty assessment within the current REACH guidelines (ECHA, 2008b), although our discussion is with respect to a wider context than just REACH. However, the tiers are in no way exclusive; ERA is often based on overlapping perspectives. Typically, a risk assessor would start at (1), and progressively utilise (2) and (3) where evidence of refinement is indicated. Inherent in this viewpoint is a continuum of conservatism; one would expect (1) to be conservative in the sense that assumptions made are protective, and this would be gradually relaxed as the assessment is refined. In this thesis we are particularly interested in (2) and (3); (1) will be pertinent to situations where scientific experts have relevant knowledge to indicate whether safety is very likely, or where one might need to refine a specific aspect of an ERA.

The deterministic procedure (listed (2) above) is to calculate a pointwise value of the RCR. If the RCR is less than unity, then this indicates that the chemical is likely to be safe. On the other hand, if the RCR is greater than unity, then the assessed substance may not be authorised for use unless a more appropriate higher tier risk assessment shows that it will cause no unacceptable impact (EC, 2002). The RCR exceeding unity does not automatically suggest the chemical is unsafe, but rather that a more refined assessment is required in order to better establish the risk. ECHA (2008b) also remarks that if the RCR is less than but still close to unity, then a more refined assessment should also be considered. There are different ways in which the PNEC and PEC might individually be calculated in order to yield a RCR: again, deterministically or probabilistically; which highlights why perspectives (1)-(3) are overlapping, since for example, the RCR might be deterministically calculated based on a probabilistically estimated and evaluated PNEC and PEC. In the context of

PNEC estimation, we refer to the respective methods as the lower and intermediate quantitative tiers; and it is together this general tier of assessment we focus upon in this thesis.

In any given ERA there may be multiple exposure paths that need to be considered. In estimating PECs for these exposure scenarios, a conservative estimate, or multiple estimates each with differing degrees of conservatism might be determined (recalling that the risk assessment is a proxy for many ecosystems). For example, assuming exposure has been probabilistically modelled for a relevant scenario (e.g. the rate of spray drift), then EUFRAM (2006) reports that a high (typically 90-th) percentile of this distribution is advocated. Consideration of exposure distributions and PEC estimation are beyond the remit of this thesis; for further information, consult Aldenberg et al. (2002), EUFRAM (2006) and references therein.

Evaluation of the RCR from a completely probabilistic viewpoint is a relatively recent tool; for a review consult EUFRAM (2006). Strictly speaking, this is the definition of Level 3 in ECHA (2008b), although as indicated above, probabilistic risk assessment is not limited to this viewpoint. By treating the hazard and exposure components of the assessment as probabilistic, one can subsequently characterise the risk. One such field of research (Aldenberg et al., 2002; Warren-Hicks et al., 2002; Verdonck et al., 2003) has been the use of joint probability curves: parametric plots of the exposure exceedance distribution and a distribution for the PNEC. The area under this curve is often used as a summary statistic, called the expected ecological risk. More recently, Aldenberg (2007, A.2) and Aldenberg et al. (2009) has researched the relation between joint probability curves and the RCR (distribution), highlighting the insufficiency of the former as a tool to make adequate risk assessment decisions and suggesting appropriate improvements.

2.1.1 A General Perspective on ERA

The definitions of 'risk' and 'uncertainty' are ambiguous, and the former is highly contentious in different fields of research. For example, the International Programme on Chemical Safety (IPCS, 2004) define uncertainty to be:

> 'imperfect knowledge concerning the present or future state of an organism, system, or (sub)population under consideration.'

Additionally, IPCS (2004) defines risk as:

> 'The probability of an adverse effect in an organism, system, or (sub)population caused under specified circumstances by exposure to an agent.'

In the context of ERA, the latter is equivalent to the probability that the RCR breaches unity. We will hesitantly use the definitions in IPCS (2004) for generic risk assessment terminology unless redefined elsewhere in this thesis.

A misleading heuristic, which often features in chemical risk assessment is: 'Risk = Toxicity × Exposure'. This principle should not be directly interpreted, but the definition does implicitly suggest the need to incorporate the degree of effect along with the corresponding probability in order to adequately define risk. By considering the fully probabilistic tier of ERA, the degree to which the RCR is greater than (or less than) unity can be interpreted as a measure of the aforementioned degree of impact. However, the probability of adverse affects is complicated by the fundamental uncertainty in both the exposure distribution and PNEC distribution. Within the deterministic procedure, the impact might be interpreted as simply 'all or nothing'.

2.2 Toxicity Data

In order to quantitatively estimate the PNEC, whether deterministically or probabilistically, for a chemical to non-target species — those which we wish to protect — in potentially exposed ecosystems, a risk assessor uses an experimentally determined sample of species tolerance values. The set of distinct species tolerance values determined for an ERA are collectively referred to as *toxicity data*. A point of contention among scientists is that the species assessed are non-randomly selected, and are often non-representative of ecological communities they are supposed to represent; we discuss this point further in the context of probabilistic PNEC estimation within Section 2.4.

A tolerance value, or more precisely, laboratory effect value, for a species is the concentration of a substance yielding an observable and measurable effect to a population (i.e. of a single type) of species. This is often estimated by means of laboratory experimentation to individual species, i.e., without interaction of other species which may be present in the ecosystem. Tolerance values are also referred to as a 'sensitivity values' — although this is a misnomer. For aquatic compartments, concentration is usually defined in terms of milligrams (mg/L) or micrograms (μg/L) per litre of water; in terrestrial ('land') compartments the corresponding measure is a dosage in milligrams per kilogram of body weight (mg/kg). One can also determine data from scientific studies known as meso- or micro-cosm experiments which aim to recreate, or at least partially mimic the ecosystem; such experiments are rarely performed due to intensive resource costs.

2.2.1 Classification

Single species tolerance values can be loosely categorised into two groups: acute [short-term] and chronic [long-term]. In particular, we discuss two standard reported tolerance values: the EC_x and NOEC.

An $\mathbf{EC}_x$ is defined as the concentration that affects x% of the tested species population over a stated period of time. It is a statistically estimated summary value, in this case the x-th percentile of a modelled concentration-response curve; experimental data is generally not incorporated into the risk assessment. It is important to note that species can be affected in different ways to toxicant exposure, for example common measured effects include growth and mortality. The latter is often of particular concern and is specially denoted as the lethal concentration to x% of the tested species (LC_x). The standard choice of x for short-term studies is $x = 50$, i.e. defining the median effect concentration. Common time periods for such studies are: 48–96 hours for fish (in denominations of 12 hours) and 24 hours for algae and plant species. It is generally required for all toxicity data collectively used in the risk assessment to be of the same endpoint, or some collection of endpoints with a firm biological justification made by scientific experts and agreed upon by regulators; it is not the role of the statistician to select endpoints. Long-term studies

(e.g. for reproduction) frequently report the EC_{10} value.

A **NOEC** (no observed effect concentration) is defined by ECHA (2008a, Table R.10-1) to be the maximal test concentration at which the substance is observed to have no statistically significant effect ($P < 0.05$) when compared with the control [group], within a stated exposure period. NOECs are generally derived for long-term studies and are subsequently preferred, as well as EC_{10} values by ERA technical guidelines. This is because they can be used to estimate PNECs which are protective of the ecological community over entire life cycles (ECHA, 2008a). NOECs are highly sensitive to experimental design (Chapman et al., 1996); Emans et al. (1993) briefly discusses some common NOEC estimation procedures. Consequently, the endpoint has been subject to much criticism, and a proposed 'no effect concentration' advocated by some instead. To elucidate, a statistically estimated threshold parameter could be obtained from a suitably fitted model; see Pires et al. (2002) and references therein. Similarly to the case of EC_xs, NOECs are in reference to a collection of specific effects (e.g. reproduction, growth). It is usually the most sensitive NOEC for each species that is used in an ERA unless there is scientific reasoning not to include certain endpoints, for example, if deemed not to be ecologically relevant.

Long-term studies are generally much more expensive to conduct relative to short-term studies, consequently there is much less available historical chronic toxicity data available. Where only short-term data is available, it will be required that any extrapolations made in determining a PNEC accounts for this. There has been research which attempts to determine the required extrapolation — separate from all others that need be considered — which estimates chronic tolerance values from acute tolerance values. The simplest of these proposals is the acute-to-chronic ratio (ACR) factor which is a fixed multiplicative factor (cf. assessment factors in Section 2.3) applied to acute tolerance values yielding chronic counterparts; see for example Duboudin et al. (2004a). EFSA (2005) reports that current EU guidance (EC, 2002) has implicitly set this at 10; something which Roex et al. (2000) suggests is in good agreement, on average, with empirical data. However, Roex et al. (2000); Forbes and Calow (2002a) and Forbes et al. (2008) report that the range

of empirically derived ACRs is very large, weakening the defensibility of the current used value of 10. Alternative statistical extrapolation methods are proposed by Duboudin et al. (2004b); Craig (2006) and Raimondo et al. (2007). It is not always the case that chronic toxicity data is relevant for chemical risk assessment; some pesticides might only have short term impacts; e.g. if they biodegrade rapidly, or disperse in a fast flowing stream — acute data would probably serve as appropriate in these situations.

The current technical guidance documents for ERA within the EU instruct risk assessors on how to deterministically calculate PNECs based on very small sample sizes of toxicity data; in some cases using only a single long-term tolerance value. A result of this has been that very low amounts of toxicity data are publicly available for existing substances.

2.2.2 Harmonisation

Individual tolerances values reported in risk assessment dossiers and elsewhere are not precise; they are recorded with measurement error. The reasons for this can be to do with the different laboratories used, different sources of test species, natural intra-species variation and experimental error. Standard so-called 'harmonisation' techniques are routinely applied in ERA to simplify assessment, however this has been criticised by Duboudin et al. (2004a) because information regarding intra-species variation and measurement error is discarded. Nonetheless, prepared databases used in scientific research have usually already been harmonised in this way.

Where toxicity data originates from multiple (historical) sources, for existing substances (i.e. retrospective ERA) there may be species which have multiple recorded tolerance values. As stated earlier, preference is often given to long-term studies. Beyond this, qualitative evidence is used to assess which out of the toxicological endpoints has the most weight regarding reliability and relevance; this often involves reviewing the experiment reports (where possible). In certain cases the historical data may be decades old where technology was less advanced for conducting scientific experiments. If multiple differing endpoints still remain after this data quality

review, then the industry and regulatory standard rule-of-thumb is to take the geometric mean; see De Zwart (2002) and ECHA (2008a) for further guidance. Where a tolerance value is reported as a censored value, it is generally not acceptable to be included in the currently accepted PNEC extrapolation methods; although ECHA (2008a) list certain cases where harmonisation is allowable. Notwithstanding this restriction in general, techniques for inclusion of censored values have been discussed by Kefford et al. (2005) using non-parametric methods, O'Hagan et al. (2005) and Hickey et al. (2008) using Bayesian methods.

Other factors which can affect whether data is permitted include: whether the species experimentally assessed represent different habitats (e.g. for aquatic compartments species habitats are freshwater or marine [saltwater]); and geographic locations (e.g. temperate or tropical). Maltby et al. (2005) found that these two factors did not have a statistically significant effect on the HC_5 estimates (which was defined as the measure for assessment of hazard) in the case of insecticides, however differences between the SSDs as a whole were not compared. These issues and others are discussed further in De Zwart (2002) and Solomon and Takacs (2002).

2.3 Assessment Factors

Deterministic calculation of PNEC values is based on the application of *assessment factors* (AF). For a standard class of ecosystem, e.g. a freshwater compartment, which is to act as a proxy for all ecosystems belonging to this class. The procedure is to divide the lowest observed tolerance value by a fixed positive scalar — called an assessment factor — considered large enough that when applied it will extrapolate to the PNEC. Assessment factors are typically 10- or 5-fold ranging from 10,000 to 5 for strictly deterministic procedures. The application of assessment factors is justifiable based on the 'precautionary principle' (Forbes and Calow, 2002a) which is expected to extrapolate to a conservative estimate of the PNEC.

The prescription of an assessment factor depends on a number considerations, for example, a non-exhaustive list includes: the compartment for which the ERA is representative (e.g. marine, freshwater, soil, birds, etc.); the quantity of toxicity

data; and whether the toxicity data is acute or chronic. Assessment factors are listed in a number of official documents for use in pesticide and chemical safety assessment. For examples, see the EU technical guidance documentation EC (2002, 2003); ECHA (2008a); and the United States Environment Protection Agency Office of Pollution Prevention and Toxics (Zeeman, 1995). Note that assessment factors are also referred to by a number of other titles in the ERA literature and official documentation, for example: safety factor, extrapolation factor, application factor, uncertainty factor.

EFSA (2005) notes that assessment factors are intended to account for:

- intra- and inter-laboratory variation of toxicity data;
- intra- and inter-species variation (biological variance);
- laboratory data to field impact extrapolation;
- short-term to long-term toxicity extrapolation.

It is envisaged, although without a firm scientific basis, that each assessment factor is a multiplicative product of smaller assessment factors each representing these components of uncertainty (EFSA, 2005, p. 10). However, there is no firm scientific understanding to their individual or overall magnitudes. Thus assessment factors are more-or-less arbitrary, with Forbes et al. (2008) describing them as 'rough, order-of-magnitude guesstimates'. If it can be assumed that the overall assessment factor is the multiplicative product of individual smaller assessment factors representing the aforementioned uncertainties, then one can immediately deduce that the final article (above) corresponds to the ACR described in Section 2.2.1. In conclusion, the degree of uncertainty and conservatism in the PNEC estimate is unknown, meaning that they are not scientifically defensible.

The summary statistic which one divides by an assessment factor is currently defined as the lowest tolerance value. It has recently been suggested (EFSA, 2005, 2008) that the geometric mean of the available toxicity data may be an adequate summary statistic with application of the current prescribed assessment factors. In fact, where regulations only require a single species to be tested (e.g. for risk assessment to birds), the geometric mean provides *at least* the same level of protection —

although officially undefined — for all sample sizes, as offered by application of the same assessment factor to the minimum tolerance value (EFSA, 2005, Appendix A).

ECHA (2008a) states that assessment factors are in fact only general guidance values, and may be lowered if sufficient justifications are presented; e.g. increased sample sizes, evidence from 'similar' substances, etc. One approach is to apply variable assessment factors which are a function of the geometric standard deviation of the toxicity data. For example: $\hat{\mu}(\mathbf{Y})/\hat{\sigma}(\mathbf{Y})^{\kappa}$, where $\hat{\mu}(\mathbf{Y})$ and $\hat{\sigma}(\mathbf{Y})$ are the geometric sample mean and standard deviation of the toxicity data $\mathbf{Y}$; and κ is some constant value independent of the data which can be tabulated for risk managers. The parameter κ later transpires to be a key component of tractable probabilistic PNEC estimators. Confusingly, within the probabilistic literature, κ is occasionally referred to as an assessment factor (e.g. Aldenberg and Jaworska 2000), whereas technically the assessment factor in this case is $\tilde{\sigma}(\mathbf{Y})^{\kappa}$ — a function of the dispersion of the toxicity data. We later redefine this parameter from within the probabilistic framework.

2.4 Species Sensitivity Distributions

Recently, considerable attention has been given to probabilistic techniques in order to derive PNECs (EUFRAM, 2006), which falls within the realm of intermediate tier quantitative ERA due to their more stringent requirements. The fundamental concept underlying this viewpoint is the 'species sensitivity distribution' (SSD; Posthuma et al. 2002a), which for a specific chemical, is a distribution modelling the interspecies variability of tolerance [sensitivity] in an assemblage of different biological species with respect to certain observable toxicological endpoints. SSDs thus provide a way, separate from any use of assessment factors for other purposes, to formally relate the tolerances of tested species to those of other untested species.

The SSD concept is now scientifically accepted within the regulatory arena for intermediate and higher tier probabilistic ERA. A seminal paper by Kooijman (1987) led to adoption of SSDs by Dutch regulators; now they are in use for regulatory ERA of: new and existing chemicals (US EPA, 1998; ECHA, 2008a); setting water quality

guidelines (Stephan et al., 1985; ANZECC and ARMCANZ, 2000); and pesticide assessment (US EPA, 2004) — this list is by no way exhaustive. Consult Suter (2002) and Van Straalen and Van Leeuwen (2002) for a history of SSDs.

The data requirements of utilising the SSD laid out in modern ERA technical guidance documents are much stricter than those for strictly deterministic PNEC derivation. For example, the current REACH guidance (ECHA, 2008a) stipulates that long-term study NOECs must be determined for a minimum of 10 species (preferably 15) spanning 8 taxonomic groups. A guidance document by the Society of Environmental Toxicology and Chemistry (Campbell et al., 1999) suggested in the context of acute pesticide exposure to aquatic systems that this minimum sample size may be 8 species; 5 species for fish-only assessment. Such sample sizes might still be deemed insufficient by mainstream statisticians; nonetheless, such samples are likely to be considered impractical for many substances by scientists in the ecotoxicology arena. In fact, Aldenberg et al. (2002) report that sample sizes much lower than 10 are not exceptional. A further criticism of the SSD concept in ERA is the sample of species used to fit the distributions being selected non-randomly due to financial, practical and socio-ethical restraints. This is potentially introducing bias into the model fitting of SSDs (Forbes and Calow, 2002c). In fact, most SSDs are populated with some standard test species — typically those easily cultivated and manageable laboratory species with well understood life cycles.

The application of SSDs in the research literature is usually subjected to less stringent conditions; see for example Posthuma et al. (2002a). Additional uncertainties, some of which were described in Section 2.3, would need to be accounted for using either alternative statistical constructs or deterministic assessment factors before one can appropriately determine a PNEC; ECHA (2008a) require an assessment factor between 5 and 1 to be applied *post hoc*.

In the following sections we provide details on the SSD as a predictive tool and the physical relevance of SSDs. A fuller description of SSDs is presented in Posthuma et al. (2002a); and a detailed critique of SSDs is presented in Forbes and Calow (2002b) in which the most significant assumptions made in SSD-theory are reported and appraised.

2.4.1 SSDs for Prediction

An SSD, of which we denote the cumulative distribution function (CDF) by $F_{\text{SSD}} : x \in \mathbb{R}^+ \mapsto [0,1]$, may be *directly*[2] used as a predictive tool in one of two ways (see Figure 2.1): the *forward* and the *inverse* method (Van Straalen, 2002). It is the inverse method which is pertinent to estimation of the PNEC, however the forward method complements this. Direct application does not preclude fully probabilistic tiers of uncertainty analysis as foreseen under Tier [Level] 3 of ECHA (2008b).

Forward Method

In the forward method one specifies an environmental concentration $x \geq 0$ and uses the SSD to estimate the fraction of species $p_x \stackrel{\Delta}{=} F_{\text{SSD}}(x) \in [0,1]$ within the assemblage that will have their toxicological endpoints exceeded at or below this concentration; blue arrows in Figure 2.1. This may also be viewed from a statistical perspective as the probability that a species randomly selected from the assemblage has its unknown tolerance value (measured with respect to some toxicological endpoint) lie below this exposure concentration. We will describe this proportion of species in the assemblage as the *potentially affected fraction* (PAF) (Traas et al., 2002) since it best matches the correct intended physical interpretation; other terminology used includes: fraction affected (Aldenberg and Jaworska, 2000) and fraction exceeded (EFSA, 2005). It would be the case that one might use the forward method in situations such as an unexpected toxic discharge so that the immediate and/or long term impacts can be evaluated. This might then be used, for example, to decide on the prioritisation of clean-up operations between competing sites with limited available resources. Such a principle has been adopted by the United Nations Flash Environmental Assessment Tool (UN, 2009) which is applicable for safety assessment in hazardous substance facilities which are severely compromised by natural hazards, e.g. earthquakes.

[2]We take the term 'directly' to mean that the SSD is used alone, without joint probabilistic consideration of other risk assessment components such as exposure.

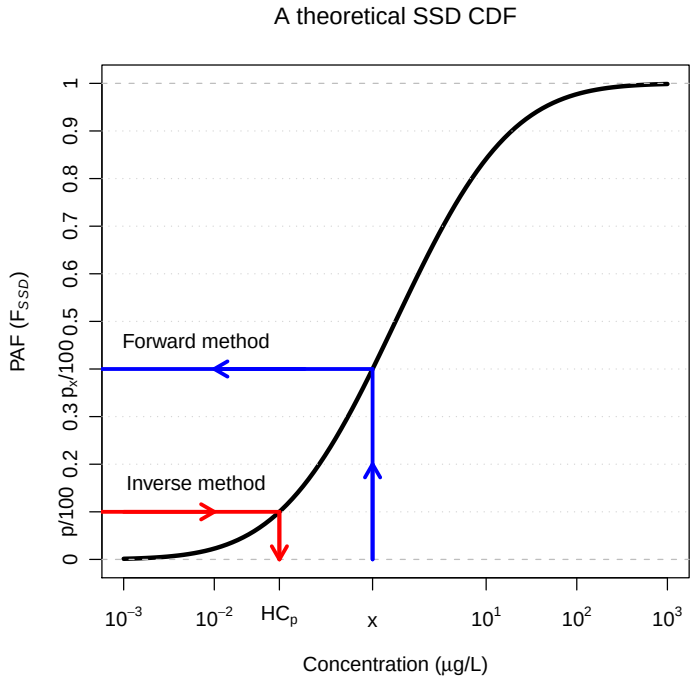


Figure 2.1: A hypothetical SSD with arrows indicating the use in the *forward* and *inverse* usage. The red arrow indicates the inverse method which yields the HC_p for given $p \in (0, 100)$; the blue arrow indicates the forward method which yields an estimate of the PAF (denoted p_x; a percentage) for environmental concentration x μg/L.

Inverse Method

In the inverse method one uses the SSD in order to estimate a concentration below which greater than $(100 - p)\%$ of the community are likely to be protected from adverse effects, whereby the risk manager selects the maximally permissible or some conservative substitute, PAF level in advance; red arrows in Figure 2.1. By definition, a PAF threshold of $p\%$ applied to a chronic effects SSD leads to a threshold concentration referred to as the *hazardous concentration* to $p\%$ of species (HC_p). The standard choice of p is $p = 5$ based on historical usage, but does not preclude the choice of other values, e.g. $p = 1, 10$, for examples see Alexander and Fairbridge (1999, p. 235), Van Straalen and Van Leeuwen (2002), and references therein. We will also use this definition and notation in the context of SSDs fitted with only acute data, something which has become standard practice due to the limited availability of reliable chronic toxicity data. Statistically, estimation of the HC_p is analogous

to estimating a percentile of a probability distribution, i.e. $F_{\text{SSD}}(\text{HC}_p) = p$. The inverse method is of use when one wants to set environmental safety limits (e.g. water quality guidelines); decide whether to register a new chemical substance before allowing it onto the open market; or perhaps as a trigger for a higher tier chemical safety risk assessment. Current guidance allows for the PNEC to be set as the HC_5 value estimated using long-term NOECs, subject to conditions on: sample size; taxonomic representativeness; and an additional deterministically applied assessment factor (as discussed above). We will survey current proposals of HC_p estimation based on the inverse method later on in this chapter.

Borrowing Strength

It has been suggested that SSDs might borrow strength from additional information to lend weight to prediction (Luttik and Aldenberg, 1997; Aldenberg and Luttik, 2002; EFSA, 2005; Grist et al., 2006; Hickey et al., 2008). The additional information is usually a collection of toxicity data from other substances considered either similar to the one being assessed, or having corresponding assessed species. Alternatively, it might refer to the inclusion of expert judgements. Defining the term *similar* is not easy nor has it ever been properly defined before. We might assume the definition to mean that substances within a group are known, or subjectively believed, to be of the same chemical class, i.e. acting on species in similar ways. However, we would remark that we are not qualified to make a firm statement about what it means for a substance to be similar. Furthermore, it is beyond the scope of this thesis. It is perhaps also useful to consider the context of the definition to be in reference to similar taxonomic groups. For example, if we have information on a range of similar substances which only relates to invertebrates, then an assessment which features fish may be distorted by incorporating the aforementioned additional information. There is currently no consensus in the ERA arena regarding the inclusion of additional information.

2.4.2 Distributional Assumptions & Interpretation

To maintain tractability which reflects the degree of parsimony in the risk assessment required, a simple continuously semi-infinite parametric distribution is often employed as the SSD; this is criticised by Forbes and Calow (2002b). Choice of such a distribution is a much researched area with many accepting that it is reasonable to fit a log-normal or log-logistic distribution when communities are suitably partitioned into taxonomic groups; e.g. see Aldenberg et al. (2002); Solomon and Takacs (2002) and Maltby et al. (2005). This is logically appropriate for many chemical risk assessments since receptors to the toxicant will likely be reasonably similar within taxonomic groups. In the context of ANZECC and ARMCANZ (2000), the Burr Type III distribution is regularly used to represent the SSD and although it offers flexibility, it relies on computationally intensive methods for estimation (Shao, 2000).

Fitting individual SSDs per taxonomic group is typically not practical due to the data constraints discussed. Therefore, as recommended in the REACH guidance (ECHA, 2008a), data from different taxonomic groups are amalgamated. While this might lead to a more 'realistic' community-representative SSD, Duboudin et al. (2004a) describes this situation as inappropriate and suggests that a more practical risk assessment strategy would be to estimate individual PNEC values separately for each taxonomic group; similar conclusions were drawn by Kefford et al. (2005). For example, in simplistic aquatic risk assessments one typically considers a vertebrate, an invertebrate, and algae species so that all trophic levels are assessed: a predator, a herbivore, and a primary producer, respectively. Taking the minimum of the three estimated PNECs would, subject to additional uncertainties being discounted, be the only way to ensure that *at least* the minimum level of protection (e.g. 95% of species) is maintained for the entire ecosystem. This is generally not done in ERA, and is a key criticism made by Forbes and Calow (2002b,c) who note that certain taxonomic groups are over- and under-represented in SSD model fitting, in some cases quite seriously; a point echoed also in Baird and Van den Brink (2007).

Duboudin et al. (2004a) reported that taxonomic grouping plays a substantial role in SSD based prediction and suggested that data should be weighted accord-

ing to 'true' taxonomic weights. O'Hagan et al. (2005); Grist et al. (2006) and Hickey et al. (2008) extend this by suggesting mixture distributions constructed using weighted per-taxon SSDs be applied. Notwithstanding the general widespread application of simple parametric SSDs, such as the log-normal distribution, objections have been raised regarding the usage by Newman et al. (2000, 2002); Grist et al. (2002) and Duboudin et al. (2004a), with the authors advocating non-parametric methods such as bootstrapping. Aldenberg and Luttik (2002) note that the log-normality assumption offers greatest mathematical tractability in risk calculations. Furthermore, they showed that the majority of substances in a toxicity database held by The Dutch National Institute for Public Health and the Environment (RIVM) with sufficiently large sample sizes did not reject the assumption of log-normality based on the Anderson-Darling test (Stephens, 1974; Aldenberg et al., 2002) at the 5% critical significance level.

There is currently debate about whether SSD based inferences have any direct ecological interpretation (e.g. see Forbes and Calow 2002b, Van den Brink et al. 2006 and references therein), since it fails to incorporate other factors, e.g. food-chains, species interactions and physiological resilience. In other words, SSDs commonly do not reflect the ecological community they are designed to represent. The definition of this *community* is described by Aldenberg et al. (2002) as being 'the Achilles heel of the *SSDeology*'. However, the estimated risks from certain methods have been shown to be well calibrated to reality in the form of field-studies and (semi-) mesocosm studies, by acting as good, yet typically protective, indicators of risk from chemical stressors; see for example Emans et al. (1993); Hose et al. (2004); Schroer et al. (2004); Maltby et al. (2005); and references therein. A more recent use of SSDs has been in the field of radiation risk assessment whereby one determines the biological endpoint as a fixed measure of irradiation and proceeds in a similar manner; see Garnier-Laplace et al. (2006) for a discussion.

2.5 Notation, Definitions & Assumptions

We assume that for a particular substance $\mathcal{S}$ under current risk assessment a limited set of toxicity data is available. The log-tolerance values (base 10) obtained for $\mathcal{S}$ are denoted y_j for $j \in J_{\mathcal{S}}$, where $J_{\mathcal{S}}$ is the collection of species experimentally tested to determine (distinct) tolerance values with $\mathcal{S}$. Logarithms are used for many reasons, including: (i) toxicity data is then close to normal; (ii) variation is stabilised; and (iii) it is established as conventional in ecotoxicology. For notational convenience, we define $\mathbf{Y} = (y_j;\, j \in J_{\mathcal{S}})$ and $|J_{\mathcal{S}}| = n$, i.e. so that the sample size of toxicity data is n. Also, we denote $\bar{y}$ and s^2 to be the unbiased sample mean and variation of the log-tolerance values for $\mathcal{S}$, given by

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{j\in J_{\mathcal{S}}} y_j;\ \text{and} \\ s^2 &= \frac{1}{n-1}\sum_{j\in J_{\mathcal{S}}} (y_j - \bar{y})^2. \end{aligned} \tag{2.1}$$

Unless stated otherwise, we will utilise the common assumption (see Section 2.4) that the toxicity data can be envisaged as being independent realisations from a log-normal distribution. Using a straightforward property of the log-normal distribution, the logarithmic transformed toxicity data is normally distributed:

$$y_j \,|\, \mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2) \ \text{ for } \ j \in J_{\mathcal{S}}, \tag{2.2}$$

where μ and σ^2 are the (unknown) location and scale parameters of the SSD over log-transformed concentration respectively. Note that if $\theta = (\mu, \sigma^2)$ were known precisely, then the log-hazardous concentration to $p\%$ of the non-target species would be $\psi_p(\theta) \overset{\Delta}{=} \mu - K_p\sigma$, where $\psi_p : \mathbb{R} \times \mathbb{R}^+ \mapsto \mathbb{R}$; and K_p is the $(100-p)$-th percentile of the standard normal distribution, e.g. $K_5 = 1.6449$. It is by convention in the relevant SSD literature that K_p is defined in this way.

In accordance with EFSA (2005), we assume assessment factors are applied divisibly to some summary statistic of the toxicity data; with the minimum order tolerance value being the normal choice for strictly deterministic calculations. Ten-

tatively assuming that assessment factors for separate sources of uncertainty combine multiplicatively, we define the adjusted toxicity statistic (ATS) as:

$$\text{ATS} = \frac{\text{Toxicity Statistic}}{\text{AF}_{\text{spec}}},$$

where AF_{spec} is the part of the overall assessment factor (AF) which extrapolates for interspecies variation in tolerance and sampling variation; such that $\text{AF} = \text{AF}_{\text{spec}} \times \text{AF}_{\text{other}}$, where AF_{other} is the assessment factor which accounts for the additional uncertainties, some of which were listed in Section 2.3. In this sense, the ATS is related to the PNEC via $\text{PNEC} = \text{ATS}/\text{AF}_{\text{other}}$. In certain probabilistically derived HC_p values, it may be the case that AF_{other} is discounted at the discretion of the risk manager based on qualitative judgements regarding additional uncertainties.

On the log-scale the role of the log-assessment factor is additive, i.e.:

$$\log_{10}(\text{ATS}) = \log_{10}(\text{Toxicity Statistic}) - \text{AS},$$

where AS is denoted as the *assessment shift*. From the probabilistic perspective under a number of different modelling and risk quantification perspectives, the AS is often assumed (see for example Aldenberg and Jaworska 2000) to be variable — dependent on the toxicity data via the standard deviation s of the log-transformed toxicity data — such that $\text{AS} = \kappa_p s$. We denote κ_p to be the *assessment shift-factor*, chosen to provide a pre-determined average level of risk $p/100$ subject to a particular risk measure and modelling assumptions. For convenience we denote $\delta_p(\mathbf{Y})$ to refer to an estimator of $\psi_p(\theta)$ — the $\log_{10}(\text{HC}_p)$ under the log-normal SSD model.

Where a toxicity database $\mathcal{G}$ of N additional substances is available, and is used to lend support in estimating $\psi_p(\theta)$ for $\mathcal{S}$, we will refer to this data by denoting y_{ij} to be a log-tolerance value; where j indicates the species which was tested with substance $i \in \mathcal{G}$. Collectively this toxicity data is denoted by $\mathbf{Y}_G$. We also define J_i to be the collection of species experimentally assessed with substance i and denote $|J_i|$ by n_i. In addition, $\bar{y}_i$ and s_i denote the usual sample mean and standard deviation of the log-transformed toxicity data for substance i.

2.6 The State of the Science

Recently there have been many suggestions put forward regarding the estimation of the HC_p using single species tolerance values when it is required that the level of protection be specified. This level is debatably implied by the choice of p, and remains undefined for fully deterministic PNEC estimation methods. Methods vary based on summary quantification and behavioural model selection. Despite this issue, a few of these methodologies have become widely accepted and commonly practiced. Some of the proposals we discuss are limited in their practical applicability. Nonetheless, we present an overview of them here, along with any assumptions made. Since this research is driven by current (EU regulatory) risk assessment guidance, we limit discussion to the most prominent estimators which do not substantially diverge from the current requirements. Unless stated otherwise, we will focus on estimating $\psi_p(\theta)$ due to the added clarity offered from the additive assessment-shift viewpoint.

Although all estimators in this thesis, unless stated otherwise, are derived under the assumption of a log-normal SSD, all methods can be extended to other tractable distributions (e.g. the log-logistic distribution) with only minor additional effort. More complicated distributions will require sophisticated numerical approaches (e.g. Hickey et al. 2008).

2.6.1 Moment Estimator

A simple but slightly crude estimator of $\psi_p(\theta)$ is obtained by the method-of-moments [M]. This method fits the SSD to the data by plugging in $\bar{y}$ and s^2 for the unknown location and scale parameters $\theta = (\mu, \sigma^2)$ respectively, yielding an estimator by directly reading off the p-th percentile of the model fit, i.e.

$$\delta_p(\mathbf{Y})_{[\mathrm{M}]} = \bar{y} - K_p s,$$

where K_p was defined earlier. The estimator is defined in general for $n \geq 2$. This method completely ignores uncertainty in the parameter estimates which is qualitatively expected to be high for small sample sizes and/or small p. Furthermore, it is not difficult to show that this estimator overestimates $\psi_p(\theta)$ for $p < 50$, which

might be unsatisfactory to a conservative risk manager. Aldenberg et al. (2002) discuss this estimator in-depth. This is a plausible estimator for admission into a risk assessment dossier (see ECHA 2008a for example), even though guidance requires that a confidence interval should be determined so that uncertainty can be assessed.

2.6.2 Luttik and Aldenberg Estimator

Luttik and Aldenberg (1997) proposed a class of estimators which were similar to the [M] estimator with two noticeable differences: (i) uncertainty in $\bar{y}$ is accounted for, and (ii) the sample standard deviation s is replaced by the pooled standard deviation s_p, calculated from a database of toxicity data for other similar substances $\mathcal{G}$. The estimator was explicitly constructed to estimate hazardous concentrations for very small sample sizes, defined to be those satisfying $n \leq 3$. Therefore, the authors advocate the inclusion of additional toxicity datasets for similar substances only when $n_i \geq 4$ is satisfied; in Chapter 3 we will relax this restriction. Under the assumption that σ is known precisely and fixed as $\sigma = s_p$, thus discounting a potentially large source of uncertainty, the [LA] estimator which underestimates $\psi_p(\theta)$ in 50% of samples is

$$\delta_p^{(0.50)}(\mathbf{Y}, \mathbf{Y}_G)_{[\text{LA}]} = \bar{y} - K_p s_p,$$

where

$$s_p^2 = \frac{\sum_{i \in \mathcal{G}} (n_i - 1) s_i^2}{\sum_{i \in \mathcal{G}} (n_i - 1)}. \tag{2.3}$$

Notice that s_p^2 does not utilise the toxicity data for $\mathcal{S}$; this is because it is assumed $|J_{\mathcal{S}}| \leq 3$ — the rationale for the estimator — which violates the requirement of inclusion, i.e. $n_i \geq 4$.

Luttik and Aldenberg (1997) also derived the 90% confidence interval for $\psi_p(\theta)$ under the same assumptions:

$$\left[\bar{y} - \left(K_p + \frac{K_5}{\sqrt{n}} \right) s_p,\ \bar{y} - \left(K_p - \frac{K_5}{\sqrt{n}} \right) s_p \right]. \tag{2.4}$$

The derivation is immediate from reasoning that the sampling distribution of $\bar{y}$ conditional on σ known (in this case equal to s_p) is normal with mean μ and variance $\sigma/\sqrt{n}$. We should note that the underlying hypothesis of homogeneity was never examined.

The work was followed up in Aldenberg and Luttik (2002), except that in this case the authors assumed σ to be fixed by any means; not necessarily as the pooled standard deviation. They suggested methods such as expert opinion or worst-case scenario options might be applied, as well as 'conservative' estimates of the pooled standard deviation; consequently we distinguish these estimators as [AL] estimators. It was suggested by Aldenberg and Luttik (2002) that the [AL] estimator be constructed using the lower bound of the confidence interval, as described by Equation 2.4, instead of $\delta_p^{(0.50)}(\mathbf{Y}, \mathbf{Y}_G)_{[\mathrm{LA}]}$. A further trivial distinction between the original [LA] and [AL] estimators is that the former assumed a log-logistic SSD, whereas the latter assumed a log-normal SSD; we restrict discussion to the latter assumption.

2.6.3 Aldenberg and Jaworska Estimator

An estimator (class) was proposed by Aldenberg and Jaworska (2000), and is now one of the most scientifically accepted estimators of $\psi_p(\theta)$ in use. In fact, application of the [AJ] estimator is permitted under REACH guidance (ECHA, 2008a), and is a frequently used tool of many scientists. The estimator derives from earlier research by Wagner and Løkke (1991) and Aldenberg and Slob (1993), contingent on log-logistic and log-normal SSD assumptions respectively. Although current guidance does not explicitly stipulate how to calculate the HC_p, the requirement of a confidence interval is straightforwardly met by the [AJ] estimator class.

From the frequentist perspective, the [AJ] method selects $\delta_p^{(\gamma)}(\mathbf{Y})_{[\mathrm{AJ}]}$ such that

$$\begin{aligned} \mathbb{P}\left[\mathrm{PAF}\left(\delta_p^{(\gamma)}(\mathbf{Y})_{[\mathrm{AJ}]}\right) \leq p\right] &\equiv \mathbb{P}\left[\delta_p^{(\gamma)}(\mathbf{Y})_{[\mathrm{AJ}]} \leq \psi_p(\theta)\right] \\ &= \gamma. \end{aligned} \tag{2.5}$$

We write $\mathrm{PAF}\left(\delta_p^{(\gamma)}(\mathbf{Y})_{[\mathrm{AJ}]}\right)$ to emphasise that the PAF is dependent upon $\delta_p^{(\gamma)}(\mathbf{Y})_{[\mathrm{AJ}]}$,

such that

$$\text{PAF}(\delta) = \Phi\left(\frac{\delta - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal CDF function.

Thus one is estimating the probability that our estimator $\delta_p^{(\gamma)}(\mathbf{Y})_{[\text{AJ}]}$ is less than the actual p-th percentile of the SSD, which turns out to be via rearrangement, equivalent to

$$\mathbb{P}\left(\left[\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} + K_p\sqrt{n}\right] \Big/ [s/\sigma] \leq \left(\bar{y} - \delta_p^{(\gamma)}(\mathbf{Y})_{[\text{AJ}]}\right)\frac{\sqrt{n}}{s}\right) = \gamma. \tag{2.6}$$

Since $\bar{y}$ and s are independent, the pivotal quantity in the left hand side of Equation 2.6 is by definition a non-central t-distributed random variable $T_{n-1,\eta}$ with $n-1$ degrees of freedom and non-centrality parameter $\eta = K_p\sqrt{n}$. The general [AJ] estimator is then defined as

$$\delta_p^{(\gamma)}(\mathbf{Y})_{[\text{AJ}]} = \bar{y} - \kappa_p(n,\gamma)s, \tag{2.7}$$

where

$$\kappa_p(n,\gamma) = \frac{1}{\sqrt{n}}F^{-1}_{T_{n-1,\eta}}(\gamma)$$

and $F^{-1}_{T_{n-1,\eta}}$ is the quantile function of $T_{n-1,\eta}$. Hence, the [AJ] estimator is defined in general for $n \geq 2$. Note that $\kappa_p(n,\gamma)$ is what we earlier coined as an assessment shift-factor (Section 2.3) used to provide an average level of impact of $p\%$ to the ecological community; moreover it does not depend on the toxicity data for $\mathcal{S}$, thus allowing for look-up tables to be produced which is highly appealing to risk managers. Aldenberg and Jaworska (2000) described assessment shift-factors satisfying this property as *universal*. The value γ can be interpreted as setting the choice of the one-sided underestimate confidence limit; but it is sometimes referred to as an 'uncertainty parameter'. Hence, setting $\gamma = 0.50$ admits a median estimator of $\psi_p(\theta)$; the estimates obtained by setting $\gamma = 0.95$ and 0.05 would together constitute a 90% confidence interval.

The [AJ] estimator was also the first to be analysed from within the Bayesian

paradigm. Aldenberg and Jaworska (2000) showed that under the prior distribution

$$p(\mu, \sigma^2) \propto \sigma^{-2} \text{ for } \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+, \tag{2.8}$$

the Bayesian estimator, which is the $100(1-\gamma)$-th percentile of the posterior distribution of $\psi_p(\theta)$, coincided with its frequentist counterpart. The prior distribution applied in this case is recognised as being the product of independent non-informative Jeffreys' priors for μ and σ^2. This is the 'practical' Jeffreys prior for (μ, σ^2) and is reported as being the recommended choice by Berger (1985, p. 89) in comparison to the standard Jeffreys prior for (μ, σ^2): $p(\mu, \sigma^2) \propto \sigma^{-3}$. For a discussion of the Jeffreys prior consult Berger (1985, pp. 87–90). From the Bayesian perspective, the role of γ is that of choosing the credible limit as opposed to the confidence limit. Aldenberg et al. (2002) note that the Bayesian calculations substantially simplify the mathematical interpretation when used in the SSD forward manner as opposed to the inverse manner, i.e. estimating the PAF conditional upon an exposure concentration; see Section 2.4.

The choice of p and γ remain a separate issue here. The choice of p, as discussed earlier, is generally decided in advance through policy decision making, with $p = 5$ the standard elective; however the choice of γ remains arbitrary. Maltby et al. (2005) suggested that setting $\gamma = 0.95$ is acceptable based on a comparison to mesocosm studies, whereas Emans et al. (1993) and Versteeg et al. (1999) empirically validate the choice of $\gamma = 0.50$ for a limited number of long-term studies. ECHA (2008a) currently requires that $\gamma = 0.50$, but also that a 50% confidence interval is calculated as a measure of uncertainty. We review this estimator, and the choice of γ, in greater detail in the following chapter.

2.6.4 The EFSA Estimator

EFSA (2005) described an approach to define a decision rule to adjust the current guidance for ERA which allowed for a specified acceptable level of protection to be achieved. The aforementioned decision rule is, for all intents and purposes, comparable to an elective for unknown $\psi_p(\theta)$ with which to evaluate the RCR; we

refer to the [EFSA] decision rule as an estimator from here onwards. In addition to the risk control approach, three behavioural models were proposed which can be used in conjunction with the methodology, each yielding a different estimator. Two estimators that did not allow the level of protection to be controlled were also discussed; one of these estimators is discussed further in Chapter 6 and the other in EFSA (2008). We limit discussion here to estimators which allow for p to be specified.

The proposal made is to choose an estimator $\delta_p(\mathbf{Y})$ such that one controls the expected PAF to be near some suitable value p. Then one chooses the [EFSA] estimator $\delta_p(\mathbf{Y})_{[\text{EFSA}]}$ such that

$$\mathbb{E}\left[\text{PAF}\left(\delta_p(\mathbf{Y})_{[\text{EFSA}]}\right)\right] = p, \tag{2.9}$$

where the expectation is taken with respect to either: the sampling distribution of the toxicity data for $\mathcal{S}$ from a frequentist perspective, or the posterior distribution of $\theta \,|\, \mathbf{Y}$ from the Bayesian perspective. The quantity in Equation 2.9 was denoted as the mean fraction exceeded (MFE) by EFSA (2005). For all intents and purposes, the risk manager would control the MFE by setting it to be p, as per the other estimators.

The frequentist calculation can be made by noting that Equation 2.9 can also be written as

$$\begin{aligned}\mathbb{E}^{\mathbf{Y}\,|\,\theta}\left[\text{PAF}\left(\delta_p(\mathbf{Y})_{[\text{EFSA}]}\right)\right] &= \mathbb{E}^{\mathbf{Y}\,|\,\theta}\left[\mathbb{P}\left(Y < \delta_p(\mathbf{Y})_{[\text{EFSA}]} \,|\, \delta_p(\mathbf{Y})_{[\text{EFSA}]}\right)\right] \\ &= \mathbb{P}\left(Y < \delta_p(\mathbf{Y})_{[\text{EFSA}]}\right),\end{aligned}$$

where Y is a random variable drawn from the SSD *independent* of the data used to calculate $\delta_p(\mathbf{Y})_{[\text{EFSA}]}$; and expectation is taken with respect to $\mathbf{Y} \,|\, \theta$. One can loosely appeal to the Markov inequality which implies that if the risk is made small, which is often the case, then the probability of exceedance is small.

The Bayesian calculation can be made by noting that Equation 2.9 is equivalent

to

$$\mathbb{E}^{\theta\,|\,\mathbf{Y}}\left[\mathrm{PAF}\left(\delta_p(\mathbf{Y})_{[\mathrm{EFSA}]}\right)\,|\,\mathbf{Y}\right] = \mathbb{E}^{\theta\,|\,\mathbf{Y}}\left[\Phi\left(\frac{\delta_p(\mathbf{Y})_{[\mathrm{EFSA}]}-\mu}{\sigma}\right)\,\middle|\,\mathbf{Y}\right], \tag{2.10}$$

where the expectation is taken with respect to the posterior distribution of θ, $p(\mu, \sigma^2\,|\,\mathbf{Y})$. This is equivalent to

$$\int_0^\infty \int_{-\infty}^\infty \Phi\left(\frac{\delta_p(\mathbf{Y})_{[\mathrm{EFSA}]}-\mu}{\sigma}\right) p(\mu, \sigma^2\,|\,\mathbf{Y})\, d\mu\, d\sigma^2.$$

In order to achieve coverage matching properties with the frequentist [EFSA] estimators, one can use non-informative prior distributions to reflect the hierarchy in the suggested behavioural model; for example, the independent product Jeffreys prior as used in the derivation of the [AJ] estimator (see Equation 2.8). From here onwards, we work solely within the Bayesian paradigm; frequentist analogies are discussed briefly later on.

We next describe the three [EFSA] estimators and the corresponding behavioural model for each, denoted as $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. For a derivation of these three estimators from a frequentist perspective, consult EFSA (2005, Appendix A.2). For an outline of the Bayesian derivation (unavailable in EFSA 2005), please consult Appendix A.2. The prior distributions considered are listed below, with the corresponding posterior distribution derivations shown in Appendix A.1.

EFSA $\mathcal{M}_1$

Following the standard model described by Equation 2.2, and the independent product Jeffreys prior for θ; solving Equation 2.9 for δ_p yields

$$\delta_p(\mathbf{Y})_{[\mathrm{EFSA}]} = \bar{y} - t_{n-1,p}\xi s, \tag{2.11}$$

where $t_{n-1,p}$ is the $(100-p)$-th percentile of a Student-t distribution with $n-1$ degrees of freedom, and $\xi = \sqrt{1+1/n}$. This estimator is defined in general for $n \geq 2$.

EFSA $\mathcal{M}_2$

If toxicity data for other substances considered similar is made available, then there may be considerable benefit in exploiting this information to stabilise the estimate of σ for $\mathcal{S}$ by incorporating the evidence about variation in values of σ from the database. It is assumed that σ is *a priori* sampled from a hyper-population distributed with an inverse-gamma distribution with shape and scale hyper-parameters α and β respectively, i.e.

$$\sigma^2 \,|\, \alpha, \beta \sim \mathcal{IG}(\alpha, \beta) \text{ for } \alpha, \beta > 0, \tag{2.12}$$

such that (α, β) have been estimated from the available toxicity database. If the hyper-parameters are specified, then it may not be necessary to have full public access to the database used to estimate them. An outline of a frequentist method for calculating suitable values of α and β from the database, as well as addressing issues of uncertainty, is reproduced in Appendix B.2 from EFSA (2005, Appendix A.4.1). We also maintain the independent standard non-informative prior distribution for μ: $p(\mu) \propto 1$, because a population of means would likely be unfathomable to regulators. This is because a distribution of mean ecological community log-tolerance for multiple substances — reducible perhaps to a distribution over unknown relative potency factors — would be uninformative and need to span several orders of magnitude. The independent Jeffreys prior in this case is the practical choice. The prior distribution is therefore defined as

$$p(\mu, \sigma^2 \,|\, \alpha, \beta) \propto (1/\sigma^2)^{\alpha+1} \exp(-\beta/\sigma^2)$$

for $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. As per the prior distribution used for $\mathcal{M}_1$, this prior distribution is chosen because it leads to mathematical tractability when combined with normal distributions, which is highly advantageous, and satisfies the remit of this thesis.

Solving Equation 2.9 for δ_p with respect to this prior distribution yields

$$\delta_p(\mathbf{Y} \,|\, \alpha, \beta)_{[\text{EFSA}]} = \bar{y} - t_{[2\alpha+n-1],p} \xi s_{\text{adj}}, \tag{2.13}$$

where s^2_{adj} — the adjusted variance estimate — is defined as

$$s^2_{\text{adj}} = \frac{2\beta + (n-1)s^2}{2\alpha + n - 1}, \tag{2.14}$$

and ξ is defined as per [EFSA] $\mathcal{M}_1$. Setting $\alpha = \beta = 0$ retrieves the same estimator derived for $\mathcal{M}_1$ as expected. For $\alpha > 0$, this estimator is defined in general for $n \geq 1$.

EFSA $\mathcal{M}_3$

Assume that within an additional database of toxicity data available for similar substances $\mathcal{G}$, the variances within these substances are homogeneous, yet unknown, perhaps due to the small data samples for each substance. This database can be exploited to better inform our estimate of σ, much the same as was assumed in the derivation of the [LA] estimator. The model is then

$$\begin{aligned} y_j \,|\, \mu, \sigma^2 &\sim N(\mu, \sigma^2) && \text{for } j \in J_{\mathcal{S}}; \\ y_{ij} \,|\, \mu_i, \sigma^2 &\sim N(\mu_i, \sigma^2) && \text{for } i \in \mathcal{G} \text{ and } j \in J_i. \end{aligned} \tag{2.15}$$

Based on this model assumption, a natural non-informative prior distribution which generalises the independent product Jeffreys prior described in Section 2.6.3 for all $N + 2$ parameters is

$$p(\mu, \sigma^2, \mu_i : i \in \mathcal{G}) \propto \sigma^{-2}$$

for $\sigma^2 \in \mathbb{R}^+$; $\mu \in \mathbb{R}$; and $\mu_i \in \mathbb{R}\ \forall i \in \mathcal{G}$.

Solving Equation 2.9 for δ_p with respect to this prior distribution yields

$$\delta_p(\mathbf{Y}, \mathbf{Y}_G)_{[\text{EFSA}]} = \bar{y} - t_{(n-1)+\varsigma,p}\xi s^*_p, \tag{2.16}$$

where ξ is defined as per [EFSA] $\mathcal{M}_1$;

$$s^{*2}_p = \frac{(n-1)s^2 + \varsigma s^2_p}{n - 1 + \varsigma}; \tag{2.17}$$

s^2_p was defined by Equation 2.3, and $\varsigma = \sum_{i \in \mathcal{G}}(n_i - 1)$. Unlike the [LA] estimators,

[EFSA] estimators utilise the available toxicity data for $\mathcal{S}$ in the pooled standard deviation estimate because no overriding restrictions are placed, *per se*, on the adequacy of datasets for sample sizes with $n < 4$. Assuming there exists at least one substance satisfying $n_i \geq 2$, then the estimator is defined in general for $n \geq 1$.

EFSA Overview

The three estimators, each based on a different behavioural model, have various pros and cons, many of which are described in EFSA (2005). $\mathcal{M}_1$ requires a lot less subjectivism from risk managers, and is appropriate for substances having non-standard modes of action, or where limited toxicity data exists in the public domain for substances considered similar. $\mathcal{M}_2$ and $\mathcal{M}_3$ make further assumptions, reflected through the prior, which while most likely ensuring smaller assessment shift-factors, require more debate than $\mathcal{M}_1$ from a risk management perspective. EFSA (2005) reports that the behavioural models were used to ensure tractability; furthermore, no actual recommendation is made that the estimators be considered for application; this was beyond the scope and authority of the report. We discuss the behavioural models in more depth in Section 2.7 and Chapter 3.

2.6.5 A Further Note on the Frequentist Perspective

The Bayesian decision rule for $\mathcal{M}_1$ (Equation 2.12) coincides with its frequentist counterpart when one updates the non-informative prior distribution as described by Equation 2.8. However, the frequentist versions of the other two approaches are subject to interpretation — something we elaborate further on in this section.

Let us begin by reconsidering the [EFSA] decision rule under $\mathcal{M}_2$. Under this behavioural model we assume that σ^2 is *a priori* distributed with an inverse-gamma distribution parameterised by shape and scale parameters α and β. The frequentist calculations presented in EFSA (2005) which led to the same decision rule as described by Equation 2.13, are valid from the frequentist viewpoint when one *does not* condition the calculations on σ, as one might do in stating the coincidence for $\mathcal{M}_1$. Thus, they are valid when we repeatedly sample, for given sample size n, across a sufficiently large population of substances, since in this scenario we do not condi-

tion on known σ. However, we have now inadvertently introduced a population of substances. If we condition on σ, *i.e.* where σ is known and we repeatedly sample within the population of a single substance, then the calculations no longer hold true. However, it is arguable that the former interpretation is more useful to a risk manager than the latter. This is because the risk manager will not be in a situation of having more than one sample within the population of the single substance for obvious reasons. On the other hand, they are in the situation to benefit from the offered coverage properties across a population of substances since they will make many risk assessments over long periods of time. Wasserman (2004) states this is in fact the correct interpretation for envisagement of repeated sampling properties. It is a risk management decision as to whether it is acceptable to have an average level of risk p over many risk assessments.

Under $\mathcal{M}_3$, *i.e.* where σ is unknown, but believed homogeneous among a population of substances, the decision rule remains valid under both interpretations. This is so because the decision rule is made in reference to a well defined population of substances *a priori*; conditioning on σ allows for inter- and intra- per-substance coverage matching (within the σ-homogeneous substance population). From the Bayesian perspective, the decision rules (Equations 2.12, 2.13 and 2.16) don't require as careful an interpretation as their frequentist counterparts.

2.6.6 Other Proposals

Estimators discussed thus far derive from probabilistic arguments which lead to tractable estimators for $\psi_p(\theta)$. There are additional estimators and assessment methods, which while intractable, are nonetheless relevant and viable for modern intermediate quantitative tier ERA; we discuss a few here.

Jagoe and Newman (1997) and Newman et al. (2000, 2002) proposed a number of bootstrap procedures. This was described as being more applicable as it was believed to handle the biased and non-random species selection procedure, as well as remove the need for over-simplified modelling of the SSD which the authors strongly criticise. Newman et al. (2000) reported that one would require a sample size between 15 and 55 in order to accurately estimate the HC_5; yet it is well recognised that this is not

achievable in practice except for a minority of high volume produced substances. Bootstrapping was also considered further by Grist et al. (2002) and Duboudin et al. (2004a,b).

O'Hagan et al. (2005) and Grist et al. (2006) have proposed a Bayesian hierarchical mixture model for the SSD made up of per-taxonomic family sensitivity distributions weighted by a (relevant) species richness index. This was exemplified for different English river environments and allowed for the inclusion of expert judgements on the sensitivity of naturally present species for which data was limited. The method is time consuming and procedurally involved. Therefore while not precluded from the intermediate quantitative tier of assessment we are concerned with, it is likely to be an inefficient use of an assessors resources. Notwithstanding this issue, the more realistic modelling of toxic stress within and between taxa would likely be a valuable tool for refined higher tier risk assessment, which otherwise may inappropriately over- and under-represent the taxa present. Hickey et al. (2008) adapted this model, in conjunction with a non-standard experimental technique of obtaining tolerance values for species naturally present in the species community (Kefford et al., 2005), to analyse the potential risk of rising salinity in eastern Australia to aquatic macroinvertebrates.

Staples et al. (2008) summarises two approaches used by the United States Environmental Protection Agency (US EPA) and one used by Environment Canada. The Environment Canada method and one of the US EPA's methods are based on quantile regression of log-transformed toxicity data; a technique also used by Wheeler et al. (2002). The second US EPA method is a deterministic-probabilistic hybrid which assumes a triangular distribution over the 4 most sensitive species' modified[3] tolerance values. We can find no justification for this procedure; consult Staples et al. (2008) for further elaboration.

A much more recent approach has been to circumvent the issue of large uncertainty from small sample sizes by using predicted data of many species as proxies for the unmeasured values. In the US, this has been developed under the name

[3]The modification made is based upon the averaging of different long-term study toxicological endpoints.

interspecies correlation estimation (ICE), see for example Dyer et al. (2006, 2008) and references therein. Each ICE model is a type II linear regression model between the log-tolerance values of two species across a range of substances which they have been assessed with. The models are then used to predict acute tolerance values for untested species based on the measured tolerance value for a surrogate species. This surrogate species can be used to estimate tolerance values for many other species; subsequently one can apply the standard decision rules listed earlier. This semi- *in silico* method appears advantageous as it limits the need for *in vivo* and *in vitro* testing. It has been shown (Gosling, 2009) that there are many defects to the current application of ICE methods and therefore we do not discuss this concept any further.

2.7 Assessment Modelling

Inherent in the different estimators discussed in Sections 2.6.1–2.6.4 are different overlapping behavioural models for probabilistically modelling interspecies tolerance. The differences in behavioural models arise predominantly due to the lack of uniform agreement on the inclusion and relevance of additional information, including toxicity databases and expert judgements, especially when sample sizes are small. This is in addition to differences arising because of the different decision rule procedures.

Although improper non-informative prior distributions are used to derive the [AJ] and [EFSA] estimators, there is no requirement that one must use such prior distributions in general. However, the fact that estimators coincide between frequentist and Bayesian viewpoints under the prior distributions discussed might be seen as an advantage from a regulators perspective. Using expert judgements to construct the necessary prior distributions is unlikely in practice, at least at the current time because this may place the risk manager under pressure from commercial organisations, in addition to potential exposure to judicial review; each of which is clearly a conflict of interest. Moreover, the subjectivism may be conceived to be a reduction in transparency to stakeholders. A discussion on jurisprudence regarding

subjective Bayesianism is considered beyond the scope of this research.

The different behavioural models used in the risk assessment proposals can be classified into four groups; the first three of which were discussed in the context of [EFSA] estimators. The models are:

$\mathcal{M}_1$: μ and σ unknown and varying between substances; database not used to provide prior information about μ and σ.

$\mathcal{M}_2$: μ and σ unknown and varying between substances; σ assumed sampled from an inverse-gamma distribution with hyper-parameters α and β; database for relevant other substances available to provide information about α and β.

$\mathcal{M}_3$: μ unknown and varying between substances, σ unknown and homogeneous between substances; database for relevant other substances available to provide prior information about σ.

$\mathcal{M}_4$: μ unknown and varying between substances; σ known.

$\mathcal{M}_1$ is the basis of Wagner and Løkke (1991); Aldenberg and Slob (1993); Aldenberg and Jaworska (2000); Aldenberg et al. (2002) and EFSA (2005, Method 3), whereby each substance risk assessment is essentially independent of one another, and therefore satisfies those who are concerned of influencing the assessment via the inclusion of additional information. Moreover, $\mathcal{M}_1$ is the default model within current technical guidance documents pertaining to the registration of general chemicals (ECHA, 2008a). $\mathcal{M}_2$ is the basis of EFSA (2005, Method 4), and is clearly motivated by the frequently observed small sample sizes, thus borrows strength from the available database in an attempt to stabilise the variance. Stabilising the variance should in principle lead to decision rules, such as [EFSA], having better performance properties, especially for small sample sizes. It was shown in the context of the [EFSA] estimators, $\mathcal{M}_2$ led to an additional 2α degrees of freedom in the estimate of σ compared to the corresponding estimate under $\mathcal{M}_1$. Even for small α, where n is small this can lead to substantial improvements. Moreover, the assumption of heterogeneity among the presumed population of substances is tenable.

$\mathcal{M}_3$ and $\mathcal{M}_4$ are in some respects very similar, except in one case σ is unknown and in the other case known. $\mathcal{M}_3$ is the basis of EFSA (2005, Method 3), whereas $\mathcal{M}_4$ is the basis of Aldenberg and Luttik (2002). The proposal in Luttik and Aldenberg (1997) fits within $\mathcal{M}_3$ and $\mathcal{M}_4$, primarily overlapping $\mathcal{M}_4$ but partially overlapping $\mathcal{M}_3$ in the sense that σ is unknown and homogeneous, yet treated as known upon estimation, thus neglecting all corresponding uncertainty. Note that $\mathcal{M}_4$ might have access to additional toxicity data if required, although it is not necessarily required since one can specify σ via other methods, for example using expert opinion. It should be noted that $\mathcal{M}_3$ and $\mathcal{M}_4$ are insupportable from wholly realistic considerations; this is not to preclude them from the outset, since their more pragmatic formulations might lead to tools for efficiently conservative assessment.

No model assumption here proposes a hyper-population of means since, as discussed in the description of [EFSA] $\mathcal{M}_2$, it is likely to be very weakly informative even for strict definitions of toxic mode of action, and more contentious due to a lack of understanding within the ERA arena. Although $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ have additional assumptions which are not commonplace in regulatory guidance documentation, and in some case not discussed in the associated scientific literature, we would note that current guidance documentation indicates that alternative adjustments to the default method is acceptable where warranted subject to defensibility.

We will keep the behavioural models listed above independent from assumptions regarding the data generating mechanism, which is currently consistent among all of the estimators described thus far as being a log-normal distribution. The latter is examined in later chapters with respect to one narrowly focused and contended issue.

There are shortcomings in all of the behavioural models described here from the modelling viewpoint of statisticians, ecologists and ecotoxicologists. However, at the intermediate tier of quantitative risk assessment, the decision process must not be overly complicated, whilst being robust. Additionally, it is not apparent whether $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$, conditional on the underlying assumptions of each being true, are on a par with $\mathcal{M}_4$; by which we mean that the degree to which the uncertainty

is handled for the latter model will see its level of assessment refinement heuristically ranked in between the former three models and the strictly deterministic approach. We will explore this perspective further in Chapter 3, alongside a more comprehensive analysis and discussion of the behavioural models in actual practice.

Finally, it is useful to understand that although the role of risk managers and risk assessors is complicated, and might appear to be reasonably distinct, we must acknowledge that at some point an overlap must be accepted with regards to choosing a model; this is *in lieu* of structural model uncertainty handling which is usually loosely based on goodness-of-fit tests.

Chapter 3

HC_p Estimation *Revisited*

3.1 Introduction & Motivation

Uncertainty is a crucial element of chemical risk assessment, and is inherent in both the hazard and exposure components of the overall risk assessment, as was discussed in Section 2.1. Not only is it required that uncertainty be assessed — whether deterministically or probabilistically — in order to be incorporated by the decision maker to lower the likelihood of adverse ecological effects occurring, it is also required that there be meaningful transparency in the handling process for the different stakeholders. Just as the 'precautionary principle' is used as a tool for conservatism in the strictly deterministic lower quantitative tier risk assessments, additional tools are commonly advocated for probabilistic assessments, such as arbitrarily reported one-sided confidence limits which are used as conservative estimators.

At the tier of assessment we are concerned with, the role of the risk assessor is or at least ideally should be, to calculate an estimate of the HC_p (namely with $p = 5$ by current requirements) under uncertainty due to not testing all species, whilst facilitating the risk managers request for a certain level of conservatism (whether protective or otherwise). Such problems can be setup quite naturally within a statistical decision theoretic framework; this allows for the concept of *loss* and *prior information* to be included in the decision making process. The former is naturally helpful for purposes of incorporating the required conservatism in a transparent manner. While consideration of the subjective Bayesian perspective is abstained

from based on our earlier brief discussion, the issue of loss functions is considered further.

Chen (2003) raised an important criticism with regards to the US EPA's current risk assessment procedural strategy, although not limited to this agency exclusively, who advocate using confidence tail-limits when estimating hazardous concentrations or other threshold impacts of risky scenarios in order to err on the side of caution. ECHA (2008a, p. 20) refers to such arbitrariness as one of the 'common drawbacks' of using statistical extrapolation. Recommendations to use the lower 95% one-sided underestimate confidence limit (equivalently the 5-th percentile of the HC_p distribution) are supported in one form or another by Van Straalen and Denneman (1989); Wagner and Løkke (1991); Jagoe and Newman (1997); Newman et al. (2000), as well as indirectly by the findings of Maltby et al. (2005). This can result in formal procedures such as chemical or pesticide registration being delayed and more costly, or even false alarm clean-up decisions being taken, if approved methodology becomes too conservative in estimation. This is a strong motivation for attempting to find an estimator which remains scientifically defensible, yet exercises an appropriate degree of conservatism which is transparent, *a priori*. This call is reflected in a statement by the highly authoritative pair Forbes and Calow (2002b), who stated:

> '[Risk assessment] needs to be robust enough to be applied routinely and conservative enough to ensure that the process is not stopped prematurely.'

In this chapter we examine the HC_p (conservative-) estimation problem within an intuitive statistical decision theoretic framework. Conservatism is introduced in a straightforward manner by unifying physical impact with the concept of utility, or rather, negative utility — a more pessimistic description referred to as *loss*. Different specifications of such loss functions are introduced for application within the appropriate framework of risk assessment; in addition we demonstrate such a specification which risk managers and stakeholders alike might see as advantageous from the *protectionist* vantage point. A simple method of characterising the loss and impact in order to extract an 'optimal' decision rule, known as a Bayes rule, is used. The relation between such estimators for a well known class of loss functions

is used to highlight the relationship to the [AJ] estimator (Section 2.6.3) — the most scientifically applied estimator.

With so many estimators available to risk managers and future users, we conclude that better understanding of the problem of conservative estimation is required, rather than the usual appeal to *ad hoc* reporting of statistical inference.

3.2 Decision Theory Primer

In this section we introduce the two pivotal elements which we use to estimate HC_p values: loss functions and Bayes rules.

3.2.1 Loss Functions

Loss functions are a common tool in modern statistical decision theory and risk analysis under both statistical paradigms — frequentist and Bayesian (Berger, 1985; Bernardo and Smith, 1994). A loss function L assigns a measure of cost to different actions for each possible outcome, or *states of nature*. Following the notation of Berger (1985), Bernardo and Smith (1994), and Section 2.5, we will define $\mathcal{A}$ to be an action space and $a \in \mathcal{A}$ to be an action. The hazard assessment problem is to estimate the $\log_{10}(\mathrm{HC}_p)$, which we denoted as $\psi_p(\theta)$, so that $\psi_p : \mathbb{R} \times \mathbb{R}^+ \mapsto \mathbb{R}$, hence $\Psi_p = \mathcal{A} = \mathbb{R}$. Therefore, the loss function $L(\psi_p(\theta), a)$ will be such that $L : \Psi_p \times \mathcal{A} \mapsto \mathbb{R}$.

The specification of loss/cost needn't be monetary, for example 'moral' cost is not unfathomable, although likely difficult to appraise. However, specification of a well defined loss function is not always straightforward for a given problem. The rationale for using loss functions in ecological risk assessment is loosely determined by the context of this diverse field. For example, one such application would be to use them as a decision making tool for the purpose of prioritising remediation or further mitigation at different sites of exposure, each presenting different features of degradation. However, incorporating loss on the functional level to this problem in order to adequately make sensible decisions would be a highly challenging task. It is much more likely that decision procedures such as the United Nations Flash

Environmental Assessment Tool (FEAT) (UN, 2009) would be used; we discuss this further in Section 3.2.2. With the exception of the latter, this chapter demonstrates the application of loss functions in the statistical context of *estimation*; here the focus being the unknown HC_5 quantity.

Chen (2003) proposed a non-parametric estimator of the HC_p based on minimising the frequentist risk function — defined as the statistically expected loss for a decision rule with respect to the data (we more formally define this later on) — for a loss function which assigned loss as one of three constant values depending on whether the decision rule lied below, within, or above an interval function of the HC_p prescribed by the risk manager; this estimator is, in our opinion, more politically motivated rather than intent on recognising the severity of over- and under-estimation. Separate from whether we regard this loss function as practical or not, the aforementioned decision rule required a toxicity data sample size of $n \geq 19$ when $p = 5$; a situation unlikely for any realistic intermediate quantitative tier of hazard assessment, thus having limited practical applicability. Incorporation of parametric assumptions encased within SSD theory can remove demand for such quantities of data, however such choice as noted earlier, requires the sharp subjective selection of a model.

3.2.2 Making Decisions

The action, as a function of the observable toxicity data $\mathbf{Y}$, is known as a *decision rule*, denoted $\delta_p(\mathbf{Y})$. Letting $\mathbf{Y}$ be a sample from a probability distribution which is parameterised by θ, the *frequentist risk* function is defined to be the average of the loss function over the global data model evaluated at θ, denoted as $R(\psi_p(\theta), \delta_p) \stackrel{\Delta}{=} \mathbb{E}^{Y|\theta} L(\psi_p(\theta), \delta_p(\mathbf{Y}))$. The reasonably acceptable condition of *admissibility* for a decision rule — defined to be where no other decision rule exists such that its risk function is dominated by the one being considered — can lead to large sets of decision rules. Consequently, an additional criterion is often advocated to prescribe a specific rule, such as the following: let $\mathcal{D}$ be the complete collection of measurable decision rules, and define $\pi(\theta)$ to be a prior distribution of θ with support on $\theta \in \Theta$, then one can obtain the *Bayes risk* $r(\pi, \delta_p) \stackrel{\Delta}{=} \mathbb{E}^{\theta} R(\psi_p(\theta), \delta_p(\mathbf{Y}))$, where the expectation

is taken with respect to $\pi(\theta)$ (denoted with a superscript θ). A *Bayes rule* is then defined to be a decision which minimises the Bayes risk, i.e. $\delta_p^* = \arg\min_{\delta_p \in \mathcal{D}} r(\pi, \delta_p)$.

An alternative (frequentist) decision rule is the *minimax rule*; defined to be the decision rule which minimises $\sup_{\psi_p(\theta) \in \Psi_p} R(\psi_p(\theta), \delta_p)$ among all decision rules. It is unlikely that a risk manager would accept minimax for at the intermediate quantitative tier of hazard assessment considered here with the purpose of estimating $\psi_p(\theta)$, since it only protects against extreme events which typically have a very small probability of occurrence; focus should be on protecting against scenarios with more appreciable probabilities. In the context of FEAT, which has a different purpose in the field of ERA than what we primarily focus upon in this thesis, assessment is based on the fundamental concept of impact — defined to be a function of exposure (including quantity) and hazard — for many different pathways and scenarios of (possibly multiple) chemical release. By restricting consideration to only those hazard-exposure pairs with appreciable probabilities of occurring, the tool can basically be interpreted as basing risk management decisions on the minimisation of the maximum loss, which is measurable in different ways for different pathways and receptors (e.g. adverse effects to human life or ecosystems). Consequently, this is a form of minimax decision making. This demonstrates that different principles of setting preference over the space of decisions are relevant, perhaps to different contexts of ecological risk assessment; notwithstanding this, we do not consider minimax decision rules beyond this point.

From the strictly Bayesian perspective, we can define the *Bayesian expected loss* as $\rho(\pi, a) = \mathbb{E}^{\theta \mid \mathbf{Y}} L(\psi_p(\theta), a)$ and similarly define the Bayes action (also referred to as the Bayes estimator; see Wasserman 2004, Chapter 12) as a^* that minimises $\rho(\pi, a)$, i.e. $a^* = \arg\min_{a \in \mathcal{A}} \rho(\pi, a)$. By reversing the order of integration in the definition of the Bayes risk, it can be shown that $r(\pi, \delta_p) = \mathbb{E}^{\mathbf{Y}} \rho(\pi, a(\mathbf{Y}))$ for any *proper* prior distribution $\pi(\theta)$, where the expectation is taken with respect to the marginalised distribution of the data $\mathbf{Y}$; see Wasserman (2004, pp. 197–198) for a straightforward proof. Consequently, any decision rule $a^*(\mathbf{Y})$ which minimises the posterior expected loss will also minimise the expectation on the right-hand side of the equality; hence the Bayes rule is equal to the Bayes action.

Prior distributions for the behavioural models described in Section 2.7 are provided in Sections 2.6.3 and 2.6.4; the subsequent posterior distributions are derived in Appendix A.1. It is acknowledged that the prior distributions applied here are improper, although the corresponding posterior distributions are in fact proper. Consequently, the Bayesian expected loss is also well defined, even though the Bayes risk will *not* be in general. A distinction is sometimes made (Berger, 1985; Wasserman, 2004) via the inclusion of the modifier 'generalised' to the title; like French and Ríos Insua (2000), we omit this differentiation because whilst our decisions are based on non-informative priors, the theory is straightforwardly extendable to include expert judgements.

Operationally, throughout this thesis we will proceed by minimising the posterior risk $\rho(\pi, a)$, i.e. determining the Bayes action. However, in the interest of simplicity, we opt to follow the convention of Berger (1985) and Bernardo and Smith (1994) by referring to the result as a Bayes rule $\delta_p^*(\mathbf{Y})$; although it is understood that the result is interpretable as a Bayes action $a^*(\mathbf{Y})$. Hence, all Bayes rules discussed in this thesis will be defined as

$$\delta_p^*(\mathbf{Y}) = \underset{\delta_p(\mathbf{Y})}{\arg\min}\ \mathbb{E}^{\boldsymbol{\theta}\,|\,\mathbf{Y}}\left[L(\psi_p(\boldsymbol{\theta}), \delta_p(\mathbf{Y}))\right]. \tag{3.1}$$

Note that scaling any loss function by a positive constant C will not change the Bayes rule. Hence, without loss of generality, we will set $C = 1$.

Working within the Bayesian paradigm for decision theory problems is judged to be sensible, based on the considerable arguments in Berger (1985); Bernardo and Smith (1994) and references therein. The intuitiveness of minimising the posterior expected loss is immediate upon this declaration. This is how the risk manager (assessor) should make a decision; Berger (1985) describes the former procedure as the 'correct way to view the situation'. This standpoint is argued for by noting that one should condition upon what is known, in this case $\mathbf{Y}$, and subsequently average the loss over what is unknown, i.e. θ. Moreover, we note that under this viewpoint, *a priori*, no other decision rule will perform better over multiple risk assessments (each for different substances) other than the one we would admit as the optimal procedure

conditional on the observed toxicity data, *a posteriori*. Although the minimisation of the Bayes risk is precluded by the improperness of the non-informative prior distributions, in situations where a proper prior distribution is available, Berger (1985) still describes the procedure [of direct Bayes rule calculation] as 'bizarre' under the aforementioned standpoint because the frequentist averages over the entire model for the data, not what is unknown.

For the discussion in this chapter, we begin by first analysing behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$; $\mathcal{M}_3$ and $\mathcal{M}_4$ are deferred until Section 3.8. An additional element of discussion pertaining to the applicability of behavioural models is also a recurrent theme of discourse.

3.3 Squared Error Loss

We begin by introducing a simple and frequently applied loss function for estimation problems: the standard squared error loss (SEL) function (also known as quadratic loss); see for example Berger (1985, pp. 60–62). The SEL function, in the context of our problem, can be defined as

$$L(\psi_p(\theta), \delta_p(\mathbf{Y})) = [\psi_p(\theta) - \delta_p(\mathbf{Y})]^2 . \tag{3.2}$$

SEL is symmetric around the point $\Delta \equiv \delta_p(\mathbf{Y}) - \psi_p(\theta) = 0$, hence punishing over-estimation equally as it punishes under-estimation at a rate which is proportionally quadratic. Zellner (1986) reports that symmetric loss functions, like SEL, are usually not suitable for real life problems; a point also made by Berger (1985).

Solving Equation 3.1 yields the Bayes rule to be defined as the posterior expectation of $\psi_p(\theta)$, i.e.

$$\delta_p^*(\mathbf{Y}) = \mathbb{E}^{\theta \,|\, \mathbf{Y}} \psi_p(\theta),$$

which when substituted with the posterior distributions for $\mathcal{M}_1$ and $\mathcal{M}_2$, we obtain the following Bayes rule [SEL] estimators

$$\delta_p^*(\mathbf{Y} \mid \alpha, \beta)_{[\text{SEL}]} = \bar{y} - \kappa_p^*(n, \alpha)\hat{\sigma}, \tag{3.3}$$

where

$$\kappa_p^*(n,\alpha) = K_p\sqrt{\frac{2\alpha+n-1}{2}}\frac{\Gamma(\frac{2\alpha+n-2}{2})}{\Gamma(\frac{2\alpha+n-1}{2})} \tag{3.4}$$

and $\hat{\sigma} = s_{\text{adj}}$ as defined by Equation 2.14; see Appendix C.1 for the derivation of these estimators. It is clear that by setting $\alpha = \beta = 0$ one retrieves the estimator for $\mathcal{M}_1$; setting (α, β) to their estimates based on the additional toxicity data retrieves the estimator for $\mathcal{M}_2$. We will drop the (α, β) arguments from the estimator to indicate an estimator for $\mathcal{M}_1$ only. The estimators are defined in general for $n \geq 3$ for $\mathcal{M}_1$; and for $n \geq 2$ if $\alpha > 0$ for $\mathcal{M}_2$.

It is evident from Equation 3.3 that the Bayes rules are of the same canonical form for each behavioural model. Moreover, the assessment shift-factor $\kappa_p^*(n,\alpha)$ is independent of the toxicity data for $\mathcal{S}$, as was the situation for the [AJ] and [EFSA] estimators. It is important to emphasise that the assessment shift-factors are strictly non-comparable under the two behavioural models because they each multiply a different estimate of σ: $\hat{\sigma} = s$ ($\mathcal{M}_1$) and $\hat{\sigma} = s_{\text{adj}}$ ($\mathcal{M}_2$).

By noting that the [SEL] and [AJ] estimators are of the same form for $\mathcal{M}_1$, analysis reveals that $\delta_p^*(\mathbf{Y})_{[\text{SEL}]} < \delta_p^{(0.50)}(\mathbf{Y})_{[\text{AJ}]}$ (see Equation 2.7, Section 2.6.3) for finite $n \geq 3$ and relevant ranges of p, i.e. [SEL] is a more conservative estimator than that of the median [AJ] estimator. Figure 3.1 (left panel) indicates this by plotting the assessment shift-factors for $p = 5$ and a wide range of n. This is because $\psi_p(\theta)$ has a relocated and rescaled non-central t-distribution which is negatively skewed for $p < 50$, meaning that, in general the mean of the distribution is less than the median. When considering the [AJ] 95% one-sided underestimate confidence limit, conservatism is reversed against the [SEL] estimator such that the former is the most conservative. For $p = 5$, we also display the [EFSA] estimators which indicate conservatism relative to [SEL] and median [AJ] estimators. Similar conclusions are made when considering $\mathcal{M}_2$ where we opted to fix $(\alpha, \beta) = (1.05, 0.088)$ — the values reported in EFSA (2005) for a toxicity database of fish exposed to pesticides; see Appendix B.2 for instruction on hyper-parameter estimation applied here. A similar graph as before is displayed in Figure 3.1 (right panel).

In order to present Figure 3.1 (right panel), it was required that we derive the [AJ] estimator class for $\mathcal{M}_2$ because Aldenberg and Jaworska (2000) only considered

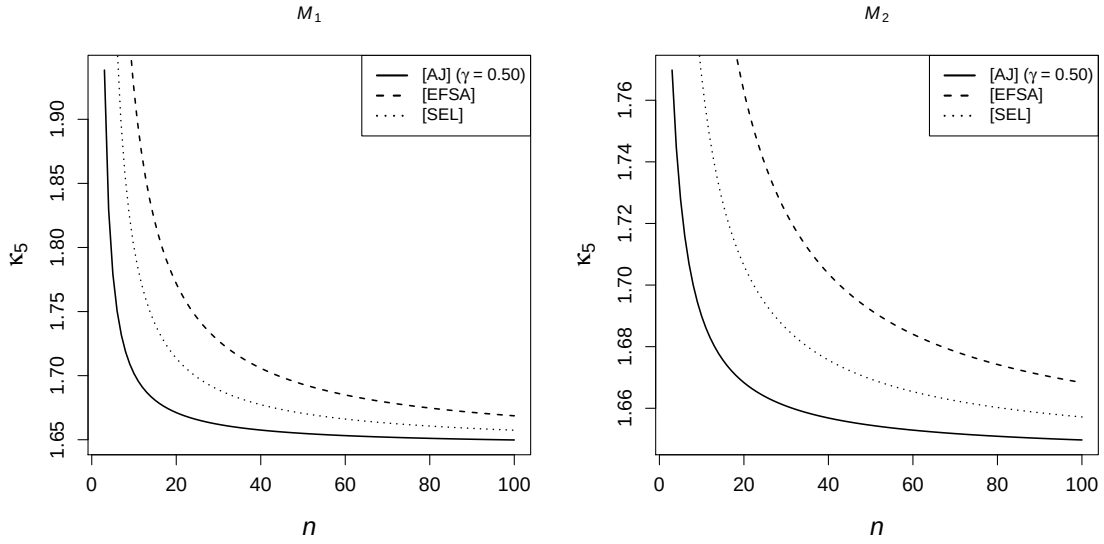


Figure 3.1: Interpolated plots of κ_5 against n for $\mathcal{M}_1$ (left) and $\mathcal{M}_2$ (right). [AJ] = solid; [SEL] = dotted; [EFSA] = dashed; $(\alpha, \beta) = (1.05, 0.088)$ $(\mathcal{M}_2)$.

$\mathcal{M}_1$. This is equivalent to determining the $100(1-\gamma)$-th percentile of the posterior distribution of $\mu - K_p\sigma \,|\, \mathbf{Y}, \alpha, \beta$, which is simply

$$\delta_p^{(\gamma)}(\mathbf{Y} \,|\, \alpha, \beta)_{[\mathrm{AJ}]} = \bar{y} - \kappa_p(n, \alpha, \gamma) s_{\mathrm{adj}}, \tag{3.5}$$

where

$$\kappa_p(n, \alpha, \gamma) = \frac{1}{\sqrt{n}} F^{-1}_{T_{2\alpha+n-1,\eta}}(\gamma)$$

and $\eta = K_p\sqrt{n}$ was defined earlier. We present the full derivation of this estimator in Appendix C.2.

Increasing p from $p = 5$, it is observed that conservatism between [SEL] and [EFSA] estimators is not consistent for all n. For example, with $p = 10$, a situation which might have ecological relevance for long-term exposure regimes (Hickey et al., 2008), the relative conservatism changes between $n = 3$ and $n = 4$ such that for $n = 3$ $\delta^*_{10}(\mathbf{Y})_{[\mathrm{SEL}]} < \delta_{10}(\mathbf{Y})_{[\mathrm{EFSA}]}$ (see Equation 2.11, Section 2.6.4); although, difference is negligible. The [AJ] median estimator still remains less conservative than the others for all 'sensible' n; see Figure 3.2.

In general, SEL may not be an appropriate loss function for risk managers because of its symmetry property, which is unattractive from a protectionist viewpoint.

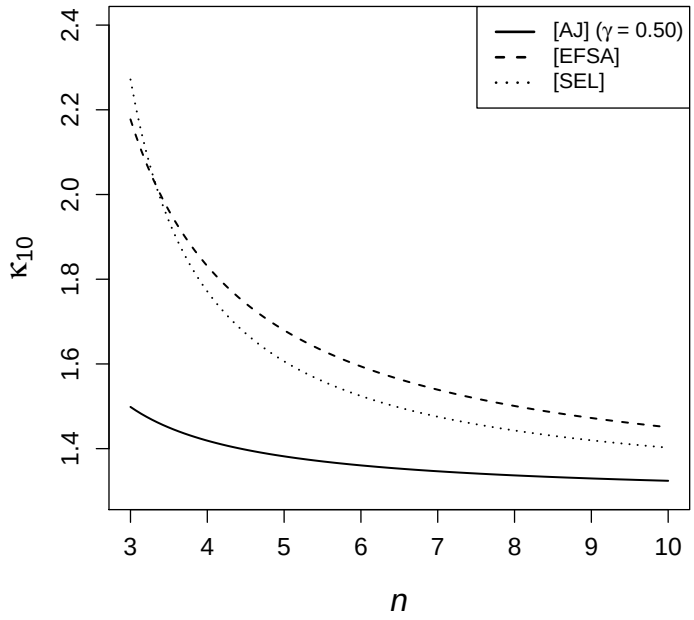


Figure 3.2: Interpolated plots of κ_{10} against n for $\mathcal{M}_1$. [AJ] = solid; [SEL] = dotted; [EFSA] = dashed. $\delta^*_{10}(\mathbf{Y})_{[\text{SEL}]} = \delta_{10}(\mathbf{Y})_{[\text{EFSA}]}$ when $n \approx 3.55$.

However, this is *not* to say that the loss function wouldn't be useful for certain risk assessment scenarios.

3.4 Generalised Absolute Loss

Generalised absolute loss (GAL) is a class of loss functions which features absolute loss as a special case. GAL is parameterised by two parameters: C_1 and C_2, which are used to fix the risk managers loss-specification. The GAL function class, in the context of our problem, can be defined as

$$L\left(\psi_p(\theta), \delta_p(\mathbf{Y})\right) = \begin{cases} C_1[\psi_p(\theta) - \delta_p(\mathbf{Y})] & \text{if} \quad \psi_p(\theta) \geq \delta_p(\mathbf{Y}) \\ C_2[\delta_p(\mathbf{Y}) - \psi_p(\theta)] & \text{if} \quad \psi_p(\theta) < \delta_p(\mathbf{Y}) \end{cases} \tag{3.6}$$

For $C_1 = C_2$ the loss function reduces to absolute loss which is symmetric about the point $\Delta = 0$. The parameters (C_1, C_2) can be interpreted as the unit cost of under- and over-estimation of the HC_p by one order of magnitude respectively. This is

because the cost increases linearly for both under- and over-estimation. Moreover, for $C_1 \neq C_2$ the loss function is asymmetric such that as $C_2 > C_1$, then over-estimation by one order of magnitude is punished more than underestimation of the same order by a factor of C_2/C_1.

By solving Equation 3.1, it can be shown that the Bayes rule [GAL] estimator is the $100C_1/(C_1 + C_2)$-th percentile of the posterior distribution of $\psi_p(\theta)$; see Appendix C.3 for the proof. Substituting the posterior distributions for $\mathcal{M}_1$ and $\mathcal{M}_2$, we determine these decision rules to be equal to

$$\delta_p^*(\mathbf{Y} \mid \alpha, \beta, C_1, C_2)_{[\text{GAL}]} = \bar{y} - \kappa_p^*(n, \alpha, C_1, C_2)\hat{\sigma}, \tag{3.7}$$

where

$$\kappa_p^*(n, \alpha, C_1, C_2) = \frac{1}{\sqrt{n}} F^{-1}_{T_{2\alpha+n-1, K_p\sqrt{n}}}\left(\frac{C_2}{C_1 + C_2}\right) \tag{3.8}$$

and $\hat{\sigma} = s_{\text{adj}}$. Additional details regarding the derivation of the [GAL] rule here is presented in Appendix C.3. The estimator for $\mathcal{M}_1$ is retrieved by setting $\alpha = \beta = 0$, for $\mathcal{M}_2$ the estimator is retrieved by setting (α, β) to their estimates based on a suitable toxicity database.

If we compare $\delta_p^*(\mathbf{Y} \mid \alpha, \beta, C_1, C_2)_{[\text{GAL}]}$ to $\delta_p^{(\gamma)}(\mathbf{Y} \mid \alpha, \beta)_{[\text{AJ}]}$ (Equation 3.5), we notice that the estimators are identical for both $\mathcal{M}_1$ and $\mathcal{M}_2$ if we let $\gamma = C_2/(C_1+C_2)$. Hence [AJ] estimators are identical to [GAL] Bayes rules, *qua* identical prior distributions. We therefore advocate the use of [AJ] to denote such estimators; although it is understood that the [AJ] estimators were strictly proposed under the context of behavioural model $\mathcal{M}_1$.

This relation implies that the two popular [AJ] estimators, i.e. the median ($\gamma = 0.50$) and one-sided 95% underestimate credible limit ($\gamma = 0.95$), correspond to [GAL] Bayes rules when $C_1 = C_2$ and $C_2 = 19C_1$ respectively. Also, the one-sided 5% underestimate credible limit ($\gamma = 0.05$) corresponds to the [GAL] Bayes rule with $C_1 = 19C_2$. However, we are less concerned with this case since we are primarily working from a protectionist viewpoint. It is clear that it is only necessary to consider the relative cost of overestimation to underestimation, i.e. C_2/C_1, since $\gamma = (1 + C_2/C_1)^{-1}$. This is because without loss of generality, we can scale any loss

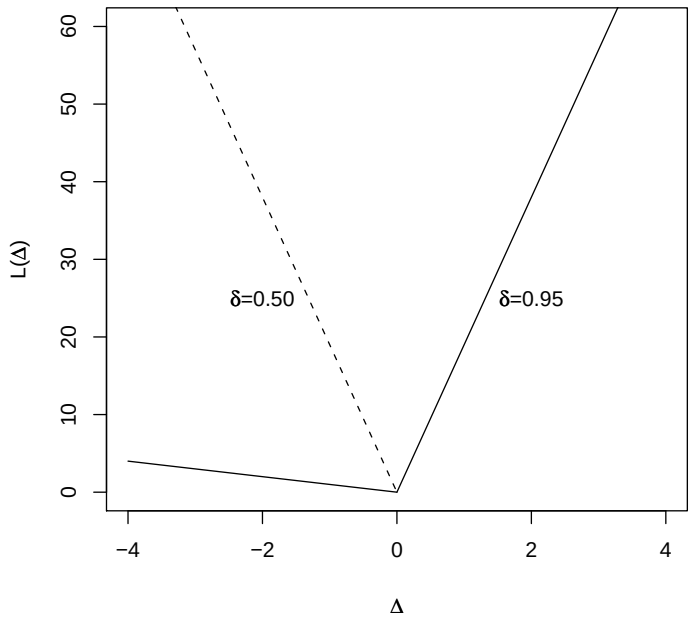


Figure 3.3: A GAL function: $C_2 = 19C_1$ (solid); $C_1 = C_2$ (symmetric union of dashed and solid). $C_2 = 19$ (without loss of generality).

function by a positive constant without altering the Bayes rule. Figure 3.3 describes the GAL loss function for the two common [AJ] point estimators of $\psi_p(\theta)$.

The revealing insight of the [AJ] estimators from a decision theoretic perspective allows one to consider the conservatism of the estimators further. Van Straalen and Denneman (1989); Wagner and Løkke (1991); Jagoe and Newman (1997) and Newman et al. (2000) suggest using $\delta_5^{(0.95)}(\mathbf{Y})_{[\mathrm{AJ}]}$ in order to err 'sufficiently' on the side of caution. However, it is acknowledged that this estimator is over conservative (Emans et al., 1993; Chen, 2003). Moreover, there is no explanation in current literature why a risk manager would not consider, say, $\delta_1^{(0.50)}(\mathbf{Y})_{[\mathrm{AJ}]}$ over $\delta_5^{(0.95)}(\mathbf{Y})_{[\mathrm{AJ}]}$; clearly both are conservative relative to $\delta_5^{(0.50)}(\mathbf{Y})_{[\mathrm{AJ}]}$.

So, assuming a risk manager can specify the maximum permissible PAF p and that they subscribe to the GAL function class, it would perhaps be more transparent for them to specify γ from considerations of C_2/C_1 on a case-by-case basis which is reflective of the assessment portfolio.

The median [AJ] estimator may be an adequate summary estimator, especially if

used with an additional deterministic assessment factor; currently this is set between 1–5. It would be a policy decision as to whether a symmetrical loss function would be appropriate. The former is currently acceptable for use in chemical safety assessment under REACH guidance (ECHA, 2008a) with an assessment factor as low as 1, i.e. where the HC_5 estimate would serve as the PNEC; Versteeg et al. (1999) tentatively supported this choice when using chronic NOEC toxicity data. However, there is little information in the technical guidance documents which specifies how the required 50% confidence interval of the HC_5 is to be used in assessing the uncertainty about the estimator in the decision making process, nor how the risk manager will select the additional assessment factor to be applied to HC_5.

Unlike SEL, GAL is a linearly increasing loss function for increasing $|\Delta|$; the appropriateness of which would need ratification by risk managers. However, on recalling that an increase in estimation error of $\Delta = 1$ would overestimate the HC_p by a factor of 10, and a further increase of $\Delta = 1$ (i.e. overall $\Delta = 2$) would increase this by a further factor of 10, one would need to consider whether the cost between these two levels of error should remain constant. A key advantage to the linearity is that a risk manger needs to only specify a single parameter in order to specify the degree of conservatism in the loss function for any level of under- or over-estimation; this is shown later to not be a property of all 'useful' loss functions.

3.5 LINEX Loss

It would appear sensible that an estimator which should be *a priori* conservative within the scope of this estimation problem would derive from an asymmetric loss function, because overestimation of the HC_p would potentially lead to greater than p% of species in the assemblage being affected by exposure. Zellner (1986) notes that asymmetrical loss is a practical perspective for many problems. Hence, for risk managers who wanted to ensure conservatism, such loss functions would be highly attractive, especially when uncertainty is large due to small toxicity data sample sizes. For many substances where risk characterisation is required, the relevant loss, whether financial or otherwise, would most likely although not always, outweigh the

loss of restricting the substance for use, whether permanently or until a higher tier assessment is performed.

Should an estimator be desired which reflects an *average*-based measure of risk then a symmetric loss function would be appropriate; this is the standard statistical prediction approach, cf. the mean, median and mode. The choice of which will be dependent on how the risk manager envisages the long-term properties of risk assessment behaving. This raises a policy based dilemma which requires addressing regarding how risk is handled when the aim is to *predict* the concentration of a substance which will be unlikely to have adverse effects.

For asymmetric loss functions which are non-linear there is a need to reconsider the error metric Δ. A simple thought experiment demonstrates this requirement. A risk manager is asked to consider two risk assessments for two separate substances $\mathcal{S}_1$ and $\mathcal{S}_2$. The potential target environments are identical, and joint-toxic effects are discounted. Then the specification of the same loss function based upon the error metric Δ means that identical loss is placed on all estimators δ. However, if the unknown interspecies variance parameters satisfy $\sigma_1 > \sigma_2$, then loss will be 'relatively' more conservative for $\mathcal{S}_2$. It is therefore desirable to incorporate this; a standardised measurement error would be one way for a risk manager to specify loss function irrespective of the SSD. We revisit this issue pending a discussion on the following loss function.

The LINear-EXponential (LINEX) loss function, first proposed and utilised by Varian (1975), conveys loss as approximately increasing linearly on one side and exponentially on the other side; hence non-linearly asymmetric. However, application of such a loss function requires thought towards its applicability since Zellner (1986) has shown that in common estimation problems, such as the estimation of the mean of a normal distribution with known variance, the traditional estimators are not always admissible under LINEX.

We use a modified LINEX loss function proposed by Zieliński (2005), which we refer to as scaled LINEX loss. The 'scaled' term refers to the scaling of $\Delta = \delta_p(\mathbf{Y}) - \psi_p(\theta)$ by σ which transfers the assignment of loss onto a 'standardised' scale. This scaling which we discuss further later on, is intuitively appealing as loss

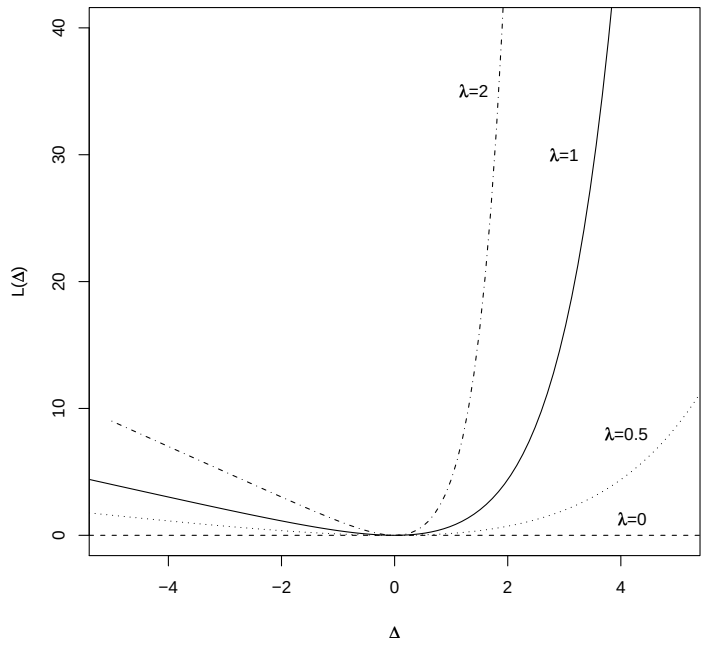


Figure 3.4: The standard LINEX loss function: $\lambda = 0$ (dashed); $\lambda = 0.5$ (dotted); $\lambda = 1$ (solid); $\lambda = 2$ (dot-dash).

is now placed on percentiles. The *standard* LINEX loss function does not feature the scaling of Δ by σ. Within the context of our problem scaled LINEX is defined as

$$L(\psi_p(\theta), \delta_p(\mathbf{Y}); \sigma) = \exp\left\{-\lambda \frac{\delta_p(\mathbf{Y}) - \psi_p(\theta)}{\sigma}\right\} - \lambda\left\{\frac{\delta_p(\mathbf{Y}) - \psi_p(\theta)}{\sigma}\right\} - 1. \quad (3.9)$$

To understand the role of the free parameter λ we have plotted the standard LINEX loss function in Figure 3.4 for different specifications of $\lambda > 0$. For $\lambda < 0$, the loss function is reflected about the point $\Delta = 0$. It is understood therefore that λ controls the asymmetry of the loss function, such that as $\lambda > 0$ increases, the conservatism increases, i.e. overestimation is punished more severely than underestimation, and *vice versa*. This feature of flexibility might increase the appeal to a risk manager, as was the case indirectly with [GAL]. In particular for $\lambda > 0$, when $|\Delta| \to \infty$, $L(\cdot)$ approximately increases exponentially when $\Delta > 0$, and approximately linearly when $\Delta < 0$. Finally, via a Taylor expansion it can be seen that for small $\lambda|\Delta|/\sigma$, $L(\cdot) \approx (\lambda\Delta)^2/2\sigma^2$. Consequently, as $\lambda \to 0$ scaled LINEX tends to a

scaled SEL function; unlike standard SEL (Equation 3.3), this also has the metric Δ scaled (see Section 3.7). Thus, scaled LINEX loss functions will not, in general, lead to identical decision rules as the standard SEL function for small λ.

Solving Equation 3.1 with the posterior distributions for $\mathcal{M}_1$ and $\mathcal{M}_2$ we obtain the Bayes rule [LINEX] estimator to be of the form

$$\delta_p^*(\mathbf{Y} \mid \alpha, \beta, \lambda)_{[\text{LINEX}]} = \bar{y} - \kappa_p^*(n, \alpha, \lambda)\hat{\sigma}, \tag{3.10}$$

where $\kappa_p^*(n, \alpha, \lambda)$ is the *unique* solution to

$$\int_0^\infty t^{(2\alpha+n-2)/2} \exp\left\{-\lambda\kappa_p^*\sqrt{t} - \left(\frac{2\alpha+n-1}{2}\right)t\right\} dt = \Gamma\left(\frac{2\alpha+n}{2}\right)\left[\frac{2\alpha+n-1}{2}\right]^{-\frac{2\alpha+n}{2}} \exp\left\{-\lambda\left[K_p + \frac{\lambda}{2n}\right]\right\} \tag{3.11}$$

for κ_p^*; and $\hat{\sigma} = s_{\text{adj}}$ as defined by Equation 2.14. See Appendix C.4 for details of the full derivation. As per the previous Bayes rules, setting $\alpha = \beta = 0$ retrieves the estimator for $\mathcal{M}_1$; setting (α, β) to their estimates based on a relevant toxicity database retrieves the estimator for $\mathcal{M}_2$. Additionally, the estimators are defined in general for $n \geq 3$ for $\mathcal{M}_1$; and for $n \geq 2$ if $\alpha > 0$ for $\mathcal{M}_2$. The [LINEX] estimators are similar to those of Zieliński (2005) who derived a frequentist decision rule as opposed to the Bayes rule.

We cannot explicitly write down a formula to calculate κ_p^* in this instance. However, manipulation shows that the left-hand side of Equation 3.11 is a parabolic cylinder function from which look-up tables (Abramowitz and Stegun, 1972) and mathematical software can be used. Alternatively, one can adopt numerical integration and solve for the singular root, usually with high precision. The fact that tables can be produced of assessment shift-factors, even for this more complex loss function, increases the interest and applicability of this estimator within the regulatory arena for the intermediate quantitative tier of hazard assessment where parsimony is often desirable.

In Figure 3.5 we show three interpolated plots (corresponding to $\lambda = 1$, 2 and 3) of [LINEX] assessment shift-factors for $p = 5$ against sample size n under the model

$\mathcal{M}_1$. In addition, we plot the assessment shift-factors κ_5 for [SEL], [EFSA] and [AJ] ($\gamma = 0.05$, 0.50 and 0.95). The left panel shows the behaviour of the assortment of assessment shift-factors for $n \leq 100$; the right panel shows this behaviour over more likely obtainable sample sizes, namely $n \leq 15$.

3.6 Fixing λ

The [LINEX] estimators derive from a novel loss function which fits in neatly with current requirements of an intermediate quantitative tier of hazard assessment. However, guidance on fixing the asymmetrical control parameter is required. In this section we propose a single strategy to fix λ; however other methods could be produced.

It is not immediately apparent whether the risk manager should, or would adjust their value for λ whilst adjusting their choice for p. Consider the following (unlikely) thought experiment: a risk manager sets $p = 5$ and fixes $\lambda = \lambda_1$; later p is re-evaluated and set to $p = 10$. Should the risk manager change their value for λ to $\lambda_2 > \lambda_1$? Increasing p will increase the PAF of species — a decision which a risk manager would only be expected to make based on scientific reasoning. Therefore, a reduction in the level of minimum protection required — the arbitrarily defined threshold of the ecological community — is perceivable as a redefinition of the PNEC for substance, albeit a counterintuitive one. However, estimation error viewed as varying the quantity Δ/σ may still be viewed as having different specifications of loss at different definitions of the PNEC when considered in a probabilistic framework. Due to regulatory standards being determined by $p = 5$ (through the inclusion of a median HC_5 for EU REACH guidance) this issue is non-pertinent at the current time. Nevertheless, where recommendations have been made to use conservative HC_p estimates in scientific literature, consideration has never been raised regarding whether the degree of conservatism should be dependent on p. This is a point for discussion with risk managers, however it is not considered further here.

Recall now that we used scaled LINEX as opposed to standard LINEX. This was the modified LINEX loss function whereby loss was assigned to a standardised

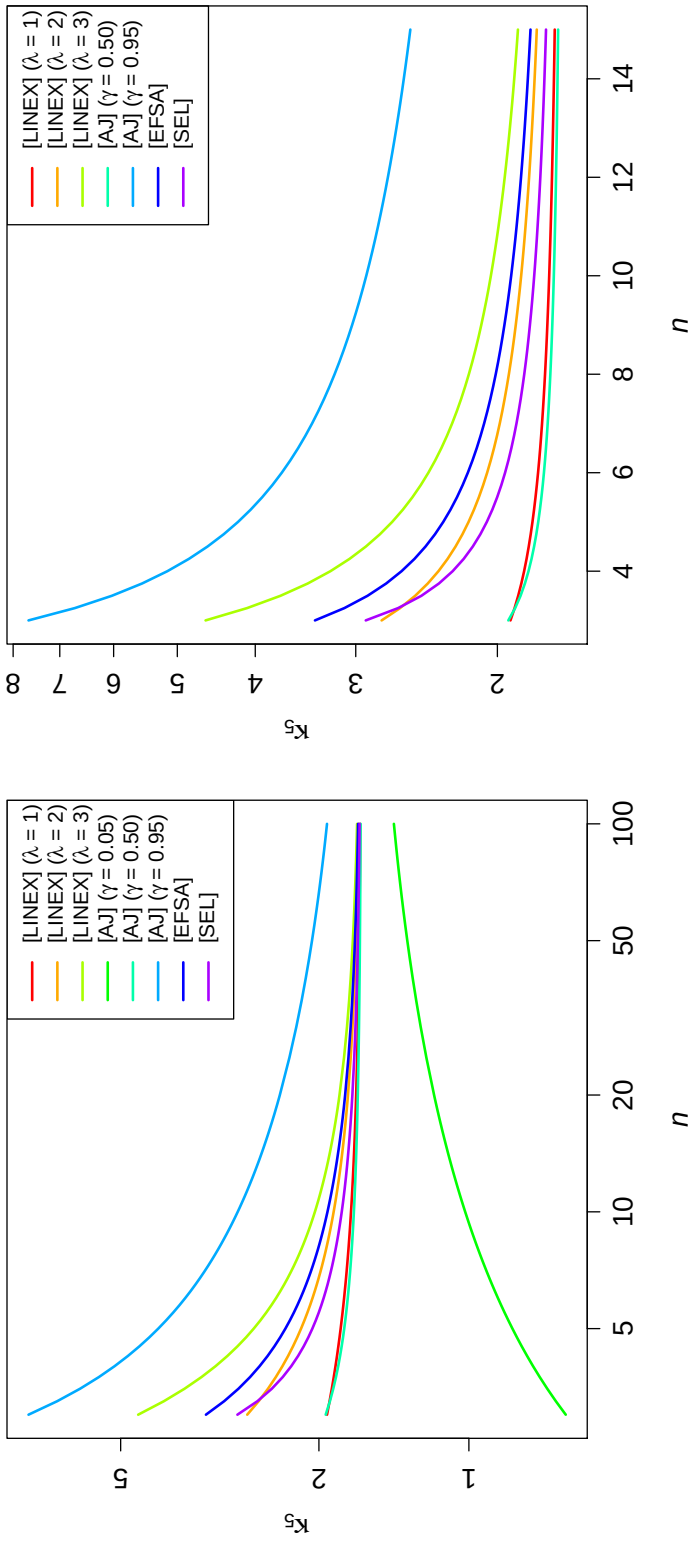


Figure 3.5: Interpolated plot of $\kappa_5(n)$ against n for $\mathcal{M}_1$. Left panel: $n \leq 100$; right panel: $n \leq 15$.

estimation error, in this problem Δ/σ. This allows for the loss function to be disentangled from knowledge of the SSD; thus allowing a suitable value of λ to be adopted, *a priori*. Essentially the specification of loss over log-toxicity — a property of the chemical — is replaced with the specification of loss over percentiles which as well as being more intuitive, is transferable and comparable in different chemical hazard assessments. This feature is likely to be influential for gaining acceptance within regulatory arena.

In accordance with Zieliński (2005), we define a measure of discrepancy between $\delta_p(\mathbf{Y})$ and $\psi_p(\theta)$ to be $t = (\delta_p(\mathbf{Y}) - \psi_p(\theta))/\sigma$, which implies that $\hat{\text{HC}}_p = \text{HC}_p \times 10^{t\sigma}$, where $\sigma > 0$. Our proposal to fix λ is to first fix the discrepancy at some value which warrants sufficient attention; $t = 2$ is one possible candidate, corresponding to the case where we overestimate the HC_p by 100 on the standardised scale. Now suppose we set this decision to have cost \$100, arbitrarily chosen to be that of cost-benefit for applying the substance. Then the following question could be posed to the risk manager: if $t = -2$, i.e. you underestimate by two standardised orders of magnitude, what *relative* cost, or equivalently percentage, would you associate with this situation? This is also approachable from the opposite direction by stating a base line for the case $t = -2$ and asking the risk manager how much 'worse' would the overestimation case be.

3.6.1 Example

Consider the classical cadmium toxicity dataset analysed in Van Straalen and Denneman (1989); Aldenberg and Jaworska (2000) and Hickey et al. (2009). The toxicity data is the NOEC tolerance values of 7 terrestrial soil species to cadmium, measured in micrograms of cadmium per milligram of soil (μg Cd/mg). This substance is common in rechargeable batteries and in plastic formulas. Table 3.1 summarises the data.

Following the procedure described above, the risk manager specifies a cost of underestimating by 100 units (on the standardised scale) to be \$100m ($m \in [0, 1]$).

Species	NOEC (μg Cd/mg)	$\log_{10}$(NOEC)
1	0.97	-0.01323
2	3.33	0.52244
3	3.63	0.55991
4	13.50	1.13033
5	13.80	1.13988
6	18.70	1.27184
7	154.00	2.18752
$\bar{y}$		0.97124
s		0.70276

Table 3.1: NOEC values for toxicity of cadmium (μg Cd/mg) of seven soil organisms.

Then we propose fixing λ to be the solution of

$$m = \frac{e^{-t\lambda'} + t\lambda' - 1}{e^{t\lambda'} - t\lambda' - 1}, \tag{3.12}$$

by solving for λ' with $t = 2$. Note that Equation 3.12 is equivalent to $L(-t)/L(t)$, which is a function of t. For SEL and GAL, $L(-t)/L(t)$ is conveniently constant. It is interesting to note that scaling SEL or GAL by σ leads to different decision rules in comparison to their standard counterparts; a possible motivating argument that they should be standardised (consult Section 3.7).

Consider the situation that three risk managers specify a value of λ. Risk manager A believes $m = 0.05$ so that the loss is consistent with the conservative indirect GAL prescription within scientific literature (e.g. Van Straalen and Denneman 1989; Wagner and Løkke 1991; Jagoe and Newman 1997; Newman et al. 2000) with $C_1/C_2 = 0.05$ (at $t = 2$ only); this corresponds to $\lambda = 2.13$. Risk manager B sets $m = 0.10$ in order to be less conservative; this corresponds to $\lambda = 1.67$. Risk manager C cautiously sets $m = 0.001$, corresponding to $\lambda = 4.49$. The loss function specification of each risk manager can then be used to derive [LINEX] estimators, as displayed in Table 3.2 for $\mathcal{M}_1$.

No estimates were derived for $\mathcal{M}_2$ because no database of toxicity data for additional substances tested with similarly related species was available. This is consis-

Risk Man.	κ_5^*	$\delta_5(\mathbf{Y})_{[\text{LINEX}]}$
A	2.02	0.3562
B	1.91	0.4256
C	2.89	0.0871

Table 3.2: HC_5 estimates and κ_5^* values for cadmium toxicity data.

tent with the statement we made earlier that $\mathcal{M}_2$ is not always an available model, especially for per-taxon risk assessments.

The starting point of $t = 2$ was a suggested point, however $t = 1, 3$ or any other t may be more suitable for elicitation; it is a policy based decision that a risk manager should take in collaboration with the risk assessor. It is important for the risk manager to remember that the loss function is constrained by its non-linear structure. As a result, setting λ at $t = t_1$ may not adequately reflect loss at $t = t_2$. Therefore, other specifications of loss might be considered to better inform the risk manager. Figure 3.6 describes the relation between m, λ and the discrepancy factor t through Equation 3.12. By virtue of the linear-exponential duality, at $t = 1$, risk managers A, B and C have assigned relative (to \$100) costs of \$8.06, \$11.06 and \$1.05 respectively; this is intuitively correct since this situation is of lower concern than the case $t = 2$. The relative fraction of costs at $t = -1$ to $t = 1$ are 0.24, 0.32 and 0.04 for risk manager A, B, and C's specification respectively. It may be prudent to assign an interval for λ by considering different starting values of t.

3.7 Scaled SEL & GAL

In Equations 3.2 and 3.6 we defined the SEL and GAL functions such that loss in each instance was placed on absolute difference: $\psi_p(\theta) - \delta_p(\mathbf{Y})$. The Bayes rule derived from the latter loss function was shown to be equivalent to the widely accepted [AJ] estimator.

LINEX loss was directly proposed based on placing loss on a discrepancy, $(\psi_p(\theta) - \delta_p(\mathbf{Y}))/\sigma$, which is arguably more intuitive. The scaled LINEX loss function was shown in Section 3.5 to approach a scaled SEL function as the non-linear asymmetry

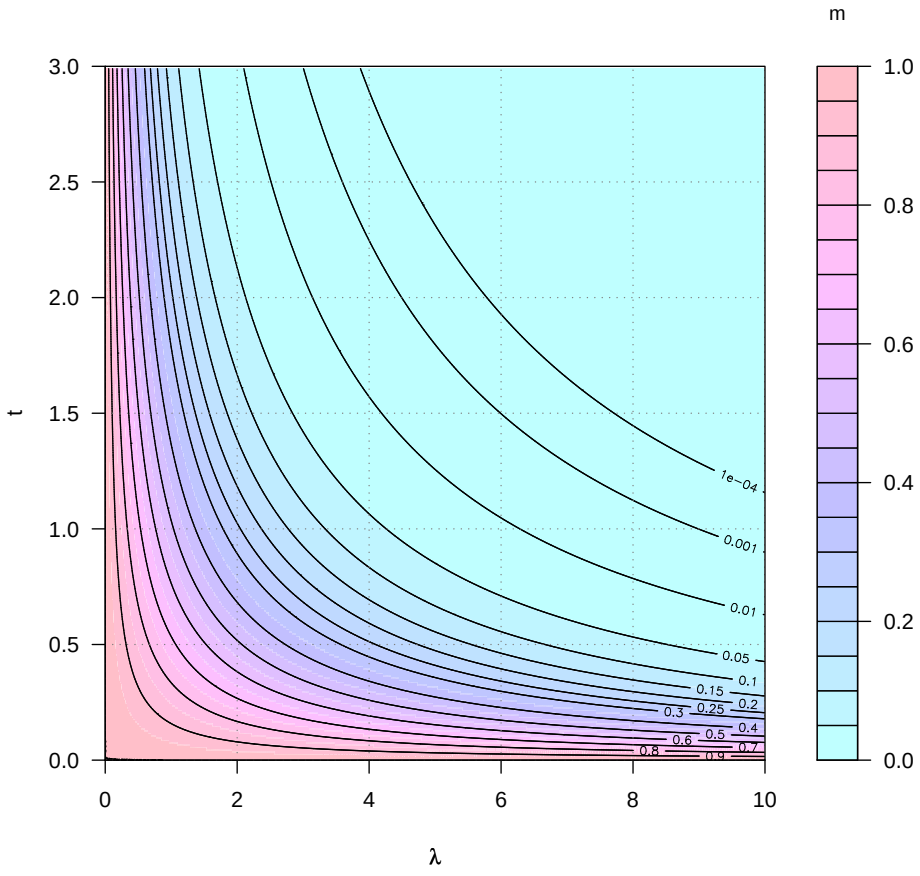


Figure 3.6: Contour plot relating λ (x-axis), t (y-axis) and m (contours). The example specifications are emphasised: $m = 0.001$ (C), $m = 0.05$ (A) and $m = 0.10$ (B).

control parameter $\lambda \to 0$. Here we provide details of the Bayes rules derived under scaled SEL and scaled GAL. To illustrate, we confine discussion to the context of behavioural model $\mathcal{M}_1$; extension to other models is straightforward. The relevant posterior distribution is described in Appendix A.1. Bayes rules for SEL and GAL with a star superscript indicate those based on scaled loss functions; those based on non-scaled loss functions were defined earlier using an asterisk superscript.

3.7.1 Scaled SEL

Scaled SEL is defined by a slight modification of standard SEL (Equation 3.2) to be

$$L(\psi_p(\theta), \delta_p(\mathbf{Y}); \sigma) = \left[\frac{\psi_p(\theta) - \delta_p(\mathbf{Y})}{\sigma}\right]^2. \tag{3.13}$$

The Bayes rule for this non-standard loss function is determined as

$$\delta_p^\star(\mathbf{Y})_{[\text{SEL}]} = \bar{y} - \kappa_p^\star(n)s, \tag{3.14}$$

where

$$\kappa_p^\star(n) = K_p\sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}.$$

A full derivation of this estimator is given in Appendix C.5. The Bayes rule follows the same canonical form as for all other estimators discussed thus far. The ratio of the assessment shift-factor for the standard [SEL] Bayes rule (Equation 3.4) to the assessment-shift factor above is $(n-1)/(n-2)$. For $n = 3$ (the minimum sample size that both estimators are defined for) the ratio is maximised at 2; decreasing uniformly with limit 1 (as $n \to \infty$). Hence, the (standard) [SEL] Bayes rule leads to a relatively more conservative decision that the (scaled) [SEL] Bayes rule.

The most startling observation is that $\kappa_p^\star(n)$ increases as n increases when $p < 50$, with a numerically identified limit of K_p (see Figure 3.7). The assessment shift-factor for the standard [SEL] Bayes rule increases as n decreases (when $p < 50$) with the same limit. However, this was not unexpected. Numerically evaluating the [LINEX] Bayes rule for λ close to zero (recalling that the scaled [SEL] estimator is retrieved for $\lambda \to 0$) also confirms this property.

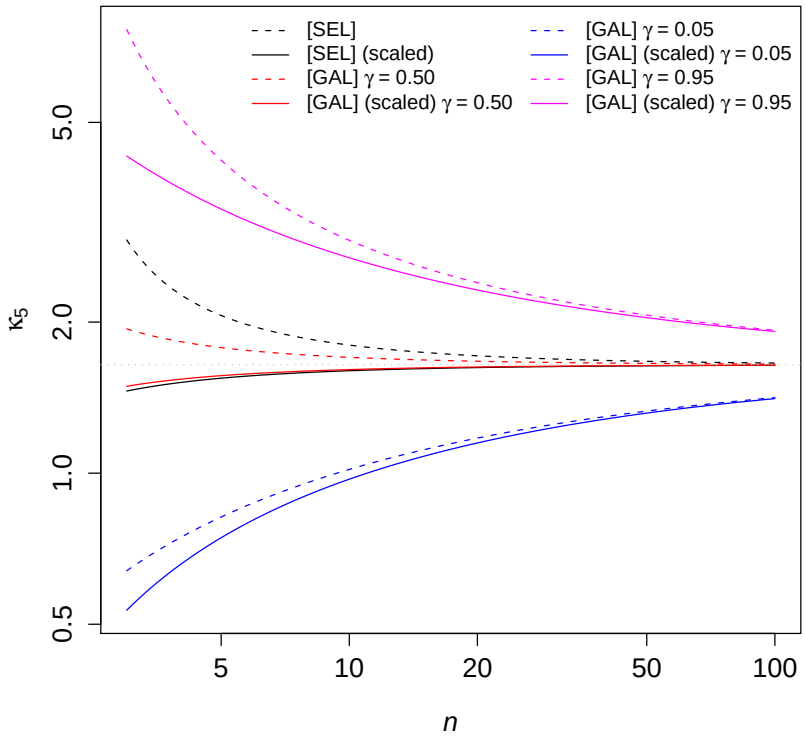


Figure 3.7: Assessment shift-factors κ_5 plotted against n for standard (solid) and scaled (dashed) [SEL] (black) and [GAL] (red) Bayes rules. Grey dashed line corresponds to K_5.

3.7.2 Scaled GAL

Scaled GAL is also a straightforward extension of standard GAL (Equation 3.6), defined as

$$L\left(\psi_p(\theta), \delta_p(\mathbf{Y}); \sigma\right) = \begin{cases} C_1 \left[\frac{\psi_p(\theta) - \delta_p(\mathbf{Y})}{\sigma}\right] & \text{if} \quad \psi_p(\theta) \geq \delta_p(\mathbf{Y}) \\ C_2 \left[\frac{\delta_p(\mathbf{Y}) - \psi_p(\theta)}{\sigma}\right] & \text{if} \quad \psi_p(\theta) < \delta_p(\mathbf{Y}) \end{cases} \tag{3.15}$$

Notice that the indicators remain unaffected since $\sigma > 0$. The Bayes rule for this non-standard loss function is

$$\delta_p^\star(\mathbf{Y} \,|\, C_1, C_2)_{[\text{GAL}]} = \bar{y} - \kappa_p^\star(n, C_1, C_2)s, \tag{3.16}$$

where

$$\kappa_p^\star(n, C_1, C_2) = \frac{\sqrt{n-1}}{n} F^{-1}_{T_{n,K_p\sqrt{n}}}\left(\frac{C_2}{C_1 + C_2}\right).$$

A full derivation of this estimator is given in Appendix C.6. As for the scaled [SEL] Bayes rule, the scaled [GAL] Bayes rule maintains the standard canonical structure. Moreover, the estimator is similar to the standard [GAL] Bayes rule (Equation 3.7) and hence the [AJ] estimator. The original multiplier decreases from $n^{-1/2}$ to $\sqrt{n-1}/n$, which is in addition to the gain of one degree of freedom for the non-central t-distribution quantile function.

For all intents and purposes it is sufficient for us to define $\gamma = C_2/(C_1 + C_2)$; this was earlier chosen so that the [AJ] estimator and [GAL] Bayes rule coincide. Comparing the assessment shift-factors of standard and scaled [GAL] Bayes rules for $\gamma = 0.50$ implied the same disparity as determined for the comparison of the standard and scaled [SEL] Bayes rule. For $p = 5$ and $\gamma = 0.05$ (0.95) the scaled [GAL] assessment shift-factors increases (decreases) as n increases (although at different rates) which is consistent with the standard [GAL] estimators.

In Figure 3.7 we plot the assessment shift-factors for $p = 5$ against sample size. This is done for: (a) the [SEL] (standard and scaled) and (b) the [GAL] (standard and scaled) with $\gamma = 0.50$. The absolute magnitude of the rate of change as n increases is noticeably larger for the Bayes rules derived from the standard loss functions compared to those derived for the scaled loss functions.

Strong objection to the use of $\delta_p^\star(\mathbf{Y})_{[\text{SEL}]}$ and $\delta_p^\star(\mathbf{Y} \mid C_1, C_1)_{[\text{GAL}]}$ (where $C_2 = C_1$ implies $\gamma = 0.50$) is anticipated because of the monotonic property. This is despite the standard [AJ] estimator (with $\gamma = 0.50$) being acceptable under REACH guidance (ECHA, 2008a), which was shown to coincide with $\delta_p^\star(\mathbf{Y} \mid C_1, C_1)_{[\text{GAL}]}$ (subject to prior distribution). We would hypothesise that regulators in this instance would advocate $\gamma > 0.50$ such that $\mathrm{d}\kappa_5^\star(n)/\mathrm{d}n < 0$. However, there is no reason, *a priori*, why a symmetric loss function would lead to a decision rule where $\kappa_p^\star > K_p$. This highlights clearly that if a risk manager is interested in conservativeness, then they should specify their loss function as accordingly. Although we can provide no intuitive explanation as to why, say, $\kappa_p^\star(n, \gamma = 0.50) \nearrow K_p$ as $n \to \infty$, it is clear that outcomes depend on different loss functions which indicate the need for risk managers

to carefully consider their choices. If risk managers considered $\delta_p^\star(\mathbf{Y} \mid \gamma = 0.50)_{[\text{GAL}]}$ as appropriate, then it may support the application of the [M] estimator (see Section 2.6.1), i.e. fixing $\kappa_p(n) = K_p$ which leads to a more conservative decision rule relative to the former.

We do not pursue this further in light of the [AJ] estimator, for intents and purposes, being the default estimator for risk assessment.

3.8 Known Variance & Homogeneity

In this section we consider the problem of HC_p estimation in the context of behavioural models $\mathcal{M}_3$ and $\mathcal{M}_4$. These two models overlap to a certain degree and for this reason we present discussions on them together. Each model can be succinctly described as the scenario where μ is unknown and varies between substances, except either σ is unknown but homogeneous among a population of substances ($\mathcal{M}_3$), or σ is known for each substance ($\mathcal{M}_4$). The premise of $\mathcal{M}_3$ and $\mathcal{M}_4$ was introduced in Sections 2.6.2, 2.6.4 and 2.7. Like $\mathcal{M}_2$, the models have been motivated by the desire to be able to perform probabilistically refined ERA when the cardinality of data available is insufficient to proceed using accepted models, e.g. $\mathcal{M}_1$, due to unacceptable levels of uncertainty. As noted by Luttik and Aldenberg (1997) and Aldenberg and Luttik (2002), it is often the case that strictly deterministic methods are applied in favour of probabilistic methods when sample sizes are deemed small.

3.8.1 $\mathcal{M}_3$

The basis of $\mathcal{M}_3$ was first proposed in EFSA (2005) and is arguably a generalisation of the model Luttik and Aldenberg (1997) proposed in deriving the [LA] estimators. Basically, the model is that σ is unknown for $\mathcal{S}$, but believed to be homogeneous among a population of additional substance SSDs, of which we have an available toxicity database for some. This allows for the variance to be estimated with higher precision when the toxicity data sample size for $\mathcal{S}$ is small, as per $\mathcal{M}_2$. A major distinction of $\mathcal{M}_3$ over that of $\mathcal{M}_4$ is with respect to the inclusion of sampling uncertainty in estimating σ. EFSA (2005) describe the foundations of $\mathcal{M}_3$ to be an

improvement over the latter practice, subject to the assumption of homogeneity; we provide details of $\mathcal{M}_4$ in Section 3.8.2.

We describe the estimators of $\psi_p(\theta)$ below for the collection of decision rule bases considered thus far; the [EFSA] estimator is already given by Equation 2.13, Section 2.6.4. A further exploratory based quantification of the difference in uncertainty between the [GAL] estimator and the [LA] estimator in the context of $\mathcal{M}_3$ is made towards the end of this section.

Following the discussion earlier, the standard model and natural non-informative prior distribution for $\mathcal{M}_3$ is discussed in Section 2.6.4, and the corresponding posterior distribution is described in Appendix A.1.

[SEL] Estimators

The [SEL] Bayes rule estimator was shown earlier to be the posterior expectation of $\psi_p(\theta)$. Hence, the Bayes rule is straightforwardly calculated in a similar manner to the derivation for $\mathcal{M}_2$ (consult Appendix C.1) to be

$$\delta_p^*(\mathbf{Y}, \mathbf{Y}_G)_{[\text{SEL}]} = \bar{y} - \kappa_p^*(n, \varsigma^*)s_p^*,$$

where

$$\kappa_p^*(n, \varsigma^*) = K_p\sqrt{\frac{\varsigma^*}{2}}\frac{\Gamma(\frac{\varsigma^*-1}{2})}{\Gamma(\frac{\varsigma^*}{2})},$$

s_p^* is defined by Equation 2.17, $\varsigma^* = \varsigma + (n-1)$ and $\varsigma = \sum_{i\in\mathcal{G}}(n_i - 1)$.

This estimator is defined, in general, when $\varsigma^* \geq 2$; consequently, the estimator is valid for $n = 1$ subject to the condition $\varsigma \geq 2$.

[GAL] Estimators

In Appendix C.3, we showed that the [GAL] Bayes rule — the decision rule which minimises the posterior expected loss of Equation 3.6 — is equal to the $100C_1/(C_1 + C_2)$-th percentile of the posterior distribution of $\psi_p(\theta)$; the loss parameters (C_1, C_2) are described in Section 3.4. Hence, based on the posterior distribution described

in Appendix A.1, one obtains the Bayes rule as

$$\delta_p^*(\mathbf{Y}, \mathbf{Y}_G \,|\, C_1, C_2)_{[\text{GAL}]} = \bar{y} - \kappa_p^*(n, \varsigma^*, C_1, C_2) s_p^*,$$

where

$$\kappa_p^*(n, \varsigma^*, C_1, C_2) = \frac{1}{\sqrt{n}} F^{-1}_{T_{\varsigma^*, K_p\sqrt{n}}}\left(\frac{C_2}{C_1 + C_2}\right).$$

For all intents and purposes we will remove the dependence of this estimator on the arguments (C_1, C_2) by substituting a control parameter γ which, as per previous behavioural models, is analogous to the role of the [AJ] uncertainty parameter. In light of the earlier connection between the [AJ] and [GAL] estimators, we denote this estimator as $\delta_p^{(\gamma)}(\mathbf{Y}, \mathbf{Y}_G)_{[\text{AJ}]}$ also from here onwards whilst working in the context of $\mathcal{M}_3$. This estimator is in general defined for $n \geq 1$ subject to the proviso that $\varsigma \geq 1$, i.e. where at least one substance toxicity dataset in the database satisfies $n_i \geq 2$.

[LINEX] Estimators

The scaled [LINEX] Bayes rule estimator can be determined in a similar fashion to that discussed in Section 3.5. Hence, the solution of Equation 3.1 for $\lambda \neq 0$ is

$$\delta_p^*(\mathbf{Y}, \mathbf{Y}_G \,|\, \lambda)_{[\text{GAL}]} = \bar{y} - \kappa_p^*(n, \varsigma^*, \lambda) s_p^*,$$

where $\kappa_p^*(n, \varsigma^*, \lambda)$ is the unique solution to

$$\int_0^\infty t^{(\varsigma^*-1)/2} \exp\left\{-\lambda\kappa_p^*\sqrt{t} - \left(\frac{\varsigma^*}{2}\right) t\right\} dt = \Gamma\left(\frac{\varsigma^*+1}{2}\right)\left[\frac{\varsigma^*}{2}\right]^{-\frac{\varsigma^*+1}{2}} \exp\left\{-\lambda\left[K_p + \frac{\lambda}{2n}\right]\right\}$$

for κ_p^*. A very brief outline on how to extend the earlier Bayes rule derivation of $\mathcal{M}_1$ and $\mathcal{M}_2$ for the model $\mathcal{M}_3$ is provided in Appendix C.4. Additionally, this estimator is well defined on the same set of sample sizes for which the [SEL] estimator is defined.

Comparing Estimators

The reliability of the homogeneity assumption is highly suspect and we have yet to explore this. However, it is likely that basis of $\mathcal{M}_3$ — which is arguably an extension of $\mathcal{M}_4$ — is to provide practical probabilistic estimators for hazard assessments where the sample of toxicity data is substantially small. We consider these estimators over a similar range of n as was explored by Luttik and Aldenberg (1997): $n \leq 4$. The decision rules are all of canonical form $\delta_p^*(\mathbf{Y}, \mathbf{Y}_G) = \bar{y} - \kappa_p^* s_p^*$. Thus it is sufficient to compare the assessment shift-factors κ_p^* only in order to analyse relative conservativeness. Figure 3.8 plots $\kappa_5^*(n, \varsigma^*, 5)$ for $n = 1, \ldots, 4$ against the total sum of additional degrees of freedom ς. In the interest of clarity the plots have been cropped which precludes the inclusion of the [AJ] upper 95% credible limit ($\gamma = 0.05$) estimators. The separation of ς^* into ς and n is a requirement of analysis due to the latter being independently influential in the estimation of μ.

It is observed that the median [AJ], [LINEX] ($\lambda = 1$) and to a lesser degree, the [EFSA] assessment shift-factors behave like a constant as ς increases for small n. The more conservative estimators, i.e. [LINEX] ($\lambda = 5, 3$) and [AJ] ($\gamma = 0.95$), change relative conservativeness as a function of n as well ς. Although deducible from the algebraic structure of the estimators, the estimators converge faster in n than ς, especially for conservative choices. Consequently, benefits in testing additional species for $\mathcal{S}$ will likely outweigh the benefits of using a toxicity database; however this suggestion discounts the clinical and ethical cost involved for conducting additional laboratory tests. On the other hand, available toxicity databases will likely either be free or relatively less expensive to obtain. This should be clearly communicated to all risk managers if such estimators were ever to be considered in a regulatory context.

3.8.2 $\mathcal{M}_4$

The behavioural model $\mathcal{M}_4$ was initially proposed by Aldenberg and Luttik (2002) (cf. the [AL] estimator; Section 2.6.2) and is also the foundation of the (log-normal SSD) [LA] estimator proposed by Luttik and Aldenberg (1997). The model states

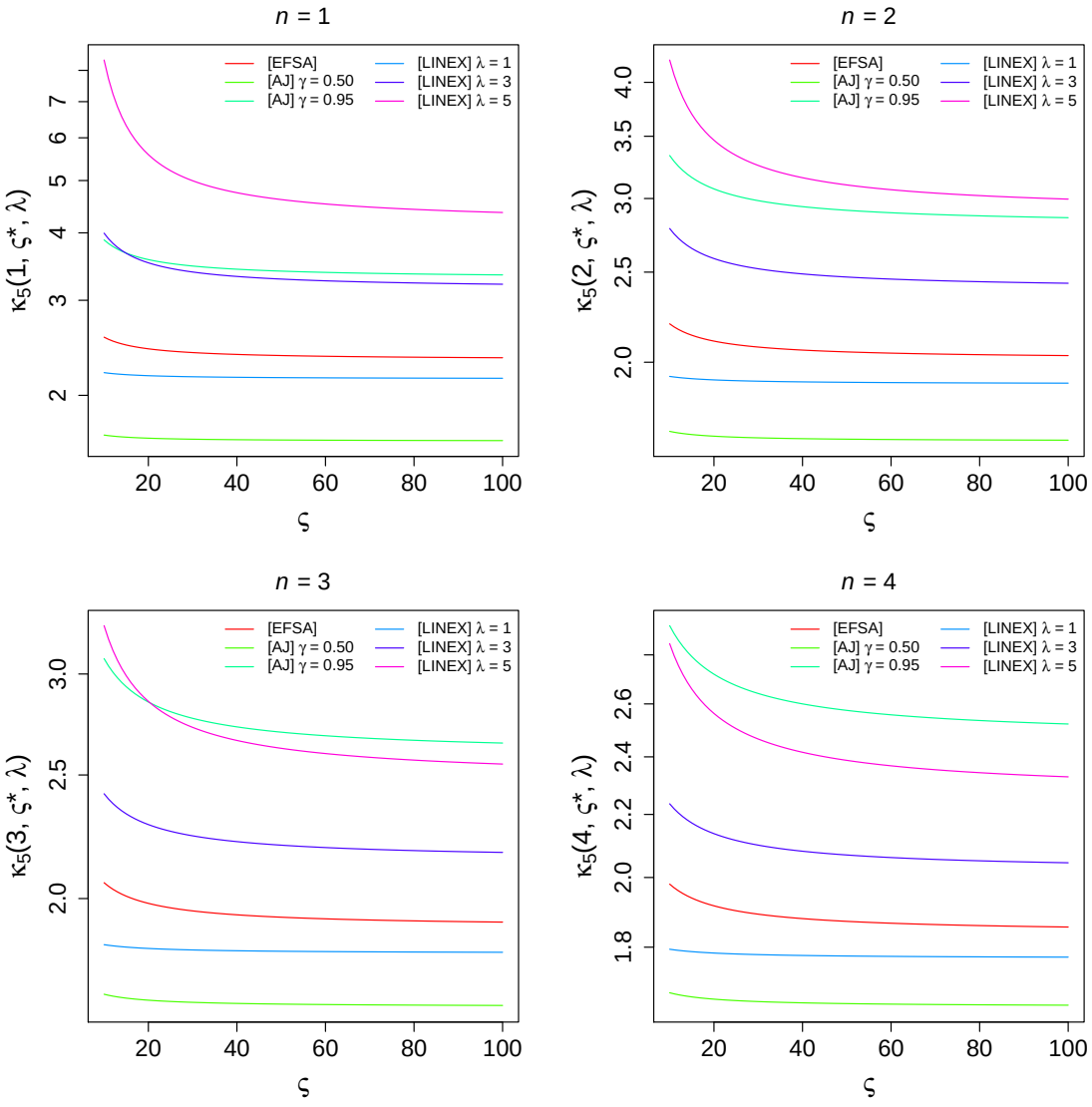


Figure 3.8: $\kappa_5(n, \varsigma^*, p = 5)$ versus ς for $n \leq 4$ ($\mathcal{M}_3$).

that the log-SSD variance σ for a substance $\mathcal{S}$ is known, or alternatively one estimates σ independently and neglects the uncertainty. In particular, the [LA] estimator by definition is a special case of the [AL] estimator obtained by fixing $\sigma = s_p$ (Equation 2.3) — estimated using a toxicity database for a collection of similar substances $\mathcal{G}$ presumed reasonably large as to warrant neglecting uncertainty. In fact the original formulation of the [LA] estimator assumed that the database used to estimate s_p satisfied $n_i \geq 4$ for each substance dataset; this restriction is lifted for the remainder of this research. Furthermore, estimators from here onwards will be interpreted as [AL] estimators, unless we apply the plug-in value of $\sigma = s_p$.

The concept of conservative HC_p estimation using the lower 5% credible limit is still valid. However, the defensibility of these estimators relative to those based on $\mathcal{M}_1$–$\mathcal{M}_3$, is unfathomable. Nonetheless, we explore these estimators by deriving them from the different risk perspectives described and attempt to determine whether they might constitute a pragmatic tool for quantitative tier hazard assessment.

In order to study the [AL] estimators in a similar manner to the [AJ] estimators, we begin by generalising their definition to be

$$\delta_p^{(q)}(\mathbf{Y} \mid \sigma)_{[\text{AL}]} = \bar{y} - \left(K_p + \frac{1}{\sqrt{n}} K_{100(1-q)} \right) \sigma,$$

where it is assumed that σ is known. This structure is consistent with the previous estimators discussed, such that setting $q = 0.95, 0.50, 0.05$ admits the $100q$-th one-sided underestimate confidence limit (when derived in the frequentist framework originally employed). Hence, q is analogous to the parameter γ in the [AJ] estimator with recommendations to set $q = 0.95$ (Luttik and Aldenberg, 1997). There are multiple options for direct specification of σ, some of which use a database of additional toxicity data for similar substances $\mathcal{G}$, for example: (i) $\sigma = s_p$; (ii) $\sigma = s$; (iii) $\sigma = s_p^*$. Notice that cases (i) and (ii) imply that $\delta_p^{(q)}(\mathbf{Y})_{[\text{AL}]} = \delta_p^{(q)}(\mathbf{Y}, \mathbf{Y}_G)_{[\text{LA}]}$ and $\delta_p^{(0.50)}(\mathbf{Y})_{[\text{AL}]} = \delta_p(\mathbf{Y})_{[\text{M}]}$ respectively. Case (iii) would be counterintuitive in hindsight of its derivation; this is because the small sample size would be accounted for in the confidence interval for μ, but not σ — which we now invoke is only dependent

on the toxicity data for $\mathcal{S}$.

Estimation of the HC_p has already been argued for as a natural problem from within the Bayesian decision theoretic framework. Since our operational procedural stance remains the same, we propose the following prior distribution: $p(\mu) \propto 1$ for $\mu \in \mathbb{R}$. The prior distribution is found to be the Jeffreys prior for the unknown mean of a Gaussian quantity with known variance (Berger, 1985, p. 88), which is therefore consistent with the specifications for behavioural models $\mathcal{M}_1$–$\mathcal{M}_3$ where it was assumed the prior distribution was in fact the product of independent Jeffreys' priors for μ and σ. The posterior distribution is straightforwardly determined to be $\mu \,|\, \mathbf{Y}, \sigma^2 \sim N(\bar{y}, \sigma^2/n)$ for $\mu \in \mathbb{R}$. Hence, the posterior distribution is proper unlike the associated prior distribution, thus enabling one to minimise the posterior expected loss. The following is a description of the estimators based on foundations discussed thus far.

[GAL] and [SEL] Estimators

From the posterior distribution described above it is simple to deduce that $\psi_p(\theta) \,|\, \mathbf{Y}, \sigma \sim N(\bar{y} - K_p\sigma, \sigma^2/n)$ which means that the [GAL] Bayes rule estimator is equivalent to

$$\delta_p^*(\mathbf{Y} \,|\, \sigma, C_1, C_2)_{[\mathrm{GAL}]} = \bar{y} - \left(K_p + \frac{1}{\sqrt{n}} \Phi^{-1} \left(\frac{C_2}{C_1 + C_2} \right) \right) \sigma.$$

Similarly to before setting $q = C_2/(C_1 + C_2)$ demonstrates that the [AL] estimator is a [GAL] Bayes rule conditional upon the non-informative prior distribution suggested. Consequently, the frequentist and Bayesian estimators have coverage matching properties. Unlike the earlier connection between the [GAL] and [AJ] estimator, the [AL] estimator was not originally discussed by Aldenberg and Luttik (2002) within the Bayesian paradigm. Moreover, the Bayes rule derived under absolute loss (i.e. $C_1 = C_2$) is identical to the [SEL] Bayes rule, since the latter is just $\mathbb{E}^{\theta \,|\, \mathbf{Y}, \sigma^2}[\psi_p(\theta)] = \bar{y} - K_p\sigma$ by definition.

[LINEX] Estimators

Minimisation of the posterior expected loss is relatively straightforward compared to the former estimators. In fact, one can show that the Bayes rule derivation effectively reduces to consideration of the moment generating function for a normal random quantity (Zellner, 1986). This is because since σ is known one can appeal to the standard LINEX function via a re-parameterisation of the loss function by scaling λ accordingly. Hence, the scaled [LINEX] Bayes rule is determined as

$$\delta_p^*(\mathbf{Y} \mid \sigma, \lambda)_{[\text{LINEX}]} = \bar{y} - \left(K_p + \frac{\lambda}{2n}\right)\sigma.$$

We provide some brief details on how to derive the Bayes rule which extends from the derivation of $\mathcal{M}_1$ and $\mathcal{M}_2$ in Appendix C.4.

[EFSA] Estimators

Recalling from Section 2.6.4 that the [EFSA] decision rules defined by Equation 2.9 are effectively posterior predictive estimators, it is straightforward to confirm that the estimators are given by

$$\delta_p(\mathbf{Y} \mid \sigma)_{[\text{EFSA}]} = \bar{y} - K_p\xi\sigma.$$

Derivation is trivial upon replacement of the probability density function for σ^2 in Appendix A.2 by the known value with point mass.

Comparing Estimators

It is interesting to consider how the [GAL] ([AL]) Bayes rules (upon substituting the parameters (C_1, C_2) by q accordingly) — a simplification of a scientifically established estimator — relate to the scaled [LINEX] Bayes rules — those advocated here. We begin by defining

$$\begin{aligned} k_p'(n, q) &= K_p + \frac{1}{\sqrt{n}}K_{100(1-q)}; \\ k_p''(n, \lambda) &= K_p + \frac{\lambda}{2n}, \end{aligned}$$

such that an estimator will take the form of $\delta_p^*(\mathbf{Y} \mid \sigma) = \bar{y} - k_p(\cdot)\sigma$; with k' yielding [LINEX] and k'' yielding [AL]. We conclude that $k'_p = k''_p$, i.e. [LINEX] and [AL] decision rules coincide under $\mathcal{M}_4$, when

$$\lambda = 2\sqrt{n}K_{100(1-q)}. \tag{3.17}$$

Although we focus our discussion to the Bayesian perspective it can be shown via the manipulation of results provided by Zellner (1986), that Equation 3.17 is pivotal in determining the admissibility of decision rules, such that replacement of equality by a greater-than comparative ($>$) in Equation 3.17 implies that [LINEX] dominates [GAL]. A consequence of this which is useful even from the Bayesian perspective, is that [LINEX] will always dominate the median [GAL] estimate when $\lambda > 0$. This was observed also in the context of $\mathcal{M}_1$ and $\mathcal{M}_2$. No simple relation like Equation 3.17 can be determined for the latter two behavioural models due to the intractability of κ_p^* for [LINEX] estimators (cf. Equation 3.11) and for the non-central t-distribution quantile function, thus precluding an analogous analytical interpretation.

Although (scaled-) LINEX loss functions are not well defined at $\lambda = 0$, by considering $\lim_{\lambda \to 0} k''_p(n, \lambda)$, one can show that the [LINEX] estimators coincide with the [AL] and [SEL] estimators when $q = 0.50$. This is intuitively correct since the LINEX loss function approaches the SEL function in the limit, up to a positive scaling constant dependent on σ, which is known by virtue of $\mathcal{M}_4$.

Consider the value λ must take if [LINEX] and [AL] rules should coincide; as either n or q increases, λ must increase. The relation given by Equation 3.17 is intuitively correct since as q increases and λ increases, both the GAL and LINEX loss functions become more asymmetrically conservative. The former relation derives from the posterior distribution of μ; as the precision increases, k''_p is down-weighted at a rate which is faster by a factor of $n^{-1/2}$.

A final noteworthy comment is that if we had derived the standard [LINEX] (i.e. placing the loss on difference in log-toxicity) Bayes rule, parameterised by asymmetry parameter λ' (independent of σ), then this would coincide with [AL] (as

derived above) if and only if $\lambda'/\sigma = 2\sqrt{n}K_{100(1-q)}$. Thus indicating that the form of the decision rule is identical to above with λ' scaled accordingly.

3.8.3 $\mathcal{M}_4$ in Practice

In this example we re-examine a general risk assessment methodology for chemical risk assessment to birds (and mammals) proposed by Aldenberg and Luttik (2002) for small sample sizes ($n \leq 5$). The methodology was fundamentally based on $\mathcal{M}_4$, but yielded fixed assessment factors for general application rather than assessment shift-factors.

The estimator for σ is based on two large toxicity databases for pesticides (satisfying $n_i \geq 4$): (1) for birds only ($N = 55$); and (2) for mammals only ($N = 69$); a description of the databases in provided in Luttik and Aldenberg (1997) and references therein. Pooled standard deviations of the log-toxicity values (measured in $\log_{10}$ mg/Kg) were reported as well as 5-th and 95-th percentile estimates (in parenthesis) as: $s_p = 0.465$ $(0.197, 0.752)$ for birds; $s_p = 0.36$ $(0.095, 0.768)$ for mammals. The percentile estimates are determined from the empirical cumulative distribution function of the standard deviation estimates; the upper 95-th percentile estimate might act as a substitute for a 'worse-case' estimate of σ.

The authors recommended that the [AL] estimators be based on $\sigma = 0.760$ for conservative estimation. This decision was arrived at by taking the mean of the upper 95-th percentile estimates for birds and mammals. The [LA] estimators, according to our definition, require $\sigma = s_p$. Since homogeneity for birds and mammals is not scientifically justifiable, *a priori*, we restrict the [LA] estimator discussion to bird risk assessment only; it is straightforward to modify the assessment for mammals.

In Table 3.3 we provide the fixed assessment factors (on the standard scale) that would be applied divisibly to the geometric mean, $10^{\bar{y}}$, of the toxicity data for the current substance $\mathcal{S}$ under assessment. This is done for $q = 0.95$, 0.50, 0.05, to yield an acute HC_5 estimate, and are listed for both [AL] (birds and mammals) and [LA] estimators (birds). In addition, we also determine the [LINEX] parameter λ^* that would be required to identically estimate the lower [AL] HC_5 based on the Bayes

	$\sigma = 0.760$ [AL]			$\sigma = 0.465$ [LA]			
n	$q = 0.95$	$q = 0.50$	$q = 0.05$	$q = 0.95$	$q = 0.50$	$q = 0.05$	λ^*
1	316.4	17.8	1.0	33.9	5.8	1.0	3.29
2	136.2	17.8	2.3	20.2	5.8	1.7	4.65
3	93.7	17.8	3.4	16.1	5.8	2.1	5.70
4	75.0	17.8	4.3	14.0	5.8	2.4	6.58
5	64.4	17.8	4.9	12.8	5.8	2.5	7.36

Table 3.3: Fixed AFs for bird and mammal HC_5 (mg/kg) extrapolation; left: [AL] conservative σ value; right: [LA] s_p for bird database. [LINEX] λ^* estimators coincide with lower [AL] estimators.

rule determined under scaled LINEX loss (described above).

Exploring the assessment factors in Table 3.3 we notice that for $n = 1$, the median ($q = 0.50$) and lower (confidence limit) estimate ($q = 0.95$) are: 17.8 and 316.4 respectively for $\sigma = 0.760$, and 5.8 and 33.9 for $\sigma = s_p$ (birds). The current guidelines in ECHA (2008a, pp. 44–46) stipulate that the assessment factor be AF = 3000 (Section 2.5) for a single (i.e. $n = 1$) acute bird toxicity value. Moreover, they implicitly suggest that included in this factor is an acute-to-chronic ratio of ACR = 100. By assuming that $\text{AF} = \text{AF}_{\text{spec}} \times \text{AF}_{\text{ACR}}$, it suggests that the estimates here of the required AF_{spec} are larger than current guidelines by factors of: 4.8 and 10.5 ($\sigma = 0.760$) and 1.1 and 0.2 respectively. Hence, the [AL] conservative estimators are currently more conservative than current guidelines, but the lower [LA] estimate is very close to guidelines. Therefore, this might provide empirical support for the use of lower [LA] estimators (at least in the case of birds) when the sample size is small for $\mathcal{S}$, which includes the option of assessment factor reduction for increasing sample sizes as not to be overly conservative. Additionally, from a protectionist standpoint, the [AL] estimators may be tenable.

The relatively large values of λ^* which lead to the lower [AL] and [LINEX] estimators coinciding are not particularly surprising in respect of the results in Section 3.6.1, since the 'worst-case' σ choice is arbitrarily conservative. Figure 3.9 displays the associated scaled LINEX loss functions corresponding to the λ^* values.

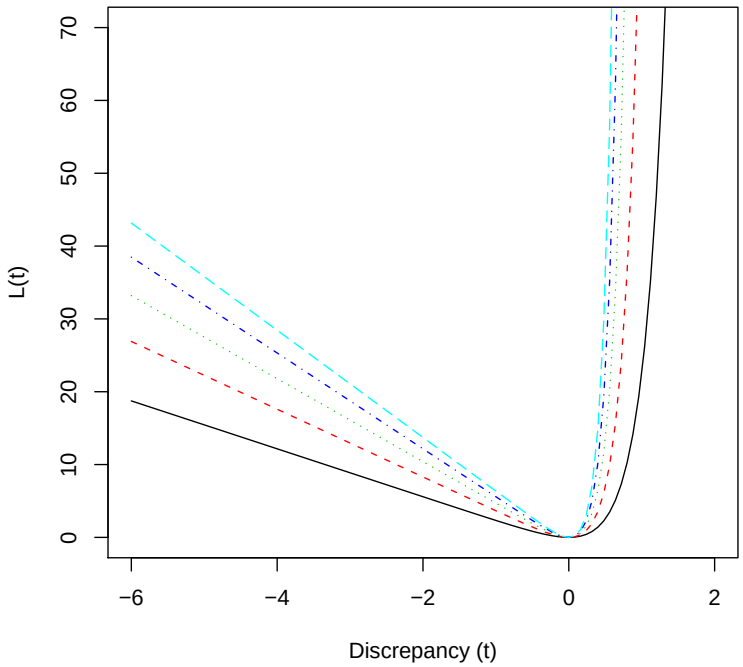


Figure 3.9: Scaled LINEX loss function with: $\lambda^* = 3.29$ (black/solid); $\lambda^* = 4.65$ (red/dashed); $\lambda^* = 5.70$ (green/dotted); $\lambda^* = 6.58$ (blue/dot-dash); $\lambda^* = 7.36$ (light-blue/long-dash). Results in $\delta_5^*(\mathbf{Y} \mid \sigma, \lambda^*)_{[\text{LINEX}]} = \delta_5^{(0.95)}(\mathbf{Y} \mid \sigma)_{[\text{AL}]}$ for $n = 1, \ldots, 5$ under $\mathcal{M}_4$.

3.8.4 Full vs. Partial Uncertainty

In order to assess the absolute consequence of accounting for uncertainty in σ, we compare the [AJ] estimator under $\mathcal{M}_3$ with the [LA] estimator under $\mathcal{M}_4$. In particular, we consider this for the contentious analysis of $n = 1$, which has application in bird and mammalian ecotoxicological risk assessment. When $n = 1$, $s_p^* = s_p$ (and $\varsigma^* = \varsigma$), consequently the assessment shift-factors are directly comparable.

Define the difference in estimators $d(p, \gamma, \varsigma)$, evaluated at a common percentile of the respective posterior distribution for $\psi_p(\theta)$, to be

$$\begin{aligned} d(p,\gamma,\varsigma) &\triangleq \kappa_p(1,q)_{[\mathrm{LA}]}\big|_{q=\gamma} - \kappa_p(1,\varsigma,\gamma)_{[\mathrm{AJ}]} \\ &= \left[K_p + K_{100(1-\gamma)} - F^{-1}_{T_{\varsigma,K_p}}(\gamma)\right]. \end{aligned}$$

It is straightforward to confirm that $\lim_{\varsigma\to\infty} d(p, \gamma, \varsigma) = 0$. When $p = 5$, $d(5, \gamma, \varsigma)$ is negative for most of the γ domain implying that incorporation of uncertainty about s_p leads to the [AJ] estimator being relatively more conservative. However, the difference is likely to be negligible; for $\varsigma = 100$, $d(5, 0.50, 100) = -0.004$ and $d(5, 0.95, 100) = -0.05$ which means an additional extrapolation (on the original toxicity scale) of assessment factors 1.01^{s_p} and 1.13^{s_p}, respectively is required to translate the [LA] estimator to the [AJ] estimator.

This comparison is subject to the modelling assumption governing $\mathcal{M}_3$ and accessibility to a relevant and sufficiently large toxicity database; for ς small the difference is substantially greater.

3.8.5 Distinguishing $\mathcal{M}_3$ & $\mathcal{M}_4$ for Application

In this section, we have discussed two competing behavioural models for application in hazard assessment, in particular that of PNEC estimation. Moreover, similarities between the decision rules derived under these models were highlighted. However, from a risk managers perspective, the array of models and intent for application might be confusing, particularly between $\mathcal{M}_3$ and $\mathcal{M}_4$.

The assumptions governing $\mathcal{M}_3$ are contentious and little scientific research exists in exploring or defending them. Where homogeneity between substance variation is

appropriate, then the approach is sensible and tantamount to the level of uncertainty refinement offered by use of $\mathcal{M}_1$ or $\mathcal{M}_2$ over lower tier quantitative methods. This is despite EFSA (2005) even suggesting, although likely erroneously, that when N is small the estimators will be better reliable than those based on $\mathcal{M}_2$, even when the assumption of homogeneity is violated. A risk assessment based on $\mathcal{M}_3$ would undoubtedly require that the assumptions are assessed prior to acceptance for application. Moreover, scientific and regulatory agreement would be required for the adoption of $\mathcal{M}_3$ for reusable decision rules.

Exploring the key assumption of homogeneity closer (which also has an overlapping bearing on the validity of the [LA] estimator) we note is perhaps an unlikely property to be satisfied in reference to predicting multi-taxonomic protection, as is currently required for aquatic communities. For birds and mammals, which were the focus of Luttik and Aldenberg (1997), the classifications of communities are more closely taxonomically defined. This may lead to the assumption being valid. In addition, the assumption of homogeneity is made in reference to some suitable population of substances, possibly requiring refinement by some suitable property, e.g. chemical class, before application of $\mathcal{M}_3$. We assessed the assumption of homogeneity based on a toxicity database for the fish taxon exposed to pesticide stressors; this database is described in-depth in Section 4.1. Standard hypothesis tests (Levene's and Bartlett's) for the null hypothesis of homogeneity were rejected at critical levels of 5% and 10%. To explore this further an analysis of the posterior distributions of σ^2 from the perspective of $\mathcal{M}_1$ and $\mathcal{M}_3$ was conducted in accordance with Aldenberg (2005). This was performed by considering only the largest twenty datasets contained in the available database with $14 \leq n_i \leq 47$.

Under $\mathcal{M}_3$ the posterior distribution of σ^2 is an inverse-gamma distribution with shape and scale parameters: $\sum(n_i - 1)/2$ and $\sum(n_i - 1)s_i^2/2$ respectively. Under $\mathcal{M}_1$ each independent σ_i^2 has an inverse-gamma posterior distribution with shape and scale parameters: $(n_i - 1)/2$ and $(n_i - 1)s_i^2/2$ respectively. Figure 3.10 (left panel) plots the independent posterior distributions (under $\mathcal{M}_1$). Figure 3.10 (right panel) sequentially plots the posterior distribution of σ^2 (under $\mathcal{M}_3$), starting from the dataset corresponding to the largest sample size (red curve) to the smallest

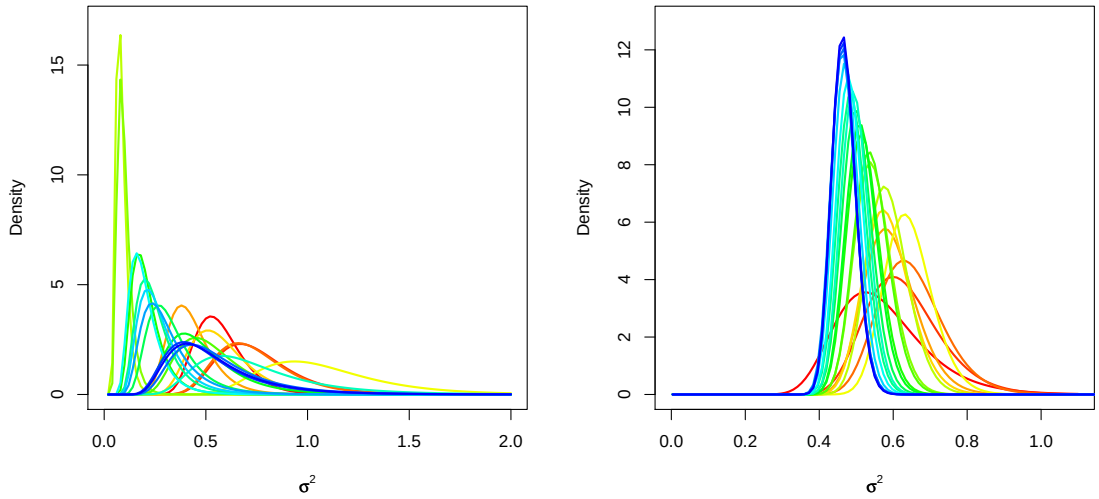


Figure 3.10: Posterior distributions of: $\sigma_i^2 \,|\, \mathbf{Y}_i$ (left); $\sigma_i^2 \,|\, \mathbf{Y}_1, \ldots, \mathbf{Y}_i$ (right); for $i = 1, \ldots, 20$ and n_1 (red) $> \cdots \geq n_{20}$ (blue) .

sample size (blue curve). Examination of Figure 3.10 indicates that the assumption of homogeneity is not supported in this instance. This analysis is limited to a database which contains many substances, with potentially varying modes of action (of which are unknown to us). However, we know of no evidence supporting the assumption that chemicals with similar modes of action will in fact have homogeneous interspecies variation parameters.

Under $\mathcal{M}_4$ the direct HC_p estimates are unlikely to be permitted by regulatory standards due to the lack of uncertainty refinement about σ. However, the demonstration of their use in Section 3.8.3 indicated that they may have validity at lower tier quantitative PNEC estimation. The current lower tier procedure is to apply predetermined assessment factors which are very ambiguous with regards to the level of protection they offer. The current REACH guidance for uncertainty analysis (ECHA, 2008b) indicates that reasonable worst-case assumptions are valid for low tier risk assessment, hence $\mathcal{M}_4$ is not precluded from application within the regulatory arena. As a consequence, we would not consider estimators derived under $\mathcal{M}_4$ as being applicable for the intermediate quantitative tier of assessment that $\mathcal{M}_1$–$\mathcal{M}_3$ apply to, but rather the lower tier where a robust yet defensible — at least from the protectionist viewpoint — estimator that is easily derivable is essential.

This is despite that there is 'higher tier' (ECHA, 2008b) probabilistic considerations involved. However, this assertion is entirely subject to a suitably conservative and permissible proposal value of σ.

The premise of $\mathcal{M}_4$ may also be used by scientists under the auspices of regulators to review the magnitude of currently recommended assessment factor values which have not changed in recent times; see EFSA (2005) for a summary of historical proposals. For example, it was shown that for some taxonomic groups there was reasonable alignment with current regulatory guidance.

3.9 Discussion

In this chapter we have calculated a number of decision rules which are readily applicable and tractable for incoming risk assessments. The approach of this chapter was about estimation based on the principle of decision making which transparently allows for incorporation of conservatism via the adaptation of loss and expert judgements. The former allows for conservatism (whether protectionist or otherwise) to be incorporated in a meaningful manner, which replaces the need for *ad hoc* conservatism, *a posteriori*.

Arguments were made for working within a Bayesian decision theoretic framework, from which we advocated the Bayes rule — a decision rule based on minimisation of the posterior expected loss. In addition, decision rules were juxtaposed to the Bayesian statistical inferential estimators — the [EFSA] estimators. The scientifically accepted [AJ] estimator — *de facto* in many applications — was interpreted under this framework, allowing for consideration regarding recent recommendations on levels of conservatism. The framework was straightforward enough that an alternative estimator was proposed based on a loss function reflective of the nature of ecotoxicological impact from a protectionist viewpoint.

Whilst $\mathcal{M}_1$ is the common regulatory and scientific model used for deriving HC_p values at the intermediate quantitative tier of hazard assessment, other behavioural models were introduced which feature in scientific and executive literature. However, it is likely that the risk management arena will view these model proposals

cautiously with respect to their applicability. In particular, the premise of $\mathcal{M}_1$–$\mathcal{M}_3$ was established as not being on par with $\mathcal{M}_4$. For very small sample sizes where assessment factors are required, then in fact $\mathcal{M}_4$ would probably serve as the only potentially viable model from a regulatory risk management perspective since full probabilistic hazard assessments are restricted to minimum sample sizes ($\approx$ 10, ECHA 2008a). Nevertheless, $\mathcal{M}_2$ and $\mathcal{M}_3$ serve to counter the reluctance of stakeholders in using smaller sample sizes in probabilistic hazard assessments. We would note, however, that the primary assumption of $\mathcal{M}_3$ were determined as inappropriate for the example database we explore in this thesis. Furthermore, the primary assumption of $\mathcal{M}_2$ has not been explored and is non-trivial to do so because the estimates of σ^2 combine different sampling uncertainties. One might consider exploring other two parameter models to model heterogeneity, however overall decision rules would likely be intractable. A wide variety of the estimators discussed in this chapter lowered the assessment shift-factor for increased samples of distinct species tested. Despite this observation, it is likely to be a conflict of interest with current regulatory aspirations regarding the reduction of *in vivo* testing.

Current guidelines on how to utilise probabilistic tools for purposes of PNEC estimation are not entirely strict, at least within the current REACH guidance (ECHA, 2008a,b), thus arguments for alternative estimators are conceivable. We could not endorse one estimator or one behavioural model over another, since a single black-box function yielding a conservative HC_p value is an unattainable aspiration, even by the standards of assessment factors. It is with this point that we would recommend estimators be based on reconciled choices between risk assessors and risk managers.

Chapter 4

On Non-Exchangeability

Our discussion up until now has been restricted to the foundations of probabilistically derived toxicant concentrations of concern. Underpinning this research is the SSD concept, with the inherent assumption of *species exchangeability*, i.e. information about relative positions of species in SSDs for other chemicals is uninformative about their relative positions for the chemical being assessed $\mathcal{S}$. As the following is an exploratory (and later on in Chapter 5 a modelling) based exercise we do not present a strict mathematical definition of exchangeability, instead we direct readers to Bernardo and Smith (1994, Section 4.2). An important statistical consequence of species exchangeability is that any measurements made for the new chemical under assessment may be considered to be a random sample from its uncertain SSD regardless of which species are to be measured. Failure of this assumption is what we refer to here as *non-exchangeability*. ECHA (2008a) states that for statistical extrapolation methods to be used, one must assume that:

> 'the group of species tested in the laboratory is a random sample of this distribution.'

The current guidance for quantitative levels of chemical safety assessment requires the experimental assessment of certain species; from here onwards we refer to such species as 'standard dossier test species'. For example, when assessing aquatic invertebrates, *Daphnia magna* is the usually required test species under Directive 91/414/EEC (EC, 1991, 2002) and REACH (EC, 2006; ECHA, 2008a) guidelines.

The former directive also requires that the rainbow trout (*Oncorhynchus mykiss*) — a species of salmonid (i.e. belongs to the taxonomic family *Salmonidae*) — is assessed when addressing the acute risk assessment for fish exposed to plant protection products. For the latter guidance document the rainbow trout is also commonly assessed for a multitude of reasons, some of which are discussed later.

The requirement or allowance without penalty of standard test species may be of practical consequence in ERA, especially if the standard test species is non-exchangeable. It is recognisable that certain species, for example, the rainbow trout are 'typically sensitive species' relative to other species in the broad taxon of fish for a wide range of chemicals (Dwyer et al., 2005). It has also been recommended that an assessment factor of three be applied to the rainbow trout's tolerance value as opposed to the usual value in recognition of this sensitivity (Ibid). Raimondo et al. (2008) reported that this leads to a relatively under-protective decision rule relative to the use of a HC_5 based on empirical evidence of 59 chemicals. Raimondo et al. (2008) also issues caution about conducting ERAs based on the surrogacy of non-salmonid species due to the apparent demonstration of relative lower sensitivity. We shall later discuss formal methods to identify such a phenomenon. Whilst this *might* be valuable knowledge for the strictly deterministic assessment factor based procedures, it is neglected when considering probabilistic intermediate tier approaches which aim to refine the hazard assessment and uncertainty. Non-exchangeability may be accounted for using alternative methods such as bootstrapping (for example, consult Jagoe and Newman 1997; Newman et al. 2000, 2002; Grist et al. 2002 and Duboudin et al. 2004a) since no distributional assumptions are made, *per se*. Luttik and Aldenberg (1997) did report that one might use the median [LA] estimator when estimated using a 'sensitive species', and the 95% one-sided underestimate confidence limit otherwise, although the justification was qualitative only. One might consider adopting the 'precautionary principle' and proceed as usual. However if one actually wants to refine hazard assessment, then better modelling is required. This point is ratified by the authoritative pair Forbes and Calow (2002a), who state:

> '... extrapolation may mislead and thereby hinder environmental protection. Because the tiered approach allows the risk assessment to stop when and if risk is deemed to be negligible it is essential that [assessment] factors applied at each tier lead to neither over- nor under-conservative estimates of risk.'

An intuitive model from a statistical viewpoint point would be to abandon the concept of 'non-exchangeability' and use a fully hierarchical model. Such ideas have been discussed recently using properties of chemicals and a measure of species vulnerability (Jager et al., 2007), or perhaps even functional models using species traits (Baird and Van den Brink, 2007; Rubach et al., 2009). However, such models require much more research, in particular more data, before having predictive utility. Furthermore, the point which motivates our modelling rationale is that risk managers are less likely to adopt methodology which is complicated and requires advanced statistical software. This stems from a lack of experience with relatively sophisticated quantitative techniques. The focus of this chapter is the detection and possible quantification of non-exchangeability. In later chapters we look at *tractable* ways to adapt current risk assessment methodology to allow for *pragmatic* and *parsimonious* risk assessment. In particular, we focus on the case of a single non-exchangeable species because our goal is tractable decision rules rather than modelling.

4.1 RIVM Fish Database

RIVM is a recognised leading centre of expertise in the field of environmental protection. For the purposes of this thesis, they have granted permission to use a toxicity database which they have compiled. However, due to proprietary rights, species and chemical names are coded and are not available for dissemination without permission from the RIVM. The actual database we use is a subset of much larger database held by the RIVM which was fully described in De Zwart (2002). We will analyse data on acute fish toxicity (EC_{50} values) to a range of pesticides. For a general discussion on toxicity data, consult Section 2.2. The full database has been amalgamated from a wide range of sources, including: the freely available US EPA

ECOTOX database (US EPA, 2007); academic literature; and internal reports. In addition, the database has been prepared in the following way:

- If more than one toxicity value was available for a species-substance combination, then a geometric mean value was calculated.
- If more than one toxicity value was available for a species-substance combination, and for a certain species a censored value was present, then this value was removed unless it exceeded the concentration interval of observed point tolerance values, in which case the limit of the value was applied.
- If in a set of available toxicity values for a species-substance combination only a censored value was present, then this value was only used if it exceeded the concentration interval of observed point tolerance values.

Further information on data pre-processing is provided in De Zwart (2002); the preparation techniques applied here are common among the applied SSD literature.

Although the database contains tolerance values for 172 species across $N = 379$ pesticide substances $\mathcal{G}$, many values are missing. The sparsity is such that the EC_{50} has only been measured for 1903 of the possible 65188 species-substance pairs. Figure 4.1 summarises the structure of the database graphically. The left panel of Figure 4.1 displays a line plot for the number of chemicals which have n_i ($1 \leq n_i \leq 47$) toxicity records. For example, there are 143 chemicals for which $n_i = 2$, but only 7 with n_i ranging from 21 to 47. By letting m_j be the number of substances species j is assessed with, the right panel of Figure 4.1 displays a line plot for the number of species which have been distinctly assessed with m_j ($1 \leq m_j \leq 344$) chemicals. For example, there are 74 species with $m_j = 1$, and only a single species with $m_j = 344$ — the rainbow trout.

Plotting the sample means against the sample standard deviations on the original scale for the EC_{50} data indicates a strong linear correlation (Figure 4.2, left panel). Whereas plotting the sample means of the log-transformed data $\bar{y}_i$ against the sample standard deviations of the log-transformed data s_i stabilises the variances (Figure 4.2, right panel). This supports the log-transformation applied to the toxicity data.

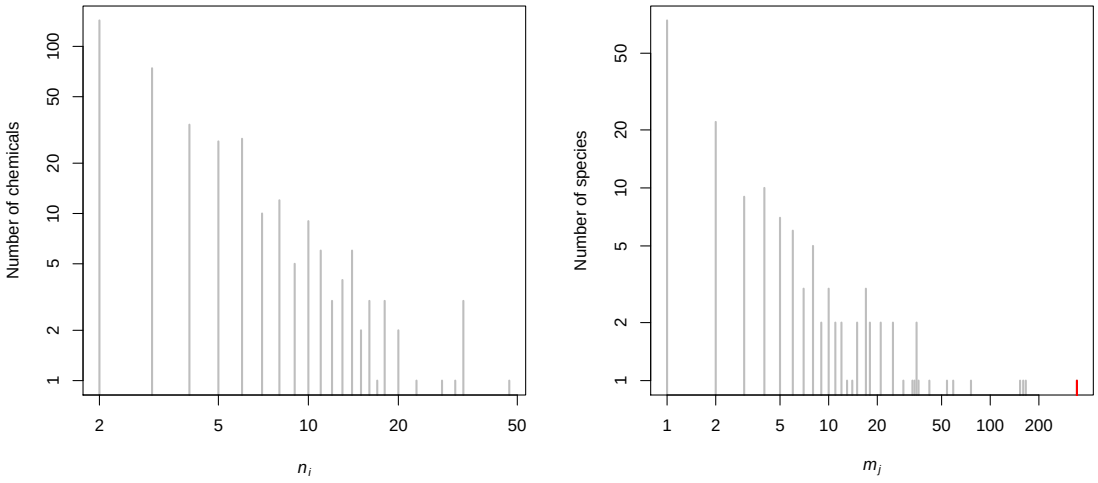


Figure 4.1: Left: line plot of n_i vs. substance count for ($i \in \mathcal{G}$); right: line plot of m_j vs. species count ($j \in \cup_{i \in \mathcal{G}} J_i$). Red line indicates rainbow trout.

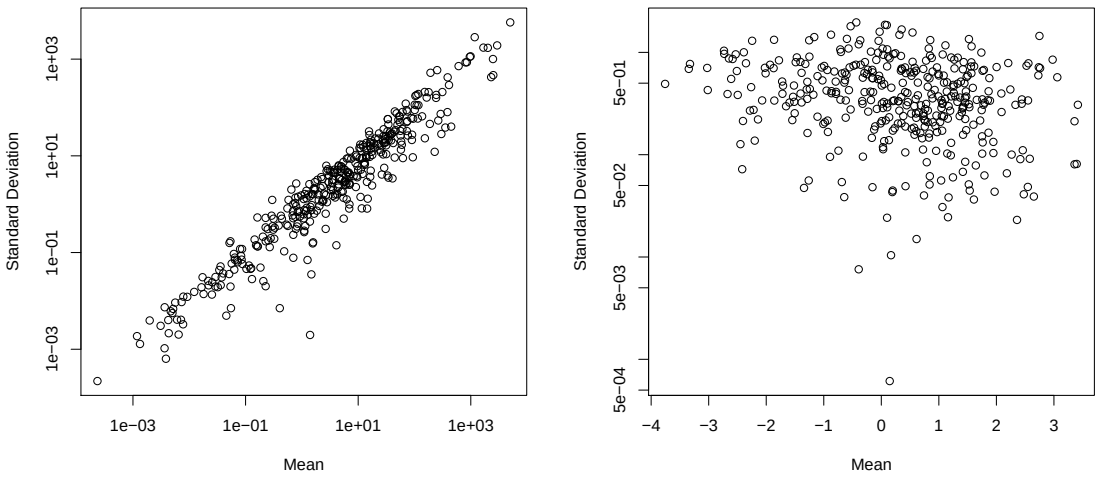


Figure 4.2: Scatterplot of the sample mean versus sample standard deviation for substances in the RIVM fish database. Left: original scale; right: log (base 10) transformed data.

It is necessary to explore the suitability of parametric distribution fits to toxicity data. ECHA (2008a) stipulates that although the log-normal is the most pragmatic choice, it should be evaluated with: (i) the Anderson-Darling [AD] goodness-of-fit test and (ii) the Kolmogorov-Smirnov [KS] goodness-of-fit-test. Each of the respective tests are based on a comparison of the reference cumulative distribution to the empirical distribution function (Stephens, 1974). The results of these tests can influence the magnitude of the *post hoc* assessment factor placed on the HC_5 estimate (see Section 2.4). The [AD] test is suggested by ECHA (2008a) because it gives more weight to the tails of the SSD, which is the region of typical interest. The [KS] test is implemented through an adaptation known as Lilliefors' goodness-of-fit test which corrects the null distribution for being estimated with the unbiased sample estimates of the data; this has less power in general. For a discussion of these goodness-of-fit tests and many others, consult D'Agostino and Stephens (1986). Alternatively, within the context of SSDs, consult Aldenberg et al. (2002). The latter paper suggests that an ideal approach would be to seek a goodness-of-fit test localised to the lower tail region. Notwithstanding the availability of such goodness-of-fit tests, Farrell and Rogers-Stewart (2006) note that no omnibus test for detecting departures from normality appears to exist. The Shapiro-Wilk goodness-of-test fit was, however, noted as being particularly noteworthy in performance over a wide range of alternative distributions for small sample sizes. This test, which is not based on the empirical cumulative distribution function, is noted by Aldenberg et al. (2002) as being comparable in performance to the [AD] test.

In accordance with current recommendations the [AD] and [KS] goodness-of-fit tests were applied to the RIVM fish toxicity database using the `ad.test` and `lillie.test` functions contained in the `R` (2006) `nortest` package respectively. It is required that $n_i > 7$ and $n_i > 4$ respectively for the two tests. For the [AD] test applied to 63 datasets: 55 had a P-value greater than 0.01; 42 had a P-value greater than 0.05; and 39 had a P-value greater than 0.10. For the [KS] test applied to 128 datasets: 115 had a P-value greater than 0.01; 103 had a P-value greater than 0.05; and 99 had a P-value greater than 0.10. A P-value of 0.05 is the typically assigned critical value (Newman et al., 2000; Newman, 2008); a P-value below this critical

value would lead to one rejecting the null hypothesis of normality. The latter proportions are consistent with Luttik and Aldenberg (1997); the former with Newman et al. (2000) who applied the Shapiro-Wilk goodness-of-fit test. We acknowledge the on-going debate regarding the validity of the log-normal assumption; however, working with large datasets is an exceptional situation. Therefore non-parametric methods are typically not effective, nor are more sophisticated probabilistic models.

Note that any SSDs we describe in this chapter are representative of fish assemblages. However, in many situations regulatory guidance requests a multi-taxa SSD be determined unless one taxon is identified as being particularly sensitive (e.g. a herbicide will likely be *a priori* more sensitive to plants than fish), in which case it is sufficient to look at the per-taxon SSD. This was discussed in Section 2.4.

4.2 The Rainbow Trout

The rainbow trout (*Oncorhynchus mykiss*, coded as S119 in the RIVM fish database) is a species of salmonid used frequently in laboratory ecotoxicological experimentation (Alexander and Fairbridge, 1999). There are a multitude of reasons for this: (i) ease of cultivation; (ii) extensive scientific knowledge of its life history (`www.FishBase.org`; accessed 24/06/2009); (iii) it is an approved indicator species (US EPA, 2002). Raimondo et al. (2008, p. 2601) report that a subspecies of rainbow trout (as well as other common test species) is federally listed in the United States.

Rainbow trout was acknowledged by EFSA (2005) to be a particularly sensitive species to chemical stressors, and this was demonstrated using a non-parametric plot, similar to that of Figure 4.3. In this figure, each point on the plot represents a different substance from the RIVM fish database. The geometric mean of the EC_{50} values for all fish (other than the rainbow trout) is divided by the EC_{50} for rainbow trout. There are 344 substances in the RIVM fish database that has a measured tolerance value for the rainbow trout. If the plotted points for these substances lies above 1 much more than 50% of the time, which they appear to do, then this crudely implies sensitivity of the rainbow trout. The EFSA (2005) figure only considered

the 220 substances satisfying $n_i \geq 3$; we demonstrate it for all substances where the rainbow trout was tested. However, the highly unbalanced incomplete factorial structure of the database means this figure can only be vaguely interpreted. In Figure 4.4 a histogram is shown which summarises the aforementioned EC_{50} ratios plotted in Figure 4.3.

An alternative exploratory demonstration is to consider the estimated PAFs evaluated at environmental concentrations equal to the rainbow trouts' EC_{50} using estimated [log-normal] SSDs. To do this, we parametrically forward estimate (see Section 2.4.1) the fraction of species potentially affected using: (i) a moment estimator; and (ii) the median [AJ] forward estimator (see Aldenberg and Jaworska 2000). The former is basically a normal cumulative distribution function evaluated at the tolerance value of the rainbow trout and the latter is straightforwardly determinable through considerations of Equation 2.5. Each SSD is estimated independently of the tolerance value of the rainbow trout in order to remove bias. Furthermore, for computational reasons, analysis is restricted to datasets satisfying $n_i \geq 3$ ($N = 220$). In Figure 4.5, we display histograms of the substances analysed; the left panel shows the moment estimators and the right panel shows the median [AJ] forward estimates. There is not a substantial difference between the two histograms, with both suggesting that the estimated PAF evaluated at the rainbow trouts' EC_{50} would be less than 50% on average. As per the previous demonstration, Figure 4.5 can only be vaguely interpreted, but may be a useful diagnostic tool for the purposes of risk communication.

4.3 Hypotheses Tests

In this section we test the null hypothesis that species tolerance values are *a priori* exchangeable within the fish taxon for each new chemical assessed. Two non-parametric hypothesis tests are proposed based on the ranks of species tolerance values available within the RIVM fish database. We denote r_{ij} to be the rank of species $j \in J_i$ with $r_{ij} = 0 \ \forall j \notin J_i$. The first test is motivated by the familiar sign-test which is less powerful than the second test — a rank-sum test — but more

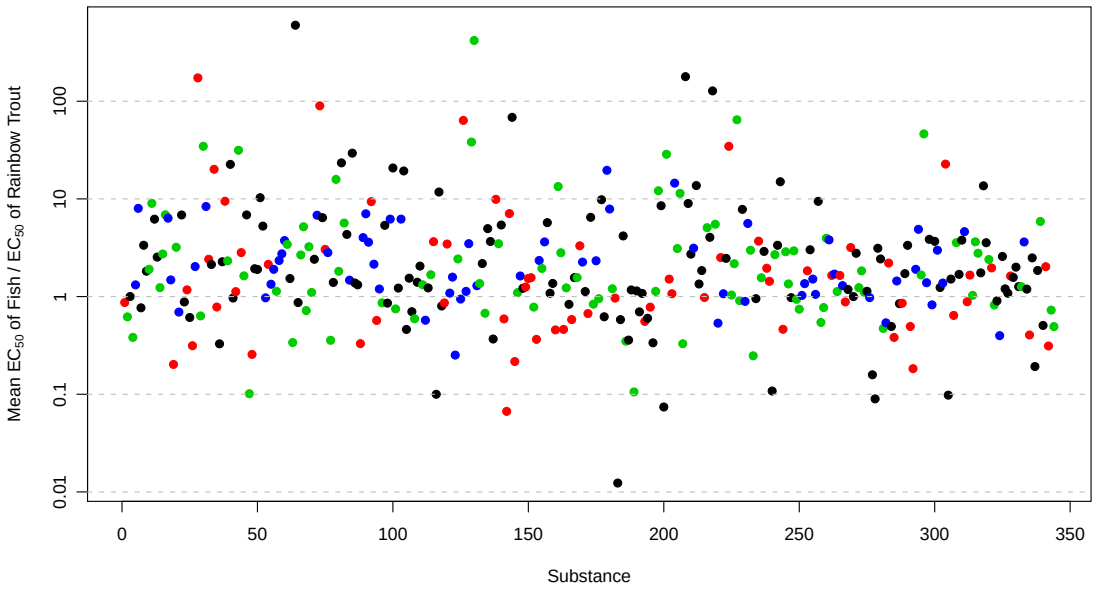


Figure 4.3: Reciprocal of ratios of rainbow trouts' EC_{50} value to geometric mean other fish species EC_{50} values. Each point represents a single substance; $n_i = 2$ (black points); $n_i = 3$ (red points); $4 \leq n_i \leq 7$ (green points); $n_i \geq 8$ (blue points).

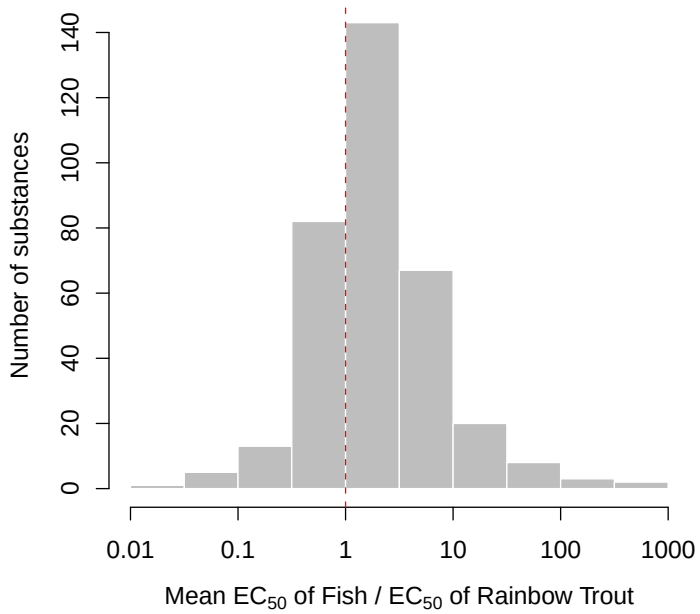


Figure 4.4: Histogram of the reciprocal of ratios of rainbow trouts' EC_{50} value to geometric mean other fish species EC_{50} values.

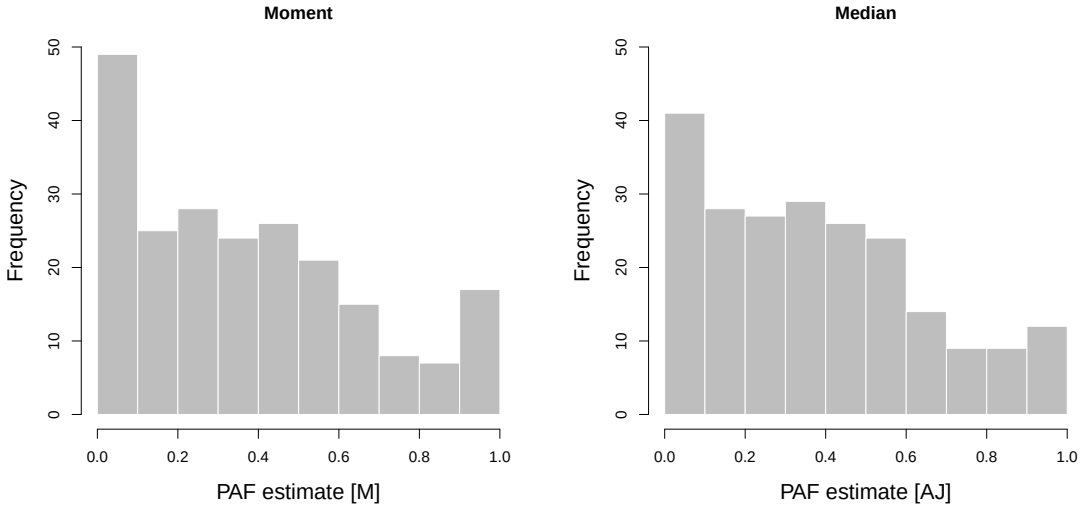


Figure 4.5: Histograms of the estimated PAF evaluated at the rainbow trouts' EC_{50} for 220 pesticides. Left: moment based estimator; right: median [AJ] estimator.

robust because the second test is sensitive to outcomes for individual chemicals.

4.3.1 Hypothesis Test 1: Sign Test

Under the null hypothesis, a species should be equally likely to appear in the left-half or the right-half of the data for each compound. We therefore apply the binomial distribution to determine whether a species occurs too often on one side or the other.

For every species j, we calculate the following quantities:

$$\begin{aligned} m_j^+ &= \sum_{i \in \mathcal{G}} \mathbf{1}\left\{r_{ij} > \tfrac{1}{2}(n_i+1)\right\}; \\ m_j^\pm &= \sum_{i \in \mathcal{G}} \mathbf{1}\left\{r_{ij} \neq \tfrac{1}{2}(n_i+1)\right\}, \end{aligned}$$

where $\mathbf{1}\{\mathcal{A}\}$ is an indicator function for a Boolean response $\mathcal{A}$ taking value 1 if $\mathcal{A}$ is true, and 0 if $\mathcal{A}$ is false. Therefore m_j^+ records how many times species j was in the right half of the data and $m_j^\pm$ records how many species j was in the right or left half; central positions receive a score of zero.

Under the null hypothesis, for each species j we have

$$n_j^+ \,|\, m_j^\pm \sim \text{Bin}\left(m_j^\pm, \tfrac{1}{2}\right),$$

Species	Code	m	$m^{\pm}$	m^{+}	$m^{+}/m^{\pm}$	P-value
Oncorhynchus mykiss	S119	344	301	83	0.28	3.9×10^{-15}
Carassius auratus	S023	76	69	56	0.81	1.7×10^{-07}
Cyprinus carpio	S051	166	150	103	0.69	5.6×10^{-06}
Heteropneustes fossilis	S071	36	36	31	0.86	1.3×10^{-05}
Oncorhynchus clarki	S118	42	41	10	0.24	1.5×10^{-03}
Pimephales promelas	S132	160	147	93	0.63	1.6×10^{-03}
Carassius carassius	S024	25	23	19	0.83	2.6×10^{-03}
Channa punctatus	S034	17	16	14	0.88	4.2×10^{-03}
Clarias batrachus	S040	17	16	14	0.88	4.2×10^{-03}
Salvelinus namaycush	S153	35	33	8	0.24	4.6×10^{-03}

Table 4.1: Species with the smallest P-values based on hypothesis test 1.

allowing us to perform a two-sided Binomial test to derive a P-value based on the associated hypothesis test. Results from the application of this hypothesis test to the RIVM fish database are displayed in Table 4.1 for the ten species with the smallest P-values.

Interpretation beyond the first row is not recommended because it is theoretically possible that the significant values are simply an artefact of the possible non-exchangeability of the rainbow trout. However, according to Rand (1995, p. 78), four of the species listed are standard test species, including: rainbow trout (*O. mykiss*), goldfish (*C. auratus*), common carp (*C. carpio*) and fathead minnow (*P. promelas*). The latter three species are all members of the *Cyprinidae* taxonomic family. Whilst there is strong evidence against exchangeability, this test is not sufficient to say whether it is only the rainbow trout which exhibits this phenomenon, and whether it presents the largest, for want of a better word, 'bias'. However, the results do indicate that the rainbow trout tends to be present in the lower half of the data, which based on discussion thus far was not unexpected.

It is interesting to consider the results of this hypothesis when tolerance values for the rainbow trout are excluded. Removing such a standard test species has strong consequences for the RIVM fish database; it reduces the number of database entries by 18%, as well as removing a significant number of chemicals currently of order $n_i = 2$ from having influence on the test statistic. It is for this reason that we do not perform the rank-sum test (Section 4.3.2) on this reduced database. Applying

Species	Code	m	$m^{\pm}$	m^{+}	$m^{+}/m^{\pm}$	P-value
Heteropneustes fossilis	S071	36	33	28	0.85	6.6×10^{-05}
Salvelinus namaycush	S153	35	32	5	0.16	1.1×10^{-04}
Carassius auratus	S023	76	68	50	0.74	1.3×10^{-04}
Oncorhynchus clarki	S118	42	40	9	0.23	6.8×10^{-04}
Channa punctatus	S034	17	16	14	0.88	4.2×10^{-03}
Clarias batrachus	S040	17	16	14	0.88	4.2×10^{-03}
Esox lucius	S054	18	17	3	0.18	1.3×10^{-02}
Pimephales promelas	S132	160	129	79	0.61	1.3×10^{-02}
Carassius carassius	S024	25	22	17	0.77	1.7×10^{-02}
Ctenopharyngodon idella	S049	15	15	12	0.80	3.5×10^{-02}

Table 4.2: Species with the smallest P-values based on hypothesis test 1 excluding rainbow trout.

the hypothesis test to the RIVM fish database which excludes the rainbow trout as a test species yields analogous results to Table 4.1, as displayed in Table 4.2.

Thus it would appear that there is evidence of non-exchangeability among the remaining species even when one has discounted the rainbow trout. It is particularly interesting to note that the ranking of species by P-values in Table 4.2 has not been simply shunted up by one place. This reinforces the point that we do not have sufficient evidence to state how many species might be non-exchangeable, and to what degree.

4.3.2 Hypothesis Test 2: Rank Sum Test

A non-parametric rank sum test, which directly uses the rank scores as opposed to indicator functions of them should be more powerful. For each species j we calculate the test statistic

$$\hat{R}_j = \sum_{i \in \mathcal{G}} r_{ij}.$$

Under the null hypothesis, for each species j we have

$$\mathbb{P}[r_{ij} = x \,|\, n_i] = \frac{1}{n_i} \text{ for } x \in \{1, 2, \ldots, n_i\}$$

and r_{ij} is independent for different values of i.

Species	Code	m	$\tilde{r}$	P-value
Oncorhynchus mykiss	S119	344	−0·42	8.6×10^{-12}
Heteropneustes fossilis	S071	36	0·83	1.9×10^{-7}
Carassius auratus	S023	76	0·68	3.1×10^{-5}
Salvelinus fontinalis	S152	33	−0·58	1.3×10^{-4}
Carassius carassius	S118	25	0·85	1.6×10^{-4}
Oncorhynchus clarki	S040	42	−0·61	3.6×10^{-4}
Clarias batrachus	S024	17	0·91	4.0×10^{-4}
Salvelinus namaycush	S153	35	−0·59	2.4×10^{-3}
Channa striata	S035	10	0·73	3.9×10^{-3}
Perca flavescens	S127	29	−0·38	6.5×10^{-3}

Table 4.3: Species with the smallest P-values based on hypothesis test 2.

The exact null sampling distribution of $\hat{R}_j$ is mathematically intractable, however the distribution for each species j can be approximated using either Monte Carlo sampling or a central limit theorem based normal distribution approximation using the theoretical expectation and variance. These are respectively determined to be $\sum_{i:j\in J_i}\frac{1}{2}(n_i+1)$ and $\sum_{i:j\in J_i}\frac{1}{12}(n_i^2-1)$. The former method is particularly difficult because the small P-values would require a very large amount of Monte Carlo simulations; the latter method requires m_j to be sufficiently large in order to be an effective approximation. As a consequence of the discretised nature of $\hat{R}_j$, we apply the standard continuity correction of $1/2$ before we apply the normal approximation. The species with the 10 smallest P-values are listed in Table 4.3. In addition we show a standardised measure of 'bias', denoted as $\tilde{r}_j$, which we calculate as the average *standardised rank* for each species j, given by

$$\tilde{r}_j = \frac{1}{m_j}\sum_{i:j\in J_i}\frac{r_{ij}-\frac{1}{2}(n_i+1)}{\sqrt{\frac{1}{12}(n_i^2-1)}}.$$

One might argue that the test statistic should be determined via a weighted sum of statistics. However, in effect this is already the case. To place r_{ij} on the scale $(0,1]$ for all i, one might consider r_{ij}/n_i. The test statistic would then be obtained for each species j as some weighted sum over all i. Selecting the weight to be n_i, so that more influence is achieved from datasets with large sample sizes, would result in the original test statistic $\hat{R}_j$. We do not consider alternative weights in this research.

In addition to the central limit normal approximation used to calculate the P-values in Table 4.3, we also calculated the P-values using Monte Carlo simulation with $10,000$ samples. For the 10 species listed in Table 4.3 there was no measurable difference in P-values, nor was there any change in the rankings. However, the Monte Carlo output did allow us to check the validity of the normal approximation. In Figure 4.6 we show the Q-Q plots of 10 species highlighted in Table 4.3 which confirm approximate convergence to normality.

Interpretation of Table 4.3 is subject to the same difficulties as Table 4.1. However, it has provided further evidence against exchangeability.

4.4 Sensitivity of Data Points

A result from Aldenberg et al. (2002) implicitly suggests that we should be more concerned with the potential non-exchangeability of a 'sensitive' species, say the rainbow trout, rather than a more 'tolerant' species, say the goldfish. This is because it is the lower ranked tolerance values which have the strongest influence when estimating the HC_p for small p, say $p = 5$ — the usual value of interest.

Let $\hat{\sigma}$ be a measure of the standard deviation estimated from the log-toxicity data for substance $\mathcal{S}$ based on $\mathcal{M}_1$, $\mathcal{M}_2$ or $\mathcal{M}_3$ (or an alternative behavioural model) with λ_0 corresponding degrees of freedom. For example, for $\mathcal{M}_1$, $\hat{\sigma} = s$ with $\lambda_0 = n - 1$. Now let us consider a decision rule under the context of exchangeability, which was shown to have general form $\delta_p(\mathbf{Y}) = \bar{y} - \kappa_p \hat{\sigma}$. We define the dimensionless sensitivity quotient $\mathcal{Q}_j$ for $j \in J_{\mathcal{S}}$ to be the rate of change of the estimator with respect to a tolerance value y_j, which is straightforwardly shown to be equivalent to

$$\mathcal{Q}_j \triangleq \frac{\partial \delta_p(\mathbf{Y})}{\partial y_j} = \frac{1}{n} - \frac{\kappa_p}{\lambda_0}\left(\frac{y_j - \bar{y}}{\hat{\sigma}}\right). \tag{4.1}$$

For small p and $\kappa_p > 0$, any standardised data point $(y_j - \bar{y})/\hat{\sigma} < \lambda_0/(n\kappa_p)$ has positive influence on the estimator $\delta_p(\mathbf{Y})$. Moreover, the magnitude of influence increases asymmetrically as one moves towards the lower order statistics. For realistic sample sizes under $\mathcal{M}_1$, the threshold increases steadily from 0 to approximately 0.4 for [AJ] with $p = 5$ and $\gamma = 0.95$ and for [EFSA] with $p = 1$. Under $\mathcal{M}_2$ the

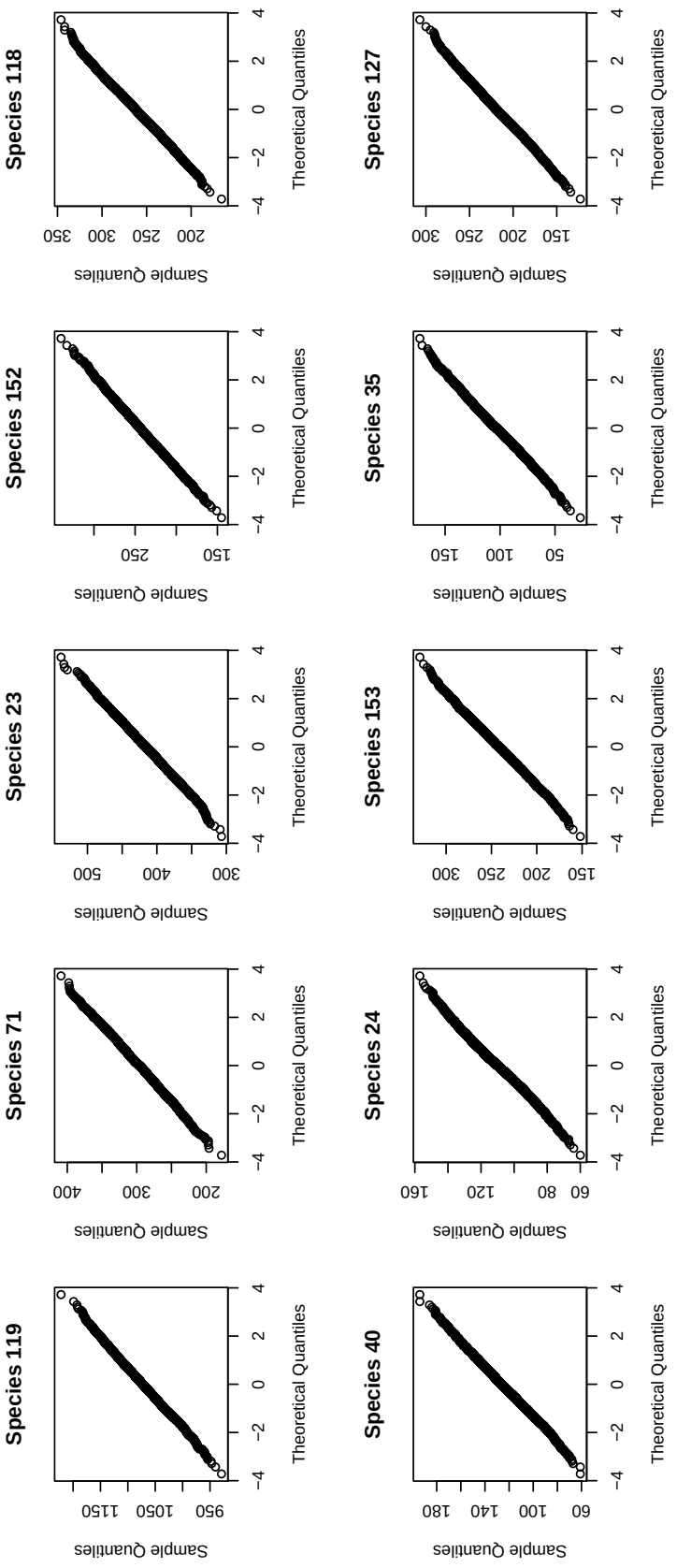


Figure 4.6: Q-Q plots of Monte Carlo output for null distribution of $\hat{R}_{j\cdot}$.

threshold slightly exceeds 0.4 for the same two estimators based on an estimate of $\alpha = 1.51$, which we obtain later by reconsidering the HC_p estimator in the context of a single non-exchangeable species.

As for the deterministic approach of calculating a PNEC by dividing the lowest observed tolerance value by a fixed assessment factor, the probabilistically derived estimators are also most influenced by the lower tolerance values. The implication for risk assessment is that systematic 'bias' which leads to a species typically lying in the lower half of an SSD will more strongly influence the estimator. Adjusting for such effects is more likely to have an effect on the resulting PNEC estimate than adjustments for 'bias' which lead to a species typically lying in the upper half of an SSD. This will be especially the case if as highlighted above, the size of the 'bias' is approximately half a standard deviation.

4.5 Initial Modelling: Species Effects

It is quite plausible that the concept of exchangeability is untenable from a statistical modelling viewpoint, and that all species are in fact non-exchangeable. In fact, if all the rainbow trout toxicity data from the analysis is eliminated, one still finds evidence of non-exchangeability for the remaining species. A statistician would naturally fit some model incorporating both chemical and species effects, ideally with careful consideration of distributional assumptions. However, the non-factorial nature of the database means that estimates would be highly confounded. Moreover, the constraints on tractability from the regulatory sector would likely not be achieved. Notwithstanding this point, we explore this modelling perspective further so that we gain vital insight into the measure of 'bias' on the log-concentration scale, for the rainbow trout and other species.

The basic model for the response variable (log-tolerance value) for substance i and species $j \in J_i$ can be succinctly written as

$$y_{ij} = \mu_i + \epsilon_{ij},$$

where μ_i is the unknown mean of the SSD over log-concentration for substance i

(measured in $\log_{10}$ mg/L), and ϵ_{ij} is the error in the model. For the behavioural models proposed earlier alongside the normality assumption, we have

$\mathcal{M}_1$: $\epsilon_{ij} \,|\, \sigma_i^2 \overset{iid}{\sim} N(0, \sigma_i^2)$ such that $(\sigma_i^2 \,\forall i \in \mathcal{G})$ are unknown;

$\mathcal{M}_2$: $\epsilon_{ij} \,|\, \sigma_i^2 \overset{iid}{\sim} N(0, \sigma_i^2)$ and $\sigma_i^2 \,|\, \alpha, \beta \sim \mathcal{IG}(\alpha, \beta)$ where α, β are known;

$\mathcal{M}_3$: $\epsilon_{ij} \,|\, \sigma^2 \overset{iid}{\sim} N(0, \sigma^2)$ such that σ^2 is unknown;

$\mathcal{M}_4$: $\epsilon_{ij} \,|\, \sigma_i^2 \overset{iid}{\sim} N(0, \sigma_i^2)$ such that σ_i^2 is known $\forall i$.

We now augment the model by including *species effects*, i.e.

$$y_{ij} = \mu_i + \zeta_j + \epsilon_{ij},$$

where ζ_j is the species effect for species j, measured on the log-concentration scale. Our additive model is consistent with Jager et al. (2007) who also explored models such that response variables were additive sums of chemical 'potency' (measured as a functional relation to the octanol-water partition coefficient and molecular weight) and species 'vulnerability' and extends research by Craig (2005).

4.5.1 Homogeneity

Although the premise of $\mathcal{M}_3$ was criticised earlier in Section 3.8, it is a sensible starting point because the fitting of such models is reasonably straightforward using conventional two-way fixed effects ANOVA (without interactions) (Stuart et al., 1999).

Setting all species effects to zero, i.e. $\zeta_j = 0 \;\forall j$, leads to the usual minimum variance unbiased estimators: $\hat{\mu}_i = \bar{y}_i$ and $\hat{\sigma}_i^2 = s_p^2$. A histogram of the substance effects for this case is displayed in Figure 4.7.

It would not be sensible to estimate species effects for all species in the RIVM database due its sparsity; recall there are 172 species and 379 substances, but only 1903 available tolerance values. Moreover, the 'design' is such that the database is substantially unbalanced. Stuart et al. (1999, pp. 632–634) refer to this as an 'unbalanced two-way incomplete block design'. A logical compromise is to include

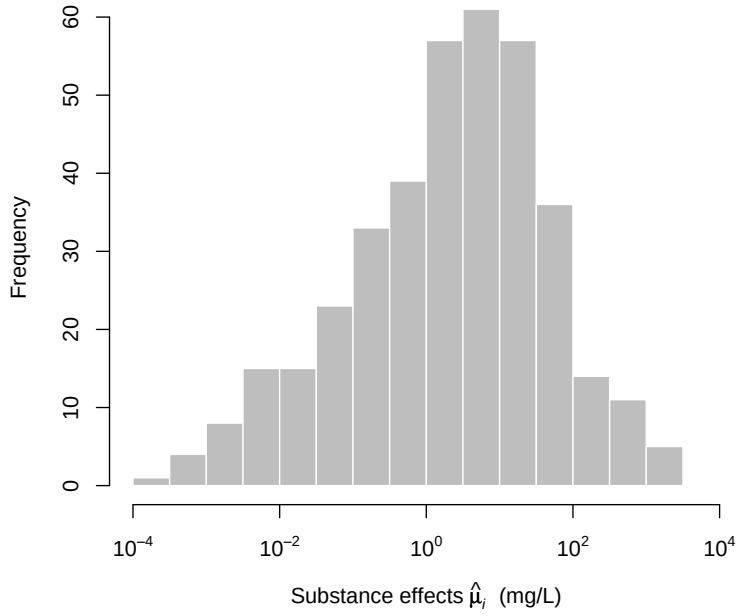


Figure 4.7: Histogram of compound effects for zero species effects.

only some species effects subject to suitable criteria. Since our model is only for exploratory analysis and not intended to act as a full predictive model, which is usually a motivating factor of such research, the rationale of not fitting all species appears satisfactory.

In order to decide which species effects to include in the model, we apply a forward model selection routine, starting from the baseline of no species effects, using the Bayesian information criterion (BIC) as our model selection criteria. The BIC is defined to be

$$2 \log \ell \left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\zeta}}, \hat{\sigma} \right) - p \log d,$$

where ℓ is the likelihood function for the full data model; $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\zeta}}$, and $\hat{\sigma}$ are the maximum likelihood estimates of $\boldsymbol{\mu} = (\mu_i, \forall i \in \mathcal{G})^T$, $\boldsymbol{\zeta}$ (the vector of species effects included in the model), and σ respectively. The number of parameters in the model is denoted as p, in this case $p = N + |\boldsymbol{\zeta}| + 1$; and d is the total number of data points used to derive the maximum likelihood estimates, in this case $d = \varsigma + N \equiv \sum_{i \in \mathcal{G}} n_i$.

The algorithm for this selection routine assuming ϵ_{ij} are conditionally indepen-

dently identically distributed normal with mean 0 and variance σ_i^2, is as follows:

1. Fit the *current* model $y_{ij} = \mu_i + \epsilon_{ij}$ using least squares, and calculate the BIC.

2. (a) For each species $j' \in \cup_{i \in \mathcal{G}} J_i$ fit the model $y_{ij} = \mu_i + \zeta_{j'} \mathbf{1}_{\{j=j'\}} + \epsilon_{ij}$ using least squares.

   (b) If the lowest BIC from the fitted models in (a) is less than the current BIC, then corresponding model is set as the updated current model.

3. Repeat step (2) for the remaining species effects until no further reduction in BIC is achieved. Return this model.

The stepwise selection procedure added 24 species effects in total; these are listed in the order of which they were included in Table 4.4. In addition we also display the maximum likelihood species effect estimates; standard error and corresponding P-value from the standard t-test under the hypothesis that the species effect is zero. Notice that there is a strong correlation of species added in Table 4.4 to species listed in Tables 4.1 and 4.3. However, some species effects added were species which were tested with only one or two substances. Typically these species account for the more appreciable species effects estimates. The standard unbiased estimate of σ was $\hat{\sigma} = 0.59$, and no significant change in substance effects was observed from the model with no species effects.

Another frequently used model selection (Burnham and Anderson, 2002) is Akaike's information criteria (AIC) defined as

$$2 \log \ell \left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\zeta}}, \hat{\sigma} \right) - 2p.$$

The forward selected model using this criterion added 48 species effects; these are listed in Table 4.5 in the order the species were included in the model. Interestingly, the first 24 added chronologically match those added using the BIC. However, the species effect estimates have altered marginally, and whilst not significantly, an observable pattern is that they have all decreased. The estimate of $\hat{\sigma} = 0.57$ remains close to the BIC selected model and no significant change was observed for the substance effect estimates. We will later briefly discuss model selection criteria.

Species	$\hat{\zeta}$	S.E.	P-value	m
S119	−0·20	0.042	1.0×10^{-06}	34
S071	0·67	0.104	1.8×10^{-10}	36
S023	0·42	0.076	3.8×10^{-08}	76
S118	−0·46	0.099	2.9×10^{-06}	42
S105	−1·92	0.436	1.2×10^{-05}	2
S015	−1·34	0.345	1.1×10^{-04}	3
S040	0·70	0.153	4.6×10^{-06}	17
S011	1·86	0.460	5.5×10^{-05}	2
S049	0·65	0.165	8.5×10^{-05}	15
S024	0·50	0.133	1.5×10^{-04}	25
S153	−0·34	0.107	1.4×10^{-03}	35
S152	−0·34	0.110	2.3×10^{-03}	33
S145	−1·42	0.433	1.1×10^{-03}	2
S085	0·95	0.274	5.1×10^{-04}	5
S160	−0·51	0.171	2.9×10^{-03}	14
S054	−0·40	0.148	6.9×10^{-03}	18
S034	0·49	0.150	1.2×10^{-03}	17
S035	0·61	0.194	1.6×10^{-03}	10
S139	1·82	0.610	2.9×10^{-03}	1
S132	0·18	0.056	1.1×10^{-03}	160
S051	0·17	0.056	2.7×10^{-03}	166
S041	1·03	0.357	4.0×10^{-03}	3
S062	−1·58	0.610	9.6×10^{-03}	1
S016	0·77	0.312	1.4×10^{-02}	4

Table 4.4: Summary of final model selected by BIC.

S119	S071	S023	S118	S105	S015	S040	S011
−0.28	0.59	0.34	−0.57	−2.04	−1.43	0.64	1.81
S049	S024	S153	S152	S145	S085	S160	S054
0.57	0.44	−0.44	−0.44	−1.48	0.90	−0.62	−0.49
S034	S035	S139	S132	S051	S041	S062	S016
0.40	0.53	1.74	0.10	0.11	0.94	−1.77	0.69
S009	S122	S127	S155	S148	S088	S161	S055
−0.81	−0.75	−0.35	−1.45	−0.6	−1.38	−0.87	−1.51
S048	S138	S102	S168	S046	S135	S101	S052
−1.29	−1.36	−0.21	−0.60	−1.28	−0.65	−0.40	0.92
S110	S060	S039	S001	S151	S014	S124	S121
−0.95	−0.49	1.11	1.14	−0.42	−0.55	−0.43	−0.41

Table 4.5: Species effects estimates based on the AIC selected model.

Consider the following model

$$y_{ij} = \mu_i + \zeta_{119}\mathbf{1}_{\{j=119\}} + \epsilon_{ij}, \tag{4.2}$$

where ζ_{119} denotes the species effect for the rainbow trout. This model is proposed to include only a single species effect for the rainbow trout, which is a sensible approach because of its prevalence in chemical safety assessment. The standard estimators for ζ_{119} and σ are $\hat{\zeta}_{119} = -0.29$ and $\hat{\sigma} = 0.64$ respectively. This model is only a slight deviation from what is currently accepted by practitioners within the scope of $\mathcal{M}_3$, yet it offers flexibility to account for the evidential effect the rainbow trout yields on the tolerance response variable, and subsequently the HC_p for small p. We explore this parsimonious model again from other behavioural modelling perspectives later.

EFSA (2005) proposed a model for incorporating a single non-exchangeable species in the context of $\mathcal{M}_1$ although, it was restricted to the viewpoint of deriving decision rules with non-specified levels of protection; we discuss this in detail in Sections 5.1 and 5.7. Basically, the model was that conditional upon μ and σ, the log-tolerance value for a single non-exchangeable species may be envisaged as being a realisation from a normal distribution with mean $\mu - k'\sigma$ and standard deviation $\phi'\sigma$, where (k', ϕ') are the 'non-exchangeability parameters'. We describe modelling in the following chapter. The expectation of the log-tolerance value for this special species is shifted from the exchangeable model mean by $-k'\sigma$, with EFSA (2005) providing a maximum likelihood based estimate of k' (see Appendix B.3 for additional information) as $\hat{k}' = 0.45$. Substituting k' by this estimate and σ by s_p, we obtain a crude approximation to the shift: $-0.45 \times 0.65 = -0.29$ — similar to $\hat{\zeta}_{119}$.

4.5.2 Heterogeneity

The introduction of species effects into the model when either $\mathcal{M}_1$ or $\mathcal{M}_2$ is adopted makes calculations more difficult. However, the latter model can offer additional insights. Thus, in this section we explore the premise of the fixed effects model primarily from the perspective of the behavioural model $\mathcal{M}_2$; conditional heterogeneity such that we assume $(\sigma_i^2 \, \forall i \in \mathcal{G})$ is a random sample from an inverse-gamma

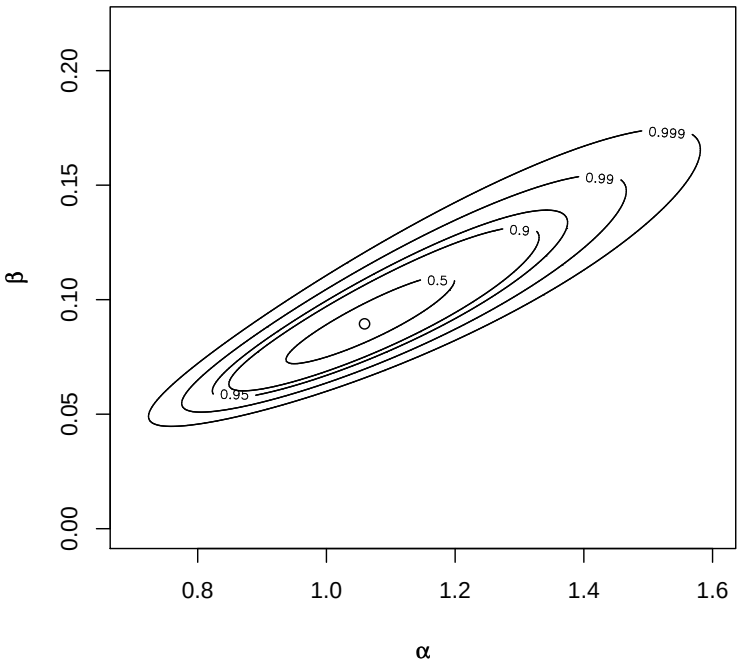


Figure 4.8: $100x\%$ confidence regions for (α, β).

distribution parameterised by shape and scale parameters α and β respectively. A method of estimating α and β by maximum likelihood methods is provided in Appendix B.2. For the RIVM fish database, it was determined that the maximum profile marginalised likelihood estimates of α and β are $(\hat{\alpha}, \hat{\beta}) = (1.05, 0.088)$. Furthermore, a method of calculating a $100x\%$ joint confidence region is shown in Appendix B.2.1, which we use to produce those displayed in Figure 4.8.

To introduce species effects in the heterogeneous model we proceed as per the homogeneous model selection by first assuming all species effects to be zero, i.e. $\zeta_j = 0 \ \forall j$. We then employ a forward stepwise BIC model selection procedure to add species effects into the model until there is no further gain. In order to be able to calculate the BIC for each sub-model, we need to obtain the likelihood function. For any given model, define $\mathcal{P}$ be a set of indices for species effects to be included in the model. Then the full data likelihood function is

$$\ell(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\zeta}) = \prod_{i \in \mathcal{G}} \prod_{j \in J_i} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_i} \exp\left\{ -\frac{1}{2\sigma_i^2} \left[y_{ij} - \mu_i - \zeta_j \mathbf{1}_{\{j \in \mathcal{P}\}} \right]^2 \right\},$$

S119	S023	S071	S055	S015	S024	S118	S153	S151
−0.22	0.29	0.45	−1.39	−2.04	0.28	−0.42	−0.33	−0.51
S145	S105	S054	S040	S035	S049	S160	S148	S152
−1.82	−3.41	−0.50	0.70	0.49	0.33	−0.54	−0.50	−0.31
S122	S041	S124	S127	S101	S139	S060	S102	S012
−0.66	1.04	−0.55	−0.22	−0.31	1.76	−0.36	−0.16	1.25

Table 4.6: Species effects estimates based on the BIC selected model.

where $\boldsymbol{\mu} = (\mu_i : i \in \mathcal{G})^T$; $\boldsymbol{\sigma^2} = (\sigma_i^2 \, \forall i \in \mathcal{G})^T$; $\boldsymbol{\zeta}$ is an $|\mathcal{P}| \times 1$ vector of species effects as indexed by $\mathcal{P}$.

Augmentation of the model, such that $(\sigma_i^2 \, \forall i \in \mathcal{G})$ is an *unobserved* random sample from an inverse-gamma distribution parameterised by shape α and scale β, requires us to work with the marginal likelihood, given by

$$\begin{aligned} \ell(\boldsymbol{\mu}, \boldsymbol{\zeta}, \alpha, \beta) &= \int_{\boldsymbol{\Sigma}} \ell(\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\zeta}) \prod_{i \in \mathcal{G}} f(\sigma_i^2 \,|\, \alpha, \beta) d\boldsymbol{\sigma^2} \\ &\propto \prod_{i \in \mathcal{G}} \frac{\beta^\alpha \Gamma(\alpha + n_i/2)}{\Gamma(\alpha)[\beta + \frac{1}{2}\sum_{j \in J_i}(y_{ij} - \mu_i - \zeta_j \mathbf{1}_{\{j \in \mathcal{P}\}})^2]^{(\alpha + n_i/2)}}, \end{aligned} \tag{4.3}$$

where $\boldsymbol{\Sigma} = (\mathbb{R}^+)^N$. Maximisation of this marginal likelihood is achieved by taking logarithms and usage of suitable software. Note that the maximum likelihood estimate of $\boldsymbol{\mu}$ is clearly a function of $\boldsymbol{\zeta}$ implying an underlying profile likelihood function. The high dimensionality of the maximisation problem requires careful handling of the maximisation routines for convergence to take place. A highly efficient strategy is to use initial starting values of $\boldsymbol{\zeta}$ derived from the corresponding homogeneous model for $\mathcal{P}$ which is very fast to calculate using standard statistical software.

Based on this stepwise model selection routine, 27 species effects were added to the model. Of the 27 species, 16 featured in the list of 24 species effects added in the homogeneous model version of the algorithm. Yet again S119 was added first to the model followed by S23 and S71; see Table 4.6 for species effect estimates and the order in which they were included into the model.

As was the situation previously, the substance effects were not greatly different from the homogeneous model analyses earlier. In addition, with the exception of

S119	S071	S023	S118	S105	S015	S040	S011
−0.15	0.51	0.37	−0.28	−3.40	−1.76	0.76	2.22
S049	S024	S153	S152	S145	S085	S160	S054
0.40	0.31	−0.22	−0.24	−1.46	1.06	−0.43	−0.37
S034	S035	S139	S132	S051	S041	S062	S016
0.33	0.55	1.84	0.13	0.09	1.08	−1.59	0.42

Table 4.7: Species effects estimates based on the homogeneous BIC selected model.

S105 and possibly S049, the species effects estimates are also reasonably similar to the homogeneous model estimates. The reason for S105 differing significantly between behavioural models $\mathcal{M}_3$ and $\mathcal{M}_2$ is likely because this species only features in two substance datasets. The maximum likelihood estimates of α and β for the final model are $\hat{\alpha} = 1.01$ and $\hat{\beta} = 0.066$, which are similar to those estimated for the exchangeable species model.

It is useful to consider that by considering σ^2 to be distributed with an inverse-gamma distribution with shape and scale parameters set to their estimates, then a heuristic estimate for the average of σ for the population of substances would be $\sqrt{\beta}\Gamma(\alpha - \frac{1}{2})/\Gamma(\alpha)$, obtained by considering the expectation of the square root of an inverse-gamma random variable. This yields 0.45 which is marginally lower than the expected value of 0.49 yielded under the species exchangeable model.

In Section 4.5.1 we identified a model in the context of $\mathcal{M}_3$ selected via a stepwise BIC model selection procedure which admitted 24 species effects into the model. If we naively assume these species effects to be present, but analyse the model from the heterogeneous $\mathcal{M}_2$ perspective, we retrieve species effects estimates as presented in Table 4.7. The compound effects have not greatly changed from the homogeneous estimates. Additionally we notice that the species effects appear to be on average reduced in absolute value, again with the noticeable difference of S105.

In this case, we retrieve $\hat{\alpha} = 1.02$ and $\hat{\beta} = 0.071$ with corresponding standard errors 0.11 and 0.013 respectively. Estimation and future predictive capabilities with respect to species effects are clearly sensitive to behavioural model choice.

Finally, consider the parsimonious model given by Equation 4.2 which allowed for the rainbow trout to be the only species effect in the model. Then the maximum

likelihood estimates of ζ_{119}, α and β are $\hat{\zeta}_{119} = -0.21$, $\hat{\alpha} = 1.00$ and $\hat{\beta} = 0.075$ respectively. Whilst the estimates of α and β have not changed significantly from the default model of zero species effects, the estimate of ζ_{119} has increased slightly from the estimate yielded under the context of $\mathcal{M}_3$ (Section 4.5.1). Considering each of the 24 stepwise models (ranging from the model including only the species effect for the rainbow trout up to the model including all 24 species effects as listed in Table 4.7), it was found that the estimate of ζ_{119} remained robust, with all estimates within the limits $(-0.15, -0.21)$. Repeating the earlier calculation of a heuristic average estimate of σ for the population of substances, we obtain $\hat{\sigma} = 0.49$ which is the same as for the species exchangeable model. Additionally, repeating the crude approximation of the average shift of the exchangeable model mean, i.e. $-k'\sigma$, with $k' = 0.45$ for the rainbow trout as described earlier, then we obtain $-0.45 \times 0.485 = -0.22$ — consistent with the value obtained here.

4.5.3 On the Choice of Model Selection Criteria

The aim of this section has been to exploratively analyse a statistical approach which can incorporate the presence of species effects which may drive the hazard assessment; this is useful for (re-) modelling in the following chapter. In particular, the approach of including all species in any proposed model is not considered sensible due to sparsity in the available database. Hence we sought a more simplified model which included only a subset of the species effects; not deviating substantially from the current scientifically accepted model. Model selection criteria are often used to choose among models by balancing adequacy of fit and model complexity. Clearly it would have been preferred to have explored all models, however this is often computationally intensive. It took us many hours to analyse approximately 4000 out of a possible 2^{172} (taking account of species effects only; substance effects are a prerequisite in the model). Moreover, databases such as the one used here are being continually updated as scientists attempt to better utilise computational power (cf. Dyer et al. 2006, 2008), hence this task will be non-trivial in future years.

We take the view of Spiegelhalter et al. (2002) who state that 'an overformal approach to model 'selection' is inappropriate'. Nonetheless stepwise model selection

routines are now commonplace tools for problems of this sort (Draper and Smith, 1998), with AIC or BIC as the typically invoked selection criterion. Spiegelhalter et al. (2002) discusses (hierarchical) model selection from the perspective of a 'focus'; the parameters of which 'should ideally depend on the purpose of the investigation' (Ibid., p. 613). Basically, we want to refine the level of hazard assessment for a *future* assessment of substance $\mathcal{S}$ not contained in the database. For example, within the context of $\mathcal{M}_2$ it is $(\alpha, \beta, \boldsymbol{\zeta})$ which is the focus. Upon defining the focus, the likelihood can be defined (Equation 4.3). As no hierarchy for $\boldsymbol{\mu}$ is provided, we included these parameters in the likelihood. Although a profile marginal likelihood approach allows us to circumvent the need for the restructuring of the likelihood. If the future assessment is to be based on $(\alpha, \beta, \boldsymbol{\zeta})$ fixed as we envisage, in order to retain tractability and reusability the application of AIC seems appropriate. From the Bayesian perspective with negligible prior knowledge, AIC has been shown to be comparable to the Deviance information criterion (Spiegelhalter et al., 2002).

BIC — an approximation of the log marginalised likelihood with an uninformative prior distribution — was the criterion applied in this section, and is implicitly criticised by Burnham and Anderson (2002) because it is reported that only where we expect a few large effects will BIC perform better than AIC in finding models 'closest to the truth'. There is debate about the choice of AIC or BIC as a model selection criterion based on the condition of consistency, which the latter satisfies but not necessarily the former. Burnham and Anderson (2004) state that the target model of AIC and BIC differs, and that BIC attains this only for asymptotic sample size. Nonetheless, it is deemed more appropriate in our opinion because a sensible degree of parsimony is required if risk managers are to adopt any proposals for refinement of the current risk assessment methodology. The inclusion of many species effects would undoubtedly lead to the 'SSDeology' becoming degenerative (discussed in Section 5.9), which would be an unacceptable scenario from the current regulatory state of the science. Note that had we been concerned with predicting within the database itself — thus changing the 'focus' — then other model selection criteria may have in fact been more appropriate. Furthermore, had we actually been concerned with seeking the (in some sense of the word) 'best' predictive model, then

AIC (or a small correction version) may be preferable. We would emphasise that our justification for BIC is informal and heuristic; for an extensive discussion of AIC and BIC we direct the reader to Burnham and Anderson (2002, 2004). We also acknowledge that other model selection criteria exist beyond AIC and BIC, some of these are discussed in Burnham and Anderson (2004), although their use is not commonplace.

4.6 Conclusions

Exploration of an RIVM fish toxicity database for pesticides has provided evidence based on two non-parametric hypothesis tests that certain fish species may be non-exchangeable with respect to others registered in the database. It is not clear how many species may be 'non-exchangeable', or to what degree. In particular, evidence was most observable for the rainbow trout possibly being non-exchangeable, which is in accordance with EFSA (2005) who also report that this is a species of particular interest because it is a frequently tested dossier species which influences many risk assessment decisions. Moreover, correction for the typical 'bias' demonstrated by the rainbow trout was shown to have a stronger influence on the estimation of HC_5 values than other less sensitive species.

Fixed effect modelling gave an indication that there may be sufficient reason to include multiple species effects within a model with even the most parsimonious model including 24 species effects terms. However, extending the current scientifically accepted models to yield more pragmatic, yet still parsimonious decision rules is a difficult task with respect to the current regulatory risk assessment process. Given the current data shortage, it seems sensible that any model which incorporates non-exchangeability — based on valid evidence — should only be driven by one or two species; most sensibly the rainbow trout for fish assessments. Whilst the modelling framework discussed in Section 4.5 forms a natural procedural tool for statistical modellers, it may only be of limited practical benefit to end users due to the requirements of more sophisticated statistical knowledge and continual review as databases are updated. More tenable approaches which do not substantially deviate

from the scientifically accepted methods would clearly be required to account for non-exchangeability.

The exploratory analysis in this chapter is for the fish taxon only. Analysis of other major taxa, e.g. birds, insects and macroinvertebrates, may also provide evidence for other non-exchangeable species. In particular, when other taxa have typical test species, e.g. *Daphnia magna* for macroinvertebrates; the Bobwhite Quail (*Colinus virginianus*) and mallard duck (*Anas platyrhynchos*) for birds, it is important that this be assessed. EFSA (2005) found no evidence of the presence of non-exchangeable species for other taxa based on the analysis of graphs similar to that of Figure 4.7. However, even the presence of just a single non-exchangeable fish species warrants further research into how one can incorporate such evidence into a risk assessment. This is especially the case as risk assessment to fish is one of the key regulatory requirements of modern chemical safety assessment. Stephan et al. (2002) recommends that one might purposefully populate estimated SSDs with recognisably sensitive species to ensure conservatism. This *ad hoc* procedure violates the statistical assumptions of the SSD definition and so we seek an alternative approach in the following chapter.

Chapter 5

Modelling Non-Exchangeability

Current probabilistic risk assessment proposals for a chemical safety assessment generally adhere to the assumption that species represented by the underlying SSD are *a priori* exchangeable. Yet, this is inconsistent with an informally recognised observation that certain species assessed are typically more sensitive than others, for example the rainbow trout (Dwyer et al., 2005; EFSA, 2005). This observation is consistent with the findings in our earlier exploratory analysis.

EFSA (2005) coined the term 'non-exchangeability' within the context of this field, although, discussion within the report was only made with respect to the adaptation of deterministic estimators and not practically discussed for methods requiring specification of the maximally permissible PAF. The focus of this chapter is therefore to describe a method for estimating the HC_p, which on the log scale we denoted ψ_p, that accounts for interspecies variability in addition to the presence of a non-exchangeable species. Since we require additional assumptions to be invoked, inclusive of those currently advocated by SSD practitioners, we explore their acceptability to end users. Emphasis is given to models which offer tractability for future risk assessments and details provided for how to separately calculate input parameters, thus allowing for reusability of decision rules.

For any model proposed that allows for the inclusion of species non-exchangeability, it is important that its complexity does not deviate excessively from the current scientifically accepted ERA modelling principles. That is why we focus on a single non-exchangeable species and exemplify our discussion using the rainbow trout.

This species has already been shown to be a special species in the regulatory arena, and is likely to be continually tested for future risk assessments. We discuss the plausibility of relaxing this restriction on a single non-exchangeable species later on.

5.1 Re-modelling

In this section we develop changes to the standard SSD concept and application by including knowledge of a single identified non-exchangeable species. We refer to such a species as a *special species* from here onwards. The title of special species is to bring attention to the possibility to adopt a slightly modified version of this work as to account for another problem of SSDs — the inability to adequately protect endangered/desirable species. However, we will not consider the extension here.

The notation used in earlier chapters is modified such that a log-transformed (base 10) tolerance value for species $j \in J^*_{\mathcal{S}}$ assessed with substance $\mathcal{S}$ is denoted y_j, where $J^*_{\mathcal{S}}$ is the collection of non-special species tested. We denote $y^\dagger$ to be the single log-tolerance value for the special species; in our example the rainbow trout. In addition, we suppose $|J^*_{\mathcal{S}}| \triangleq n^* = n - 1$, so that decision rules developed are comparative to their exchangeable counterparts for fixed sample sizes.

We begin by first describing the model initially proposed by EFSA (2005). The model is that for the n^* species in $J^*_{\mathcal{S}}$, y_j is conditionally independently distributed normal with mean μ and variance σ^2, which is consistent with the current SSD model in the context of species exchangeability. For the special species, $y^\dagger$ is conditionally independent normal with mean $\mu - k'\sigma$ and standard deviation $\phi'\sigma$. Thus when $k' = 0$ and $\phi' = 1$ we retrieve the completely exchangeable model among all tested species. In situations other than this we say that the special species is *non-exchangeable* with respect to the other *exchangeable* species.

The predictive distribution of the special species' tolerance value is such that the usual log-SSD mean has been shifted by $-k'\sigma$ and the usual log-SSD standard deviation multiplied by ϕ'. The parameters k' and ϕ' need to be defined with reference to some suitable population of substances. By allowing k' to be the same across substances, the shift of $-k'\sigma$ maintains that the expected position of the

special species in the log-SSD is to be unaffected by the variability of σ across the population of substances. The application of this model for estimating the HC_p by EFSA (2005) assumed the parameters k' and ϕ' are known.

Although the predictive distribution for the single special species is appealing, when incorporated into the methods of risk measure/control discussed within this research, tractability is lost. This is unsatisfactory and so we modify the predictive distribution of the special species so that it has expectation $\mu - k$ over log-concentration. The expected position of the special species in the SSD is now affected by the variability of σ across the population of substances. However we later show that evidence does not overwhelmingly support the hypothesis that the shift should be proportional to σ. It should be noted that the amount of data available for testing this assumption maybe potentially masking any distinction. The standard deviation of the special species' predictive (log-)distribution is $\phi\sigma$, although taking the same form, ϕ is now different from ϕ' by virtue of the change in model. One should view k as representing the 'bias' of the special species (on the log-concentration scale), and ϕ as the allowance for a different variance (dimensionless).

It is interesting to note that this model is very similar to the model described by Equation 4.2, such that only one species effect (the rainbow trout's) was included in the model and $\phi = 1$. A generalisation of this model is

$$\begin{aligned} y_j &= \mu + \zeta_{119}\mathbf{1}_{\{j=S119\}} + \epsilon_j, \\ \epsilon_j \,|\, \sigma, \phi &\sim \begin{cases} N(0, \sigma^2) & \text{for } j \neq 119 \\ N(0, [\phi\sigma]^2) & \text{for } j = 119 \end{cases} \end{aligned}$$

where $\zeta_{119} \equiv -k$.

For the remainder of this chapter we will restrict our discussion to the behavioural models: $\mathcal{M}_1$ — which has scientific and regulatory acceptance — and $\mathcal{M}_2$, the reason for this is two-fold. First, the basis of $\mathcal{M}_3$ has not been established and its assumption remains unvalidated for application, at least in the context of aquatic ERA of fish. Second, as discussed in Sections 2.7 and 3.8.2, it is not clear whether $\mathcal{M}_4$ is equivalent to the other behavioural models with regards to the level of uncertainty refinement. Accounting for non-exchangeability under these perspectives

might therefore be inappropriate. However, we note that it is relatively straightforward to extend the concepts here to the perspective of $\mathcal{M}_3$ and $\mathcal{M}_4$ if so warranted.

5.2 Posterior Distributions

In this section we give details of the posterior distribution for the log-SSD mean and variance parameters based on observing some toxicity data for a substance $\mathcal{S}$ under assessment. The current approach thus far has been to update non-informative prior distributions in order to obtain a posterior distribution. It seems logical to maintain the same prior distributions, since the role of subjectively elicited prior distributions for probabilistic risk assessment of chemical safety has yet to explored for its merit except for in a few reports, for example Grist et al. (2006), O'Hagan et al. (2005) and Hickey et al. (2008). Furthermore, there would likely be hesitation or resistance regarding the introduction of prior knowledge at the lower-intermediate quantitative tier of risk assessment we are concerned with; this does not preclude its use, nor limit potential acceptance at higher tiers.

For now we shall assume that all hyper-parameters k, ϕ, α and β are precisely known or specified for the computation of calculating the posterior distributions. The assumption of non-exchangeability ideally would not change a risk managers beliefs should they have specified any on the prior distributions. Therefore for $\mathcal{M}_1$ and $\mathcal{M}_2$, we assume precisely the same non-informative prior distributions we used in the species exchangeable context since these parameters hold the same operational interpretation in both exchangeable and non-exchangeable models. Hence, for $\mathcal{M}_1$ we assume, *a priori*, that $p(\mu, \sigma^2) \propto \sigma^{-2}$ for $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$. For $\mathcal{M}_2$, the prior distribution of $\sigma^2 \in \mathbb{R}^+$ is given by the hierarchical model, i.e. an inverse-gamma distribution with shape α and scale β, and for μ we assume $p(\mu) \propto 1$ for $\mu \in \mathbb{R}$. We now describe the posterior distributions of these parameters for the SSD of substance $\mathcal{S}$ under both behavioural models; derivation is a straightforward extension of the posterior distributions described in Appendix A.1 for the species exchangeable modelling context.

5.2.1 $\mathcal{M}_1$

$$\begin{aligned} \mu \,|\, \sigma^2, k, \phi; \mathbf{Y} &\sim NOWN\left(\hat{\mu}, \frac{\sigma^2}{\phi^{-2}+n-1}\right) \text{ for } \mu \in \mathbb{R}; \\ \sigma^2 \,|\, k, \phi; \mathbf{Y} &\sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{n-1}{2}\hat{\sigma}^2\right) \text{ for } \sigma^2 > 0, \end{aligned} \tag{5.1}$$

where

$$\begin{aligned} \hat{\mu} &= \frac{\phi^{-2}(y^\dagger + k) + \sum_{j \in J_{\mathcal{S}}^*} y_j}{\phi^{-2} + n^*} \\ &= \frac{\phi^{-2}(y^\dagger + k) + n^* \bar{y}^*}{\phi^{-2} + n^*}; \end{aligned} \tag{5.2}$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1}\left[\phi^{-2}(y^\dagger + k - \hat{\mu})^2 + \sum_{j \in J_{\mathcal{S}}^*} (y_j - \hat{\mu})^2\right] \\ &= \frac{1}{n-1}\left[\phi^{-2}(y^\dagger + k - \hat{\mu})^2 + n^*(\hat{\mu} - \bar{y}^*)^2 + (n^*-1)s^{*2}\right] \end{aligned} \tag{5.3}$$

and $\bar{y}^*$ and s^* are the usual mean and standard deviation of the n^* tolerance values for the tested species in $J_{\mathcal{S}}^*$.

Note that $\hat{\mu}$ and $\hat{\sigma}^2$ are the usual weighted least squares estimators from the frequentist perspective. Hence, sampling distributions can easily be determined and frequentist based decision rules derived; we focus solely on the Bayesian viewpoint.

5.2.2 $\mathcal{M}_2$

$$\begin{aligned} \mu \,|\, \sigma^2, k, \phi; \mathbf{Y} &\sim N\left(\hat{\mu}, \frac{\sigma^2}{\phi^{-2}+n-1}\right) \text{ for } \mu \in \mathbb{R}; \\ \sigma^2 \,|\, k, \phi, \alpha, \beta; \mathbf{Y} &\sim \mathcal{IG}\left(\frac{2\alpha+n-1}{2}, \frac{2\alpha+n-1}{2}\tilde{\sigma}^2\right) \text{ for } \sigma^2 > 0, \end{aligned} \tag{5.4}$$

where

$$\tilde{\sigma}^2 = \frac{2\beta + (n-1)\hat{\sigma}^2}{2\alpha + n - 1}. \tag{5.5}$$

It can be determined that from the frequentist viewpoint, if one incorporates sampling σ^2 from an inverse-gamma population then $\mathbb{E}[\tilde{\sigma}^2/\sigma^2] = 1$; see for example Section 2.6.5.

5.3 Estimators

In this section we describe some of the estimators which derive from the proposed model for non-exchangeability. It is important to emphasise that it is only the assumptions surrounding the data generating mechanism which have changed; the behavioural models and methods for measuring or controlling risk remain consistent with earlier chapters. Another important observation is that estimators are all of canonical form. In the exchangeable context, $\delta_p(\mathbf{Y})$ was a linear combination of the mean of the toxicity data and a multiple of the standard deviation, i.e. $\bar{y} - \kappa_p\tilde{\sigma}$. The situation under non-exchangeability here remains the same except adjustments have been made to each part; $\bar{y}$ is replaced by $\hat{\mu}$, $\tilde{\sigma}$ is modified accordingly for $\mathcal{M}_1$ and $\mathcal{M}_2$ separately and κ_p is adjusted to reflect the uncertainty in the SSD. In all cases, κ_p maintains its property of being independent of the toxicity data for $\mathcal{S}$. For the remainder or this chapter, unless specified otherwise, notations such as δ_p, κ_p and $\hat{\sigma}$ will all be made with reference to the non-exchangeable model.

5.3.1 [EFSA] Estimators

The theory behind [EFSA] estimators was discussed in Section 2.6.4. As per the exchangeable context, we consider the posterior predictive distribution of $(Y - \hat{\mu})/(\tilde{\sigma}\sqrt{1+\psi^{-2}})$, where $\psi^2 = \phi^{-2} + n^*$ (recalling that setting $\alpha = \beta = 0$ results in $\tilde{\sigma} = \hat{\sigma}$, i.e. $\mathcal{M}_1$; otherwise $\mathcal{M}_2$), which is determined to be a Student t-distribution, with $\pi = 2\alpha + n - 1$ degrees of freedom. Hence, the general [EFSA] decision rule for $\mathcal{M}_2$ is

$$\delta_p(\mathbf{Y} \mid k, \phi, \alpha, \beta)_{[\text{EFSA}]} = \hat{\mu} - \kappa_p(n, \alpha)\tilde{\sigma},$$

where

$$\kappa_p(n, \alpha) = \sqrt{1+\psi^{-2}}t_{\pi,p}.$$

To obtain the decision rule for $\mathcal{M}_1$, set $\alpha = \beta = 0$. Additionally, as before, values of κ_p for fixed n and p are non-comparable between behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$ because the actual assessment shift-factors are obtained by multiplying each by different standard deviation estimates. Setting $k = 0$ and $\phi = 1$ yields the

estimators derived in the exchangeable context.

An interesting observation can be made by considering the ratio of κ_p for the exchangeable and non-exchangeable models for fixed n and p. For $\phi < 1$ we have that $(\phi^{-2} + n - 1) > n$, so one can argue that ϕ decreasing represents an increase of $\phi^{-2} - 1$ additional species. Since the degrees of freedom in the Student t-percentile are uninfluenced by ϕ this interpretation is limited.

5.3.2 [AJ] Estimators

The [AJ] estimators have been extensively discussed throughout this report and have been shown to be important for many reasons; see Sections 2.6.3 and 3.4 for example. Essentially the class of estimators is defined to be the $100(1-\gamma)$-th percentile of the posterior distribution of $\psi_p(\theta)$. Since $\psi_p(\theta)$ is defined in reference to the exchangeable species SSD, it is required that we be specific about what we are protecting; to aid in the flow of this section we delay this discussion until Section 5.4. The [AJ] decision rule for $\mathcal{M}_2$ is defined to be

$$\delta_p^{(\gamma)}(\mathbf{Y} \mid k, \phi, \alpha, \beta)_{[\mathrm{AJ}]} = \hat{\mu} - \kappa_p(n, \alpha, \phi, \gamma)\tilde{\sigma}.$$

By observing that the pivotal random quantity

$$\left[\frac{\hat{\mu} - \mu}{\sigma/\sqrt{\phi^{-2} + n^*}} + K_p\sqrt{\phi^{-2} + n^*}\right] \Bigg/ [\tilde{\sigma}/\sigma]$$

has a non-central t-distribution with $\pi = 2\alpha + n - 1$ degrees of freedom and non-centrality parameter $\eta = K_p\psi$, it is straightforward to deduce that $\kappa_p\sqrt{\phi^{-2} + n^*}$ is equal to the 100γ-th percentile of this distribution. Hence,

$$\kappa_p(n, \alpha, \phi, \gamma) = \psi^{-1}F^{-1}_{T_{\pi,\eta}}(\gamma). \tag{5.6}$$

Again, setting $\alpha = \beta = 0$ yields the decision rule for $\mathcal{M}_1$. As was the case under the exchangeable model description, the decision rule is valid subject to the same interpretation as before under the frequentist viewpoint. Moreover, as $k = 0$ and $\phi = 1$, the decision rules coincide with their frequentist counterparts.

5.3.3 [LINEX] Estimators

Although we do not wish to detract from the key theme of this chapter, we additionally present the [LINEX] Bayes rule estimator deriving from the scaled LINEX loss function described in Section 3.5. This has many appealing properties, and although superfluous relative to current requirements, it is easily adapted as the other forms of risk measurement and control (above) have been. The [LINEX] estimator for $\mathcal{M}_2$ is

$$\delta_p^*(\mathbf{Y} \mid k, \phi, \alpha, \beta, \lambda) = \hat{\mu} - \kappa_p^*(n, \alpha, \phi, \lambda)\tilde{\sigma},$$

where $\kappa_p^*(n, \alpha, \phi, \lambda)$ is the unique solution to

$$\int_0^\infty t^{\alpha+(n^*-1)/2} \exp\left\{-\lambda\kappa_p^*\sqrt{t} - \left(\alpha + \frac{n-1}{2}\right)t\right\} dt = \Gamma\left(\alpha + \frac{n}{2}\right)\left[\alpha + \frac{n-1}{2}\right]^{-\alpha+\frac{n}{2}} \exp\left\{-\lambda\left[K_p + \frac{\lambda}{2[\phi^{-2}+n-1]}\right]\right\}$$

for κ_p^*. Again, setting $\alpha = \beta = 0$ yields the decision rule for $\mathcal{M}_1$. Derivation of this estimator follows the same method as in the species exchangeable context (consult Appendix C.4) in conjunction with the revised posterior distributions defined by Equations 5.1–5.4.

Note that the specified value of λ which fixes the loss function as some level of preference for the risk manager will not change between the exchangeable and non-exchangeable contexts.

5.3.4 Discussion

Given values for k and ϕ in addition to any other necessary parameters, then it is a simple exercise to tabulate assessment shift-factors for a range of sample sizes n and maximum permissible PAF levels p. These can then be used by risk assessors alongside estimates of $\hat{\mu}$ and $\tilde{\sigma}$, which are straightforwardly calculated using Equations 5.2 and 5.5 respectively to yield an estimate of the HC_p. Otherwise, it is simple to produce simple software applications, or perhaps modify current software applications such as the RIVM's E_TX (Van Vlaardingen et al., 2004) program,

which risk assessors and managers can use. It is beyond the remit of this research to provide such tables for risk managers, however we acknowledge the appeal and use of them which strengthens the potential for adoption of these decision rules for use in regulatory risk assessment.

Since [AJ] and [EFSA] estimators are of canonical form, a comparison of them can be made by considering the assessment shift-factors κ_p for fixed n. In Figure 5.1 we plot interpolated κ_5 values corresponding to the [EFSA] (dashed curves) and (median, $\gamma = 0.50$) [AJ] (solid curves) estimators for $\mathcal{M}_1$ and a sample size range of $3 \leq n \leq 20$. We speculate that sample sizes above $n = 20$ are unlikely to be achievable in practice for risk assessment especially for individual taxon SSDs. Although both estimators are valid for $n = 2$, we could not achieve adequate accuracy in the calculation of κ_5, hence they are not included in the comparison. It is concluded from Figure 5.1 that the estimators exhibit the same properties as those derived within an exchangeable context, i.e. for fixed sample size the [EFSA] estimator is more conservative than the median [AJ] estimator.

The effect of the non-exchangeability parameter ϕ (corresponding to different coloured curves in Figure 5.1) on assessment shift-factors is far more pronounced for [EFSA] than for median [AJ], even at a sample size of $n = 20$. The effect of ϕ on conservative [AJ] estimators ($\gamma = 0.95$) — which yield more conservative estimators relative to [EFSA] — was still observable even at larger values of n, although at small values the differences were within the order of the effects observed for [EFSA]. In the interest of visual clarity, corresponding curves for $\gamma = 0.95$ are not overlaid in Figure 5.1.

A particularly interesting observation is that for fixed n, as ϕ increases, so does the [EFSA] rule for κ_5, yet the [AJ] rule for κ_5 decreases minutely. This is better visualised in Figure 5.2 (left panel) which plots κ_5 against ϕ ($0.1 \leq \phi \leq 2.1$) for a range of different sample sizes. The magnitude of the rate of change with respect to ϕ is greater for [EFSA] values than for the median [AJ] values, again reflecting the relative insensitivity of the latter on changes in ϕ. The right panel of Figure 5.2 shows a magnified region of the left panel indicating this point.

This phenomenon of the [AJ] κ_p values is perplexing and might appear paradox-

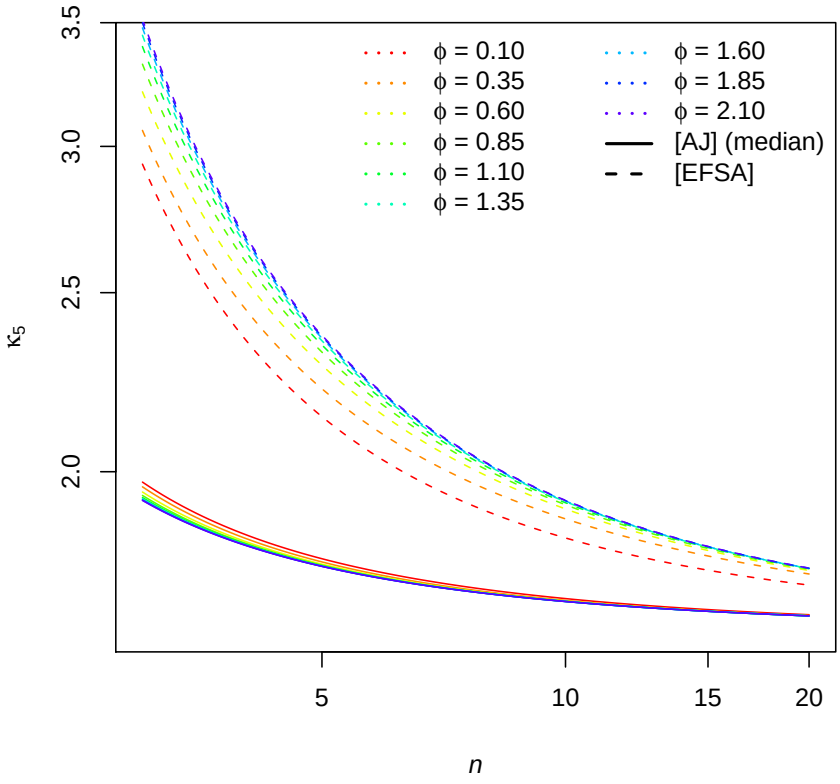


Figure 5.1: Plot of κ_5 values extrapolating to [EFSA] and median [AJ] estimators against n for varying ϕ ($\mathcal{M}_1$).

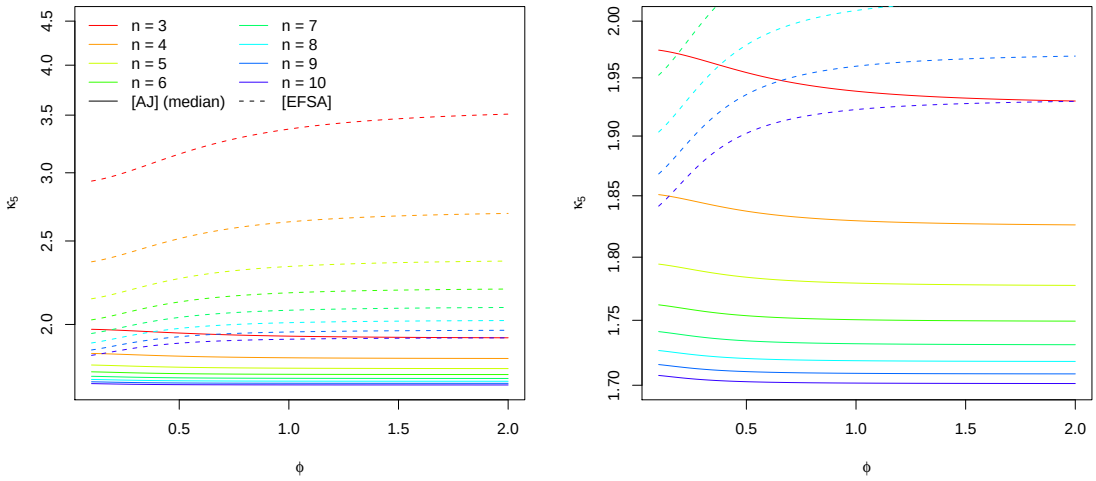


Figure 5.2: Plot of κ_5 values extrapolating to [EFSA] and median [AJ] estimators against ϕ for varying n ($\mathcal{M}_1$). Right panel: magnified region of κ_5.

ical. Whilst this property is observable for all p, it only occurs for certain ranges of γ. For $\gamma = 0.95$ — which has been suggested for application — the κ_5 values increase as ϕ increases for both the [EFSA] and [AJ] estimators. This is immediate by plotting Equation 5.6 against γ for different values of ϕ with fixed n and p; for example see Figure 5.3 for the case of $n = 8$, $p = 5$. These scaled non-central t-quantile functions intersect each other at different points (although not apparent from the figure), such that the numerical ordering changes as γ does. In particular, at $\gamma = 0.50$ (middle grey dashed line) — yielding κ_5 which extrapolates to the median [AJ] estimator — the ordering is counterintuitive. We can additionally deduce from this figure that the effects on κ_p by ϕ are much larger for the tail ends of γ in comparison to $\gamma = 0.50$, including $\gamma = 0.95$.

Rather than setting $\kappa_p\sqrt{n^* + \phi^{-2}}$ to be the 100γ-th percentile of a non-central t-distribution with $n-1$ degrees of freedom and non-centrality parameter $K_p\sqrt{n^* + \phi^{-2}}$, we could fix $\kappa_p\sqrt{n^* + \phi^{-2}}$ to be the expectation of a random variable with this distribution. This would correspond to the [SEL] Bayes rule estimator; $\kappa_p(n)$ is then defined, independently of ϕ, according to Equation 3.4 with $\alpha = 0$. In the example provided by Figure 5.3, this assessment shift-factor is indicated as a pink dashed line. It is observed that for any value of γ exceeding the corresponding value where $\kappa_p(n, \phi, \gamma)$ and the [SEL] assessment shift-factor intersects, the ordering is consistent with the [EFSA] assessment shift-factors. For a wider range of choices of n and $p < 50$, we have numerically confirmed this re-ordering holds true. Thus we would conjecture that it is true for all $n > 3$ and $p < 50$, or equivalently, that $F_{T_{\nu,\eta}}(\mathbb{E}[T_{\nu,\eta}])$ is monotonically increasing for η where $T_{\nu,\eta}$ is a random variable that follows a non-central t-distribution with ν degrees of freedom and non-centrality parameter $\eta > 0$.

We note, however, that that since the HC_p estimators derived here depend on ϕ through $\hat{\mu}$ and $\hat{\sigma}$, as well as κ_p, then overall interpretation of ϕ is somewhat limited.

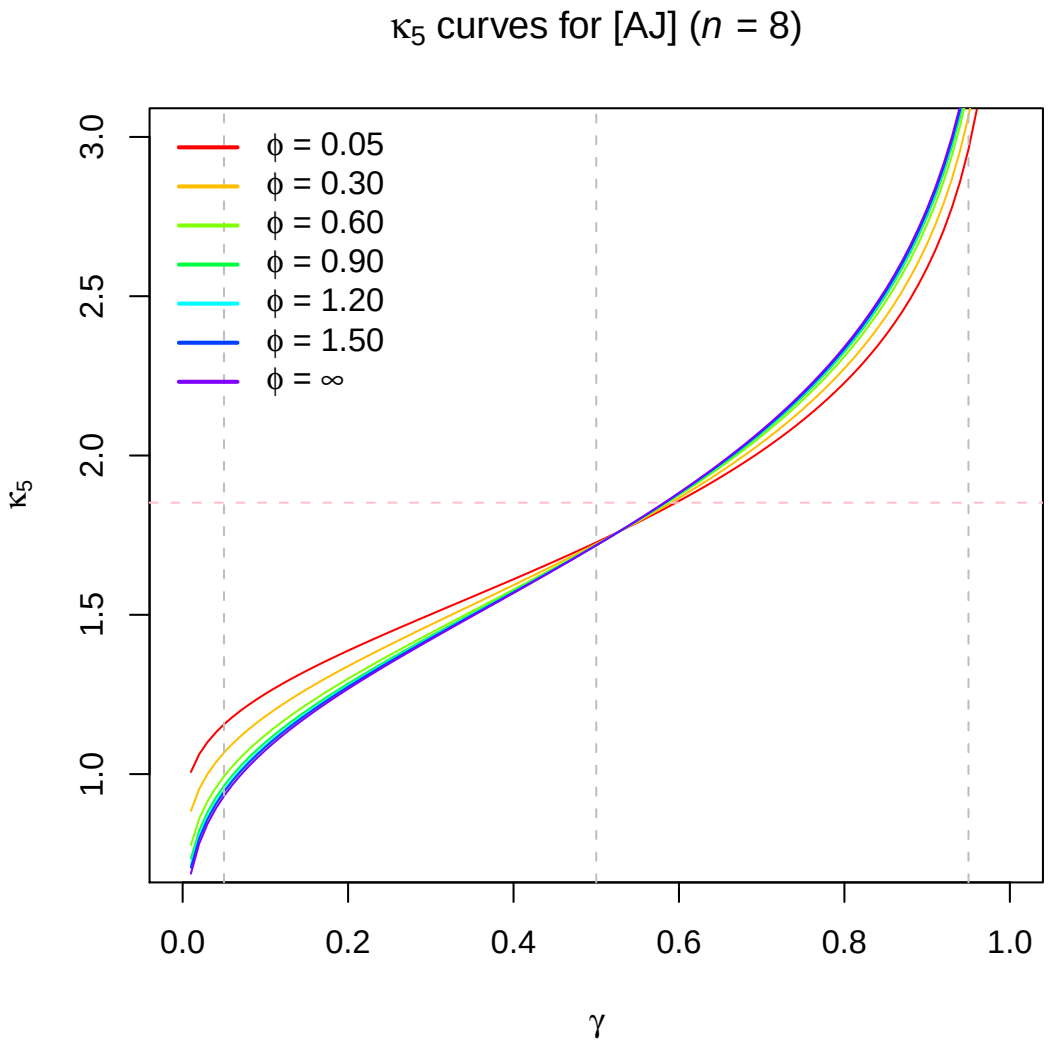


Figure 5.3: Plot of [AJ] $\kappa_p(8, \phi, \gamma)$ (Equation 5.6) against γ for different ϕ values under $\mathcal{M}_1$. Grey dashed lines correspond to values for extrapolating to the upper (5%), median (50%), and lower (95%) [AJ] estimators. Pink dashed line is defined by Equation 3.4 (with $\alpha = 0$).

5.4 Interpretation Under Non-Exchangeability

When we allow for non-exchangeability, clarity is required about what exactly we mean by the 'SSD' and 'HC_p'. In the strictest sense, the SSD is the empirical cumulative distribution function of sensitivity (tolerance) for a well-defined population of species. The inability to observe the toxicological endpoint for every species in this population calls for the use of statistical constructs, namely the probabilistic SSD.

In this framework the HC_p is interpreted as the concentration of the given toxicant which will affect a randomly selected species from the assemblage with probability $p\%$ (Posthuma et al., 2002b). In the context of species exchangeability, the interpretation of the SSD is contentious since the statistical population (species community) it represents is poorly defined; Aldenberg et al. (2002) referred to this as 'the Achilles heel of SSDeology'. Moreover, the standard statistical assumptions neglect that the population is finite. Despite this, the non-exchangeable species perspective requires additional consideration.

From the modelling of a single non-exchangeable species, not every species is identically distributed. Hence, the 'true' SSD which we assume to be *a priori* normal with mean μ and variance σ^2 under species exchangeability differs under the non-exchangeability model since it only represents those species considered exchangeable. Thus, the SSD in this case is the empirical cumulative distribution function of tolerance for a re-defined population of species. The quantity we have sought to estimate through this research is $\psi_p(\theta) = \mu - K_p\sigma$ — the p-th percentile of a normal distribution of tolerance over log-concentration — and we still attempt to estimate this quantity within this chapter. One viewpoint would be that we are utilising the information from the special species tolerance value to increase accuracy in the estimation of the SSD for the ordinary species, in which sense the estimator δ_p would represent the ordinary species. Alternatively, one may take the viewpoint that we are correcting the special species' tolerance value based on historical evidence so that on average it is exchangeable, and thus accounted for in the estimator δ_p. The two viewpoints are effectively the same in the context of this research.

Returning to the issue of population definition, we have already discussed (cf. Section 2.4) that the SSD concept is at least in the regulatory context intended

to represent multi-taxa communities. From the exchangeable species viewpoint, the adequacy of the unimodal log-normal SSD for representing such species communities is contentious. Forbes and Calow (2002b) and Duboudin et al. (2004a) report that there are significant over- and under-representations of taxonomic groups used in ERA. Even when considering a relatively broad single species group, e.g. macroinvertebrates, Hickey et al. (2008) found that weighting according to taxonomic order can noticeably influence the HC_p estimate. This issue is pertinent to the non-exchangeable viewpoint also. For practical reasons we define non-exchangeability in reference to a well defined population; in this research, namely the exposition of fish to pesticide stressors.

5.5 Hyper-parameter Specification

In this section we give details of the estimation of the hyper-parameters used in the models described: k and ϕ for $\mathcal{M}_1$; k, ϕ, α and β for $\mathcal{M}_2$. As discussed, we assume these parameters are fixed precisely in advance, so that they can be used as plug-in values for the estimators described in Section 5.3. We have not treated these hyper-parameters as uncertain because this will cause tractability to be lost, however we will later revisit whether this has any important effects on the estimators.

EFSA (2005) provided a method for: (a) estimating (α, β) for estimators of $\psi_p(\theta)$ derived under the model $\mathcal{M}_2$ within a species exchangeable context; and (b) estimating (k', ϕ') for a model used to account for a single non-exchangeable species which we described in Section 5.1. Although ϕ' bears resemblance to ϕ, the EFSA plug-in value is not valid in the model we propose here. Details of the methods used by EFSA (2005) for calculating (α, β) and (k', ϕ') are described in Appendices B.2 and B.3 respectively.

We assume we can represent a toxicity database $\mathcal{G}$, such as the RIVM fish toxicity database used throughout this research, as $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$. $\mathcal{G}_1$ is the group of substances deemed to be relevant to the substance under current assessment $\mathcal{S}$, which have all been assessed with the special species. $\mathcal{G}_2$ is the group of substances which are relevant for the estimation of α and β, hence relevant to $\mathcal{S}$. In EFSA (2005), $\mathcal{G}_2$

is effectively defined to be all pesticides tested for each distinct taxonomic group, e.g. fish, birds, etc. Decisions and agreement regarding the relevance of additional substances to that of $\mathcal{S}$ falls under the remit of the risk managers role and scientific experts. Note that when working in the context of $\mathcal{M}_2$ it will be necessary to estimate all hyper-parameters (k, ϕ, α, β) simultaneously as they are linked in the likelihood function, which we describe later.

We will make the assumption that $\mathcal{G}_1 \subseteq \mathcal{G}_2$, i.e. the group of substances used to estimate the non-exchangeability parameters are a subset of those used to estimate the heterogeneity parameters. We believe this to be a reasonable assumption, especially when the single special species is a frequently assessed dossier species of which a relatively large amount of data is available. Whilst it might be possible that one would want to estimate the non-exchangeability parameters using substance datasets not relevant to the heterogeneity behavioural model, it seems unlikely.

For our illustrative example we define $\mathcal{G}_2$ to be the RIVM fish toxicity database, which is in keeping with EFSA (2005). We define $\mathcal{G}_1$ to be the collection of substances within the RIVM fish toxicity database which have been tested on the special species subject to the condition that $n_i^* \geq 2$ for each $i \in \mathcal{G}_1$, i.e. at least two ordinary species have been assessed with each chemical in addition to the special species. The latter condition was adopted by EFSA (2005) and so in the interests of comparison we maintain the restriction. Furthermore, the condition removed computational issues encountered when $n_i^* = 1$ regarding the maximisation routines applied to the posterior distributions. By defining v_i to be the number of substances contained in $\mathcal{G}_i$, we have $v_1 \leq v_2 \leq N$ substances used in estimation procedure overall.

For each substance $i \in \mathcal{G}_1$, we denote y_{ij} as the log-tolerance value for species $j \in J_i^*$, with $|J_i^*| = n^*$. Each y_{ij} is assumed to be a realisation from a normal distribution with mean μ_i and standard deviation σ_i. In addition, we denote $y_i^\dagger$ as the log-tolerance value of the special species tested with substance i, and assume it is a realisation from a normal distribution with mean $\mu_i - k$ and standard deviation $\phi\sigma_i$. For each substance $i \in \mathcal{G}_2 \backslash \mathcal{G}_1$, we denote the log-tolerance value of species $j \in J_i$ as y_{ij}, which we assume is a realisation from a normal distribution with mean μ_i and standard deviation σ_i. Additionally, $|J_i| = n_i$.

In the interests of clarity we only provide background details of the estimation procedures in the following sections. Technical details are provided in Appendices B.4 and B.6.

5.5.1 $\mathcal{M}_1$

The method for estimation of k and ϕ is to calculate the joint posterior mode of the marginalised posterior distribution of unknown parameters; such an estimator is known as the *maximum a posteriori* ($\mathcal{MAP}$) estimator. To do this we first construct the likelihood function for all data in $\mathcal{G}_1$ and then multiply this by the product of independent prior distributions for $(\mu_i, \sigma_i^2)\ \forall i \in \mathcal{G}_1$, namely $p(\mu_i, \sigma_i^2) \propto \sigma_i^{-2}$ for $\mu_i \in \mathbb{R}$ and $\sigma_i^2 \in \mathbb{R}^+$. In addition, we multiply by the prior distribution of k and ϕ. We will assume *a priori*, $p(k, \phi) \propto 1$ for $k \in \mathbb{R}$ and $\phi > 0$. The anticipated hesitation of regulators and stakeholders in adopting subjective prior distributions has already been noted earlier on. Nevertheless, one might argue that expert knowledge is useful for the specification of prior distributions for the non-exchangeability parameters. Practical suggestions on this which might overcome such hesitation would likely only yield very wide uniform distributions. We do not consider this case any further.

The un-normalised posterior distribution is then obtained, upon which integration with respect to the nuisance parameters $(\mu_i, \sigma_i^2)\ \forall i \in \mathcal{G}_1$ leaves one with the marginalised posterior distribution for k and ϕ. Maximisation of this distribution yields the $\mathcal{MAP}$ estimator which is what we use as our fixed plug-in values. In addition, joint modal estimators allow us to calculate the Hessian and subsequently approximate the joint posterior distribution using a Laplace approximation (Schervish, 1995, pp. 446–448) which we describe later on.

5.5.2 $\mathcal{M}_2$

The assumption that $\mathcal{G}_1 \subseteq \mathcal{G}_2$ makes the specification of prior distributions simpler under the behavioural model of $\mathcal{M}_2$ (since the specification of a prior for SSD parameters of substances in the latter set will account for those in the former set). The prior distribution for each μ_i remains as per before, i.e. $p(\mu_i) \propto 1$ for $\mu_i \in \mathbb{R}$

$\forall i \in \mathcal{G}_2$. The same applies for each σ_i^2, i.e. distributed with an inverse-gamma distribution with shape α and scale β for $\sigma_i^2 \in \mathbb{R}^+ \; \forall i \in \mathcal{G}_2$. If we had not assumed $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then this might complicate the estimation procedure since different prior distributions would need to be specified for substances in $\mathcal{G}_1$ and not in $\mathcal{G}_2$. In addition to the priors for each μ_i and σ_i^2, we also apply the same prior distribution for k and ϕ as per $\mathcal{M}_1$, and apply $p(\alpha, \beta) \propto 1$ for $\alpha > 0$, $\beta > 0$. Hence, deriving the marginalised posterior distribution for the hyper-parameters, as described for $\mathcal{M}_1$ and maximising will yield the joint $\mathcal{MAP}$ estimator.

5.5.3 On the Reusability of Hyper-parameter Estimates

It is reasonable based upon the structure of estimation procedures described above, that a risk manager would not need to be in possession of the databases used to estimate the hyper-parameters. Subject to the identification of the single special species and class of relevant substances, it should be sufficient to simply specify the hyper-parameters. EFSA (2005) published values of α and β for five different taxonomic groups; they also published values of k' and ϕ' for the rainbow trout. However, conditional on the hyper-parameters for the given behavioural model, the prior distribution for each substance is independent by construction, and so the posterior distribution is a sufficient summary for the assessment of $\mathcal{S}$. This sufficiency allows for the posterior distribution to be published without the need for access to the raw data within the database. This would be attractive to risk managers who might otherwise be unsure of the practicality of the methodology here.

The disclosure of databases is an issue of concern in modern day risk assessment. This has motivated many researchers to propose methods for estimating tolerance values of untested species based on the tolerance value of a surrogate species (Dyer et al., 2006, 2008; Jager et al., 2007), or even based on knowledge of the chemical structure (Cronin et al., 2003). However it is vital that true observations be used to estimate parameters in our models especially the non-exchangeability parameters, otherwise unidentified systematic errors may compound. Large databases are currently publicly available such as the US EPA ECOTOX database (US EPA, 2007), however the process of cleaning such data is time consuming and has led to organi-

	k	ϕ	α	β
$\mathcal{M}_1$	0.195	0.702	–	–
$\mathcal{M}_2$	0.205	0.656	1.523	0.315

Table 5.1: Posterior hyper-parameter $\mathcal{MAP}$ estimates.

sations protecting their investment. As an example, the RIVM have kindly given us permission to use their database, but subject to agreement that it is not disclosed publicly. Hence the sufficiency of the posterior distribution discussed above would also be attractive to those with proprietary rights.

Making the posterior distributions publicly available is permitted. Moreover, posterior distributions can be used to assess whether fixed plug-in hyper-parameter estimates are plausible. Of course, the latter requires additional analysis, which we discuss later on in this chapter. In addition, the posterior distributions can be updated as more data is made available; this in principal would be the ideal situation. However, lack of formal cooperation and impartial resources between competing chemical companies means this is unlikely to be achievable in practice. It is also important to realise that if the substance $\mathcal{S}$ being assessed cannot be described as similar to others held in the database, for example due to a highly specific or unknown mode of action, then it may not be possible to utilise this methodology.

5.5.4 Example: RIVM Fish Database

In this section we apply the methodology to the species identified in Chapter 4 as presenting strong evidence of non-exchangeability — the rainbow trout. Joint modal estimates for each behavioural model are given in Table 5.1.

Initial interpretation of the parameter estimates indicates that the estimates for $\mathcal{M}_1$ and $\mathcal{M}_2$ of k and ϕ are similar. The estimate of ϕ' which was determined in EFSA (2005) using frequentist methods was $\phi' = 0.62$, and although it isn't comparable, *per se*, to the ϕ estimates presented here, it is reasonably close. In addition, the incorporation of non-exchangeability into our model also limits the validity of juxtaposition of the heterogeneity parameter estimates for $\mathcal{M}_2$ to their

species-exchangeable counterparts. However, one does notice quite a large difference between those in Table 5.1 and those displayed in EFSA (2005) (derived in Appendix B.2). Yet this is unlikely to be a consequence of the change in modelling, but rather the method of estimation; here we have a posterior modal estimate, where as in EFSA (2005) a frequentist maximum likelihood estimate was sought. Ignoring non-exchangeability, the Bayesian posterior distribution differs by a shift of $-1/2$ to some of the shape and scale parameters in the function to be maximised when compared to the structure of the marginalised profile likelihood function (Equation B.2); support regions for the heterogeneity parameters are shown in Figure 4.8. This is a result of gaining an additional factor of σ_i after integrating out μ_i conditional upon σ_i.

Comparing the estimate of k for $\mathcal{M}_2$ to the estimate of the species effect estimate $-\hat{\zeta}_{119} = 0.206$ given in Section 4.5.2, we realise that the estimates coincide. However, the model in Section 4.5.2 assumed $\phi = 1$, whereas here it has been estimated to be 0.656.

5.6 Hyper-parameter Uncertainty

It is necessary to report uncertainty in a risk assessment. If a risk manager is to consider adopting adjusted estimators, it is required that the uncertainty in the hyper-parameters is formally evaluated, as well as the consequences for treating them fixed in respect to the estimation of the HC_p. In this section we explore how to assess these uncertainties with our running example of the rainbow trout representing the special species.

Although EFSA (2005) separately considered the estimation of: (i) (α, β) in a species exchangeable context of $\mathcal{M}_2$; and (ii) (k', ϕ') for a different version of modelling a single non-exchangeable species for $\mathcal{M}_1$; no evaluation of the uncertainty was made in the case of (ii). Quantification of such uncertainty and its consequences for decision making needs to be communicated effectively to stakeholders if such procedures are to be adopted, such an exercise requires a considerable increase in resources.

5.6.1 Posterior Distribution Summaries

In Section 5.5 we described the method of hyper-parameter 'specification'. The values suggested were the joint modes of the marginalised posterior distributions; referred to as $\mathcal{MAP}$ estimators. Given a multivariate posterior probability distribution, it is relatively easy to determine modal estimates than expectations. However, $\mathcal{MAP}$ estimators also allow us to easily simulate from the joint posterior distributions. We will illustrate how for both models below. The procedure involves two steps: (1) approximating the joint posterior with a more easily simulated-from distribution; (2) use an appropriate method to efficiently simulate from the approximated distribution until the sample has converged to the true posterior distribution.

We begin by deriving the *Laplace approximation* to the posterior distribution. In Section 5.2 we described the posterior distribution of the hyper-parameters $\vartheta = (k, \phi)$ $[\mathcal{M}_1]$ or $\vartheta = (k, \phi, \alpha, \beta)$ $[\mathcal{M}_2]$, denoted $p(\vartheta \,|\, \mathbf{Y})$. If we Taylor-expand the logarithm of $p(\vartheta \,|\, \mathbf{Y})$ around the $\mathcal{MAP}$ estimate $\hat{\vartheta}$, we get

$$\ln p(\vartheta \,|\, \mathbf{Y}) \approx \ln p(\hat{\vartheta} \,|\, \mathbf{Y}) + \tfrac{1}{2}(\vartheta - \hat{\vartheta})^T B(\hat{\vartheta})(\vartheta - \hat{\vartheta}) + \cdots ,$$

where $B(\hat{\vartheta})$ is the Hessian matrix of $\ln p(\vartheta \,|\, \mathbf{Y})$, evaluated at $\vartheta = \hat{\vartheta}$, defined as

$$B_{ij} = \left. \frac{\partial^2}{\partial \vartheta_i \partial \vartheta_j} \ln p(\vartheta \,|\, \mathbf{Y}) \right|_{\vartheta = \hat{\vartheta}}$$

The linear term in the expansion is zero because the vector of first order partial derivatives evaluated at the $\mathcal{MAP}$ estimator $\hat{\vartheta}$ is zero by definition of a maximum. Ignoring terms higher than second order and transforming back to the standard scale, we obtain

$$p(\vartheta \,|\, \mathbf{Y}) \approx p(\hat{\vartheta} \,|\, \mathbf{Y}) \exp\left\{ -\tfrac{1}{2}(\vartheta - \hat{\vartheta})^T \left[-B(\hat{\vartheta}) \right] (\vartheta - \hat{\vartheta}) \right\}.$$

This is proportional to a multivariate normal distribution with mean $\hat{\vartheta}$ and covariance matrix $\left[-B(\hat{\vartheta}) \right]^{-1}$.

We can use this approximation as a *proposal distribution* for simulating values from the posterior distribution using a standard Metropolis-Hastings random walk

Markov chain Monte Carlo (MCMC) procedure. The procedure for generating T samples is as follows

Step 0: Choose a valid initial starting vector of $\hat{\vartheta}$; denote it $\hat{\vartheta}_0$.

$\vdots$

Step j: To obtain the j-th sample $\hat{\vartheta}_j$,

1. sample a new value of $\hat{\vartheta}_{\text{new}}$ as $\hat{\vartheta}_{\text{new}} = \hat{\vartheta}_{j-1} + Z_j$ where $Z_j \sim N\big(0, t\big[-B(\hat{\vartheta})\big]^{-1}\big)$ and t is a 'tuning' parameter;
2. set
$$\hat{\vartheta}_j = \begin{cases} \hat{\vartheta}_{\text{j-1}} & \text{if } U_j < \frac{p(\hat{\vartheta}_{\text{new}} \,|\, \mathbf{Y})}{p(\hat{\vartheta}_{j-1} \,|\, \mathbf{Y})}; \\ \hat{\vartheta}_{\text{new}} & \text{otherwise} \end{cases}$$
where $U_j \sim \text{Unif}(0, 1)$.

$\vdots$

Step $T + 1$**:** Stop when a sample $\hat{\vartheta}_1, \hat{\vartheta}_2, \ldots, \hat{\vartheta}_T$ is obtained.

The 'tuning' parameter t is selected to improve the rate of acceptance of $\hat{\vartheta}_{\text{new}}$ at each iteration. Setting it as $t = 1$ would imply the proposal distribution is identical to the Laplace approximation; we set $t = 2$ which resulted in an acceptance rate of 34% and 12% for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Acceptance rates were too high for $t = 1$ which resulted in slow mixing of the Markov Chain.

In order to improve the quality of the MCMC sample, we used two standard techniques. The first is to discard the first few samples — called a 'burn-in' period — to ensure that we have reached convergence to the posterior distribution. The second is to only accept every i-th sample — called 'thinning' — to remove the effects of autocorrelation from the chain. In each model we obtained a total of $10,000$ samples after having first applied a burn-in of $5,000$ samples and a thinning rate of 20 for $\mathcal{M}_1$ and 40 for $\mathcal{M}_2$. No diagnostic evidence of convergence failure was observed.

In Figures 5.4 and 5.6 we plot the posterior distributions of each hyper-parameter for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. We also plot the joint posterior distribution of (α, β)

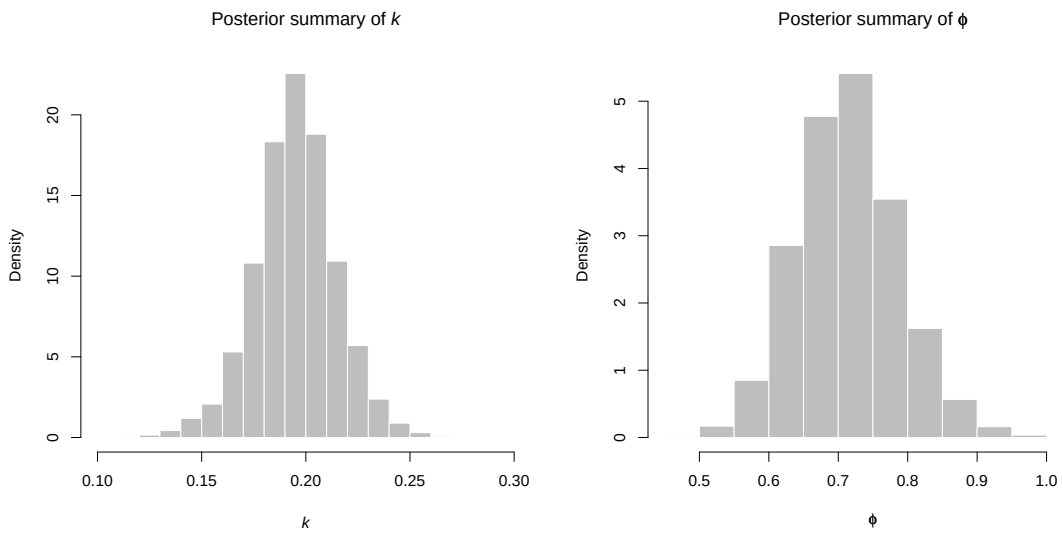


Figure 5.4: Histograms of MCMC sample for hyper-parameters ($\mathcal{M}_1$).

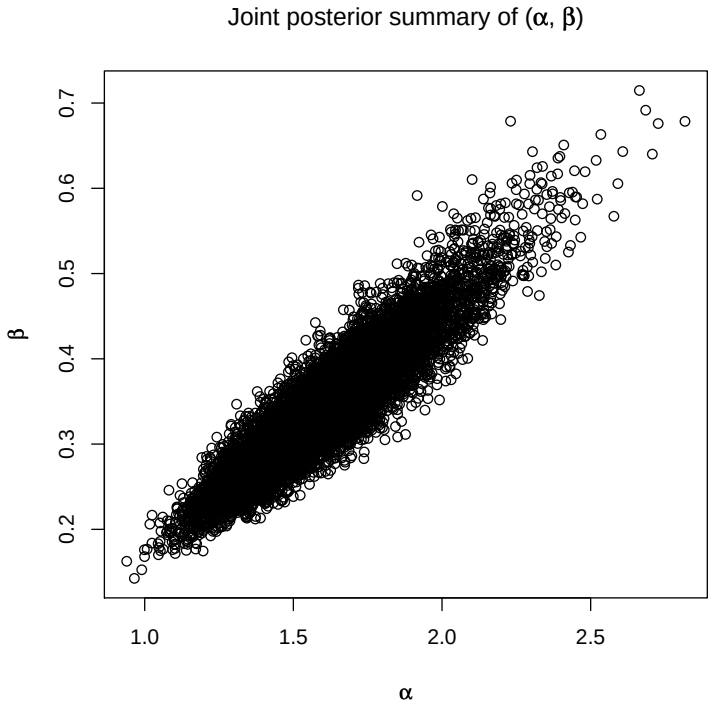


Figure 5.5: Joint distribution of hyper-parameters (α, β) ($\mathcal{M}_2$).

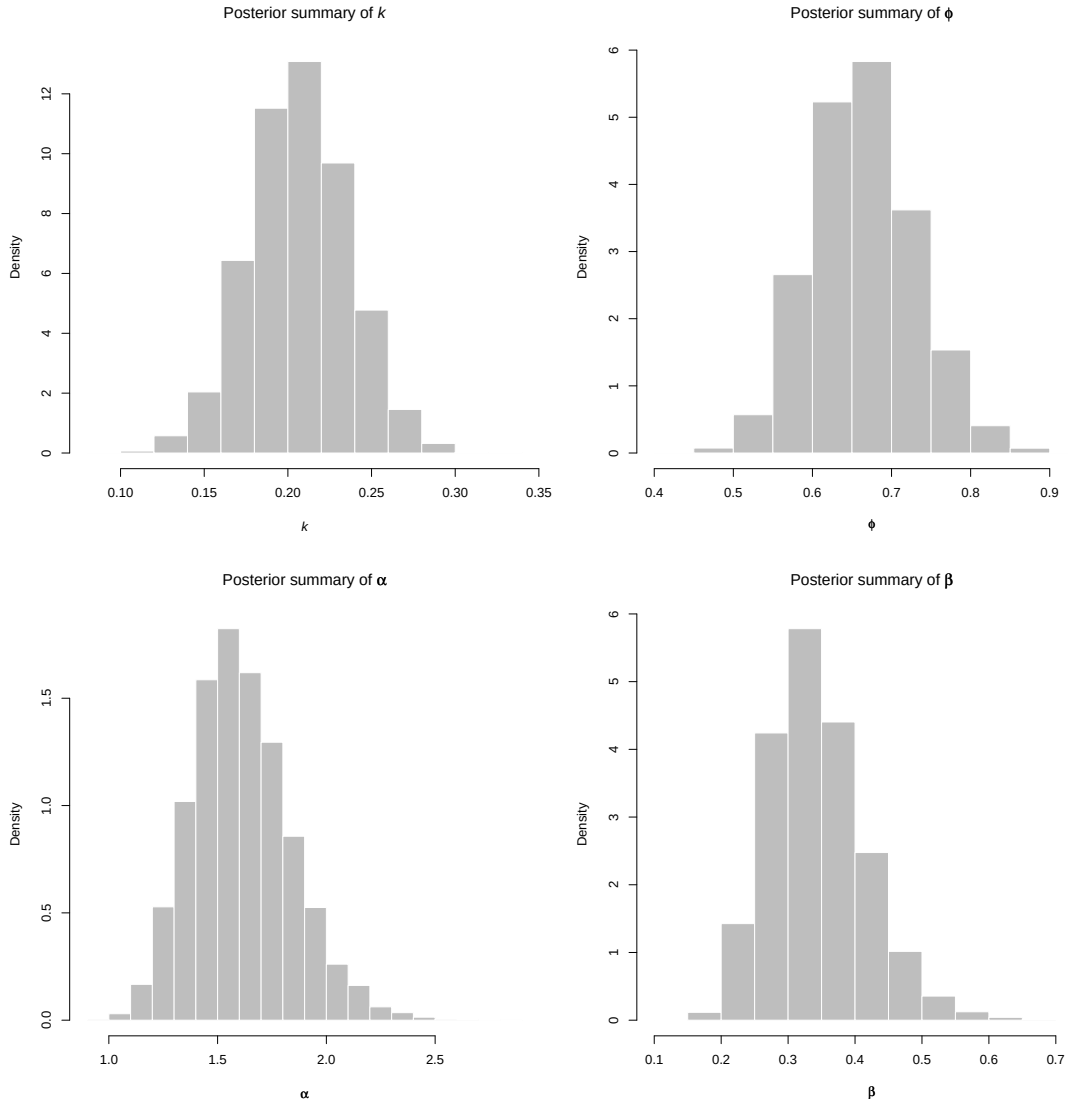


Figure 5.6: Histograms of MCMC sample for hyper-parameters ($\mathcal{M}_2$).

	Mean	S.D.	Median	95% Cred. Int.
k	0.195	0.0195	0.195	(0.154, 0.234)
ϕ	0.714	0.0732	0.713	(0.579, 0.866)

Table 5.2: Summary statistics of hyper-parameter posterior distribution ($\mathcal{M}_1$).

	Mean	S.D.	Median	95% Cred. Int.
k	0.209	0.0296	0.207	(0.148, 0.265)
ϕ	0.664	0.0654	0.662	(0.543, 0.799)
α	1.611	0.2292	1.591	(1.215, 2.112)
β	0.343	0.0717	0.336	(0.224, 0.504)

Table 5.3: Summary statistics of hyper-parameter posterior distribution ($\mathcal{M}_2$).

in Figure 5.5. Analysing these distributions shows that posterior distributions for k and ϕ are symmetric and unimodal. There is a small amount of evidence that the posterior distributions of α and β might be weakly asymmetric. This is also suggested in Figure 5.5 which additionally indicates that *a posterior*, α and β are strongly positively correlated. In Tables 5.2 and 5.3 we display relevant summary statistics which might be used to address the issues of uncertainty handling.

It is observable from the tables that in each model, the mean and median estimate of k and ϕ are very close to the $\mathcal{MAP}$ estimates which we advocated using earlier on, supporting the comment earlier regarding the symmetry of the distribution. The $\mathcal{MAP}$ estimates of α and β are less than the median estimates, which are less than the mean estimates also, giving support to the a slight positive skewness in the posterior distribution of α and β.

It is a risk management decision as to whether our choice of $\hat{\vartheta}$ is sufficient and whether the uncertainty is sufficiently small as to be neglected. It is unlikely a risk manager would be able to interpret the information in Figures 5.4–5.5 directly, nor Tables 5.2–5.3 appropriately in order to make a recommendation. Therefore, it is necessary that we explore the consequences of uncertainty for the decision rules.

5.6.2 Fixed versus Stochastic

The information in the preceding section is limited in ascertaining the adequacy of using a fixed estimator for $\hat{\vartheta}$ as applied to a stochastic one. Our decision rules are of the form $\delta_p(\mathbf{Y} \mid \hat{\vartheta})$ in order to preserve tractability which is lost when allowing for uncertainty in ϑ; we explore this further here.

To assess how $\delta_p(\mathbf{Y} \mid \hat{\vartheta})$ differs from the decision rule obtained when hyper-parameter uncertainty is accounted for, we conduct a 'sensitivity analysis'. First define the following *performance measures*:

$$\begin{aligned} P_1(\delta_p \mid \vartheta) &= \frac{1}{1-\gamma}\mathbb{P}\left[\psi_p(\theta) \leq \delta_p \mid \vartheta; \mathbf{Y}\right] - 1; \\ P_2(\delta_p \mid \vartheta) &= \frac{100}{p}\mathbb{E}\left[\Phi\left(\frac{\delta_p - \mu}{\sigma}\right) \,\middle|\, \vartheta; \mathbf{Y}\right] - 1. \end{aligned}$$

$P_1(\delta_p \mid \vartheta)$ is a scaled posterior distribution function of $\psi_p(\theta)$, conditional on ϑ, evaluated at δ_p (independent of ϑ), and subsequently shifted by 1. The [AJ] estimator, conditional on ϑ would be obtained by equating this measure to 0 and solving for δ_p. Similarly, $P_2(\delta_p \mid \vartheta)$ is a scaled posterior expectation of the PAF (which was denoted the MFE in Section 2.6.4) at δ_p, again, shifted by 1. Equating this measure to 0 and solving for δ_p would yield the conditional [EFSA] estimator. Since each measure is centred about zero, $P_l(\delta_p \mid \vartheta)$ $(l = 1, 2)$ can be envisaged as being a standardised comparable measure of performance for decision rule δ_p.

If we fix $\hat{\delta}_p = \delta_p(\mathbf{Y} \mid \hat{\vartheta})$, we can evaluate the *performance discrepancy* $P_l(\hat{\delta}_p \mid \vartheta)$ for an individual value of ϑ. By considering the expectation of this quantity — denoted as the *marginal performance discrepancy* for decision rule $\hat{\delta}_p$: $P_l(\hat{\delta}_p)$ — with respect to the posterior distribution of ϑ, we can assess the consequence of hyper-parameter uncertainty. We therefore re-use the MCMC samples of ϑ for each model and calculate $P_l(\hat{\delta}_p \mid \vartheta_i)$ individually for each sample $\vartheta_1, \vartheta_2, \ldots, \vartheta_{10,000}$. Averaging this sample then yields an estimate of the marginal performance discrepancy $P_l(\hat{\delta}_p)$.

In Figure 5.7 we show boxplots of the marginal performance discrepancy $P_l(\hat{\delta}_p)$ across all substances in the RIVM fish toxicity database for which the special species was present. This is done for $\mathcal{M}_1$ and $\mathcal{M}_2$, $p = 1$ and $p = 5$, and for lower ($\gamma = 0.95$) and median ($\gamma = 0.50$) [AJ] estimators ($l = 1$), and [EFSA] estimators ($l = 2$).

When $P_l(\hat{\delta}_p)$ is greater than zero, it implies that the maximum permissible level of (average) risk is exceeded.

In Figure 5.8 we show the corresponding boxplots of the relative variation in conditional performance discrepancy of each substance assessed with the special species in the RIVM fish toxicity database. For each substance, this is calculated as the variation of the sample $P_l(\hat{\delta}_p \,|\, \vartheta_i)$ for $i = 1, \ldots, 10,000$, for each performance measure l.

There is a clear outlying substance observed in both figures; this is substance 170 in the RIVM database. The summary statistics of toxicity data for this substance are: $n^* = 2$, $y^\dagger + \hat{k}_{\mathcal{M}_1} = 0.098$, $y^\dagger + \hat{k}_{\mathcal{M}_2} = 0.108$, $\bar{y}^* = 0.097$, $s^* = 0.025$. Consequently, varying k in either direction from $\hat{k}$, most noticeably for $\mathcal{M}_1$ which satisfies $y^\dagger + \hat{k}_{\mathcal{M}_1} \approx \bar{y}^*$ and $s^* \ll \hat{k}_{\mathcal{M}_1}$, will increase $\tilde{\sigma}$ and lead to an increase in $P_l(\hat{\delta}_p)$. Apart from this exceptional substance, the relative marginal performance discrepancy is less than 25% for the remaining substances, which is unlikely to be significant to a risk manager. However, the corresponding variation in the performance discrepancy is more substantial. A possible explanation for this is that the conditional performance measure is approximately linear for ϑ and also that $\hat{\vartheta} \approx \mathbb{E}\vartheta$, so that the marginal performance discrepancy is approximately equal to the conditional performance evaluated at $\hat{\vartheta}$.

In addition to this measure of discrepancy, we also explore the error in the [AJ] estimators by comparing them to those which account for hyper-parameter uncertainty. Using MCMC output, a posterior sample of $\psi_p(\theta)$ is obtained, and the [AJ] estimator is determined as the $100(1-\gamma)$-th percentile of this sample; we denote this as $\tilde{\delta}_p$. An evaluation of the consequences of discounting the uncertainty about $\hat{\vartheta}$ in the estimator is made by considering the relative error for each HC_p estimator on the original concentration scale in order to avoid ambiguity. The error is thus defined as

$$\text{Relative error} = \frac{\left|10^{\hat{\delta}_p} - 10^{\tilde{\delta}_p}\right|}{\left|10^{\tilde{\delta}_p}\right|}.$$

In order to gain sufficient accuracy it was required that more MCMC samples were taken than the $10,000$ obtained for the hyper-parameters earlier. For this exercise we

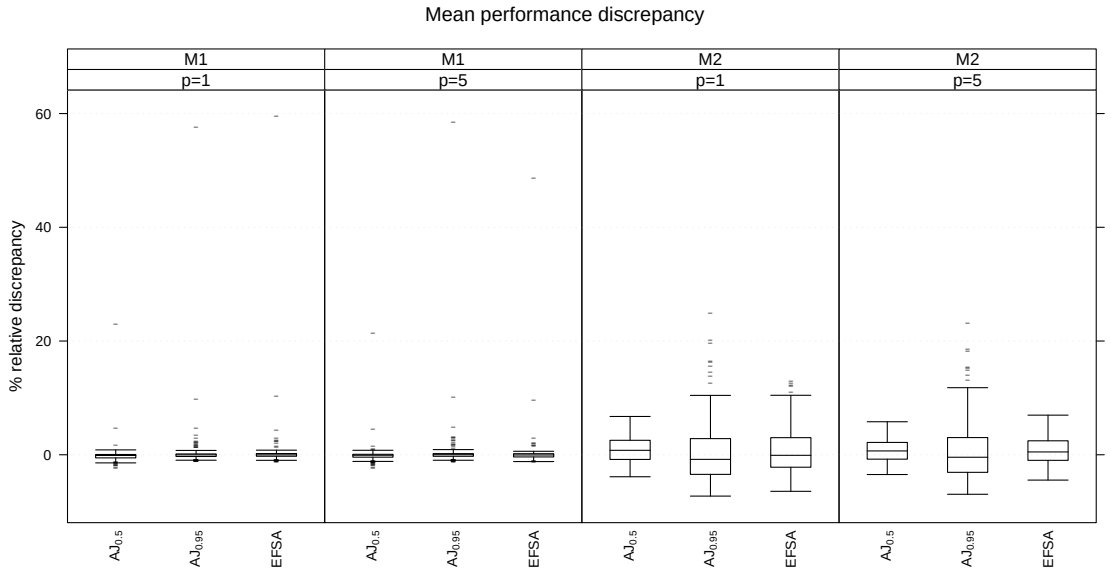


Figure 5.7: Boxplots of per-substance marginal performance discrepancies. For each of the four panels, the [AJ] and [EFSA] discrepancies are evaluated using $P_1(\hat{\delta}_p)$ and $P_2(\hat{\delta}_p)$ respectively.

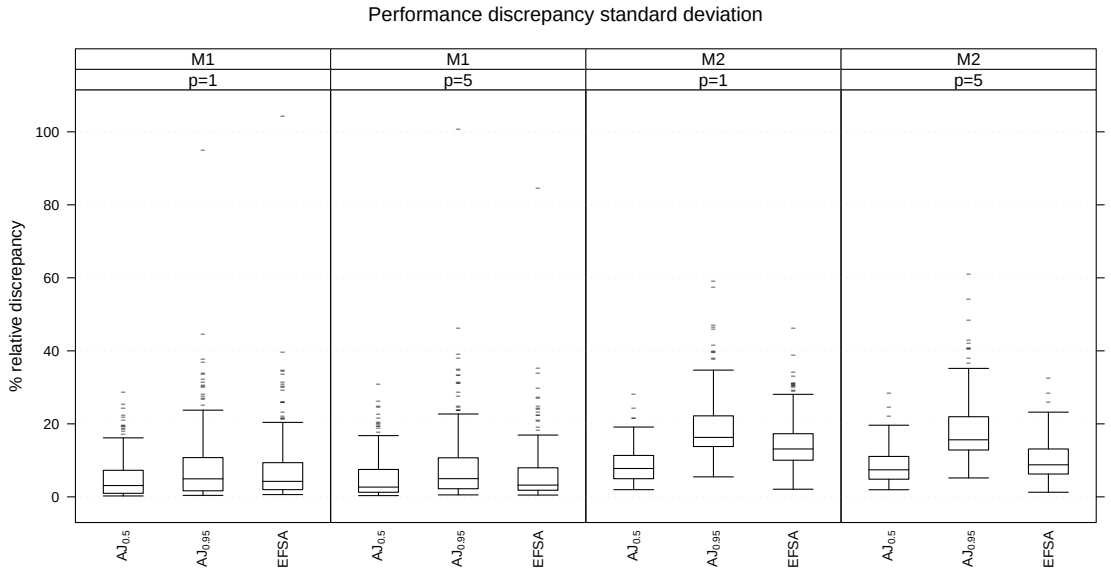


Figure 5.8: Boxplots of per-substance performance discrepancy variation. For each of the four panels, the [AJ] and [EFSA] discrepancies are evaluated using $P_1(\hat{\delta}_p)$ and $P_2(\hat{\delta}_p)$ respectively.

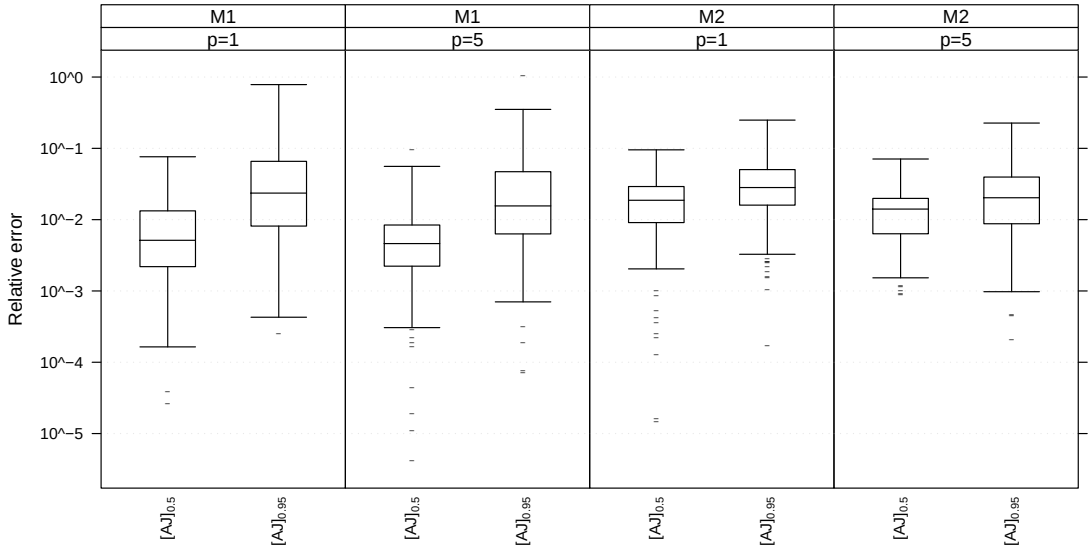


Figure 5.9: Boxplots of relative errors of $\hat{\text{HC}}_p = 10^{\hat{\delta}_p}$ for [AJ] estimators. Error axis is on log (base 10) scale.

took a sample of 45,000 for each model, which was achieved after applying the same burn-in and thinning rates as before. The requirement of a larger sample stems from the posterior uncertainty around σ_i for certain substances in the database where the corresponding sample size of toxicity data n_i is very small.

In Figure 5.9 we show the boxplots of the relative errors in estimators for $p = 1$, 5, $\gamma = 0.50$, 0.95, and $\mathcal{M}_1$ and $\mathcal{M}_2$, for substances in the RIVM fish toxicity database assessed with the special species.

Figure 5.9 indicates that the relative error is never more than 100%; a case associated with the conservative [AJ] estimator with $p = 5$ under $\mathcal{M}_1$. The median estimators' relative error is bounded by 10% for both choices of p and both behavioural models. It would be a regulatory decision as to whether this constitutes robustness, but there is no overwhelming evidence of sensitivity. As was the case in the performance analysis (see Figures 5.7 and 5.8), relative error seems to be worse overall for behavioural model $\mathcal{M}_2$.

5.7 Comparing Non-Exchangeability Models

5.7.1 The Modelling Problem

As discussed in Section 5.1, EFSA (2005) proposed an adaptation of the current SSD model to handle a single non-exchangeable species for a substance $\mathcal{S}$ being assessed. This model is given by

$$\begin{aligned} y_j \,|\, \mu, \sigma^2 &\sim N(\mu, \sigma^2) \text{ for } j \in J^*_{\mathcal{S}}; \\ y^\dagger \,|\, \mu, \sigma^2, k', \phi' &\sim N\left(\mu - k'\sigma, [\phi'\sigma]^2\right). \end{aligned} \tag{5.7}$$

However, when utilised it leads to non-tractable decision rules based on the methods listed in this thesis where one seeks to control the PAF in some particular manner; this is not ideal for gaining acceptance in ERA.

An alternative model was proposed in Section 5.1 by making the bias shift fixed rather than proportional to σ. This model is given by

$$\begin{aligned} y_j \,|\, \mu, \sigma^2 &\sim N(\mu, \sigma^2) \text{ for } j \in J^*_{\mathcal{S}}; \\ y^\dagger \,|\, \mu, \sigma^2, k, \phi &\sim N\left(\mu - k, [\phi\sigma]^2\right). \end{aligned} \tag{5.8}$$

Unlike the EFSA (2005) model, this version does lead to tractable decision rules.

In Section 5.1, we extensively discussed these two models. In particular we described how the EFSA (2005) model (Equations 5.7) has the property that the expected position of the special species in the SSD is unaffected by the variability of σ across the (presumed) population of substances whilst allowing k' to be the same across substances; this is not so when considering the latter model proposal (Equations 5.8).

In this section we attempt to assess the support of each of the models proposed for non-exchangeability. We do not, however, consider one model as being the 'best', if such a notion is even justifiable. From here onwards, we refer to the model for the tolerance data which includes a single special species proposed by EFSA (2005) as $\mathcal{D}_1$ (Equations 5.7) and the model proposed in this research $\mathcal{D}_2$ (Equations 5.8).

5.7.2 Bayes Factors

Model selection is a widely discussed topic in both Bayesian and frequentist statistics (cf. Section 4.5.3). Here we give information about the model selection criteria used in this section for comparing models $\mathcal{D}_1$ and $\mathcal{D}_2$.

Model selection is sometimes complicated by the assumption that there exists a *true model*; what Aitkin (1999) describes as 'oxymoronic'. A discussion on this topic is well beyond the scope of our research; for an in-depth discussion consult Bernardo and Smith (1994) and references therein. We will work in what Bernardo and Smith (1994) call an $\mathcal{M}$-*closed* framework whereby one model is assumed to be true without knowing which. This is difficult to accept in a literal manner, nonetheless it can be neatly exploited for setting model selection criteria.

In Section 4.5 we made use of two commonly applied model selection tools — AIC and BIC — for an exploratory analysis of the RIVM fish toxicity database. However, in the context of comparing models for individual substances which are usually based on very small sample sizes, the applicability of AIC and BIC comes under doubt since we cannot appeal to asymptotics. Moreover, the models are non-nested which additionally restricts other options for model comparison and selection.

The aim here is with regards to making a comparison of the models rather than actually selecting one; a Bayes factor (BF) approach is convenient for this problem. Bayes factors are central to Bayesian hypothesis testing, and are dominantly used as a model selection criteria. Under certain conditions, it can be shown (e.g. Kass and Raftery 1995) that comparison by BIC is asymptotically equivalent to comparison by Bayes factors; we typically have small n, so this result is not applicable here. For the remainder of this section we will describe the calculation of Bayes factors for the purposes of comparing the two competing models for non-exchangeability and discuss their intrinsic features which add justification to their use. Our review and application is based on reviews by Bernardo and Smith (1994) and Kass and Raftery (1995).

Bayes factors are effectively the Bayesian extension of the frequentist likelihood ratio test, but are applicable in a wider sense. Given two opposing scientific hypotheses H_1 and H_2, corresponding to alternate models $\mathcal{D}_1$ and $\mathcal{D}_2$ with data $\mathbf{Y}$,

$\log_{10}(B_{ij})$	B_{ij}	Evidence against H_j
< 0	< 1	(Negative) In favour of H_j
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Table 5.4: Jeffreys' guide to interpretation of Bayes factors.

then the Bayes factor in favour of H_1 (against H_2) is simply

$$B_{12}(\mathbf{Y}) = \frac{\mathbb{P}(\mathbf{Y} \,|\, \mathcal{D}_1)}{\mathbb{P}(\mathbf{Y} \,|\, \mathcal{D}_2)} = \left\{ \frac{\mathbb{P}(\mathcal{D}_1 \,|\, \mathbf{Y})}{\mathbb{P}(\mathcal{D}_2 \,|\, \mathbf{Y})} \right\} / \left\{ \frac{\mathbb{P}(\mathcal{D}_1)}{\mathbb{P}(\mathcal{D}_2)} \right\}. \tag{5.9}$$

The right hand side of Equation 5.9 is the ratio of posterior odds to prior odds, so that if the evidence provided has increased the odds *a posteriori*, then the Bayes factor will be greater than unity. The Bayes factor makes no requirements that the models need be nested. Moreover, the structure of the Bayes factor is such that for the composite-versus-composite hypothesis test (with respect to $\mathcal{D}_1$ and $\mathcal{D}_2$) we average the likelihood over the parameter space as opposed to maximising it which is the basis of the frequentist likelihood ratio test. A Bayes factor therefore only measures the evidence in the data, thus it must be interpreted relative to the prior evidence (Kass and Raftery, 1995). We opt to set $\mathbb{P}(\mathcal{D}_1) = \mathbb{P}(\mathcal{D}_2) = \frac{1}{2}$, which would mean the Bayes factor reduces to the posterior odds in favour of H_1. For completeness, given model $\mathcal{D}$, one calculates $\mathbb{P}(\mathcal{D} \,|\, \mathbf{Y})$ as

$$\mathbb{P}(\mathcal{D} \,|\, \mathbf{Y}) = \int \ell(\mathbf{Y} \,|\, \theta_{\mathcal{D}}; \mathcal{D}) p(\theta_{\mathcal{D}} \,|\, \mathcal{D}) \mathrm{d}\theta_{\mathcal{D}}, \tag{5.10}$$

where $\theta_{\mathcal{D}}$ is the parameter vector for model $\mathcal{D}$ with prior distribution $p(\theta_{\mathcal{D}} \,|\, \mathcal{D})$, and $\ell(\mathbf{Y} \,|\, \theta_{\mathcal{D}}; \mathcal{D})$ is the likelihood function given the model for the data.

By considering the logarithm of a Bayes factor, it is found that the prior weight of evidence and the Bayes factor combine additively. Kass and Raftery (1995) present values as shown in Table 5.4 as a commonly used approximate interpretation of Bayes factors on the logarithmic (base 10) scale.

In order to calculate the Bayes factor for a substance $\mathcal{S}$, we fix the hyper-parameters ϑ as the $\mathcal{MAP}$ estimates because in the future risk assessment the decision rules proposed also treat these parameters as fixed, and so it is the performance of these rules that we ultimately wish to determine. The hyper-parameters (k', ϕ') and (k, ϕ) were defined earlier as the non-exchangeability parameters for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively. The fixed estimates of the hyper-parameters will *differ* between the two different behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$; it is required that one apply the correct values for each context. In addition, we define (α', β') as the corresponding variance heterogeneity hyper-parameters for $\mathcal{D}_1$; in $\mathcal{D}_2$ they remain denoted as (α, β).

For both behavioural models discussed in this section decision rules developed for the assessment of a substance $\mathcal{S}$ are derived using independent non-informative prior distributions for the SSD parameters μ and σ^2; the reasoning for this has already been discussed. Since these prior distributions are improper, i.e. do not integrate to a finite value, the normalising constant is undefined. However, as observed in the structure of Equation 5.10, this is required for the calculation of Bayes factor, implying that in general the Bayes factor is defined on an arbitrary scale with no reference for comparison. This can lead to paradoxical results, as discussed in Kass and Raftery (1995). There are a number of recent proposals for remediation of this, most of which involve using additional data as a device to update the improper prior to a proper one, and apply this as the prior distribution in the Bayes factor calculation.

We argue that the Bayes factors we define for comparison of $\mathcal{D}_1$ and $\mathcal{D}_2$ are indeed well defined. This is because the SSD parameters μ and σ are identically operationally defined for both non-exchangeability models upon fixing the behavioural model perspective with respect to a hypothetically infinite population of exchangeable species in the SSD. In such contexts, Bernardo and Smith (1994, p. 422) argue that one can envisage the undefined normalising constants in the numerator and denominator as the ratio of proper (and equal) constants obtained in the limit of prior distributions tending to the improper versions.

The Bayes factor in favour of $\mathcal{D}_1$ over $\mathcal{D}_2$ for $\mathcal{M}_1$ is given by

$$B_{12} = \frac{\phi\sqrt{\phi^{-2}+n^*}}{\phi'\sqrt{\phi'^{-2}+n^*}}\frac{\hat{\beta}^{\hat{\alpha}}}{\Gamma(\hat{\alpha})}\int_0^\infty \tau^{\hat{\alpha}-1}\exp\left\{-\frac{\tau}{2}[(n-1)\hat{\sigma}^2(\tau)]\right\}d\tau, \tag{5.11}$$

where

$$\hat{\sigma}^2(\tau) = \frac{1}{n-1}\left[\phi'^{-2}\left(y^\dagger + k'\tau^{-1/2} - \hat{\mu}(\tau)\right)^2 + n^*\left(\hat{\mu}(\tau) - \bar{y^*}\right)^2 + (n^*-1)s^{*2}\right];$$

$$\hat{\mu}(\tau) = \frac{\phi'^{-2}(y^\dagger + k'\tau^{-1/2}) + n^*\bar{y^*}}{\phi'^{-2}+n^*};$$

$\hat{\alpha} = \frac{1}{2}(n-1)$ and $\hat{\beta} = \hat{\alpha}\hat{\sigma}^2$.

Similarly, the Bayes factor in favour of $\mathcal{D}_1$ over $\mathcal{D}_2$ for $\mathcal{M}_2$ is given by

$$B_{12} = \frac{\beta'^{\alpha'}}{\beta^\alpha}\frac{\Gamma(\alpha)}{\Gamma(\alpha')}\frac{\phi\sqrt{\phi^{-2}+n^*}}{\phi'\sqrt{\phi'^{-2}+n^*}}\frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})} \times \int_0^\infty \tau^{\tilde{\alpha}'-1}\exp\left\{-\frac{\tau}{2}[2\beta' + (n-1)\hat{\sigma}^2(\tau)]\right\}d\tau, \tag{5.12}$$

where $\tilde{\alpha} = \alpha + \hat{\alpha}$; $\tilde{\beta} = \beta + \hat{\beta}$; and $\tilde{\alpha}' = \alpha' + \hat{\alpha}$. A sketch of the derivation of both Bayes factors is provided in Appendix D.

5.7.3 Analysis

For $\mathcal{D}_2$, the hyper-parameter $\mathcal{MAP}$ estimates are provided in Table 5.1. EFSA (2005) provided estimates for $\mathcal{D}_1$ and $\mathcal{M}_1$, but they were frequentist profile-marginalised maximum likelihood estimates not $\mathcal{MAP}$ estimates, thus inconsistent with those derived here. Moreover, there was no consideration of the joint modelling of $\mathcal{D}_1$ and $\mathcal{M}_2$. Therefore, we derive $\mathcal{MAP}$ estimates based on identical prior distributions as used to calculate hyper-parameter estimates for $\mathcal{D}_2$, which we provide in Table 5.5; see Appendices B.5 and B.7 for further details on the hyper-parameter estimation procedures.

The values for (α', β') are identical to those for $\mathcal{D}_2$ to the first 3 decimal places, but differ substantially from the EFSA profile marginalised maximum likelihood estimates. The estimates of ϕ' are much more different to the corresponding estimates

	k'	ϕ'	α'	β'
$\mathcal{M}_1$	0.458	0.642	–	–
$\mathcal{M}_2$	0.452	0.604	1.523	0.315

Table 5.5: Posterior hyper-parameter $\mathcal{MAP}$ estimates ($\mathcal{D}_1$).

of ϕ for $\mathcal{D}_2$, emphasising our comment that these parameters have different roles; the $\mathcal{M}_1$ estimate is reasonably close to the EFSA frequentist maximum likelihood estimate of 0.625. The estimates of k' differ from the EFSA estimate by 0.009 and 0.003 respectively.

In order to compare the models we first examine the relative support for individual substances in the set $\mathcal{G}_1$ of the RIVM fish toxicity database, each of which has been assessed with the special species — the rainbow trout — and at least two additional ordinary species. In order words, we individually treat each substance in $\mathcal{G}_1$ as $\mathcal{S}$. This allowed us to calculate 220 per-substance Bayes factors. There is a slight issue of independence because for each per-substance Bayes factor calculated, the data was also used in the calculation of the hyper-parameters. We do not believe this to be contentious since the hyper-parameter estimates would be unlikely to differ much if the single dataset used to calculate the respective Bayes factor was omitted from the procedure described earlier because the database is reasonably large. Moreover, it is the value we report here that a risk manager would most likely apply in a future intermediate quantitative tier risk assessment, and it is these decision rules we wish to evaluate.

Figures 5.10 and 5.11 show plots of the Bayes factors for the individual substances contained in $\mathcal{G}_1$ of the RIVM fish toxicity database for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. The horizontal axis represents an arbitrary indexing by the RIVM of the 220 substances tested, and the vertical axis represents the logarithm (base 10) of the Bayes factor which is in favour of $\mathcal{D}_1$ against $\mathcal{D}_2$ for the substances. Dashed lines indicate critical regions of determination as laid out in Table 5.4.

The Bayes factors for $\mathcal{M}_1$ indicate that for the majority of substances, neither $\mathcal{D}_1$ or $\mathcal{D}_2$ are more strongly favoured than the other. Only 15 of the per-substance Bayes factors present evidence which Kass and Raftery (1995) would describe as

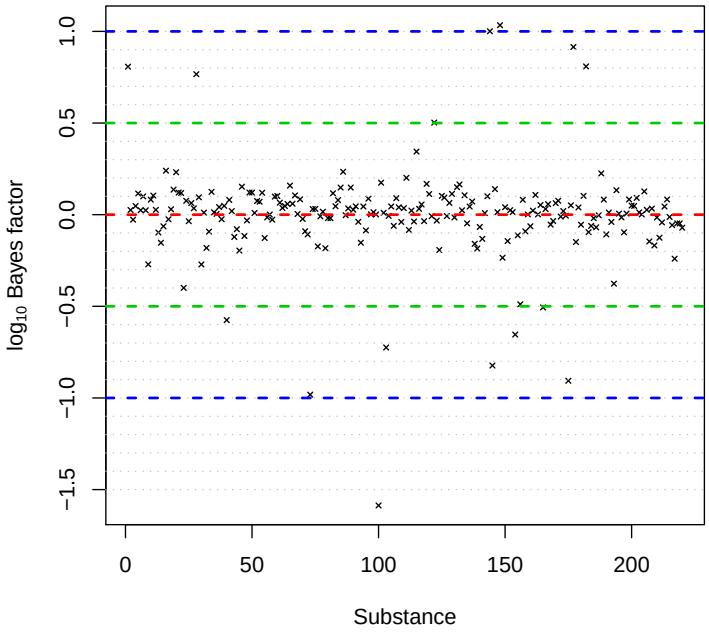


Figure 5.10: Bayes factors for $\mathcal{D}_1$ versus $\mathcal{D}_2$ for substances in $\mathcal{G}_1$ ($\mathcal{M}_1$).

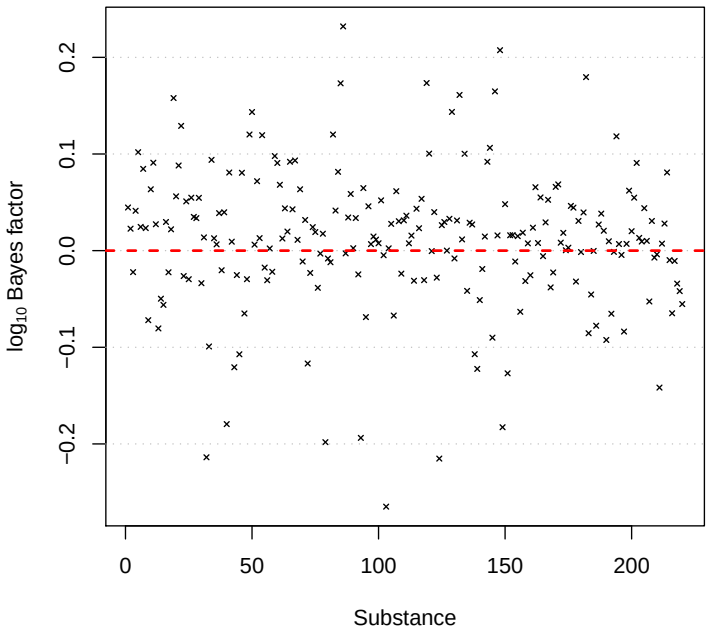


Figure 5.11: Bayes factors for $\mathcal{D}_1$ versus $\mathcal{D}_2$ for substances in $\mathcal{G}_1$ ($\mathcal{M}_2$).

'at least substantial' for either $\mathcal{D}_1$ or $\mathcal{D}_2$; see Table 5.4. Otherwise the rest were within a region which the former authors suggest is 'not worth more than a bare mention'. Note that of the two Bayes factors which are classified as 'decisive' by Kass and Raftery (1995), the one in favour of $\mathcal{D}_2$ (consider $\log_{10}(B_{12}) = -\log_{10}(B_{21})$) corresponds to substance 170 in the RIVM fish toxicity database; this substance was already shown to be exceptional in the performance of its decision rule(s) (see Section 5.6.2).

The situation for $\mathcal{M}_2$ is slightly different such that all Bayes factors lie in the region declared as 'not worth more than a bare mention' according to Table 5.4. However, 141 Bayes factors were positive, compared to 131 for $\mathcal{M}_1$ indicating some support for $\mathcal{D}_1$ over $\mathcal{D}_2$. In all cases, the Bayes factors are made relative to null prior knowledge, i.e. we naively set $\mathbb{P}(\mathcal{D}_1) = \mathbb{P}(\mathcal{D}_2) = \frac{1}{2}$ for each substance. However, we don't have reason to necessarily believe one model is *a priori* more likely than the other.

In addition to analysis of the per-substance Bayes factors, we also calculated the overall Bayes factor of evidence for $\mathcal{D}_1$ against $\mathcal{D}_2$. For any given behavioural model, this is simply the product of the Bayes factors since the per-substance SSDs are independent conditioned upon the fixed hyper-parameters. Under $\mathcal{M}_1$, this is 2.5, and under $\mathcal{M}_2$ it is 425. The former is classed as 'not worth more than a bare mention' by Table 5.4, whilst the latter is classed as 'decisive'. Whilst for $\mathcal{M}_2$ there is strong support for $\mathcal{D}_1$, the Bayes factor is severely undermined by the fact that we have used fixed hyper-parameter values estimated from the same data used to evaluate it. If we fully defined the Bayes factor by incorporating the uncertainty in the hyper-parameters using the prior distributions which subsequently yielded the $\mathcal{MAP}$ estimates, then we could no longer argue that the Bayes factor is suitably defined. The reason for this is that the operational role of k' and k is inherently different for $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, so that the undefined normalising constants will essentially be of different representation. One could arguably specify proper prior distributions, however the Bayes factor would most likely be sensitive to choice of prior specification and any such priors would be non-comparable because of the different operational meaning of k' and k.

Species	$\log_{10} \mathrm{EC}_{50}$	Rank
S025	0.462398	1
S051	1.316967	8
S061	1.240549	6
S087	1.301030	7
S119	0.939519	3
S144	1.204120	5
S151	0.973128	4
S169	0.919078	2
$\bar{y}$	1.044599	
s	0.286518	

Table 5.6: Toxicity data for substance 6 of the RIVM fish toxicity database.

In conclusion, individual substances do not show overwhelming evidence of preference between the two models $\mathcal{D}_1$ and $\mathcal{D}_2$. There is indication of an overall support for $\mathcal{D}_1$, which has the advantage of maintaining the expected position of the special species in the future risk assessment for $\mathcal{S}$, but application of this model is at the expense of tractable decision rules. Hence, the modest support for $\mathcal{D}_2$ in conjunction with the tractability is most likely the pragmatic choice for a risk manager.

5.8 Example Assessment

In the interest of risk communication, it is perhaps helpful for a risk manager to see an example hazard assessment using the newly proposed methodology. For this example we consider an analysis of substance 6 from the RIVM fish toxicity database, essentially treating it as the substance $\mathcal{S}$. However, one must recall that hyper-parameter estimates were also constructed from datasets including the former. Although this introduces a minor degree of duplicity in the use of the toxicity data, it is unlikely to be significant since the database used for hyper-parameter estimation is relatively large. Table 5.6 shows the $n = 8$ log-tolerances values for this substance, in addition to the standard mean and standard deviation (on log-concentration).

The rainbow trout is observed to lie in lower half of the SSD according to both non-parametric rank and empirical position in the method-of-moment fitted SSD. The assumption of normality is not rejected by an Anderson-Darling goodness-of-fit

test at the 5% or 10% critical value with a P-value of 0.1557.

Current guidance suggests estimating the HC_p based on a maximum permissible PAF of $p = 5\%$, however we will concentrate on the trio $p = 1$, 5 and 10. Within the standard species exchangeable context, we display estimators $\delta_p(\mathbf{Y})$ for $\mathcal{M}_1$ (the standard approach) and $\mathcal{M}_2$. For $\mathcal{M}_2$ it is required we specify fixed estimates of the hyper-parameters $\vartheta = (\alpha, \beta)$. Although EFSA (2005) provide these values for fish based upon the joint maximum profile marginalised likelihood estimate, we consider it more appropriate to apply the $\mathcal{MAP}$ estimates for estimation purposes so that the exchangeable model HC_p estimates can be meaningfully compared to the non-exchangeable model HC_p estimates. These values are obtained by first setting $k = 0$ and $\phi = 1$ in Equation B.7 (see Appendix B.6) and then maximising with respect to its remaining arguments; in this case $\mathcal{G}_1$ is taken to be all substances which have been tested with the rainbow trout (regardless of sample size) so that $\mathcal{G}$ is effectively the *entire* RIVM fish database. The values retrieved are: $\hat{\vartheta} = (1.528, 0.267)$ which yields $s_{\text{adj}} = 0.332100$. [EFSA] and [AJ] ($\gamma = 0.95, 0.50, 0.05$) estimates of $\psi_p(\theta)$ under $\mathcal{M}_1$ and $\mathcal{M}_2$ are respectively shown in Tables 5.7 and 5.8 for $p = 1$, 5 and 10.

In addition to the estimators displayed for the species exchangeable model, we also display the [EFSA] and median [AJ] ($\gamma = 0.50$) estimators, with 90% credible interval, which were derived in this thesis and take account of the non-exchangeability of the rainbow trout. These estimates are given in Tables 5.9 and 5.10 for $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. It was found that $\hat{\mu} = 1.076$ and $\hat{\sigma} = 0.2856$ for $\mathcal{M}_1$; $\hat{\mu} = 1.080$ and $\tilde{\sigma} = 0.3463$ for $\mathcal{M}_2$. These estimators are based on the $\mathcal{MAP}$ estimates of the hyper-parameters as defined in Table 5.1.

Examining the estimates from the usual exchangeable model indicates that the [EFSA] estimators are relatively more conservative than the median [AJ] estimators, which is consistent with findings in Chapter 3. Moreover, the relative level of conservatism decreases as p increases. Interestingly, the estimators based on the behavioural model of $\mathcal{M}_2$ lead to relatively more conservative estimators than their counterparts under $\mathcal{M}_1$. For $p = 5$ the absolute difference in median [AJ] HC_5 estimates between $\mathcal{M}_1$ and $\mathcal{M}_2$ is $10^{0.5522} - 10^{0.4821} = 0.5316$ mg/L, equivalent to the $\mathcal{M}_1$ estimate being 18% larger than $\mathcal{M}_2$. For the lower credible limit [AJ] estimate

p		$\delta_p(\mathbf{Y})$
1	[EFSA]	0.1335
	[AJ]	0.3467
		$(-0.2029, 0.6168)$
5	[EFSA]	0.4688
	[AJ]	0.5522
		(0.1314, 0.7701)
10	[EFSA]	0.6146
	[AJ]	0.6616
		(0.3048, 0.8569)

Table 5.7: [EFSA] and [AJ] estimators for the species *exchangeable* model ($\mathcal{M}_1$).

p		$\delta_p(\mathbf{Y})$
1	[EFSA]	0.0723
	[AJ]	0.2478
		$(-0.2492, 0.5312)$
5	[EFSA]	0.4067
	[AJ]	0.4821
		(0.0904, 0.7181)
10	[EFSA]	0.5615
	[AJ]	0.6068
		(0.2666, 0.8232)

Table 5.8: [EFSA] and [AJ] estimators for the species *exchangeable* model ($\mathcal{M}_2$).

p		$\delta_p(\mathbf{Y})$
1	[EFSA]	0.1742
	[A&J]	0.3807
		$(-0.1616, 0.6454)$
5	[EFSA]	0.5063
	[A&J]	0.5855
		(0.1731, 0.7967)
10	[EFSA]	0.6506
	[A&J]	0.6945
		(0.3470, 0.8820)

Table 5.9: [EFSA] and [AJ] estimators for the species *non-exchangeable* model ($\mathcal{M}_1$).

p		$\delta_p(\mathbf{Y})$
1	[EFSA]	0.0746
	[A&J]	0.2495
		$(-0.2600, 0.5373)$
5	[EFSA]	0.4206
	[A&J]	0.4939
		(0.0965, 0.7299)
10	[EFSA]	0.5809
	[A&J]	0.6240
		(0.2817, 0.8379)

Table 5.10: [EFSA] and [AJ] estimators for the species *non-exchangeable model* ($\mathcal{M}_2$).

($\gamma = 0.95$) the difference is only 0.1219 mg/L, or 10% larger; and for the [EFSA] estimators the increase is 0.3921 mg/L, or 15% larger.

We notice from comparing the non-exchangeable model estimates to the exchangeable model estimates that there has been a decrease in the level of relative conservatism. Moreover, the level of increase in the estimates is down weighted when the species non-exchangeable model is incorporated with the behavioural model $\mathcal{M}_2$. For $\mathcal{M}_1$, the increase in the median HC_5 estimates from the non-exchangeable model in comparison to the exchangeable model estimates is 0.2842 mg/L, equivalent to an increase of 8%. Similar differences were found for the other estimators. For $\mathcal{M}_2$, the corresponding increase was 0.0835 mg/L — an increase of only 3%.

The overall ranges of HC_p estimates was: $(0.5495, 4.4198)$ mg/L for $p = 1$; $(1.2314, 6.2618)$ mg/L for $p = 5$; and $(1.8476, 7.6208)$ mg/L for $p = 10$. Whether either of the differences described above is significant would be a risk management decision.

From an exploratory perspective it is interesting to consider what the general consequence of accounting for non-exchangeability through the model we propose might realistically be. In order to assess this, we repeated the analysis described above for the single substance on all substances in $\mathcal{G}_1$. Although 124 additional substances assessed with the rainbow trout are available, but not contained in $\mathcal{G}_1$ because of the requirement we imposed that $n_i^* \geq 2$, we do not consider them here either. The reasoning for this is two-fold: (i) the estimators have a tendency to heavily distort a general overview of the consequences; (ii) it is highly unlikely that regulatory procedures would ever allow for probabilistically derived estimators based upon $n = 2$ to be entered into a risk assessment dossier. The latter point is based on the fact that current recommendations and regulatory requirements are much higher, with Campbell et al. (1999) (the 'HARAP' guidelines) recommending $n \geq 5$ in the context of acute pesticide exposure to fish, which is highly pertinent to the RIVM fish database we analyse here. Since this is only an exploratory exercise, we deem it appropriate to limit the review to the 220 substances in $\mathcal{G}_1$.

In Figure 5.12 we plot the $\log_{10}(\text{HC}_5)$ estimators based on the species exchangeable model against the difference in the $\log_{10}(\text{HC}_5)$ estimators based on the species

non-exchangeable and species exchangeable models. The ordinate is displayed in this way for two reasons: (i) it increases clarity of the overall plot; and (ii) it can be interpreted as the logarithm of the relative HC_5 estimators for non-exchangeability / exchangeability. For each behavioural model, $\mathcal{M}_1$ and $\mathcal{M}_2$, there is a strong indication of log-linearity. However, the difference between species exchangeable and non-exchangeable model estimators based on $\mathcal{M}_1$ vary more widely than those based on $\mathcal{M}_2$.

The maximum ratio of a HC_5 estimate derived under the non-exchangeable model relative to a HC_5 estimate derived under a species exchangeable model was slightly over 0.5 orders of magnitude (6.3) which was seen for the [AJ] ($\gamma = 0.95$) estimator. The corresponding minimum ratio corresponded to nearly 1.5 orders of magnitude. With the exception of these cases, relative differences in HC_5 estimators between the two models were reasonably small for the majority of substances.

Figure 5.13 shows a set of similar plots for the case $p = 1$. In this case, conclusions do not substantially differ. However, the maximum and minimum ratios suggest differences of greater than 1 and 2 orders of magnitude respectively. Whilst this is likely to be significant to risk assessors and managers alike, the most noticeable differences arose from substances where sample sizes were very small. Consider the 'HARAP guidelines' (Campbell et al., 1999) which suggest within the context of data we analyse here, that n should be ≥ 5. In this case, the ratio of the [AJ] ($\gamma = 0.95$) HC_p estimates under the species non-exchangeable model to their exchangeable model counterparts lied between (0.1, 1.6) and (0.05, 1.8) for $p = 5$ and $p = 1$ respectively, for all substances. These maximum relative ratios decreased further as n increased; for $n = 10$, the corresponding factors approximately halved in comparison to $n = 5$.

Nonetheless, without large-scale field or perhaps mesocosm data there is no way of exploring the consistency of either the standard exchangeable model estimators or the species non-exchangeable model estimators.

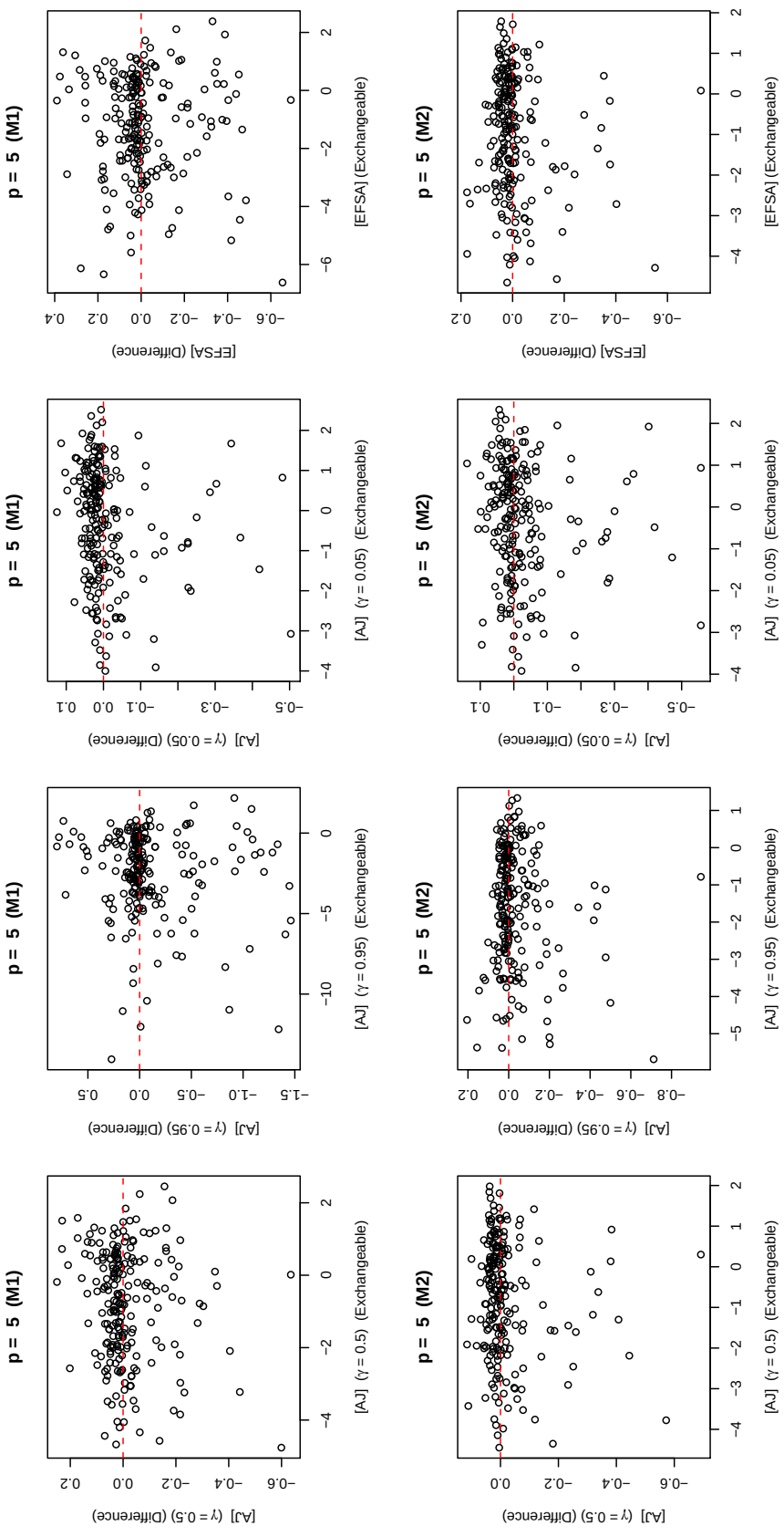


Figure 5.12: Plot of $\delta_5(\mathbf{Y} \mid \hat{\vartheta})$ estimators for species exchangeability model versus the difference in $\delta_5(\mathbf{Y} \mid \hat{\vartheta})$ between the species non-exchangeable and exchangeable models. Each point corresponds to a substance in $\mathcal{G}_1$. Red dashed line indicates where the difference is neutral.

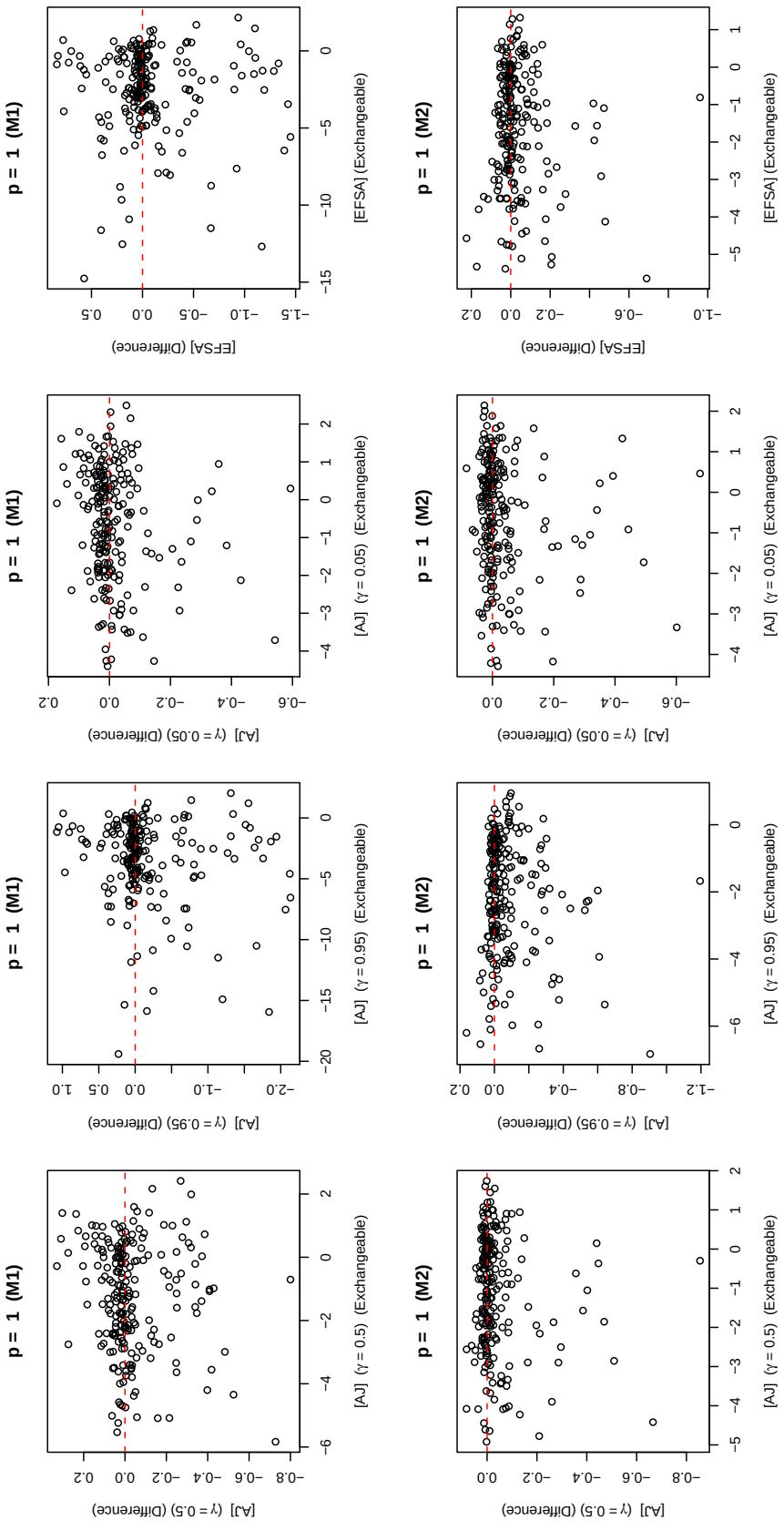


Figure 5.13: Plot of $\delta_1(\mathbf{Y} \mid \hat{\vartheta})$ estimators for species exchangeability model versus the difference in $\delta_1(\mathbf{Y} \mid \hat{\vartheta})$ between the species non-exchangeable and exchangeable models. Each point corresponds to a substance in $\mathcal{G}_1$. Red dashed line indicates where the difference is neutral.

5.9 Beyond a Single Special Species

If we accept that there is a species which is non-exchangeable with others, then we must accept the possibility of additional such species. The sources of evidence in Chapter 4 appear to support this viewpoint. For example, Table 4.2 shows the results of a hypothesis test performed on the RIVM fish toxicity database with the rainbow trout's tolerance values completely removed; there was still a large number of significant species remaining. The argument iteratively implies that perhaps the modelling of non-exchangeability would be better replaced by some kind of hierarchical model which incorporates both species and chemical effects. However, the incomplete factorial nature and high degree of sparsity of the database would require a robust method of fitting. Moreover, there is a requirement of an augmented database, otherwise the uncertainty in hyper-parameter estimation would make unsound the assumptions of treating them as fixed.

Including additional species into the non-exchangeability models described in Section 5.1 may lead to serious criticism by different stakeholders. We have no overriding reason to extend the assumption to other species, especially since intermediate quantitative tier risk assessments should encompass a degree of parsimony. Nevertheless, we give very brief coverage to having multiple 'biased' species.

Consider the situation of observing log-tolerance data for a substance $\mathcal{S}$ such that

$$y_j \,|\, \mu, \sigma^2, k_j, \phi_j \sim N\big(\mu - k_j, [\phi_j \sigma]^2\big),$$

where (k_j, ϕ_j) represent the species non-exchangeability parameters for species $j \in J_{\mathcal{S}}$, such that $|J_{\mathcal{S}}| = n$. As before, we assume they are in reference to some suitable population of substances, say pesticides or, specific modes of action.

Mathematically, it is straightforward to extend the calculations made for a single biased species to the situation here of n 'biased' species. We condition on the non-exchangeability parameters being fixed and known, as was done previously, and apply the same non-informative prior distributions for μ and σ under both behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$. This leads to the posterior distributions for μ and

σ^2 under $\mathcal{M}_2$ being defined as

$$\begin{aligned}\mu \,|\, \sigma^2, \mathbf{k}, \boldsymbol{\phi}; \mathbf{Y} &\sim N\left(\hat{\mu}_n, \frac{\sigma^2}{\sum_{j\in J_{\mathcal{S}}} \phi_j^{-2}}\right);\\ \sigma^2 \,|\, \mathbf{k}, \boldsymbol{\phi}; \mathbf{Y} &\sim \mathcal{IG}\left(\frac{2\alpha+n-1}{2}, \frac{2\alpha+n-1}{2}\tilde{\sigma}_n^2\right),\end{aligned}$$

where $\mathbf{k} = (k_j;\, j \in J_{\mathcal{S}})$; $\boldsymbol{\phi} = (\phi_j;\, j \in J_{\mathcal{S}})$;

$$\begin{aligned}\hat{\mu}_n &= \frac{\sum_{j\in J_{\mathcal{S}}} \phi_j^{-2}(y_j + k_j)}{\sum_{j\in J_{\mathcal{S}}} \phi_j^{-2}};\\ \tilde{\sigma}_n^2 &= \frac{2\beta + (n-1)\hat{\sigma}_n^2}{2\alpha + n - 1};\\ \hat{\sigma}_n^2 &= \frac{1}{n-1}\sum_{j\in J_{\mathcal{S}}} \phi_j^{-2}(y_j + k_j - \hat{\mu}_n)^2.\end{aligned}$$

The subscript n denotes that the estimator is based on n-pairs of non-exchangeability parameters. Setting $\alpha = \beta = 0$ yields the posterior distribution for $\mathcal{M}_1$.

The specification of the decision rules follows on naturally from the posterior distribution. Moreover, the highly appealing tractability of the decision rules is maintained. For example, the [EFSA] and [GAL] decision rules for estimating $\psi_p(\theta)$ are both of canonical form

$$\delta_p(\mathbf{Y} \,|\, \mathbf{k}, \boldsymbol{\phi}, \alpha, \beta) = \hat{\mu}_n - \kappa_p(n, \alpha, \boldsymbol{\phi})\tilde{\sigma}_n,$$

where

$$\begin{aligned}\kappa_p(n, \alpha, \boldsymbol{\phi})_{[\text{EFSA}]} &= \sqrt{1 + \psi^{-2}}\, t_{\pi,p};\\ \kappa_p(n, \alpha, \boldsymbol{\phi}, \gamma)_{[\text{AJ}]} &= \psi^{-1} F^{-1}_{T_{\pi,\eta}}(\gamma);\end{aligned}$$

$\psi^2 = \sum_{j\in J_{\mathcal{S}}} \phi_j^{-2}$; $\eta = K_p\psi$ and $\pi = 2\alpha + n - 1$. Setting $\alpha = \beta = 0$ yields the decision rules for $\mathcal{M}_1$, otherwise setting (α, β) according to their estimates yields the decision rule for $\mathcal{M}_2$.

Estimation of the hyper-parameters is slightly more complicated. As an example, consider two special species. In order to allow for integration with respect to

	k_1	k_2	ϕ_1	ϕ_2	α	β
$\mathcal{M}_1$	0.1779	-0.3420	0.7111	1.0245	–	–
$\mathcal{M}_2$	0.1821	-0.3438	0.6464	0.9623	1.5044	0.3040

Table 5.11: Posterior hyper-parameters $\mathcal{MAP}$ estimates for two non-exchangeable species.

the nuisance parameters to be straightforward, we must now consider groups of substances pertaining to the estimation of (k_1, ϕ_1), (k_2, ϕ_2), $(k_1, k_2, \phi_1, \phi_2)$. For $\mathcal{M}_2$, we must also consider all other substances which are to be included in the estimation of the heterogeneity parameters. We have estimated the hyper-parameters under the inclusion of the rainbow trout and the goldfish in the non-exchangeable assumption. The inclusion of the rainbow trout has already been justified; the goldfish is included purely for illustrative purposes, however it featured highly in hypotheses tests conducted in Chapter 4 (see for example Tables 4.1–4.3). Moreover, the goldfish has been assessed with a relatively large number of substances within the database. So as not to detract from the illustrative nature of this example, details of the calculations are not presented here. However, they are determinable in a similar manner to those described earlier in the context of a single non-exchangeable species. The $\mathcal{MAP}$ estimates are presented in Table 5.11 with (k_1, ϕ_1) and (k_2, ϕ_2) corresponding to the rainbow trout and the goldfish respectively.

We have briefly explored uncertainty around the $\mathcal{MAP}$ estimates and found no significant difference from the single species model, except slightly wider bounds around the parameters pertaining to the goldfish. The hyper-parameter estimates for the rainbow trout have not changed significantly, however the 'bias' parameter estimate has decreased for both behavioural models. The bias parameters for the goldfish (k_2) are observed to be negative, implying that the species is typically tolerant relative to the expected (log-) SSD median. If the goldfish is a standard dossier species, this might undermine the level of protection offered by the HC_5 when species exchangeability is assumed. In particular the shift is larger in magnitude compared to the rainbow trout's; the allowance for a different variance is effectively discounted. Using these hyper-parameter estimates in the decision rules listed above,

setting $(k_j, \phi_j) = (0, 1)$ where appropriate, would yield an estimate of $\psi_p(\theta)$.

A conceptual problem which we introduced earlier in Section 5.4 is exacerbated when including additional species into the non-exchangeable assumption. This is regarding the presupposition that the SSD is a surrogate for all ecosystems of broad similar types, e.g. aquatic systems. In the proposal made here, the SSD does not describe any of the special species, *per se*. However, protection is prescribed strictly in terms of the SSD, though the special species does contribute to the estimation. As the number of special species increases, aside from the issues of estimating the increasing number of hyper-parameters, it would be required that at some stage a procedure would need to be developed which allows for better interpretation of the SSD in reflection of multiple non-exchangeable species.

5.10 Conclusions

A proposal on how to incorporate non-exchangeability of a single species in the regulatory accepted procedure of setting PNECs based on an extrapolation of the HC_p has been described. The setting is most likely inconsistent with the viewpoint of statistical modellers, however it is pragmatic for the level of risk assessment it is intended for, namely intermediate quantitative tier hazard assessment. Moreover, the inclusion of just a single species into the assumption allows for parsimony which is effectively only a slight deviation from the currently accepted methodology. Letting this single species be the rainbow trout has been argued here and in the previous chapter, due to its prevalence in current ecological chemical safety assessment particularly in the field of pesticides. Should there exist evidence for typically biased species belonging to alternative taxa, then this proposal would also be appropriate. A possibly contentious point which a risk manager might raise would be the failure to incorporate hyper-parameter uncertainty into the decision rules. However we have demonstrated for a wide range of pesticides and the rainbow trout as the special species, that the hyper-parameter uncertainty does not significantly influence the overall performance of the estimators, especially when viewed in light of the additional neglected uncertainties.

Chapter 6

Revised Deterministic PNEC Estimators

Focus thus far has been limited to investigating methods and techniques to derive estimators of the HC_p such that setting $p = 5$ will yield a proxy for the PNEC, defined up to a further arbitrary fixed assessment factor. Such estimators fall primarily under the scope of intermediate quantitative tier risk (hazard) assessment. This level of assessment can be considered a refinement of lower quantitative tier risk assessment, i.e. using fixed assessment factors. Nonetheless, use of fixed assessment factors is an efficient and established practice in ERA.

In this chapter we revisit the procedures of calculating PNECs based on fixed assessment factors. In particular, we assess research in EFSA (2005) by evaluating a generalised deterministic decision rule through a probabilistic lens; such procedures have the added benefit of being comparatively straightforward to that of current recommendations. We extend this research by demonstrating robustness to assumptions, both empirically and analytically.

6.1 Introduction

As introduced in Section 2.3, the current strictly deterministic method for determining a PNEC is, to keep it in its most basic framework, based on dividing a summary statistic of the available toxicity data by an assessment factor (AF). The

regulatory accepted summary statistic is the lowest tolerance value of the dataset which typically has extremely low cardinality. EFSA (2005, 2008) have recently proposed the geometric mean as the summary statistic. The argument in favour of this would be that it takes full account of all the toxicity data, not just rank information. An argument against and conversely an argument for applying the minimum order statistic, is that it is less conservative than the minimum tolerance value. An additional potential hindrance is that application of the geometric mean may be open to misuse. This is because manufacturers may scientifically test additional species from those required which are *a priori* believed to demonstrate larger tolerances to the chemical in order to increase the derived PNEC. However, such misuse could also potentially be exploited in the strictly deterministic methods.

The value of the required assessment factor applied to the summary statistic depends on a number of criteria, such as: taxonomic diversity of the data sample; acute or chronic endpoint assessment; sample size and contextual interpretation. Typical values are usually powers of 10 (and intermediate 5-fold values), varying between 1 to 5 orders of magnitude. Assessment factors for regulatory application are provided in: EC (2002, 2003) and ECHA (2008a) pertaining to the EU and Zeeman (1995) pertaining to the US. In addition, Forbes and Calow (2002a) and EFSA (2005) overview different international sources of assessment factors. An example of a typical prescription of assessment factors as found in ECHA (2008a) for application with marine ecological compartments is shown in Table 6.1; similar tables are listed for other compartments (e.g. freshwater, sewage treatment plants, mammals, etc.).

The assessment factors reported in the former sources are described as accounting for a number of identified uncertainties; this list was described in Section 2.3, and includes: inter- and intra-species variation; inter- and intra-laboratory variation; temporal toxicity extrapolation and laboratory data to field impact extrapolation. Assessment factors are usually considered as multiplicative combinations of smaller assessment factors pertaining to different sources of uncertainty (EFSA, 2005). If so, then it is not explicitly clear what proportion of the identified uncertainties each assessment factor is accounting for. However, it may be inappropriate to consider

Assessment criteria	AF
Lowest short-term L(E)C_{50} from FW or SW representatives of three taxonomic groups (algae, crustaceans and fish) of three trophic levels	10,000
Lowest short-term L(E)C_{50} from FW or SW representatives of three taxonomic groups (algae, crustaceans and fish) of three trophic levels, + two additional marine taxonomic groups (e.g. echinoderms, molluscs)	1,000
One long-term result (e.g. EC_{10} or NOEC) (from FW or SW crustacean reproduction or fish growth studies)	1,000
Two long-term results (e.g. EC_{10} or NOEC) from FW or SW species representing two trophic levels (algae and/or crustaceans and/or fish)	500
Lowest long-term results (e.g. EC_{10} or NOEC) from three FW or SW species (normally algae and/or crustaceans and/or fish) representing three trophic levels	100
Two long-term results (e.g. EC_{10} or NOEC) from FW or SW species representing two trophic levels (algae and/or crustaceans and/or fish) + one long-term result from an additional marine taxonomic group (e.g. echinoderms, molluscs)	50
Lowest long-term results (e.g. EC_{10} or NOEC) from three FW or SW species (normally algae and/or crustaceans and/or fish) representing three trophic levels + two long-term results from additional marine taxonomic groups (e.g. echinoderms, molluscs)	10
SSD method	5 – 1

Table 6.1: Assessment factors for deriving a PNEC for aquatic compartments. FW = freshwater, SW = saltwater. Reproduced and extended from ECHA (2008a, Table R.10-5).

assessment factors in this manner since uncertainties will likely overlap.

Although the prescribed assessment factors remain fixed, ECHA (2008a) state that for certain (aquatic) compartments:

> 'The assessment factors presented [in ECHA 2008a] should be considered as general factors that under certain circumstances may be changed.'

Consequently, there is motivation for analysis of this tier of hazard assessment.

6.2 Comparing Deterministic and Probabilistic Estimators

Quantitative determination of a PNEC is permitted using either: (i) assessment factors applied to the minimum tolerance value; or (ii) a probabilistically derived HC_5 (with an applied assessment factor of between 1 and 5). Since these two methods are very different, it is of interest to consider the relative differences and similarities; we do this empirically in this section.

Procedure (ii) might be considered as a refinement of (i), hence the regulatory requirements for implementation are more stringent. For example, in the context of REACH (ECHA, 2008a), the minimum sample size for determining a HC_5 estimate is $n = 10$, whereas the minimum sample size for using strictly deterministic assessment factors can be as low as $n = 1$. Further conditions such as endpoint type and taxonomic classification also determine whether probabilistic methods are admittable in the chemical safety assessment dossier.

Whilst the majority of substances in the RIVM fish database (see Section 4.1) are unlikely to satisfy technical requirements for probabilistic analysis, we nonetheless compare probabilistic HC_5 values to deterministic values. It should be noted that in each case, the values are pertinent to risk assessment of fish populations, and do not constitute a community level PNEC. To perform this evaluation, we first considered setting the aforementioned summary statistic to be one of the following:

1. the minimum tolerance value;

2. the tolerance value of the rainbow trout.

The first summary statistic is explored for obvious reasons concerning the current format of EU decision rules. The second summary statistic is explored because of its prominence in pesticide ERA. Raimondo et al. (2008) also compared this summary statistic to HC_5 values because it has been suggested that the rainbow trout may be a suitable surrogate species for ERA of endangered species (Dwyer et al., 2005).

Probabilistic estimators discussed in this research only account for uncertainty about interspecies variability. Therefore, any determined estimate of the HC_5 is only comparable to the (deterministic) adjusted toxicity statistic (ATS; see Section 2.5). For all intents and purposes, deterministic assessment factors will only be considered in the context of those accounting for this aspect of uncertainty; this factor was denoted AF_{spec} in Section 2.5. A caveat of comparison is that the deterministic and probabilistic estimators shown will be subject to additional and most likely different, fixed assessment factor extrapolations before being listed as PNEC estimators.

In Figure 6.1 (top row) we plot the $\log_{10}(\text{ATS})$ (based on $AF_{spec} = 10$ being applied to the lowest observed tolerance value) against the [AJ] ($\gamma = 0.50, 0.95$) and [EFSA] estimators of $\psi_5(\theta)$ for each substance in the RIVM fish toxicity database satisfying $n \geq 3$. We only consider estimators under behavioural model $\mathcal{M}_1$ as this is the regulatory accepted model. The reason we elected to use $AF_{spec} = 10$ is because it is implicitly suggested in technical guidance documents (EFSA, 2005). Since on the log-scale the assessment factor acts as an assessment shift of $\log_{10}(AF_{spec})$ to the log summary statistic, choosing a different value of AF_{spec} is simply equivalent to shifting the abscissa by the assessment shift; the ordinate remains unchanged. A line of equality is also drawn (red line in plots) to indicate where $\delta_5(\mathbf{Y}) > \log_{10}(\text{ATS})$ (blue points) and where $\delta_5(\mathbf{Y}) < \log_{10}(\text{ATS})$ (green points). Additionally, Figure 6.1 (bottom row) is similarly constructed but with the summary statistic equal to the tolerance value of the rainbow trout; consequently there are fewer substance points in these plots. The reason for the restriction of $n \geq 3$ was because for certain datasets with $n = 2$, extremely small HC_5 estimates were obtained (up to 10^{-50}) which heavily distorted interpretation.

From Figure 6.1 it is observed that there is a strong linear correlation between

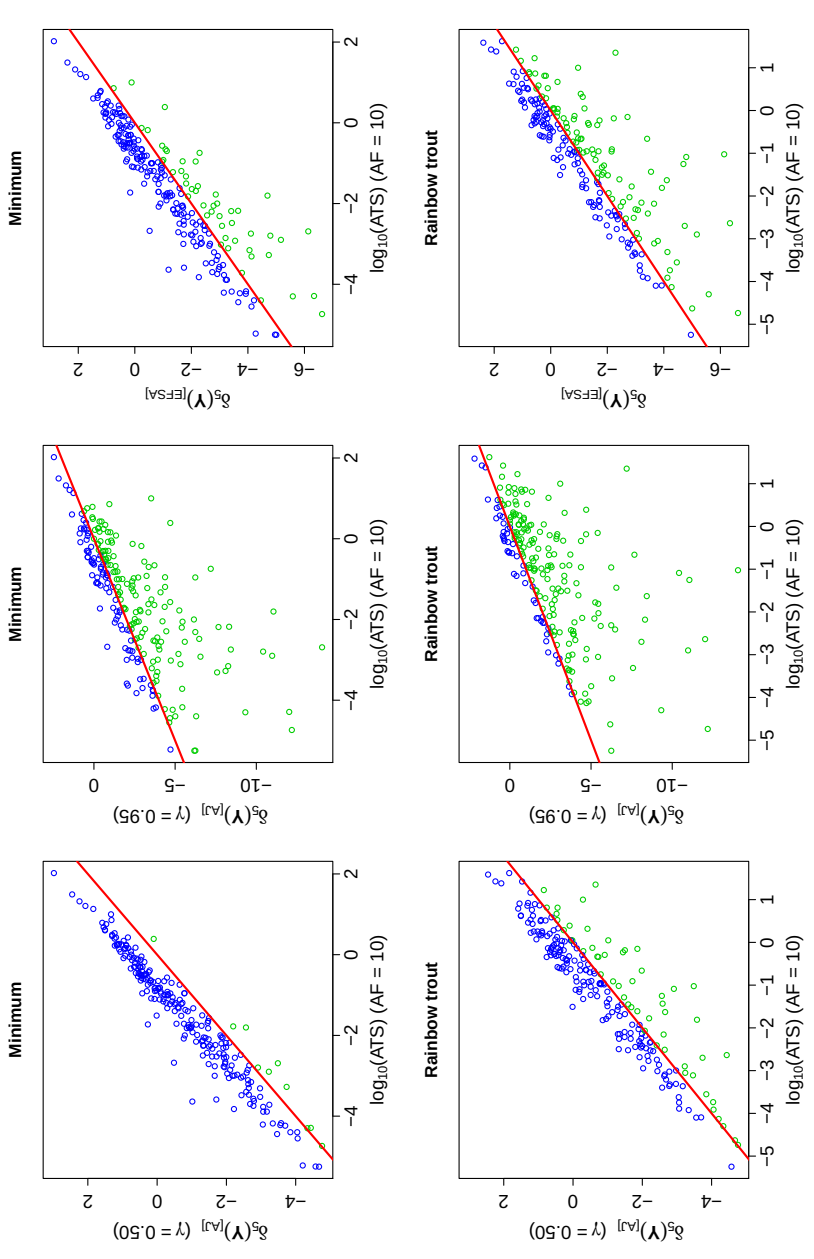


Figure 6.1: Plot of $\log_{10}(\text{ATS})$ against $\delta_5(\mathbf{Y})$. Summary statistic used to determine the ATS (with $\text{AF}_{\text{spec}} = 10$) is: (i) the minimum tolerance value (top row); (ii) the rainbow trout's tolerance value (bottom row). Red line corresponds to $\log_{10}(\text{ATS}) = \delta_5(\mathbf{Y})$. $\log_{10}(\text{HC}_5)$ estimators are: [AJ] ($\gamma = 0.50$) (left); [AJ] ($\gamma = 0.95$) (middle); [EFSA] (right).

the median [AJ] $\log_{10}(\text{HC}_5)$ estimators and the $\log_{10}(\text{ATS})$ (correlation coefficient: $\rho = 0.97$ (minimum); $\rho = 0.92$ (rainbow trout)). A high degree of correlation between the [EFSA] estimators and the $\log_{10}(\text{ATS})$ was also observed ($\rho = 0.90$ (minimum); $\rho = 0.84$ (rainbow trout)), but less so compared to the conservative [AJ] estimator ($\rho = 0.68$ (minimum); $\rho = 0.61$ (rainbow trout)). The latter is mainly due to a number of datasets with $n_i = 4$, 5. Note that these correlations hold for all AFs since linear correlation is independent of positive affine transformations. For 96% of pesticides, the median [AJ] estimator was relatively less protective than the ATS based on the minimum tolerance value summary statistic. For the conservative [AJ] and [EFSA] estimators, the corresponding fractions were 37% and 79% respectively. The corresponding fractions for the three estimators relative to the ATS based on the rainbow trout tolerance value summary statistic are 78%, 27% and 56% respectively. It is intuitively correct that the latter fractions are less than the former fractions because the rainbow trout's tolerance value will always be greater than or equal to the minimum tolerance value. Furthermore, the fractions are consistent with earlier findings that the conservative [AJ] estimator is typically more conservative than the [EFSA] estimator, which is always more conservative than the median [AJ] estimator.

6.3 A Generalised Decision Rule

In this section we explore a generalised deterministic decision rule based only the rank order of the toxicity data. Through a probabilistic lens, we focus on making recommendations regarding the application of this decision rule. Consider a chemical risk assessment which is to be based on n distinct species log-tolerance values $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$; we use capitalisation to emphasise random variables prior to observation. The SSD assumption under the usual species exchangeable context states that all log-tolerance realisations are independent and identically distributed from a distribution function F_Y. Although we have thus far adopted the SSD assumption of normality over log-concentration, we work within greater generality and only assume that F_Y is continuous for the time being. We denote the

order statistics of log-tolerance values to be $Y_{(1:n)} \leq Y_{(2:n)} \leq \cdots \leq Y_{(n:n)}$.

In Section 2.5 it was reported that $\log_{10}(\text{ATS})$ is obtained by applying an assessment shift K to the summary statistic of observed toxicity data (on original concentration scale). By defining the summary statistic to be an order statistic, we subsequently denote $T(n, i, K)$ to be $\log_{10}(\text{ATS})$, defined as

$$T(n, i, K) = Y_{(i:n)} - K. \tag{6.1}$$

This is effectively a decision rule for choosing the PNEC. Setting $i = 1$ retrieves the standard decision rule such that 10^K is equivalent to the assessment factor, and $10^{Y_{(1:n)}}$ is the lowest tolerance value (on the original concentration scale).

For purposes of evaluating this generalised decision rule, we will use the MFE, as proposed in EFSA (2005), as a measure of risk. The MFE calculates the statistically expected proportion of species whose (log-)tolerance values will be exceeded at a concentration less than some particular environmental exposure (log-)concentration; in this case the log-concentration is $T(n, i, K)$. Unlike other proposals of measuring the PAF, the MFE doesn't require specification of additional control parameters. Moreover, it is attractive from a risk management perspective because of its mathematical closed form. By definition, the MFE is equivalent to

$$\int_{-\infty}^{\infty} f_T(t) F_Y(t) dt = 1 - \int_{-\infty}^{\infty} F_T(t) f_Y(t) dt,$$

whereby the alternative expression defined by the right-hand side arises from integration by parts.

It is assumed that each Y_i can be decomposed as $Y_i = \mu + \sigma Z_i$ where Z_i is a realisation from a standardised distribution F_Z having expectation zero and variance one. The parameters μ and σ^2 therefore represent the population mean and variance over log-concentration respectively. It is straightforward to deduce that the i-th

order statistic $Y_{(i:n)}$ has a distribution function equal to

$$\begin{aligned} F_{Y_{(i:n)}}(y) &= F_{\beta(i,n-i+1)}\left(F_Y\left(y;\mu,\sigma\right)\right) \\ &= F_{\beta(i,n-i+1)}\left(F_Z\left(\frac{y-\mu}{\sigma}\right)\right), \end{aligned} \tag{6.2}$$

where $F_{\beta(i,n-i+1)}$ is the cumulative distribution function (CDF) of a Beta distribution with shape parameters i and $n-i+1$. Hence, we can extend this to determine the CDF and probability density function (PDF) of the generalised decision rule as

$$\begin{aligned} F_{T_{(n,i,K)}}(t) &= F_{\beta(i,n-i+1)}\left(F_Z\left(\frac{t-\mu+K}{\sigma}\right)\right); \\ f_{T_{(n,i,K)}}(t) &= f_{\beta(i,n-i+1)}\left(F_Z\left(\frac{t-\mu+K}{\sigma}\right)\right) f_Z\left(\frac{t-\mu+K}{\sigma}\right)\frac{1}{\sigma} \end{aligned}$$

respectively; where f_Z is the PDF of the random variable Z. Combining the definition of the MFE with the distributions above yields the following expressions for the MFE evaluated at $T(n,i,K)$

$$\begin{aligned} &\int_{-\infty}^{\infty} f_{\beta(i,n-i+1)}\left(F_Z\left(\frac{t-\mu+K}{\sigma}\right)\right) f_Z\left(\frac{t-\mu+K}{\sigma}\right) F_Z\left(\frac{t-\mu}{\sigma}\right)\frac{dt}{\sigma} \\ &\qquad = 1 - \int_{-\infty}^{\infty} F_{\beta(i,n-i+1)}\left(F_Z\left(\frac{t-\mu+K}{\sigma}\right)\right) f_Z\left(\frac{t-\mu}{\sigma}\right)\frac{dt}{\sigma}. \end{aligned}$$

Both integrals can be simplified via a change of variables to obtain

$$\begin{aligned} \text{MFE}(n,i,K^*) &= \int_{-\infty}^{\infty} f_{\beta(i,n-i+1)}\left(F_Z\left(z+K^*\right)\right) f_Z\left(z+K^*\right) F_Z\left(z\right) dz \qquad (6.3) \\ &= 1 - \int_{-\infty}^{\infty} F_{\beta(i,n-i+1)}\left(F_Z\left(z+K^*\right)\right) f_Z\left(z\right) dz, \qquad (6.4) \end{aligned}$$

where K^* is denoted as the *standardised assessment shift* in EFSA (2005, Appendix A), i.e. $K^* = K/\sigma$. From here onwards, we refer to the MFE as a function of standardised assessment shift, i.e. $\text{MFE}(n,i,K^*)$.

6.3.1 Comparing Decision Rules

Subject to certain criteria, it is plausible that a risk assessment may be based on a sample size of $n = 1$; see for example the third row of Table 6.1. However, additional conditions in the assessment guidelines may be applicable (not displayed in Table 6.1) which may preclude the use. Furthermore, assessment factors for other regulatory contexts (e.g. assessment of fish under the realm of EC 1991) may not allow for $n = 1$ in general. Since we seek a generalised decision rule, we nominally define $n = 2$ as being the minimum sample size which has wide reaching applicability to ERA. Thus, risk assessors will likely choose to use $n = 2$ unless there are other reasons to consider larger sample sizes. Henceforth, we define the *current protection threshold* to be $\text{MFE}(2, 1, K^*)$ for a given value of K^*, i.e. the MFE of applying an assessment factor to the minimum of two tolerance values. This threshold defines a reference point which can be used to evaluate other assessment procedures.

We refer to a decision rule $T(n, i, K)$ as *desirable* if it satisfies $\text{MFE}(n, i, K^*) \leq \text{MFE}(2, 1, K^*)$ for some SSD over a relevant interval of K^*, namely $(0, K^*_{\text{max}}]$. In other words, the MFE of the decision rule does not exceed the *status quo* for some suitably large interval of standardised assessment factors, which may even be the entire positive real line. Whilst desirability is a useful criterion for classifying decision rules, there may exist multiple rules for fixed order statistic index i. In order to aid a risk manager, a stronger criterion may be sought after.

Regulators would benefit from revising the decision rule which yields the current protection threshold to $T(n, 1, K)$ for any $n > 2$. In other words, fixing the order statistic to be the minimum observed tolerance value, but increasing the minimum sample to be larger than $n = 2$ would yield an MFE that is desirable. On the other hand, this revised decision would lead to a sure loss for industry (e.g. the chemical manufacturer or importer) since more laboratory tests are required. Now consider the decision rule $T(n, i, K)$ with $n \geq 2$ and $i > 1$; for $n = 2$ this will be unacceptable from a regulatory viewpoint. From an industry perspective, however, an increase in order statistic index may be beneficial. Hence, we define the decision rule $T(n, 2, K)$ to be *mutually beneficial* to both the regulators and industry if $\text{MFE}(n, 2, K^*) \leq \text{MFE}(2, 1, K^*)$ for all $K^* \in (0, K^*_{\text{max}}]$ and there exists no other

$n' < n$ for which this inequality holds true. In other words, $T(n, 2, K)$ constitutes a desirable decision rule with the smallest possible sample size. Note that this mutually beneficial decision rule would still be employed with the same assessment factor applied by the current requirements.

Once can iterate this argument to obtain the next mutually beneficial decision rule $T(n, 3, K)$. In general, if we have a decision rule $T(m, j, K)$ which is classed as mutually beneficial, then the next mutually beneficial decision $T(n, i, K)$ is obtained by seeking the smallest $n \geq m$ and largest $i > j$ which satisfies

$$\text{MFE}(n, i, K^*) \leq \text{MFE}(m, j, K^*) \leq \text{MFE}(2, 1, K^*) \; \forall K^* \in (0, K^*_{\text{max}}]. \tag{6.5}$$

This procedure yields a sequence of mutually beneficial decision rules: $T(n_1, i_1, K)$, $T(n_2, i_2, K)$, $\ldots$, $T(n_{\text{max}}, i_{\text{max}}, K)$ with $i_1 < i_2 < \cdots < i_{\text{max}}$ and $n_1 \leq n_2 \leq \cdots \leq n_{\text{max}}$. We anticipate $n_{\text{max}} \lesssim 15$ because it is likely that beyond this value one will fall into the realm of intermediate quantitative tier hazard assessment methods, i.e. direct calculation of a HC_p. Each mutually beneficial decision rule $T(n_t, i_t, K)$ is made in reference to $T(n_{t-1}, i_{t-1}, K)$ (with $t = 0$ corresponding to the decision rule which yields the current protection threshold). One might consider defining each mutually beneficial decision rule with respect to $T(2, 1, K)$ only. However, it does not seem plausible that a risk manager would be satisfied with a situation where the MFE evaluated at $T(n_1, i_1, K)$ does not uniformly bound the MFE evaluated at $T(n_2, i_2, K)$ for $i_2 > i_1$ even if both decision rules are classed as desirable. For the sake of convenience, we refer to all decision rules in the aforementioned sequence as mutually beneficial.

In constructing the 'mutual benefit' criterion for choosing among the decision rules, we have discounted two factors. First, from an industrial viewpoint we have ignored the cost of testing additional species. However, it is anticipated that testing additional species would in many contexts outweigh the cost of performing a higher tier risk assessment. Second, from a regulatory viewpoint we have ignored stakeholder utility (e.g. public consultation) which may not rationalise the trade-off between n and i for fixed K.

As an example of our discussion, consider the MFE based on the log-normal SSD assumption; this is in line with the standard model discussed throughout this research. This assumption was also the basis of the decision rules for Method 2 in EFSA (2005, p. 28). By noting that Z has a standard normal distribution, then this implies, as shown in EFSA (2005), that

$$\text{MFE}(n, i, K^*) = 1 - \int_{-\infty}^{\infty} F_{\beta(i,n-i+1)} \left(\Phi\left(z + K^*\right)\right) \phi(z) dt,$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ respectively denote the CDF and PDF of a standard normal distribution.

In order to explore the generalised deterministic decision rule $T(n, i, K)$ we numerically evaluate the MFE for: sample sizes $2 \leq n \leq 13$; order statistics $1 \leq i \leq \min(n, 4)$; and standardised assessment shift $K^* \in [0, 4]$. This interval of standardised assessment shifts is deemed sufficient for all intents and purposes. However, subject to the magnitude of the SSD variance, this may require modification and would likely be a policy decision. Figure 6.2 plots $\text{MFE}(n, i, K^*)$ against K^* for the selected sample sizes and order statistics; each line colour and line-type corresponds to a distinct pair (n, i). Note that the solid red curve corresponds to the current protection threshold. To aid in analysing Figure 6.2, we additionally plot the MFE curves relative to the current protection threshold (see Figure 6.3).

With reference to the current protection threshold, it is observed from Figure 6.2 that the decision rule corresponding to $(n, i) = (3, 1)$ is desirable over $K^* \in [0, 4]$. Further inspection indicates that $(n, i) = (5, 2)$ corresponds to a mutually beneficial decision rule since its MFE is uniformly smaller than the current protection threshold over the specified region of K^*, but not for $(n, i) = (4, 2)$. Furthermore, two additional mutually beneficial decision rules are identified corresponding to $(n, i) = (8, 3)$ and $(11, 4)$. Figure 6.4 is a de-cluttered plot of the three mutually beneficial decision rules for the log-normal SSD identified above, with the current protection threshold curve overlaid.

A number of additional observations from the analysis are worth mentioning. First, for fixed sample size the MFE decreases as the order statistic index decreases

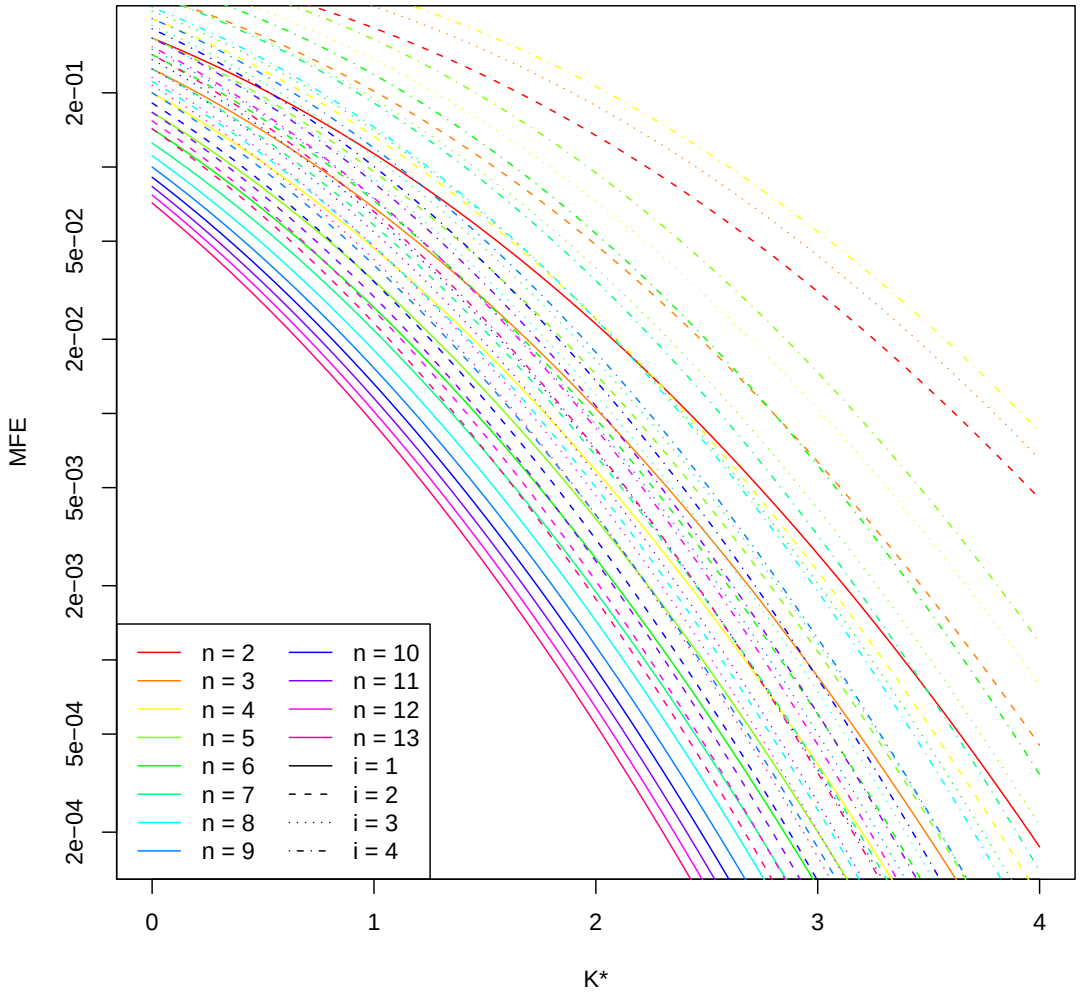


Figure 6.2: MFE(n, i, K^*) versus K^*.

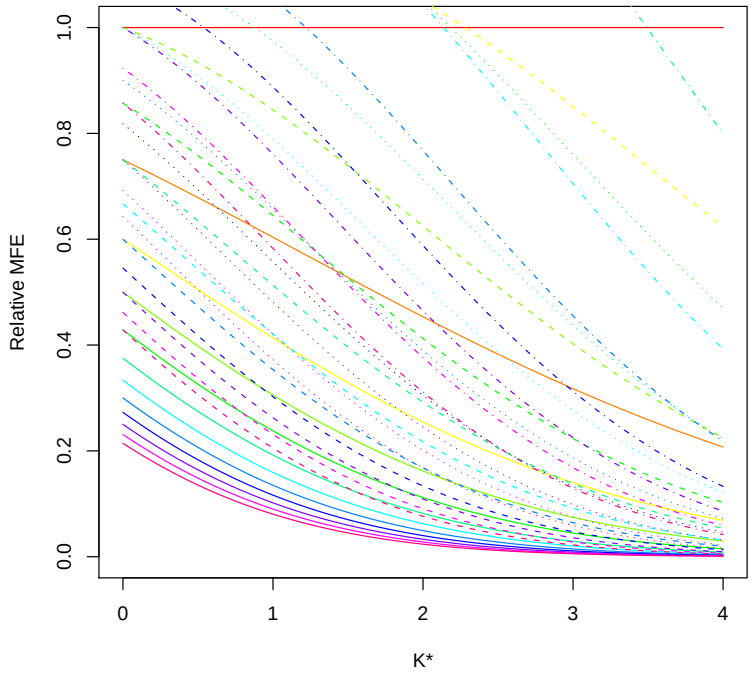


Figure 6.3: $\text{MFE}(n, i, K^*)/\text{MFE}(2, 1, K^*)$ versus K^*.

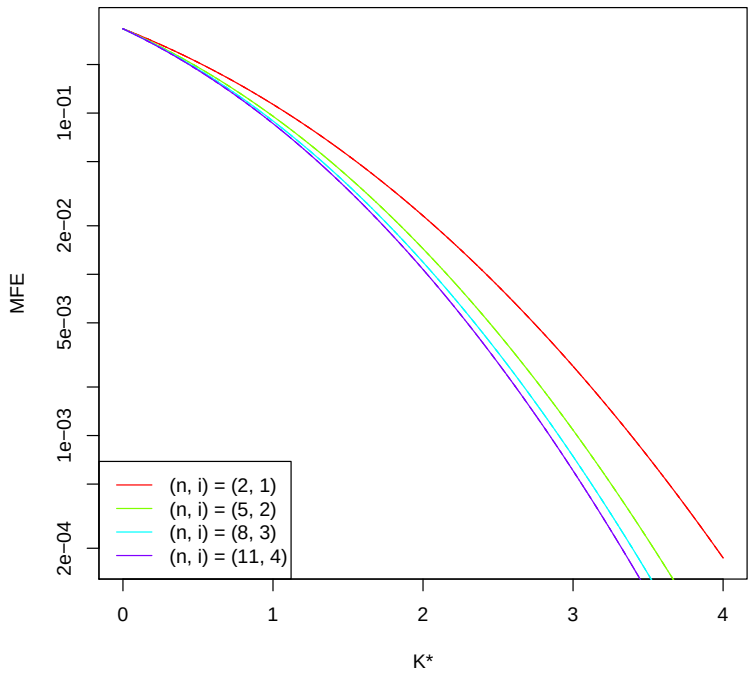


Figure 6.4: $\text{MFE}(n, i, K^*)$ versus K^* for mutually beneficial decision rules.

with strict inequality for all $K^* \neq 0$. Second, for fixed order statistic index the MFE decreases as the sample size increases. Thus, $\text{MFE}(n, i, K^*)$ is a decreasing function as either n increases, i decreases, or as K^* increases. However, a caveat is that these points are only shown to be true for the log-normal SSD; we analyse them for other SSDs later on. Additionally, by properties of the Beta distribution, it is clear that the PAF of species variability decreases as we move from mutually beneficial decision rule $T(m, j, K)$ to the next mutually beneficial decision rule $T(n, i, K)$ such that $i > j$. Both parties — regulators and industry — stand to gain from this because the probability of a high PAF is lowered which consequently implies the chance of triggering a higher tier risk assessment is lowered.

6.4 Evaluating Robustness

In this section we evaluate the generalised decision rule defined in the previous section under a range of different SSDs. In particular, we focus on SSDs over log-concentration which deviates slightly from normality. This can be observed as an *ad hoc* empirical analysis of robustness; a similar analysis was performed by EFSA (2008) for decision rules which employed the geometric mean as a summary statistic. The various families of distributions we consider is intended to capture key features of where an SSD (over log-concentration) may depart from normality, for example, skewness, fat/long tails, peakedness and bimodality. Our choice of distributions to explore these properties is not exhaustative, but rather chosen to be representative and offer flexibility (through shape parameters). With the exception of the logistic SSD (chosen because of its prevalence in some regulatory sectors), all families yield the normal distribution as a special case. Larger families of distributions, for example the Johnson SU distribution (Johnson, 1949), which offer more flexibility via larger numbers of shape parameters, might be an appropriate avenue of further research if combinations of different properties required evaluation of robustness.

6.4.1 Distributions

Here we give brief details of the distribution functions we consider, namely: the logistic distribution; the skew-normal distribution; Student's t-distribution; the exponential power distribution and a class of normal mixture distributions. Note that all distributions we describe are over log-concentration, that is if X is the tolerance value of a random species drawn from the SSD, then $Y = \log_{10}(X)$ is the log-tolerance value of this species. For this section, we will use the term SSD to refer to distributions of Y. Also, note that all distributions are fully defined on the real line.

The Logistic Distribution

The logistic distribution, parameterised by mean μ and scale s, has a PDF

$$f(y;\mu,s) = \frac{\exp\{-(y-\mu)/s\}}{s\left(1+\exp\{-(y-\mu)/s\}\right)^2}.$$

If Y has a logistic distribution, then $\mathbb{E}Y = \mu$ and $\mathrm{Var}(Y) = \pi^2 s^2/3$.

The logistic distribution is frequently assumed by many SSD proponents (Aldenberg and Slob, 1993; Traas et al., 2002; Dyer et al., 2006); this is mainly because of its complete analytically tractability. Recently, however, the normal distribution has become recognised as the more pragmatic choice (ECHA, 2008a).

The Skew-Normal Distribution

The skew-normal (SN) distribution, parameterised by location ω, scale ψ, and shape α, has a PDF

$$f(y;\omega,\psi,\alpha) = \frac{2}{\omega}\phi\left(\frac{y-\psi}{\omega}\right)\Phi\left(\alpha\frac{y-\psi}{\omega}\right).$$

If Y has a SN distribution, then $\mathbb{E}Y = \psi + \omega\sqrt{2/\pi}\delta$ and $\mathrm{Var}(Y) = \omega^2(1-2\delta^2/\pi)$, where $\delta = \alpha/\sqrt{1+\alpha^2}$.

The SN distribution is a generalisation of the normal distribution (Azzalini, 1985), and might be useful for ecological compositions which demonstrate some degree of skewness over log-transformed concentration. It is implemented in `R` (R,

2006) through the `sn` package (Azzalini, 2008). As $\alpha \to \pm\infty$, the SN distribution tends towards the half-normal distribution; when $\alpha = 0$, the normal distribution is recovered.

Student's t Distribution

Student's t-distribution, parameterised by ν degrees of freedom, has a PDF

$$f(y;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{y^2}{\nu}\right)^{-(\frac{\nu+1}{2})}.$$

If Y has a Student t-distribution, then $\mathbb{E}Y = 0$ and $\mathrm{Var}(Y) = \nu/(\nu-2)$ which is defined for $\nu > 2$.

Student's t-distribution has longer tails for small degrees of freedom, but its distribution function can be shown to tend towards a standard normal distribution function as $\nu \to \infty$.

The Exponential Power Distribution

The exponential power (EP) distribution, parameterised by mean μ, scale σ, and shape $p \geq 1$, has a PDF

$$f(y;\mu,\sigma,p) = \frac{1}{2p^{1/p}\sigma\Gamma(1+1/p)}\exp\left\{-\frac{|y-\mu|^p}{p\sigma^p}\right\}.$$

If Y has an EP distribution, then $\mathbb{E}Y = \mu$ and $\mathrm{Var}(Y) = p^{2/p}\sigma\Gamma(3/p)/\Gamma(1/p)$.

The EP distribution (also known as the generalised error distribution) is a class of distributions which allows for varying degrees of tails; see Nadarajah (2005) for further details. It is implemented in `R` (`R`, 2006) through the `normalp` package (Mineo and Ruggieri, 2005). When $p = 2$ the normal distribution is recovered; when $p = 1$ the Laplace distribution is recovered.

Bimodal Normal Mixture Distribution

The bimodal normal mixture (BNM) distribution, parameterised by locations (μ_1, μ_2), scales (σ_1, σ_2), and weight ω, has PDF

$$f(y; \mu_1, \mu_2, \sigma_1, \sigma_2, \omega) = \frac{\omega}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + \frac{(1-\omega)}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right),$$

where $\phi(\cdot)$ is the PDF of a standard normal random variable. If Y has a BNM distribution, then $\mathbb{E}Y = \omega\mu_1 + (1-\omega)\mu_2$ and $\mathrm{Var}(Y) = \omega^2\sigma_1^2 + (1-\omega)^2\sigma_2^2$.

The BNM is a non-standard distribution that we explore for purposes of assessing bimodality. This property is clearly plausible for multi-taxa species communities where the toxicant has a very specific mode of action which may target one (possibly small) group of species since it represents two clusters of species. If ω tends towards either zero or unity, then we obtain a normal distribution. The normal distribution is also recovered if one sets $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$ for any $\omega \in [0, 1]$.

6.4.2 Analysis

Of the distributions we described, we only explored certain cases in each. In all cases, we only consider the standardised distributions, i.e. having expectation 0 and variance 1, by consequence of Equation 6.3. The MFE curves are not necessarily comparable for fixed (n, i) across K^* since for fixed K, K^* may be shape dependant. This does not preclude our robustness analysis, since we are only interested in the robustness of the previously identified mutually beneficial decision rules.

For the SN distribution we considered $|\alpha| = 1, \ldots, 5$; for Student's t-distribution we considered at $\nu = 3, 4, 5$; for the EP distribution we considered $p = 1.5$, 2.5, 3, 5 and for the BNM distribution we consider $\omega = 0.1$, 0.5 and 0.9 with $\mu_2 = -\mu_1 = 2$ and $\sigma_1 = \sigma_2 = 1$. For the BNM distribution, other parameterisations provide much different results, however the key features for discussion are sufficiently captured by these choices. In Figure 6.5 we display the *standardised* PDF and CDF for each distribution across different parameterisations. For each plot we also show the standard normal PDF and CDF (bold red curve).

By considering plots similar to those of Figures 6.2–6.4, we are able to determine

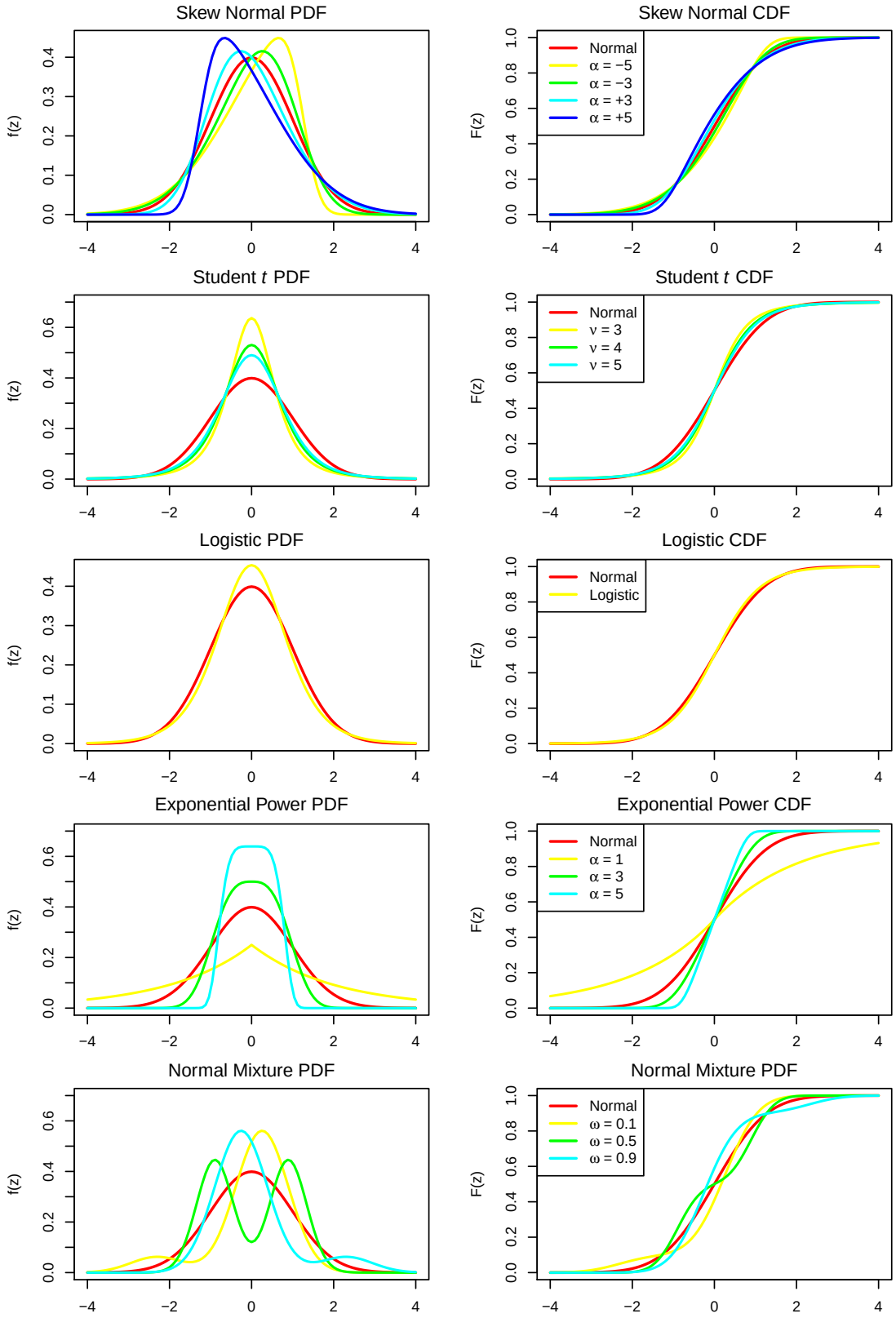


Figure 6.5: Standardised distributions of alternative SSDs.

that for an interval of $K^* \in (0, 4]$, the aforementioned mutually beneficial decision rules (see Section 6.3.1) appear to coincide with those determined for the other distributions analysed with the exception of the BNM distribution with $\omega = 0.1$.

In the interest of clarity we do not display all the corresponding versions of Figures 6.2–6.4, but instead display the MFE curves (left panel) corresponding to the mutually beneficial decision rules [based on $(n, i) = (5, 2)$, $(8, 3)$, $(11, 4)$] and the current protection threshold [$(n, i) = (2, 1)$], over $K^* = [0, 4]$, for various parameterisations in. These plots are shown in: Figure 6.6 (logistic); Figure 6.7 (SN); Figure 6.8 (Student t); Figure 6.9 (EP) and Figure 6.10 (BNM). In addition, for each plot (right panel) we also show the MFE curves relative to the current protection threshold curve.

It is clear from Figure 6.10 (right panel) that the BNM distribution with $\omega = 0.1$ is exceptional such that the decision rules based upon procedures of $(n, i) = (5, 2)$, $(8, 3)$ and $(11, 4)$ are not mutually beneficial because the MFE curves corresponding to the former procedures cross the current protection threshold curve. The first crossing occurs at $K^* \approx 1.1$, and the second crossing occurs at $K^* \approx 2.5$ at which point mutual benefit is restored. For this standardised SSD, the mutually beneficial decision rules (with $K^*_{\max} = 4$) were determined from Figure 6.11 to correspond to the procedures: $(n, i) = (6, 2)$, $(11, 3)$, and $(16, 4)$. As K^* becomes appreciably large it will be the left 'hump' of the BNM (which is a normal distribution) that will have most mass. Thus the MFE curves for very large K^* would be similar to those for the normal SSD.

An additional noticeable observation arises in the context of the Student-t SSD. The decision rules evaluated in Figure 6.8 were classifiable as mutually beneficial on the interval of standardised assessment shifts $K^* \in [0, 4]$. However, extending this interval to $K^* = [0, 8]$ (Figure 6.12) indicated that the decision rules are no longer mutually beneficial; this is similar to the case for the BNM ($\omega = 0.10$). Figure 6.12 (right panel) clearly demonstrates the crossing for $\nu = 3$ and $\nu = 4$; the crossing for $\nu = 5$ is not observable on the plotted domain. Further calculations verified that letting $\nu \to \infty$ leads to MFE curves coinciding with those determined under a normal SSD. It was further found that log-MFE curve plots (not displayed

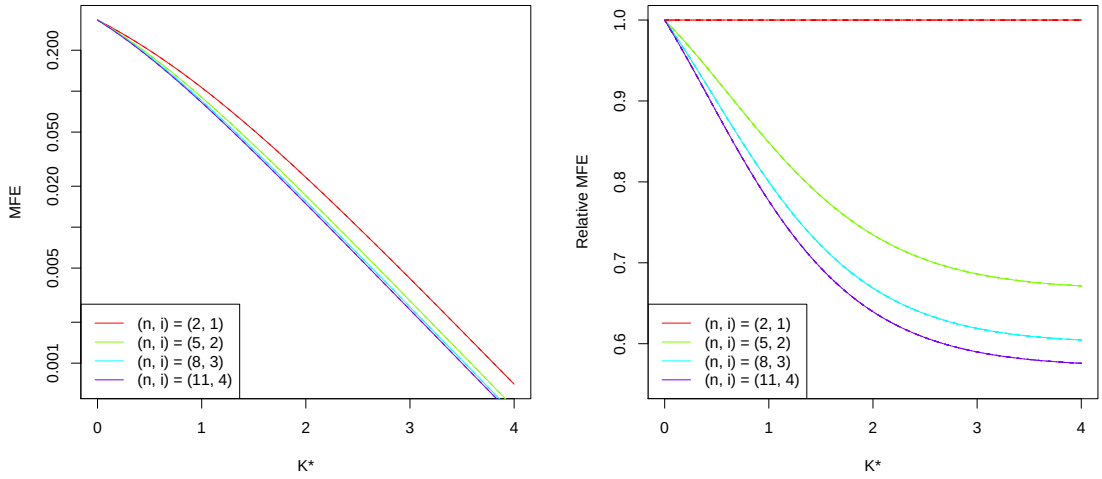


Figure 6.6: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the logistic SSD.

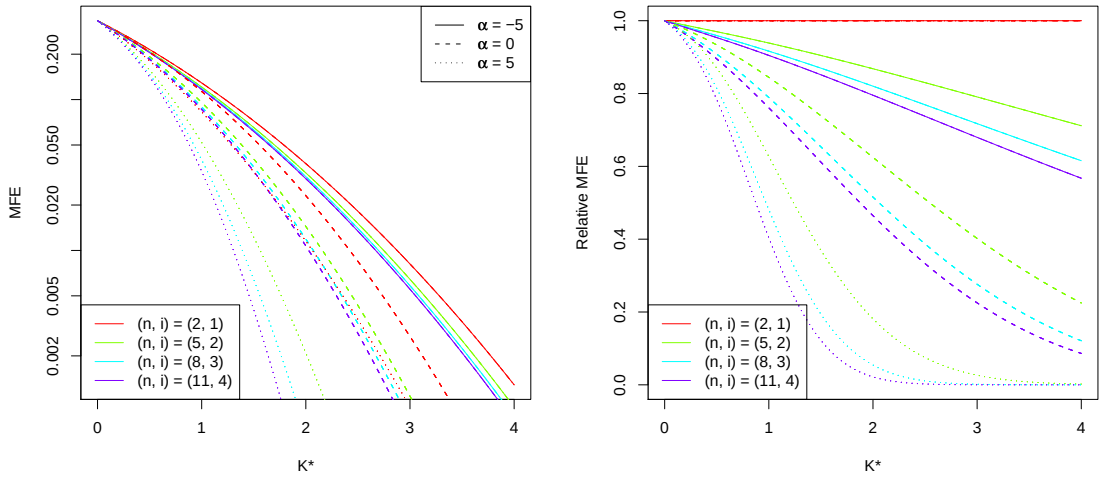


Figure 6.7: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the SN SSD. $\alpha = \pm 5$ (dotted and solid lines respectively) and $\alpha = 0$ (dashed lines; reduces to the normal distribution analysis).

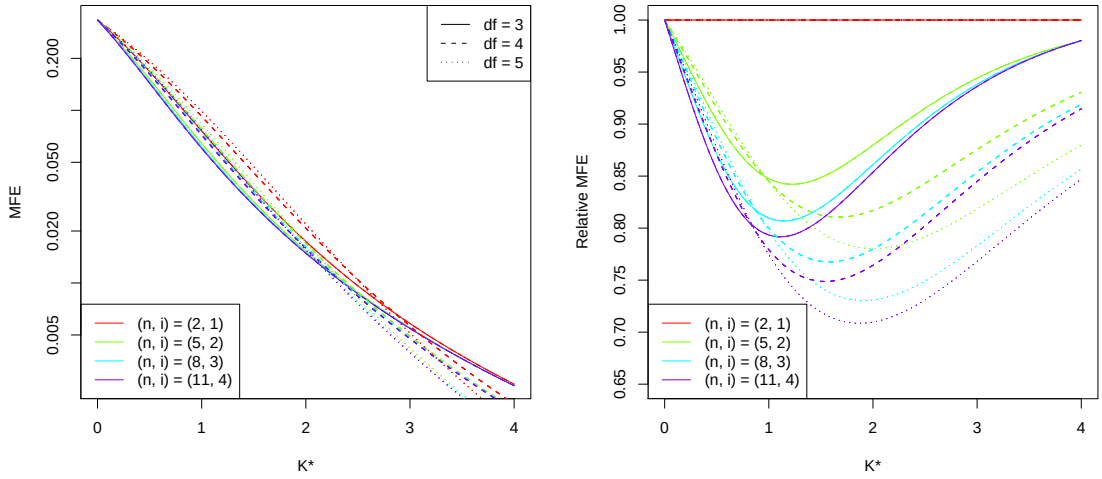


Figure 6.8: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the Student-t SSD. $\nu = 3$ (solid); $\nu = 4$ (dashed); and $\nu = 5$ (dotted).

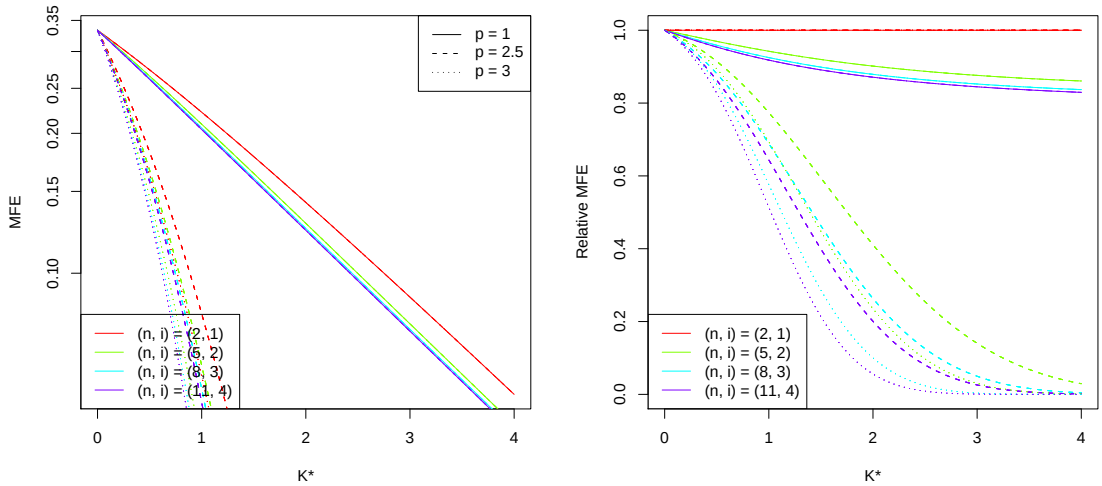


Figure 6.9: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the EP SSD. $p = 1$ (solid); $p = 2.5$ (dashed); and $p = 3$ (dotted).

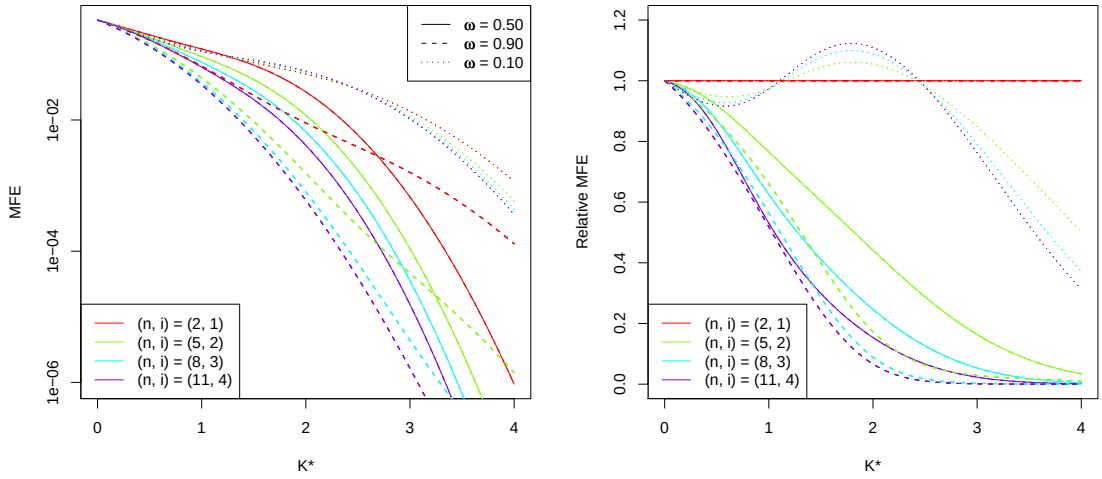


Figure 6.10: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the BNM SSD. $\omega = 0.10$ (dotted); $\omega = 0.50$ (solid); and $\omega = 0.90$ (dashed).

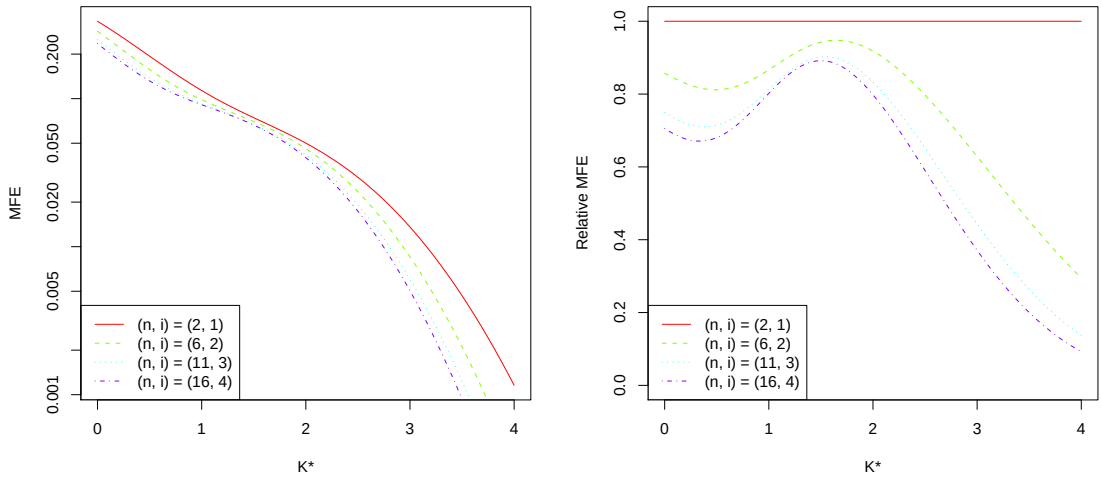


Figure 6.11: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(6, 2)$, $(11, 3)$, and $(16, 4)$ under the BNM SSD with $\omega = 0.10$.

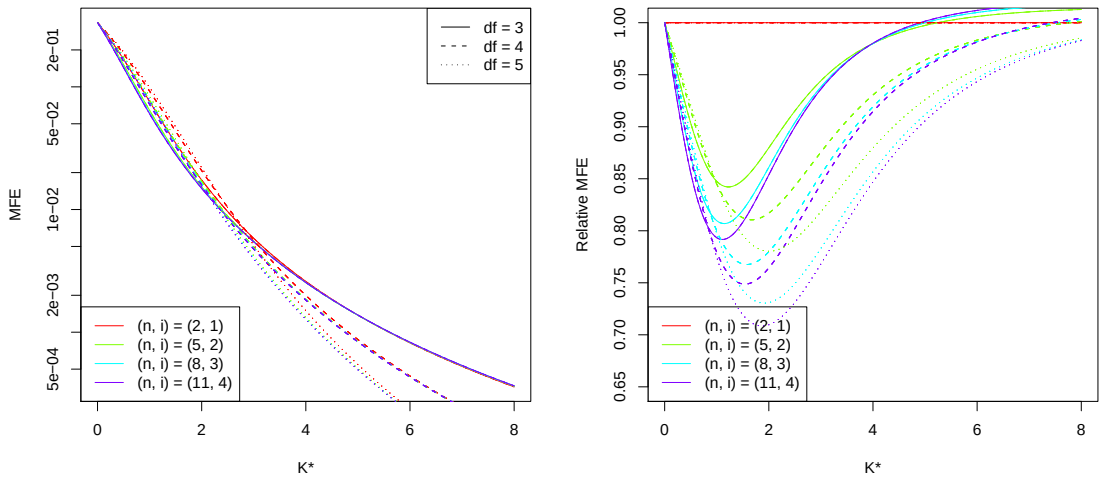


Figure 6.12: MFE (left panel) and relative MFE (right panel) curves for decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under the Student-t SSD. $\nu = 3$ (solid); $\nu = 4$ (dashed); and $\nu = 5$ (dotted). $K^* \in [0, 8]$.

here) for $\nu \lesssim 50$ were approximately convex and approximately concave for $\nu \gtrsim 50$. Inferences based on the evaluation of the aforementioned decision rules for $\nu \gtrsim 50$ were numerically consistent with other distributions including the normal SSD.

In light of this, it is plausible that for some risk assessments a domain of $K^* \in [0, 4]$ may not suffice in justification of the mutually beneficial decision rules, even when unimodality is assumed. However, heuristically, if $K^* = 4$ — the approximate value where the decision rules failed to be mutually beneficial for the Student-t SSD with $\nu = 3$ — then the actual assessment factor applied (to the summary toxicity statistic on original concentration scale) will be approximately $10^{4\sigma}$. So for $\sigma \geq 1/4$, the currently implicitly suggested assessment factor of 10 will fall within the currently explored domain. Of course, one is not in a position to know σ in advance unless one assumes the model of Aldenberg and Luttik (2002) (cf. $\mathcal{M}_4$).

Finally, whilst the aforementioned mutually beneficial decision rules also coincided with those evaluated under the EP distribution (up to the specified parameterisations), it was noticeable that for small shape parameter p, the log-MFE curves were far less concave than those for other distributions.

Taxon	s_p	95% C.I.	ς	N
Fish	0.65	$(0.628, 0.675)$	1524	169
Crustaceans	0.92	$(0.877, 0.977)$	657	86
Mammals	0.36	—	—	69

Table 6.2: Pooled standard deviations of log transformed (base 10) toxicity data (EC_{50}s) stratified by taxon. Reproduced from EFSA (2005, Table 2). C.I. = confidence interval; ς was defined earlier to be the total degrees of freedom.

6.5 Empirical Reassessment

Thus far we have evaluated a generalised deterministic decision rule under a number of different SSDs to explore robustness. In particular, a set of mutually beneficial decision rules were arrived at from these considerations. That is, a set of decision rules where a trade-off is made between regulators and industry which ensures that the current undefined mean level of protection accountable to interspecies variability is not exceeded; in fact it is lower for most decision rules.

In this section we reassess the mutually beneficial decision rules from an application viewpoint. Such a reassessment is precluded using the methodology derived above since the MFE — a summary measure of the level of protection — is specified over standardised assessment shifts. One approach to overcome this is to recall that $K = K^*\sigma$, where conditional on σ, K has an interpretation which is understood by risk managers, i.e. that of a log transformed assessment factor. Earlier analysis in Chapter 3 suggested σ was heterogeneous between substances; notwithstanding this, we adopt the basis of $\mathcal{M}_4$, in particular that of fixing $\sigma = s_p$. The hypothesis of homogeneity for σ was not supported for the taxonomic group of fish (see Section 3.8), nonetheless it has been applied in Luttik and Aldenberg (1997); Aldenberg and Luttik (2002) and EFSA (2005). EFSA (2005, Table 2) presented pooled standard deviation estimates for SSDs over log-concentration for different taxonomic groups based on a larger pesticide toxicity database; this RIVM database is a subset which is described in De Zwart (2002). We reproduce these values for three of the taxonomic groups in Table 6.2.

By substituting $\sigma = s_p$ for a specific taxonomic group listed in Table 6.2, we

can rescale the abscissa such that it reflects the log assessment factor (K), thus being comparable to those in Table 6.1 (up to a exponential transform). It should be noted that values of K are not independent of behavioural model, whereas lower tier quantitative assessment generally is without reference to a model, hence caution is required in interpreting and communicating this analysis to end users. In the interests of clarity, we limit the evaluation of the mutually beneficial decision rules to the log-normal and log-logistic SSD; extension to other distributions would be relatively straightforward. In Figure 6.13 we show MFE curves for each taxonomic group, plotted over log-assessment factor, for the aforementioned mutually beneficial decision rules; log-normal (left panel) and log-logistic SSD (right panel). Note that since our treatment of assessment factor extrapolation is pertinent to $\mathrm{AF}_{\mathrm{spec}}$ only (see Section 2.5), we only consider the domain $K \leq 2$, i.e. $\mathrm{AF}_{\mathrm{spec}} \leq 100$. This is because the relevant official sources for assessment factors appear to implicitly suggest $\mathrm{AF}_{\mathrm{spec}} = 10$ or $\mathrm{AF}_{\mathrm{spec}} = 100$; larger overall assessment factors will be relevant to additional uncertainties not incorporated in the SSD model.

There is a further caveat on the interpretation of this empirical reassessment because the standard deviation values used in the analysis were obtained using acute EC_{50} toxicity data. If one was interested in interpreting the results from a risk assessment based on chronic data, then it would need to be assumed that the pooled standard deviations are representative of those for chronic data. Evidence in Roex et al. (2000) and Craig (2006) is inconclusive as to whether this assumption is valid and further investigation is clearly required.

Since the fish taxon has been a key feature of examples presented in this research, we display the MFE values for the mutually beneficial decision rules in Table 6.3 for fixed assessment factor values. In addition, we display the MFE for decision rules which corresponds to $(n, i) = (3, 1)$ since this is admissible into many risk assessment dossiers. As per Figure 6.13, we only tabulate values of $\mathrm{AF}_{\mathrm{spec}}$ considered practical, namely, those less than two orders of magnitude.

It is clear from Figure 6.13 that if the assessment procedures were applied blindly to all taxa, then taxa which exhibit less variability in toxicological sensitivity will be relatively more protected than those taxa with higher variability. However, such

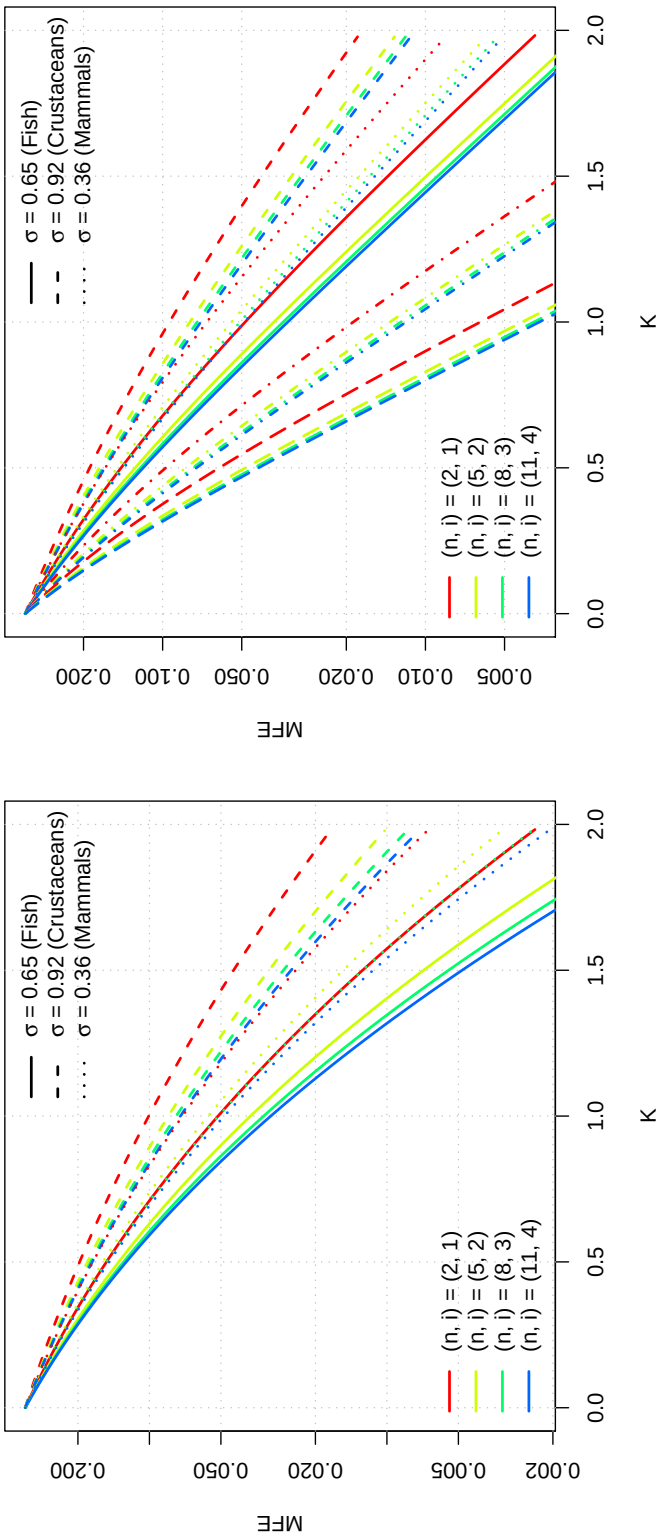


Figure 6.13: Per-taxon MFE curves for mutually beneficial decision rules corresponding to $(n, i) = (2, 1)$; $(5, 2)$; $(8, 3)$ and $(11, 4)$ under: log-normal SSD (left panel) and log-logistic SSD (right panel). $K = \log_{10}(\mathrm{AF}_{\mathrm{spec}})$.

			Normal		
AF_{spec}	(2, 1)	(3, 1)	(5, 2)	(8, 3)	(11, 4)
1	3.33×10^{-01}	2.50×10^{-01}	3.33×10^{-01}	3.33×10^{-01}	3.33×10^{-01}
2	2.15×10^{-01}	1.47×10^{-01}	2.01×10^{-01}	1.97×10^{-01}	1.94×10^{-01}
5	1.02×10^{-01}	6.06×10^{-02}	8.48×10^{-02}	7.87×10^{-02}	7.56×10^{-02}
10	5.14×10^{-02}	2.68×10^{-02}	3.75×10^{-02}	3.30×10^{-02}	3.08×10^{-02}
20	2.30×10^{-02}	1.04×10^{-02}	1.44×10^{-02}	1.18×10^{-02}	1.07×10^{-02}
50	6.61×10^{-03}	2.43×10^{-03}	3.20×10^{-03}	2.38×10^{-03}	2.02×10^{-03}
100	2.23×10^{-03}	6.86×10^{-04}	8.60×10^{-04}	5.81×10^{-04}	4.69×10^{-04}
			Logistic		
1	3.33×10^{-01}	2.50×10^{-01}	3.33×10^{-01}	3.33×10^{-01}	3.33×10^{-01}
2	2.08×10^{-01}	1.43×10^{-01}	1.94×10^{-01}	1.89×10^{-01}	1.86×10^{-01}
5	9.58×10^{-02}	5.91×10^{-02}	8.02×10^{-02}	7.53×10^{-02}	7.30×10^{-02}
10	4.85×10^{-02}	2.79×10^{-02}	3.78×10^{-02}	3.48×10^{-02}	3.34×10^{-02}
20	2.32×10^{-02}	1.27×10^{-02}	1.70×10^{-02}	1.55×10^{-02}	1.48×10^{-02}
50	8.22×10^{-03}	4.30×10^{-03}	5.75×10^{-03}	5.19×10^{-03}	4.95×10^{-03}
100	3.65×10^{-03}	1.87×10^{-03}	2.50×10^{-03}	2.25×10^{-03}	2.15×10^{-03}

Table 6.3: MFE values evaluated at decision rules of the form $T(n, i, K)$ (assuming $\sigma = s_p$) for the fish taxon. The EFSA (2005) pooled standard deviation is the used value of s_p.

an implication, as drawn here, is conditional on the assumption of homogeneity of σ for each taxon.

6.6 A Theoretical Inspection

In the following sections we theoretically explore the notion of the generalised deterministic decision rule $T(n, i, K)$ from the analytical probabilistic viewpoint in order to better defend their applicability for risk assessment. Previous sections rely upon numerical and graphical validation techniques which would most likely satisfy the requirements of risk managers, and aid in the communication of our proposal. However, theoretical elaboration can strengthen the defensibility, as well as lead to greater insight.

6.6.1 Useful Results

Here we describe a few useful results which provide extra insight into the theoretical structure of the MFE evaluated at the generalised deterministic decision rule. A trivial, but nonetheless important, observation is that the MFE evaluated at $T(n, i, K)$ for a given SSD is non-increasing as i increases.

Proposition 6.6.1 *For any SSD,* $\text{MFE}(n, i, K^*) \geq \text{MFE}(n, k, K^*)$ *for* $n \geq i \geq k$.

Proof First note that $Y_{k:n} \leq Y_{i:n}$ for $i \geq k$. Since the potentially affected fraction (PAF) of species affected is an increasing function in environmental concentration, we have that

$$\text{PAF}\left(T(n, k, K)\right) \leq \text{PAF}\left(T(n, i, K)\right).$$

Hence, by fact that the expectation operator is monotonic, the result is immediate. ∎

Having shown that the MFE is a decreasing function for decreasing order statistic index choice, we can also show that that the MFE is a decreasing function for increasing sample size, as demonstrated in the next proposition.

Proposition 6.6.2 *For any continuous SSD,* $\mathrm{MFE}(n_1, i, K^*) > \mathrm{MFE}(n_2, i, K^*)$ *for* $n_2 > n_1 \geq i$.

Proof From Equation 6.4 we have that

$$\begin{aligned} \mathrm{MFE}(n, i, K^*) &= 1 - \int_{-\infty}^{\infty} F_{\beta(i,n-i+1)}\left(F_Z\left(z + K^*\right)\right) f_Z(z) dz \\ &= 1 - \mathbb{E}^Z F_{\beta(i,n-i+1)}\left(F_Z\left(z + K^*\right)\right), \end{aligned}$$

where the expectation is taken with respect to the standardised SSD. Using a standard identity of the regularised incomplete Beta function (Abramowitz and Stegun, 1972), namely

$$F_{\beta(a,b)}(u) = F_{\beta(a+1,b)}(u) + \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)}(1-u)^b u^a,$$

we can also determine that

$$F_{\beta(a,b)}(u) = F_{\beta(a,b+1)}(u) - \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}(1-u)^b u^a.$$

This implies that

$$F_{\beta(a,b)}(u) \leq F_{\beta(a,b+1)}(u),$$

with a strict inequality if and only if $u \neq 0, 1$. It is immediate from this result that

$$F_{\beta(i,n-i+2)}\left(F_Z\left(z + K^*\right)\right) > F_{\beta(i,n-i+1)}\left(F_Z\left(z + K^*\right)\right)$$

uniformly on the finite real line; this is a type of first order stochastic dominance. Since the expectation operator is monotonic, we therefore determine that

$$\begin{aligned} 1 - \mathbb{E}^Z F_{\beta(i,n-i+2)}\left(F_Z\left(z + K^*\right)\right) &< 1 - \mathbb{E}^Z F_{\beta(i,n-i+1)}\left(F_Z\left(z + K^*\right)\right) \\ \Rightarrow \mathrm{MFE}(n+1, i, K^*) &< \mathrm{MFE}(n, i, K^*), \end{aligned}$$

and so

$$\mathrm{MFE}(n_2, i, K^*) < \mathrm{MFE}(n_1, i, K^*)$$

for any $n_2 > n_1 \geq i$. ∎

The special case of $K^* = 0$, which implies an assessment factor of $\mathrm{AF}_{\mathrm{spec}} = 1$, offers complete tractability in calculating the MFE of the generalised decision rule, moreover, it is completely independent of the SSD.

Proposition 6.6.3 *Define T_i^n to be a random variable which has a Beta distribution with shape parameters i and $n - i + 1$. Then we have that*

$$\mathrm{MFE}(n, i, K^*) = \mathbb{E}^{T_i^n} F_Z\left(F_Z^{-1}(T_i^n) - K^*\right), \tag{6.6}$$

where the expectation is taken with respect to the distribution of T_i^n. Furthermore, when $K^* = 0$,

$$\mathrm{MFE}(n, i, 0) = \frac{i}{n+1}$$

irrespective of the SSD.

Proof Apply the change of variables: $t = F_Z(z + K^*)$ to Equation 6.3 to obtain the expression given by Equation 6.6. By setting $K^* = 0$, this expression immediately reduces to

$$\mathrm{MFE}(n, i, 0) = \mathbb{E}^{T_i^n}[T_i^n] = \frac{i}{n+1}. \quad ∎$$

If we consider what we earlier defined to be the current protection threshold for any standardised assessment shift K^*, i.e. $\mathrm{MFE}(2, 1, K^*)$, then when $K^* = 0$ we have that the current protection threshold is equal to $1/3$. This means that any other decision rule $T(n, i, K)$ must satisfy $\mathrm{MFE}(n, i, 0) \leq \mathrm{MFE}(2, 1, 0) = 1/3$ if it is to be considered desirable or mutually beneficial. This result is used to state the next proposition. Note also that Equation 6.6 relates to another equation: $\mathbb{E}^Z F_Z\left(\bar{Z} - K^*\right)$, which defines the MFE when an assessment factor is applied to the geometric mean of the toxicity data. The latter equation is the focus of EFSA (2008).

Proposition 6.6.4 *For a given sample size $n \geq 2$, a necessary, but not sufficient, requirement for a generalised decision rule to be classed as mutually beneficial is for*

the order statistic index i to satisfy

$$i \leq \frac{n+1}{3}. \tag{6.7}$$

The proof is straightforward by Proposition 6.6.3.

For the decision rules corresponding to the assessment procedures of $(n, i) = (2, 1)$, $(5, 2)$, $(8, 3)$, and $(11, 4)$, mutual benefit is satisfied by equality in Equation 6.7 rather than inequality. Any value of i satisfying a strict inequality in this case would yield a decision rule classifiable as desirable to the risk manager by our earlier definition.

6.7 Stochastic Dominance

Given a set of decision rules $T(n, i, K)$, where effectively K is set in advance, then it is a risk management decision as to the choice of (n, i). Currently, regulatory guidelines require, in general, $(n, 1)$ for $n \geq 2$. As demonstrated in Section 6.3.1, other prescriptions of (n, i) can lead to decision rules which are mutually beneficial to risk managers and industry. However, the level of protection provided by each decision rule is uncertain, suggesting that a criterion is required in order to be able to adequately decide which estimator(s) should be considered; we advocated the MFE as a sufficient criterion, subject to a presumed SSD, or many SSDs.

By extending upon the definitions in Section 6.3.1, we classify a generalised decision rule $T(n, i, K)$ to be desirable over another $T(m, j, K)$, if for a given SSD, $\text{MFE}(n, i, K^*) \leq \text{MFE}(m, j, K^*)$ for all $K^* \in (0, K^*_{\text{max}}]$. Analytically showing where this property holds for pairs of decision rules is difficult and so we appeal to the theory of stochastic dominance (Hadar and Russel, 1969; Levy, 2006). In turn, we shall analytically demonstrate that the aforementioned mutually beneficial decision rules, indentified using numerical techniques do in fact satisfy stochastic dominance.

6.7.1 An Equivalent Measure of Protection

Stochastic dominance is a dominance argument used in the application of utility based decision theory (Berger, 1985). Basically, a utility function $u(\cdot)$, defined up to a positive affine transformation, acts on a set of possible outcomes obtainable via different decisions to admit real numbers used to rationalise a level of preference. To choose a decision under uncertainty, classical utility theory assumes that rational behaviour corresponds to maximising ones expected utility. For the problem here, the 'outcomes' are $T(n, i, K)$, and the 'utility function' can be vaguely interpreted as the PAF function (over log-concentration). An alternative viewpoint would be, conditional on K^* being fixed for all decision rule options, to consider the decisions as specifying (n, i) and the utility function as $F_Z(F_Z^{-1}(t) - K^*)$ (cf. Equation 6.6). The 'expected utility' in this instance would be $\text{MFE}(n, i, K^*)$. Due to the utility function literature focusing on the maximisation of expected utility, it will be more convenient for us to consider an equivalent function

$$V(u; K^*) = 1 - F_Z\left(F_Z^{-1}(1-u) - K^*\right).$$

$V(u; K^*)$ is equivalent to the former function, except that it measures the fraction of species potentially *unaffected* whilst still satisfying $V'(u; K^*) > 0$, i.e. protecting more species is preferred to protecting less. The reconsidered function allows us to straightforwardly exploit current utility theory and definitions. For what follows, we shall assume that $V(u; K^*)$ is known and well defined.

Define $U_i^n = 1 - T_i^n$, where T_i^n was defined in Proposition 6.6.3. It is straightforward to show that U_i^n has a Beta distribution with shape parameters $n - i + 1$ and i; notice that the shape parameters have interchanged from the distribution of T_i^n. Hence,

$$\begin{aligned} \mathbb{E}^{U_i^n} V(U_i^n; K^*) &= \mathbb{E}^{U_i^n}\left[1 - F_Z\left(F_Z^{-1}(1-U_i^n) - K^*\right)\right] \\ &= 1 - \text{MFE}(n, i, K^*). \end{aligned} \tag{6.8}$$

The separation of the underlying SSD and assessment procedure (n, i) in Equa-

tions 6.6 and 6.8 is highly important because it allows for the two components to be analysed separately.

To conclude our specification of $V(u;K^*)$, we also present the first and second derivatives with respect to u to be

$$V'(u;K^*) = \frac{f_Z\left(F_Z^{-1}(1-u)-K^*\right)}{f_Z\left(F_Z^{-1}(1-u)\right)}; \tag{6.9}$$

$$V''(u;K^*) = \frac{1}{f_Z^2\left(F_Z^{-1}(1-u)\right)}\left\{\frac{f_Z\left(F_Z^{-1}(1-u)-K^*\right)f_Z'\left(F_Z^{-1}(1-u)\right)}{f_Z\left(F_Z^{-1}(1-u)\right)} - f_Z'\left(F_Z^{-1}(1-u)-K^*\right)\right\}. \tag{6.10}$$

These functions are required later in demonstrating stochastic dominance.

6.7.2 Definitions

Recall our redefinition of a decision rules desirability, i.e. a decision rule $T(n,i,K)$ is desirable over another $T(m,j,K)$, if for a given SSD, $\text{MFE}(n,i,K^*) \leq \text{MFE}(m,j,K^*)$ for all $K^* \in [0, K^*_{\max}]$. By Equation 6.8, this definition is equivalent to $\mathbb{E}^{U_i^n} V(U_i^n;K^*) \geq \mathbb{E}^{U_j^m} V(U_j^m;K^*)$. Therefore, 'preference' in estimators is implied by a 'preference' of distribution for U_i^n; such a concept is referred to as 'orderability' under *stochastic dominance* (Hadar and Russel, 1969). This is better understood by expanding the final condition of desirability to show

$$\int_0^1 \left[f_{\beta(n-i+1,i)}(u) - f_{\beta(m-j+1,j)}(u)\right] V(u;K^*)du \geq 0. \tag{6.11}$$

Showing if, and when, Equation 6.11 holds is the focus of the remainder of this section. It is clear that $\left[f_{\beta(n-i+1,i)}(u) - f_{\beta(m-j+1,j)}(u)\right] \not\geq 0 \ \forall u \in [0,1]$, hence we must appeal to other methods. Two standard approaches to showing whether Equation 6.11 holds are first and second order stochastic dominance, denoted FOSD and SOSD respectively. Following the text of Hadar and Russel (1969) and Levy (2006), we provide the definitions of FOSD and SOSD below.

First Order Stochastic Dominance (FOSD)

Given two probability density functions $f(x)$ and $g(x)$ with corresponding cumulative distribution functions $F(x)$ and $G(x)$, and utility function $u(x)$, we say that f dominates g by *first order stochastic dominance* when $u'(x) > 0 \ \forall x$; and $F(x) \leq G(x) \ \forall x$ with at least one strict inequality.

To prove that FOSD does in fact satisfy stochastic dominance, we invoke integration by parts to retrieve

$$\begin{aligned} &\int_{-\infty}^{\infty} [f(x) - g(x)] \, u(x) dx \\ &= \ [\{F(x) - G(x)\} \, u(x)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} [F(x) - G(x)] \, u'(x) dx. \end{aligned}$$

The term on the left is clearly zero for both limits because $\lim_{x \to \pm\infty} [F(x) - G(x)] = 0$ by definition of cumulative distribution functions and $u(x)$ is bounded. By our hypothesis, the term on the right returns a non-negative quantity because $[F(x) - G(x)] \leq 0$ (with at least one value, x_0, for which the inequality is strict thus ensuring that f and g are not identical) under the assumption that $u'(x) > 0$ for all x.

Second Order Stochastic Dominance (SOSD)

Given two probability density functions $f(x)$ and $g(x)$ with corresponding cumulative distribution functions $F(x)$ and $G(x)$, and utility function $u(x)$, we say that f dominates g by *second order stochastic dominance* when $u'(x) > 0 \ \forall x$; $u''(x) \leq 0 \ \forall x$; and $\int_{-\infty}^{x} F(t) dt \leq \int_{-\infty}^{x} G(t) dt \ \forall x$ with at least one strict inequality.

To prove that SOSD also satisfies stochastic dominance, we follow on from first order stochastic dominance and repeat integration by parts for a second time to obtain

$$\begin{aligned} &\int_{-\infty}^{\infty} [f(x) - g(x)] \, u(x) dx \qquad (6.12) \\ &= \ - \left[\left\{ F^{(2)}(x) - G^{(2)}(x) \right\} u'(x) \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \left[F^{(2)}(x) - G^{(2)}(x) \right] u''(x) dx, \end{aligned}$$

where

$$\begin{aligned} F^{(2)}(x) &= \int_0^x F(t)dt = \int_0^x \int_0^t f(w)dwdt; \text{ and} \\ G^{(2)}(x) &= \int_0^x G(t)dt = \int_0^x \int_0^t g(w)dwdt \end{aligned}$$

are the second-order cumulative distribution functions. The assumptions listed above ensure that Equation 6.12 is non-negative.

6.7.3 Failure of FOSD

Failure to satisfy either FOSD or SOSD is not enough to say that stochastic dominance is violated. Higher n-th order stochastic dominance has been discussed in the economics literature; for examples consult Levy (2006) and references therein. There is good reason for focusing on FOSD and SOSD due to the assumptions placed on the behaviour of the utility function. Namely, (i) $u'(x) > 0$ and (ii) $u''(x) \leq 0$ are rational assumptions to make. We explained the logic behind (i) earlier; for (ii) the condition suggests risk aversion in the standard utility theory context (consult Levy 2006 for further elaboration).

FOSD is sometimes limited in its applicability, however subtle use was made of it earlier in proving that MFE strictly decreases as sample size increases for fixed order statistic index, i.e. $\text{MFE}(n, i, K^*) < \text{MFE}(m, i, K^*)\ \forall K^* > 0$ such that $n > m \geq i$. The current inspection is with regard to estimators with different sample sizes and order statistic indices. Although $V'(y; K^*) > 0\ \forall y \in (0, 1)$, which is immediately obvious from Equation 6.9, Figure 6.14 shows that FOSD fails by plotting the cumulative distributions of the random variable U_i^n for $(n, i) = (5, 2)$, $(8, 3)$, and $(11, 4)$ — the mutually beneficial decision rule procedures for many different 'near-normal' SSDs. It is observed that the CDFs intersect one another at the different points (although not apparent from the figure), thus existing points u_0 such that $F_{\beta(n-i+1,i)}(u_0) > F_{\beta(m-j+1,j)}(u_0)$.

We therefore appeal to SOSD which is a stronger argument than its predecessor. Crossings in the CDFs of $F_{\beta(n-i+1,i)}(u)$ and $F_{\beta(m-j+1,j)}(u)$ are allowed by restricting the conditions to $F^{(2)}_{\beta(n-i+1,i)}(u)$ and $F^{(2)}_{\beta(m-j+1,j)}(u)$, so long as the difference in

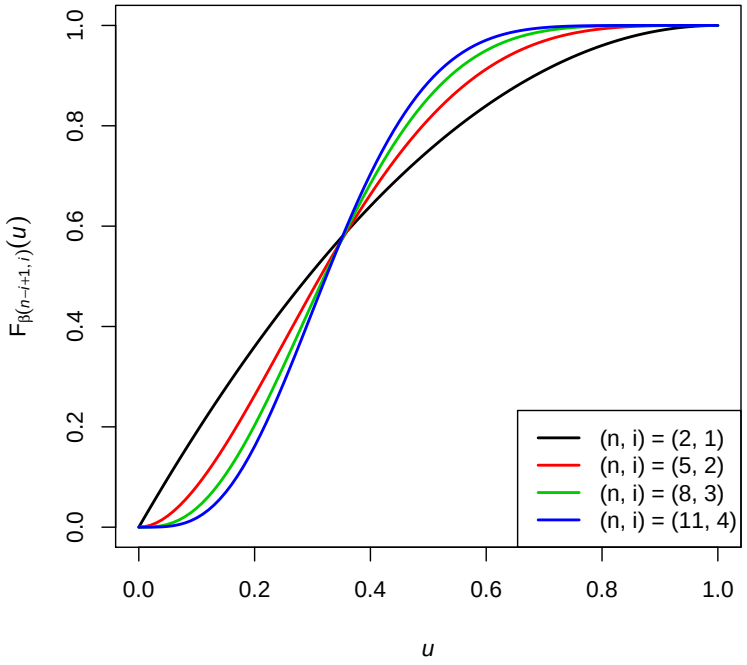


Figure 6.14: CDFs of U_i^n for some different decision rule procedures.

the area between $F_{\beta(n-i+1,i)}(u)$ and $F_{\beta(m-j+1,j)}(u)$ before they intersect is greater than the difference in area after they meet. In the following sections we examine $F^{(2)}_{\beta(n-i+1,i)}(u)$ and $V''(u;K^*) \leq 0 \ \forall y \in (0,1)$ with respect to the aforementioned decision rules.

6.7.4 Analysis of U_i^n

Here we explore whether

$$\left[F^{(2)}_{\beta(n-i+1,i)}(u) - F^{(2)}_{\beta(m-j+1,j)}(u)\right] \leq 0 \tag{6.13}$$

holds for the decision rules described earlier as mutually beneficial. If we assume that $V''(u;K^*) \leq 0 \ \forall u \in (0,1)$, then a sufficient condition for satisfying Equation 6.13 is the decision rule $T(n,i,K)$ being mutually beneficial over $T(m,j,K)$ for a given SSD. The converse argument only allows us to state that if the above condition is satisfied, then the decision rule is desirable. Notwithstanding this, we seek the

smallest $n \geq m$ and largest $i > j$ such that

$$F^{(2)}_{\beta(n-i+1,i)}(u) \leq F^{(2)}_{\beta(m-j+1,j)}(u) \; \forall u \in (0,1), \tag{6.14}$$

starting with $(m, j) = (2, 1)$ and iteratively redefining (m, j) as the current solution (n, i).

In Figure 6.15 (left panel) we plot the second-order cumulative distribution function $F^{(2)}_{\beta(2,1)}(u)$ (black curve) which corresponds to our definition of the current protection threshold, obtained by setting $(m, j) = (2, 1)$. In addition, we also plot $F^{(2)}_{\beta(n-i+1,i)}(u)$ for $i = 2$ and $n = 2, \ldots, 5$. In the interest of clarity, Figure 6.15 (right panel) displays the corresponding curves as per the left panel, relative to $F^{(2)}_{\beta(2,1)}(y)$. It is clear from these figures that the solution to Equation 6.14 is $(n, i) = (5, 2)$. This concurs with empirical analysis in Section 6.3.1.

Setting $(m, j) = (5, 2)$, we can repeat this analysis. Figure 6.16 (left panel) plots $F^{(2)}_{\beta(n-i+1,i)}(u)$ for $i = 3$ and $n = 2, \ldots, 5$; for comparison, curves corresponding to $(n, i) = (2, 1)$ (black curve) and $(5, 2)$ (red curve) are also displayed. In the interest of clarity, the middle and right panels plot the curves relative to $F^{(2)}_{\beta(2,1)}(u)$ and $F^{(2)}_{\beta(4,2)}(u)$ respectively. This figure clearly implies that the solution to Equation 6.14 is $(n, i) = (8, 3)$, again concurring with earlier numerical analysis.

Finally, iterating this argument once further yields the next solution of Equation 6.14 to be $(n, i) = (11, 4)$. In Figure 6.17 we plot the curves of $F^{(2)}_{\beta(n-i+1,i)}(u)$ for all identified decision rules which satisfy Equation 6.14 (including $(n, i) = (2, 1)$) and thus ensure SOSD conditional on $V''(u; K^*) \leq 0$, which happen to also coincide with the numerically determined mutually beneficial decision rules in Section 6.3.1. All second-order CDFs uniformly dominate one another on the interval $(0, 1)$, and are only equal at the limits of the domain.

We know of no general analytical proof for the general dominance property of second-order Beta CDFs for the risk assessment procedures listed. However a case-by-case proof basis is relatively straightforward which is acceptable since the sample size range of interest is likely to be for $n \lesssim 10$. As an example, we demonstrate the result for the first case, i.e. showing that $F^{(2)}_{\beta(n-i+1,i)}(u) \leq F^{(2)}_{\beta(m-j+1,j)}(u)$ for

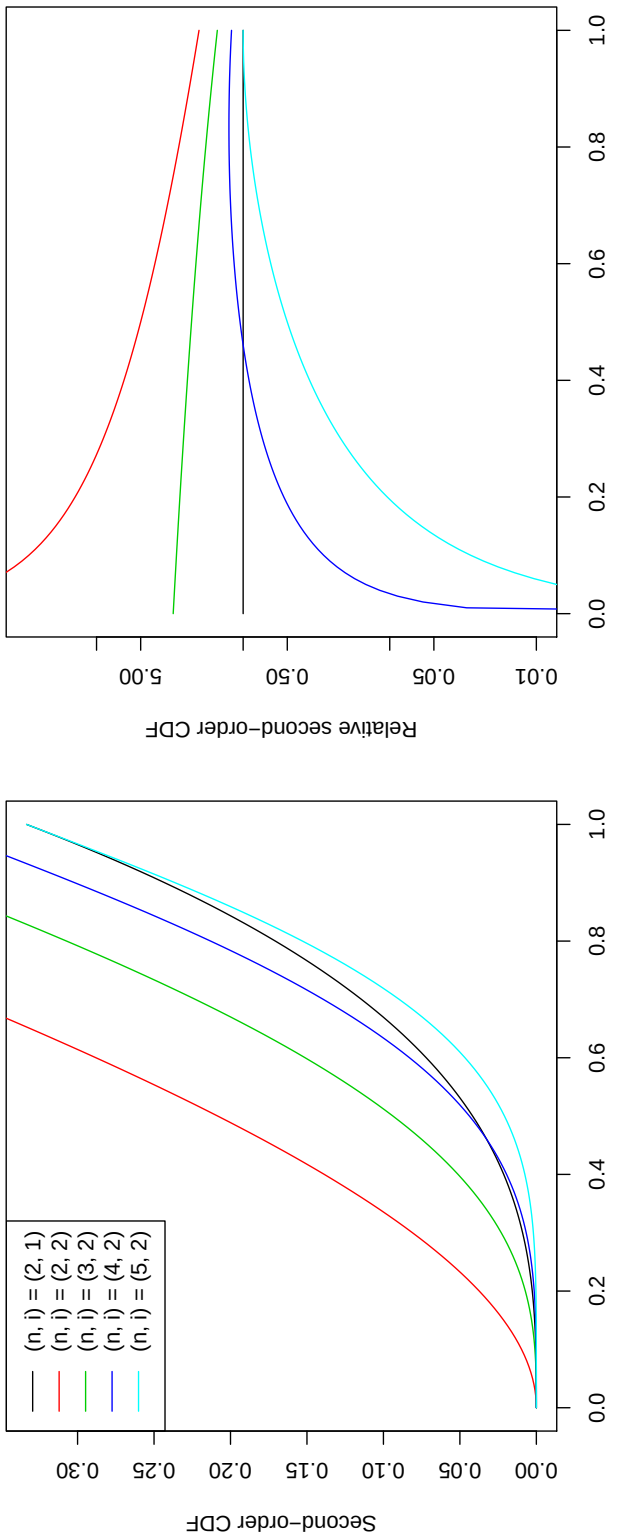


Figure 6.15: Left panel: Second-order CDFs $F^{(2)}_{\beta(n-1,2)}(u)$; right panel: relative to $F^{(2)}_{\beta(2,1)}(u)$. Legend applies to both panels.

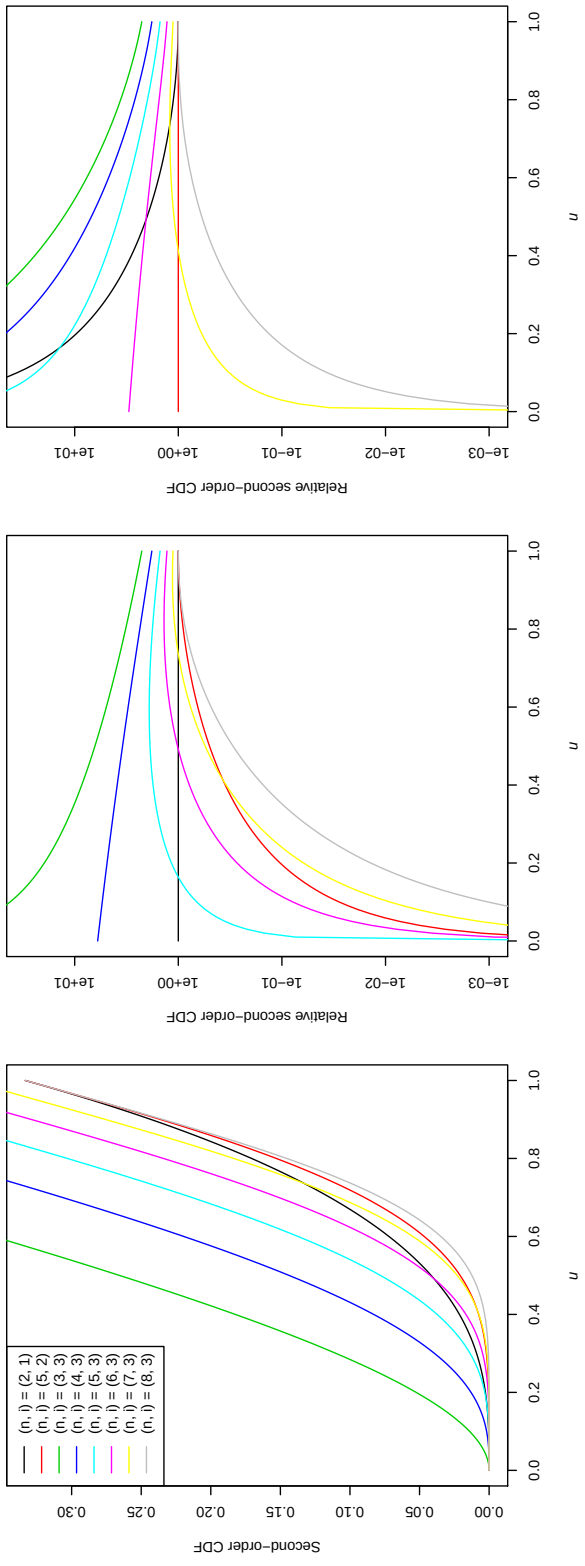


Figure 6.16: Left panel: Second-order CDFs of $F^{(2)}_{\beta(n-2,3)}(u)$; middle panel: relative to $F^{(2)}_{\beta(2,1)}(u)$; right panel: relative to $F^{(2)}_{\beta(4,2)}(u)$. Legend applies to all panels.

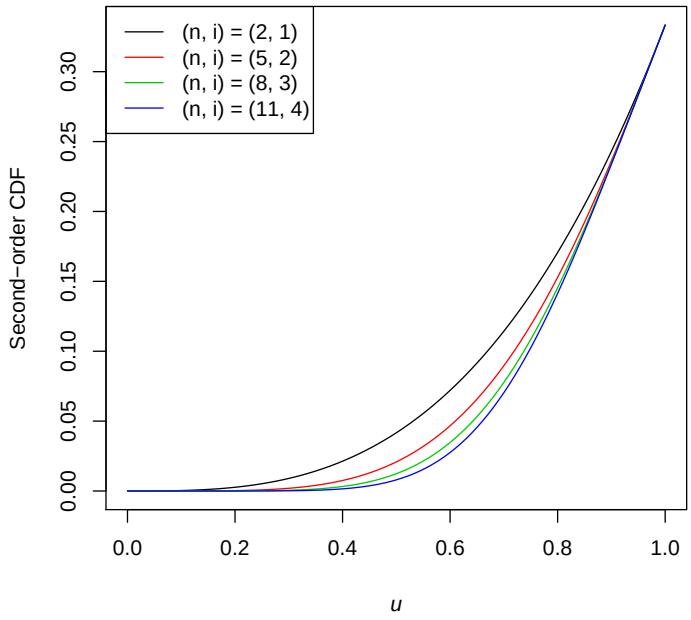


Figure 6.17: Second-order CDFs $F^{(2)}_{\beta(n-i+1,i)}(u)$ for $(n,i) = (2,1)$, $(5,2)$, $(8,3)$ and $(11,4)$.

$(n,i) = (5,2)$ and $(m,j) = (2,1)$, but not so for $(n,i) = (4,2)$. This implies that relative to the procedure $(m,j) = (2,1)$ yielding the current protection threshold, $n = 5$ is the minimum sample size that satisfies Equation 6.14 for order statistic index $i = 2$. Showing the former requires computation of

$$\begin{aligned} F^{(2)}_{\beta(4,2)}(u) - F^{(2)}_{\beta(2,1)}(u) &= \int_0^u \int_0^t \left[f_{\beta(4,2)}(x) - f_{\beta(2,1)}(x) \right] dxdt \\ &= \int_0^u \int_0^t \left[20x^3(1-x) - 2x \right] dxdt, \end{aligned} \tag{6.15}$$

where the last line of Equation 6.15 is arrived at by definition of a Beta density function. Performing the double integral on the left-hand side of the above equation yields $-\frac{1}{3}\left(2u^6 - 3u^5 + u^3\right)$. Simple analysis shows that this function takes its maxima as zero at $u = 0$ and $u = 1$ and that it has a minimum value which is less than zero. Hence, Equation 6.15 is less than or equal to zero for all $u \in [0,1]$ with at least one point satisfying a strict inequality.

Using the same arguments for comparing the procedure of $(n,i) = (4,2)$ to the

risk assessment procedure of $(m, j) = (2, 1)$ yielding the current protection threshold, we have $n = 4$ satisfying Equation 6.14 for $i = 2$ if and only if

$$\begin{aligned}\int_0^u \int_0^t \left[12x^2(1-x) - 2x\right] dxdt &= -\left(\frac{3}{5}u^5 - u^4 + \frac{1}{3}u^3\right) \\ &\leq 0.\end{aligned}$$

It is straightforward to determine this does not hold for all $u \in [0, 1]$; for example, at $u = 1$ one obtains $1/60 \not\leq 0$. Hence $(n, i) = (4, 2)$ does not satisfy Equation 6.14 relative to the default assessment procedure $(m, j) = (2, 1)$.

6.7.5 Analysis of $V(u; K^*)$ for Different SSDs

Here we explore whether $V''(u; K^*) \leq 0$ for $u \in (0, 1)$ and some region of interest $K^* \in (0, K^*_{\max}]$ based on some of the SSDs discussed in Section 6.4.1, namely the normal, logistic, Student-t and skew-normal. The first two are central to regular probabilistic ERA; the third allows further examination in light of the findings of Section 6.4.2. The final SSD examined is important because SSDs (over-log concentration) are generally assumed to be symmetric, which may not be the case. From here onwards, unless specified otherwise, we take $K^*_{\max} \to \infty$. It is easy to confirm that $V'(u; K^*) > 0$ for $u \in (0, 1)$ directly from Equation 6.9 since $f_Z(z) > 0$. In Appendix E we determine $V'(u; K^*)$ and $V''(u; K^*)$ for the aforementioned SSDs. Additionally, in Figures E.1a (normal); E.1b (logistic); E.2a (SN; $\alpha = -3$); E.2b (SN; $\alpha = +3$); and E.3a (Student-t; $\nu = 3$), we plot $V(u; K^*)$ and its first and second derivatives, exemplified with $K^* = 1$ (all figures are located in Appendix E). In the following sections we briefly describe the analysis of $V(u; K^*)$ for each of the SSDs.

Normal Distribution

It is straightforward to confirm that $V''(u; K^*) \leq 0 \ \forall u \in (0, 1)$ and $\forall K^* \geq 0$. In light of $V'(u; K^*) > 0$, one can analytically confirm the decision rules $(n, i) = (5, 2)$, $(8, 3)$ and $(11, 4)$ stochastically dominate each other, and the assessment procedure of $(n, i) = (2, 1)$ corresponding to the current protection threshold, based on the

normal SSD and MFE risk measure.

Logistic Distribution

By the properties of the hyperbolic tangent function, it is straightforward to deduce that $V''(u;K^*) \leq 0\ \forall u \in (0,1)$ and $\forall K^* \geq 0$. As per the normal SSD, when coupled with the fact that $V'(u;K^*) > 0$, the aforementioned decision rules are confirmed to stochastically dominate each other based on the logistic SSD and MFE risk measure.

Skew-Normal Distribution

The structure of $V''(u;K^*)$ is relatively more complicated to analyse. One can appeal to sophisticated software packages to analyse the functions, for example, we utilised `R` (Version 2.4.1) and Maple (Version 9.5). In Section 6.4 we numerically examined members of SN family with $|\alpha| \leq 5$. For these members, we concluded that $V''(u;K^*) \leq 0$ for all $u \in (0,1)$; hence we maintain our assertion of stochastic dominance for the aforementioned decision rules based on the SN SSD ($|\alpha| \leq 5$) and MFE risk measure.

Student's t-Distribution

It is straightforward to show that $V''(u;K^*) \leq 0$ in the case of a Student-t SSD with ν degrees of freedom if

$$F_Z^{-1}(1-u) \in \left[\tfrac{1}{2}K^* - \sqrt{\tfrac{1}{4}K^{*2}+\nu-2}, \tfrac{1}{2}K^* + \sqrt{\tfrac{1}{4}K^{*2}+\nu-2}\right],$$

and $K^* \neq 0$, which is equivalent to u being an element of the closed interval

$$\left[1 - F_{t_\nu}\left(\tfrac{1}{2}\sqrt{\tfrac{\nu}{\nu-2}}K^* + \sqrt{\tfrac{\nu}{4(\nu-2)}K^{*2}+\nu}\right), 1 - F_{t_\nu}\left(\tfrac{1}{2}\sqrt{\tfrac{\nu}{\nu-2}}K^* - \sqrt{\tfrac{\nu}{4(\nu-2)}K^{*2}+\nu}\right)\right],$$

where $F_{t_\nu}(\cdot)$ is the cumulative distribution function of a Student-t random variable with ν degrees of freedom. It is easy to demonstrate that $\nu \to \infty$ implies $V''(u;K^*) \leq 0$, as expected based on the fact that the Student-t distribution approaches a normal distribution in the limit.

A further result of this is that for finite ν, there exists $u \in (0,1)$ for which $V''(u;K^*) > 0$, thus violating the hypothesis of SOSD. This change in sign is noticeable in Figure E.3a where $V''(u;K^*) > 0$ for $u \notin (0.034, 0.82)$. Whilst very difficult to see from Figure E.3b, the corresponding region for when $K^* = 4$ is $u \notin (0.003, 0.64)$. Although in Figure 6.12 the MFE curves for $\nu = 3$ indicate that the aforementioned decision rules are mutually beneficial, and thus desirable, at $K^* = 1$, SOSD does not hold for this case. This highlights that failure to satisfy SOSD is a not a sufficient condition for failure of stochastic dominance.

In Section 6.4, we used numerical techniques to demonstrate that the generalised decision rule based on the assessment procedures $(n,i) = (5,2)$, $(8,3)$ and $(11,4)$ under the context of a Student-t SSD are not necessarily desirable (or mutually beneficial) for all K^*. Whilst this might damage the credibility of the proposed decision rules from a regulatory perspective, we should note a few counteracting points. Beyond the critical value of K^* about which the decision rules lose relative desirability, the relative margin of difference in MFE between the different assessment procedures is not particularly large. Moreover, very large values of K^*, which propel the summary statistic into the far-lower tail of the SSD, are unlikely to be practical for the dimension of uncertainty they are intended to account for. Thus the effects, exhibited in the tails of this distribution on infinite range of log-toxicological concentrations, may in fact have little practical bearing, potentially still allowing for our choice of revised assessment procedures. Moreover, the critical point of K^* where the breach in mutual benefit occurs was numerically found to be increasing rapidly as ν increases relatively slowly. Despite the t-distribution being of importance in the Bayesian updating of the normal model (with the prior distribution described by Equation 2.8) after averaging over the uncertainty about the normal variance, here the t-distribution is a model for the variability in tolerance. To our knowledge, practitioners of SSD based risk assessment have never proposed such a fat-tailed distribution, so its importance in this context is limited.

6.8 Generalised Decision Rules for Species Non-Exchangeability

The decision rules discussed throughout this chapter were discussed in the context of species exchangeability; a reasonable assumption, perhaps, contingent on the tier of assessment. However, for the sake of completeness, we briefly discuss the generalised decisions under the assumption of a single non-exchangeable species being present in the observed toxicity data, in accordance with the discussion by EFSA (2005).

In Section 5.1 we discussed two models for a single non-exchangeable species: $\mathcal{D}_1$ (Equations 5.7) and $\mathcal{D}_2$ (Equations 5.8). The model used most prominently throughout this research, $\mathcal{D}_2$, was motivated by the requirement of mathematical tractability for repeated use. However, since the purpose of this section is essentially guidance of order statistics conditional upon sample size, one can effectively use the more 'flexible' model proposed by EFSA (2005) since complicated calculations can be made in advance and only need to be made once.

In the species exchangeable context, we denoted $Y_{(i:n)}$ to be the i-th log-tolerance value for a sample of n values; the distribution function $F_{Y_{(i:n)}}(y)$ for $y \in \mathbb{R}$ is given by Equation 6.2. In the single-non-exchangeable species context under the model $\mathcal{D}_2$, Y_j are normal with mean μ and standard deviation σ for $j \in J^*_{\mathcal{S}}$, and $Y^\dagger$ is normal with mean $\mu - k'\sigma$ and standard deviation $\phi'\sigma$. By conditioning the event $\{Y_{(i:n)} \leq y\}$ on $Y^\dagger$, one can define the CDF of $Y^*_{(i:n)}$ (the asterisk indicates the inclusion of the non-exchangeable species) to be

$$F_{Y^*_{(i:n)}}(y) = \mathbb{P}[Y_{(i:n)} \leq y \,|\, Y^\dagger \leq y]\mathbb{P}[Y^\dagger \leq y] + \mathbb{P}[Y_{(i:n)} \leq y \,|\, Y^\dagger > y]\mathbb{P}[Y^\dagger > y],$$

which is equivalent to

$$F_{Y^*_{(i:n)}}(y) = F_{Y_{(i:n-1)}}(y) + \left[F_{Y_{(i-1:n-1)}}(y) - F_{Y_{(i:n-1)}}(y)\right] \Phi\left(\frac{y - \mu + k'\sigma}{\phi'\sigma}\right).$$

It is then straightforward by consideration of location-scale properties to subsequently deduce the distribution of $\log_{10}(\text{ATS}) = Y^*_{(i:n)} - K$, which accounts for species non-exchangeability.

Repeating the analysis of Sections 6.3, in particular the example laid out for the log-normal SSD in Section 6.3.1, one can derive a general expression for the MFE, for each decision rule procedure (n, i), standardised assessment factor choice K^*, and non-exchangeability parameters specific to a certain species to yield

$$\begin{aligned}\mathrm{MFE}(n, i, K^*, k', \phi') = 1 - \int_{-\infty}^{\infty} \Big\{ & F_{\beta(i,n-i)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{n \neq i\}} \\ & + \left[F_{\beta(i-1,n-i+1)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{i \neq 1\}} + \mathbf{1}_{\{i=1\}}\right. \\ & \left. - F_{\beta(i,n-i)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{n \neq i\}}\right] \Phi\left(\frac{z + K^* + k'}{\phi'}\right) \Big\} \phi(z) dz.\end{aligned}$$

The numerical quadrature involved in evaluating this function is highly sensitive to the choice of (n, i) when K^* is large. In Appendix F we re-express the integrand in order to increase the precision in evaluation. Although the evaluation of this function is difficult and requires careful handling, the rules we report from it are reusable, thus satisfying risk managers prerequisite of tractability.

In order to demonstrate this method with the rainbow trout as the special species, we fix the non-exchangeability parameters to be those provided in EFSA (2005) because the research presented in this chapter is frequentist in nature, thus not being sensible to apply the Bayesian $\mathcal{MAP}$ estimates used previously. With $(k', \phi') = (0.45, 0.62)$, one can numerically identify the mutually beneficial assessment procedures (assuming $(n, i) = (2, 1)$ is the procedure which when adjusted for species non-exchangeability provides the current protection threshold) from Figure 6.18 to be $(n, i) = (6, 2)$, $(10, 3)$, and $(14, 4)$. The latter differs slightly from the rule $(n, i) = (13, 4)$ reported in EFSA (2005). However, extrapolation shows the absolute difference in MFE between the procedures is less than 2% for all K^*, which may be insignificant to risk managers, especially when compared to the difference in MFE for the minimum threshold: 1% for $(n, i) = (14, 4)$ and 0.6% for $(n, i) = (13, 4)$.

Issues regarding robustness to other types of SSD are beyond the scope of this brief discussion, and would require more complicated estimation of non-exchangeability parameters. This is in addition to any further consideration that may be required regarding the numerical integration. A caveat for the mutually beneficial decision

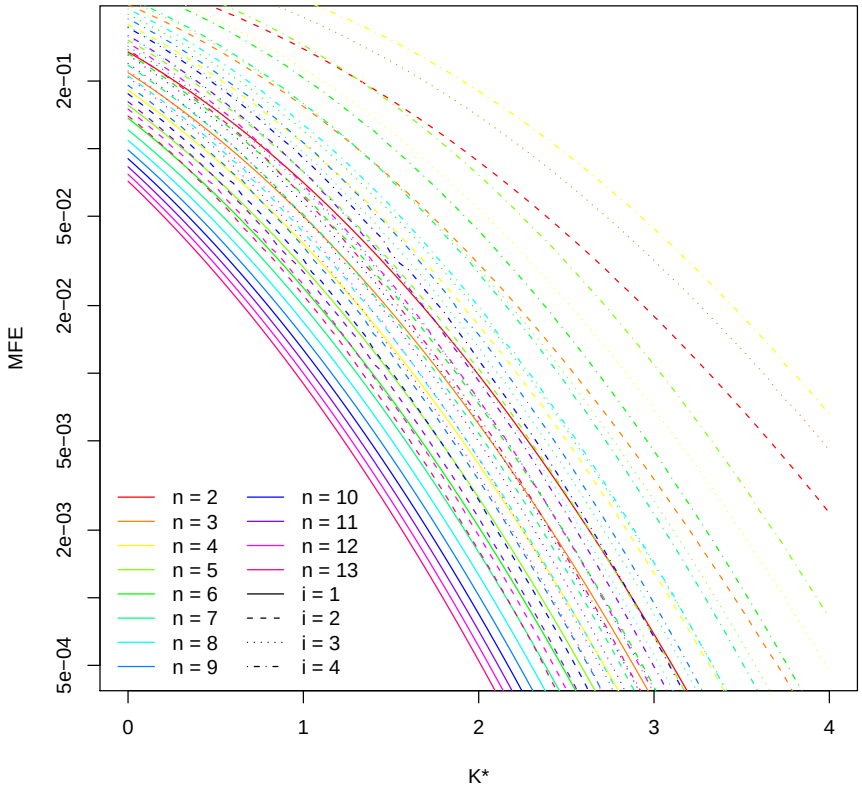


Figure 6.18: MFE$(n, i, K^*, 0.45, 0.62)$ versus K^*. Each line colour and type corresponds to a distinct assessment procedure (n, i) for $n \leq 13$ and $i \leq \min(n, 4)$.

rules listed above would be that they only strictly apply to the rainbow trout; perturbations in (k', ϕ') leads to alternative decision rules. These decision rules listed might be objected to by the chemical industry as they appear to suggest that the resultant PNECs will be more conservative (e.g. when $n = 5$, $i = 2$ for species exchangeability, whereas $i = 1$ when accounting for species non-exchangeability of the rainbow trout). However, this is just an artefact of the MFE curve corresponding to the minimum level of protection also being adjusted (by up to 6% absolute difference) to reflect the inherent sensitivity of the rainbow trout.

6.9 Conclusions

Where a qualitative assessment of the risk to an ecological community is required, the simplest regulatory accepted method for characterising the hazard component is generally based on applying an assessment factor to a summary statistic — the lowest observed value — of laboratory derived toxicity data. ECHA (2008a, p. 18) states that:

> 'The assessment factors presented should be considered as general factors that under certain circumstances may be changed.'

Justification for changing the assessment is required and some possible reasons are provided in ECHA (2008a), for example: more test species; knowledge of similar substances; knowledge of the toxic mode of action. Some of these justifications have been discussed in a probabilistic setting within this thesis. There may in addition be reason for changing the choice of assessment factor based on the conclusions of research in this chapter. However, we do not recommend this. This is because (i) the assessment factor (AF) is meant to account for additional uncertainties which are not captured by the SSD, and (ii) the current level of protection which is obtained by application of current deterministic decision rules is nowhere mentioned (EFSA, 2005).

The evaluation of the generalised decision rule from a probabilistic viewpoint, and with respect to the MFE as a summary of protection, naturally led to a dominance criterion being established. This criterion was used to imply a set of mutually beneficial decision rules, i.e. industry being incentivised by regulators to assess more species. The probabilistic modelling only accounted for uncertainty in interspecies variability. Other uncertainties, for example, acute-to-chronic extrapolation and the difference in taxonomic/trophic groups assessed, would need to be considered further, but would likely be 'accounted for' by assessment factor prescriptions in the current technical guidance documentation. However, some of the limitations here would also extend to the standard probabilistic SSD approaches.

The set of mutually beneficial estimators proposed in this chapter move beyond the current regulatory *status quo* of setting the toxicity summary statistic to be

lowest measured tolerance value, to other procedures such as: the second ordered tolerance value from a sample of five, etc. Furthermore, the reported decision rules are interpretable for either acute or chronic assessment endpoints, which enhance the appeal of the revised assessment methodology from a risk management perspective. The same assessment factors, as currently prescribed for the risk assessments, are still applied with the revised summary statistic. In the vast majority of SSDs marginally deviant from normality (over log-concentration), the rules were shown to be robust. The exceptions included the negative-skewed bimodal normal mixture SSD and the Student-t SSD; discussion was provided on each of these cases. However, a potentially important observation drawn in this report is that unimodality is not a restrictive enough condition to offer mutual benefit, although there was no reason to assume as much, *a priori*.

Alterations to the decision procedures were made with respect to suitably large (and in some cases infinite) domains of standardised assessment shift. In order to gain an appreciation of the differences in the current level of protection offered for different taxonomic groups, the standardised assessment shift scale — under the log-normal and log-logistic SSD viewpoint — was transformed under the context of $\mathcal{M}_4$, namely where σ is fixed as the pooled standard deviation estimates provided in EFSA (2005) (cf. Luttik and Aldenberg 1997). Any conclusions drawn from this brief empirical reassessment are subject to the naïve assumptions of the aforementioned behavioural model.

Chapter 7

Conclusions & Future Directions

This chapter provides a short summary of the main results from this thesis, followed by a discussion of future research potential and needs.

7.1 Conclusions

In this thesis we have extended the current state of the science for ecotoxicological risk assessment. This has been regarding the standalone issue of estimating the concentration of a toxicant — denoted as the PNEC — below which is unlikely to cause adverse effects to ecological assemblages. The key chapters of original research in this thesis, namely Chapters 3–6, describe hazard assessment decision rules which are intended to yield estimates of a proxy for the PNEC. As is the case for current guidance documentation, the proposals made in this thesis address distinct criticisms of the current state of assessment rather than a consolidation of them. However, there is a wide degree over overlap.

The initial background review discovered that there were a number of different models and estimators for the PNEC. The degree of uncertainty refinement also separated the different estimators. Of the quantitative approaches, we identified two general procedures for determining a PNEC: deterministic and probabilistic. The former is contingent on the principle of the precautionary principle, whereas the latter attempts to model the sensitivity of the biological assemblage and quantify aspects of the uncertainty.

Probabilistically derived estimators are generally founded on the basis of a model, called the species sensitivity distribution. This represents the interspecies variability of tolerance for a given community. The widely accepted principle for ensuring the long-term stability of an ecosystem is to protect at least 95% of species. The threshold concentration at which this occurs is denoted as the HC_5 — effectively the 5-th percentile of the SSD when estimated using chronic toxicity data. The lack of resource-intensive long-term toxicity data has meant that there has been considerable research in the ecotoxicological risk assessment literature that focuses on estimating the 'acute HC_p'. With the lack of toxicity data in general to suggest otherwise, the SSD is often accepted as being describable by a log-normal distribution.

From this definition of a protection goal, we sought to examine estimation of this percentile under uncertainty. Current literature makes ambiguous recommendations such as choosing the upper 95% one-sided underestimate confidence limit of the HC_p as to err on the side of caution. The current technical guidance documentation requests that the median HC_5 should be determined along with a 50% confidence interval. However there is no indication of how the confidence level will be interpreted.

We analysed the HC_p estimation problem from a decision theoretic perspective. This allowed for the inclusion of loss — a valuable concept to account for overestimation being more serious than underestimation (from a protectionist viewpoint). A range of scientifically defensible estimators was discussed and compared, each deriving from different principles. In particular, a range of behavioural models was considered. This included the default model accepted by regulators, which has been considered in recent literature and official documentation. The assumptions of the behavioural models would require further investigation. Based on considerations of the generally accepted need for conservatism, a new estimator was proposed and details given on how it might be implemented.

Aside from the theoretical considerations given to the estimation of the HC_p, a major contribution of this thesis was connected to the physicalities of the data generating mechanism. Species non-exchangeability — a term coined by EFSA (2005) —

describes the situation in which a single species' tolerance value (extendable, if necessary, to multiple species) is not simply regarded as identically distributed upon its identification. The motivating example of this is the rainbow trout which is regarded by ecotoxicologists as a 'typically sensitive species' relative to other fish species. The biological mechanics of why the rainbow trout may be more susceptible to adverse effects induced by toxicant exposure are unknown. However the implications are highly important since it may lead to overachieved protection goals, or rather over-conservatism, especially in light of the rainbow trout being a typical dossier species for chemical safety assessment. An exploratory analysis demonstrated that there is evidence to reject the null-hypothesis of exchangeability among species in a large database of fish toxicity data for pesticide exposure. In particular, the rainbow trout was highlighted as exhibiting the most prevalent and systematic bias.

Decision rules which accounted for species non-exchangeability were presented by modifying the assumptions placed on the data generating mechanism. In order to keep within the remit of the research aims of this report, a tractable modification was sought. A simpler model to that proposed in EFSA (2005) was suggested which led to HC_p estimators, and corresponding uncertainty measures, being fully tractable extensions of those currently used by risk assessors. A Bayesian model selection criterion was used to ascertain the adequacy of the parsimonious model. The introduction of additional parameters into the estimators required that we explore the assumption of discounting their associated uncertainty to ensure scientific defensibility. A performance analysis was developed and applied to a large multi-substance database. Whilst it is a regulatory decision as to whether the inferences drawn from the analysis would be satisfactory for application, we considered them to be within sufficient tolerance boundaries.

Exploratory analysis in fact indicated that there may be more than one non-exchangeable species present. The development of the estimators is fully extendable to incorporate multiple non-exchangeable species. However this is not advised because the SSD concept would begin to break down; the simple predictability of the construct would no longer function as required. It is plausible that the exchangeability assumption is untenable from a statistical modelling perspective in light of

this. The natural procedure would be to fit a model where each species and chemical has an effect. However such modelling techniques would likely over-extend the statistical skills available within the risk assessment circle. The remit of this thesis is to develop decision rules which are tractable for lower and intermediate tiers of quantitative risk assessment.

Although the vast majority of the research in this thesis deals with improvements of the probabilistic risk assessment field, deterministic decision rules are also regularly appealed to. The assessment factors underpinning these estimators are the least defensible due to the ambiguous and non-explained magnitudes. Current guidance documentation describes the factors as 'guidance values' and open to reduction provided sufficient reasoning is provided. Limited in necessary experience in the regulatory arena, we stopped short of recommending changes to absolute assessment factors, with only minor discussion in Chapters 3 and 6. We did explore the toxicity summary statistic upon which the assessment factor is applied — currently defined to be the lowest tolerance value.

Since the current guidance is described for sample sizes ranging from less than three, we explored a generalised estimator for a wider range of sample sizes which would typically be exempt from fully probabilistic analysis (cf. ECHA 2008a). Recommendations regarding the order statistic with which to use (from the minimum up to the fourth ordered tolerance value) were based upon a probabilistic evaluation of the deterministic estimator. This is so that the mean fraction of species whose endpoints are violated by each estimator is not exceeded by a preceding recommended estimator. Furthermore, the estimators were denoted as *mutually beneficial* between the chemical manufacturer and the risk manager. Similar findings were explored in EFSA (2005). However, we presented analytical verification, as well as a demonstration of the robustness of the generalised estimators to distributional assumptions.

Where appropriate, modelling was performed within the Bayesian paradigm. Unlike many other areas of risk assessment, Bayesian analysis is relatively new in ecotoxicological risk assessment. An often contentious element of Bayesian analysis is the prior distribution with two schools of thought existing: the subjectivists and

the objectivists. The former advocate the use of data to update expert judgements which have been elicited, whereas the latter appeal to non-informative priors in order to maximise the influence of the data. For purposes of being able to tractably compare our results with other established or assessment body literature, we appealed to 'standard' non-informative prior distributions. Expert judgements could be incorporated into the research presented here, straightforwardly in many cases, and may be a key to overcoming data shortages. Regulators are likely to be sceptical about the inclusion of expert judgements as they will bear the responsibility of having to scrutinise the judgements to ensure that they are unbiased.

Finally, we would note that the research elements in this thesis address weaknesses not only raised in current literature, but also in current guidance documentation. Uptake of this research would not be fast, but may facilitate additional research in the meantime. Therefore we list some additional research needs in the next section.

7.2 Future Research

The purpose of this thesis has been to develop the current technical guidance in ecotoxicological risk assessment, namely the hazard assessment. Yet our accomplishment has been restricted to only a fraction of the research needs required to make current ecotoxicological risk assessment fit for purpose. In this section we discuss some of the primary deficiencies which require further investigation.

The arrival of EC (2006) into the commercial arena of the EU means that there is mounting pressure on chemical manufacturers to assess the risk of substances whilst testing fewer species in order to align risk assessment with modern societal and ethical considerations. The current number of species required to be tested is already significantly low. Consequently more attention has been focused on data augmentation using predictive tools based on historical data. Tools such as the US EPA's ICE program (Dyer et al., 2006, 2008) were mentioned in Section 2. The validity of such methods still has to be more firmly established before acceptance will be granted. However, there is potential to better construct the predictive models

employed which underlie these tools. Moreover, the behavioural models discussed in this thesis may offer an alternative and perhaps simpler path of future risk assessment. As a consequence, demonstrating the need for immediate research into the area of *borrowing strength* from historical assessment data. It may be appropriate to explore this from the perspective of meta-analysis. Roex et al. (2000) performed a meta-analysis in trying to explore acute-to-chronic extrapolation between species across different studies. Meta-analysis could also strengthen our understanding of retrospective risk assessment where per-species tolerance values derive from different studies [scientific experiments] or for exploring measurement error further (see below) where multiple records have been recorded for the same species-chemical pair. However, the issue of publication bias, especially where reports are made to regulators only, may undermine the scope of this analysis.

In addition, a further need is to construct better hierarchical modelling for intermediate quantitative tier risk assessments. For example, databases of historical toxicity data based on many different modes of action would most likely be better modelled using a hierarchical modelling structure. Current guidance specifies that SSDs must be representative of multiple taxa, which might be inappropriate due to different species richness coefficients and per-taxa relative sensitivities. Initial research into this, for example Grist et al. (2006) and Hickey et al. (2008), has explored hierarchical modelling by taxonomic families and orders respectively.

The SSD literature predominantly focuses on interspecies extrapolation whilst accounting for sampling uncertainty. Yet there are additional uncertainties which need to be accounted for; the reason for ECHA (2008a) imposing an assessment factor to the estimated HC_5. A fundamental requirement is the need to explore measurement error; currently no distinction is made between actual and observed species tolerance. The harmonisation techniques currently practiced by risk assessors are due to varying standards of historical data deriving from scientific experiments spanning multiple decades. If data was available, then one could revise the decision rules to incorporate measurement error. This would likely lead to smaller SSD variance estimators (since noise stemming from the measurement error is removed). Hyper-parameter estimates would also change; the heterogeneity parameters would

have more weight in the estimation of SSD variance under $\mathcal{M}_2$. If one also considers non-exchangeability, then we might expect more weight to be given to the historical data (if we discount the non-exchangeability of tolerance being a complex artefact of measurement error). Although, k would not be expected to change significantly in the case of $\mathcal{D}_2$ since it acts as a shift; its role under $\mathcal{D}_1$ would be more complicated due to it is a shift-factor of the SSD standard deviation. In general, the consequences of accounting for measurement error in the decision rules we discuss are unclear as it affects so many components. Although current guidance such as ECHA (2008a) insists upon chronic data populated SSDs, there is an obvious advantage to understanding whether acute risk assessment decision rules can be adequately mapped to chronic versions, and if so, quantifying the attached uncertainty.

Greater attention is currently being given to the assessment of multiple substance risk. Laboratory conditions usually only test one substance at a time, thus discounting combined effects. As chemicals can combine independently, additively, synergistically or antagonistically and have correlated effects, the mathematics and statistics required in making such assessments is much more complex (Plackett and Hewlett, 1952), let alone the uncertainty quantification. The current approach proposed in Traas et al. (2002) makes a number of crude assumptions, thus motivating the need for further research.

A key component of any future research will be the adequate balance of pragmatism for purposes of scientific defensibility and parsimony. This would be for purposes of making non-higher tier assessment transparent to assessors and stakeholders, as well as being economical. Such research will require communication between risk managers and scientists, including those with statistical expertise. Defining protection goals will be one of the future challenges. However it will potentially allow for regulators to efficiently choose assessment factors without unnecessarily triggering higher tier assessments or leading to over-protective risk management goals.

Bibliography

Abramowitz, M. and Stegun, I. A. Eds. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* **9th ed.** New York: Dover.

Aitkin, M. (1999). Discussion Paper. In: Lindsey, J. K. (1999). Some Statistical Heresies. *The Statistician*, **48**: 1–40.

Aldenberg, T. (2005). EFSA Sigma Homogeneous? Unpublished Report Submitted to the EFSA Plant Protection Products and Their Residues (PPR) Panel. RIVM, Bilthoven, The Netherlands.

Aldenberg, T. (2007). Busy 1.1a Reference Manual and Guide. Available at: `www.busymath.googlepages.com/documents` [Accessed: 9th September 2009].

Aldenberg, T. and Jaworska, J. S. (2000). Uncertainty of the Hazardous Concentration and Fraction Affected for Normal Species Sensitivity Distributions. *Ecotoxicology and Environmental Safety*, **46**: 1–18.

Aldenberg, T., Jaworska, J. S. and Traas, T. P. (2002). Normal Species Sensitivity Distributions and Probabilistic Ecological Risk Assessment. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 49–102.

Aldenberg, T., Jaworska, J. S. and Traas, T. P. (2009). Beyond the Risk Characterisation Ratio (RCR) in Chemical Safety Assessment. *In preparation.*

Aldenberg, T. and Luttik, R. (2002). Extrapolation Factors for Tiny Toxicity Data Sets From Species Sensitivity Distributions With Known Standard Deviation.

In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 103–118.

Aldenberg, T. and Slob, W. (1993). Confidence Limits for Hazardous Concentrations Based on Logistically Distributed NOEC Toxicity Data. *Ecotoxicology and Environmental Safety*, **25**: 48–63.

Alexander, D. E. and Fairbridge, R. W. (1999). *Encyclopaedia of Environmental Science.* Amsterdam: Kluwer Academic Publishers.

ANZECC and ARMCANZ (Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand) (2000). Australian and New Zealand Guidelines for Fresh and Marine Water Quality. National Water Quality Management Strategy Paper No. 4. ANZECC and ARMCANZ, Canberra.

Azzalini, A. (1985). A Class of Distribution Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**: 171–178.

Azzalini, A. (2008). `R` Package `sn`: The Skew-Normal and Skew-t Distributions (Version 0.4–14).
Available at: `http://azzalini.stat.unipd.it/SN` [Accessed: 8th June 2010].

Baird, D. J. and Van den Brink, P. J. (2007). Using Biological Traits to Predict Species Sensitivity to Toxic Substances. *Ecotoxicology and Environmental Safety*, **67**: 296–301.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* **2nd ed.** New York: Springer-Verlag.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods and Research*, **33**: 261–304.

Campbell, P. J., Arnold, D. J. S., Brock, T. C. M., Grandy, N. J., Heger, W., Heimbach, F., Maund, S. J. and Streloke, M. Eds. *Guidance Document on Higher Tier Aquatic Risk Assessment for Pesticides (HARAP).* Brussels: Society of Environmental Toxicology and Chemistry.

Chapman, P. M., Cardwell, R. S. and Chapman, P. F. (1996). A Warning: NOECs Are Inappropriate For Regulatory Use. *Environmental Toxicology and Chemistry*, **15**: 77–79.

Chen, L. (2003). A Conservative, Non-parametric Estimator for the 5th Percentile of the Species Sensitivity Distribution. *Journal of Statistical Planning and Inference*, **123**: 243–258.

Craig, P. S. (2005). Statistical Aspects of Safety/Extrapolation Factors. Unpublished Report Submitted to the EFSA Plant Protection Products and Their Residues (PPR) Panel. The University of Durham, Durham, U.K.

Craig, P. S. (2006). Uncertainty Factors in Ecotoxicological Risk Management. Proceedings of the 4th Edinburgh Risk Conference; March 29–31, 2006. Edinburgh, UK. Available at: `http://eprints.dur.ac.uk/archive/00000254/01/41133.pdf` [Accessed: 1st March 2008].

Cronin, M. T. D., Walker, J. D., Jaworska, J. S., Comber, M. H. I., Watts, C. D. and Worth, A. P. (2003). Use of QSARs in International Decision-Making Frameworks to Predict Ecologic Effects and Environmental Fate of Chemical Substances. *Environmental Health Perspectives*, **111**: 1376–1390.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

De Zwart, D. (2002). Observed Regularities in SSDs for Aquatic Species. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 133–154.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis.* **3rd ed.** New York: Wiley.

Duboudin, C., Ciffroy, P. and Magaud, H. (2004a). Effects of Data Manipulation and Statistical Methods On Species Sensitivity Distributions. *Environmental Toxicology and Chemistry*, **23**: 489–499.

Duboudin, C., Ciffroy, P. and Magaud, H. (2004b). Acute-to-Chronic Sensitivity Distribution Extrapolation. *Environmental Toxicology and Chemistry*, **23**: 1774–1785.

Dwyer, F. J., Mayer, F. L., Sappington, L. C., Buckler, D. R., Bridges, C. M., Greer, I. E., Hardesty, D. K., Henke, C. E., Ingersoll, C. G., Kunz, J. L., Whites, D. W., Augspurger, T., Mount, D. R., Hattala, K. and Neuderfer, G. N. (2005). Assessing Contaminant Sensitivity of Endangered and Threatened Aquatic Species: I. Acute Toxicity of Five Chemicals. *Archives of Environmental Contamination and Toxicology*, **48**: 143–154.

Dyer, S. D., Versteeg, D. J., Belanger, S. E., Chaney, J. G. and Mayer, F. L. (2006). Interspecies Correlation Estimates Predict Protective Environmental Concentrations. *Environmental Science and Technology*, **40**: 3102–3111.

Dyer, S. D., Versteeg, D. J., Belanger, S. E., Chaney, J. G., Raimondo, S. and Barron, M. G. (2008). Comparison of Species Sensitivity Distributions Derived from Interspecies Correlation Models to Distributions used to Derive Water Quality Criteria. *Environmental Science and Technology*, **42**: 3076–3083.

EC (European Commission) (1991). Council Directive 91/414/EEC of the Council of the European Communities of 15 July 1991 Concerning the Placing of Plant Protection Products on the Market. *Official Journal of the European Union*, L 230 19.8.1991, p. 1.

EC (European Commission) (2002). Guidance Document on Aquatic Ecotoxicology in the Context of the Directive 91/414/EEC. SANCO/3268/2001, **4**: 62.

EC (European Commission) (2003). Technical Guidance Document on Risk Assessment in Support of Commission Directive 93/67/EEC on Risk Assessment for New Notified Substances Commission Regulation (EC) No. 1488/94 on Risk Assessment for Existing Substances Directive 98/8/EC of the European Parliament and of the Council Concerning the Placing of Biocidal Products on the Market, Part II. Luxembourg: Office for Official Publications of the European Communities. L-2985, Chapter 3 *Environmental Risk Assessment*.

EC (European Commission) (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union*, L 396 30.12.2006, p. 1.

ECHA (European Chemicals Agency) (2008a). Guidance for the Implementation of REACH: Guidance on Information Requirements and Chemical Safety Assessment Chapter R.10: Characterisation of Dose-Response for Environment. July 2008.

ECHA (European Chemicals Agency) (2008b). Guidance for the Implementation of REACH: Guidance on Information Requirements and Chemical Safety Assessment Chapter R.19: Uncertainty Analysis. July 2008.

EFSA (European Food Safety Authority) (2005). Opinion of the Scientific Panel on Plant Health, Plant Protection Products and their Residues on a Request from EFSA Related to the Assessment of the Acute and Chronic Risk to Aquatic Organisms with Regard to the Possibility of Lowering the Uncertainty Factor if Additional Species were Tested. *The EFSA Journal*, **301**: 1–45.

EFSA (European Food Safety Authority) (2008). Risk Assessment for Birds and Mammals - Revision of Guidance Document under Council Directive 91/414/EEC (SANCO/4145/2000 - final of 25 September 2002) - Scientific Opinion of the Panel on Plant protection products and their Residues (PPR) on the Science behind the Guidance Document on Risk Assessment for birds and mammals. *The EFSA Journal*, **734**: 1–181, Appendix 7.

Emans, H. J. B., Van de Plassche, E. J., Canton, J. H., Okkerman, P. C. and Sparenburg, P. M. (1993). Validation of Some Extrapolation Methods Used for Effects Assessment. *Environmental Toxicology and Chemistry*, **12**: 2139–2154.

EUFRAM (2006). The EUFRAM Framework. Introducing Probabilistic Methods into the Ecological Risk Assessment of Pesticides. **1**: Version 5. Available at: `http://www.eufram.com/` [Accessed: 9th September 2009].

Farrell, P. J. and Rogers-Stewart, K. (2006). Comprehensive Study of Tests for Normality and Symmetry: Extending the Spiegelhalter Test. *Journal of Statistical Computation and Simulation*, **76**: 803–816.

Forbes, V. E. and Calow, P. (2002a). Extrapolation in Ecological Risk Assessment: Balancing Pragmatism and Precaution in Chemical Controls Legislation. *BioScience*, **52**: 249–257.

Forbes, V. E. and Calow, P. (2002b). Species Sensitivity Distributions Revisited: A Critical Appraisal. *Human and Ecological Risk Assessment*, **8**: 1625–1640.

Forbes, V. E. and Calow, P. (2002c). Sensitivity Distributions - Why Species Selection Matters. *SETAC Globe*, **3**: 22–23.

Forbes, V. E., Calow, P. and Sibly, R. M. (2008). The Extrapolation Problem and How Population Modelling Can Help. *Environmental Toxicology and Chemistry*, **27**: 1987–1994.

French, S. and Ríos Insua, D. (2000). *Statistical Decision Theory*. London: Edward Arnold.

Garnier-Laplace, J., Della-Vedova, C., Gilbin, R., Copplestone, D., Hingston, J. and Ciffroy, P. (2006). First Derivation of Predicted-No-Effect Values for Freshwater and Terrestrial Ecosystems Exposed to Radioactive Substances. *Environmental Science and Technology*, **40**: 6498–6505.

Gosling, J. P. (2009). A Review of the U.S. Environmental Protection Agency's Interspecies Correlation Estimation Method. Unpublished Report. The Food and Environment Safety Authority, York, U.K.

Grist, E. P. M., Leung, K. M. Y., Wheeler, J. R. and Crane, M. (2002). Better Bootstrap Estimation of Hazardous Concentration Thresholds for Aquatic Assemblages. *Environmental Toxicology and Chemistry*, **21**: 1515–1524.

Grist, E. P. M., O'Hagan, A., Crane, M., Sorokin, N., Sims, I. and Whitehouse, P. (2006). Bayesian and Time-Independent Species Sensitivity Distributions for Risk Assessment of Chemicals. *Environmental Science and Technology*, **40**: 395–401.

Hadar, J. and Russell, W. R. (1969). Rules for Ordering Uncertain Prospects. *American Economic Review*, **59**: 25–34.

Hickey, G. L., Kefford, B. J., Dunlop, J. E. and Craig, P. S. (2008). Making Species Salinity Sensitivity Distributions Reflective of Naturally Occurring Communities: Using Rapid Testing and Bayesian Statistics. *Environmental Toxicology and Chemistry*, **27**: 2403–2411.

Hickey, G. L., Craig, P. S. and Hart, A. (2009). On The Application of Loss Functions in Determining Assessment Factors for Ecological Risk. *Ecotoxicology and Environmental Safety*, **72**: 293–300.

Hose, G. C. and Van den Brink, P. J. (2004). Confirming the Species-Sensitivity Distribution Concept for Endosulfan Using Laboratory, Mesocosm, and Field Data. *Archives of Environmental Contamination and Toxicology*, **47**: 511–520.

IPCS (International Programme on Chemical Safety) (2004). IPCS Risk Assessment Terminology. Geneva: World Health Organization. Available at:

`http://www.inchem.org/documents/harmproj/harmproj/harmproj1.pdf` [Accessed: 7th December 2009]

Jager, T., Posthuma, L. De Zwart, D., Van de Meent, D. (2007). Novel View On Predicting Acute Toxicity: Decomposing Toxicity Data in Species Vulnerability and Chemical Potency. *Ecotoxicology and Environmental Safety*, **67**: 311–322.

Jagoe, R. H. and Newman, M. C. (1997). Bootstrap Estimation of Community NOEC Values. *Ecotoxicology*, **6**: 293–306.

Johnson, N. L. (1949). Systems of Frequency Curves Generated By Methods of Translation. *Biometrika*, **36**: 149–176.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**: 773–795.

Kass, R. E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**: 1343–1370.

Kefford, B. J., Palmer, C. G., Jooste, S., Warne, M. and Nugegoda, D. (2005). 'What is it Meant by 95% of Species'? An Argument for the Inclusion of Rapid Tolerance Testing. *Human and Ecological Risk Assessment*, **11**: 1025–1046.

Kooijman, S. A. L. M. (1987). A Safety Factor for LC50 Values Allowing for Differences in Sensitivity Among Species. *Water Research*, **21**: 269–276.

Levy, H. (2006). *Stochastic Dominance: Investment Decision Making Under Uncertainty.* **2nd ed.** New York: Springer.

Luttik, R. and Aldenberg, T. (1997). Assessment Factors for Small Samples of Pesticide Toxicity Data: Special Focus on LD50 Values for Birds and Mammals. *Environmental Toxicology and Chemistry*, **16**: 1785–1788.

Maltby, L., Blake, N., Brock, T. C. M. and Van den Brink, P. J. (2005). Insecticide Species Sensitivity Distributions: The Importance of Test Species Selection and Relevance to Aquatic Ecosystems. *Environmental Toxicology and Chemistry*, **24**: 379–388.

Mineo, A. M. and Ruggieri, M. (2005). A Software Tool for the Exponential Power Distribution: The `normalp` Package. *Journal of Statistical Software*, **12**: 1–24.

Nadarajah, S. (2005). A Generalized Normal Distribution. *Journal of Applied Statistics*, **32**: 685–694.

Newman, M. C., Ownby, D. R., Mezin, L. C. A., Powell, D. C., Christensen, T. R. L., Lerberg, S. B. and Anderson, B-A. (2000). Applying Species Sensitivity Distributions in Ecological Risk Assessment: Assumptions of Distribution Type and Sufficient Numbers of Species. *Environmental Toxicology and Chemistry*, **19**: 508–515.

Newman, M. C., Ownby, D. R., Mezin, L. C. A., Powell, D. C., Christensen, T. R. L., Lerberg, S. B., Anderson, B-A. and Padma, T. M. (2002). Species Sensitivity Distributions in Ecological Risk Assessment: Distribution Assumptions, Alternate Bootstrap Techniques, and Estimation of Adequate Number of Species. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 119–132.

Newman, M. C. (2008). "What Exactly Are Your Inferring?" A Closer Look At Hypothesis Testing. *Environmental Toxicology and Chemistry*, **27**: 1013–1019.

O'Hagan, A., Crane, M., Grist, E. P. M. and Whitehouse, P. (2005). Estimating Species Sensitivity Distributions With the Aid of Expert Judgements (Unpublished). Available at:
`http://www.tonyohagan.co.uk/academic/pdf/SSD-stat.pdf` [Accessed: 1st July 2007].

Pires, A. M., Branco, J. A., Picado, A. and Mendonça, E. (2002). Models for the Estimation of a 'No Effect Concentration'. *Environmetrics*, **13**: 15–27.

Plackett, R. L. and Hewlett, P. S. (1952). Quantal Responses to Mixtures of Poisons. *Journal of the Royal Statistical Society.* Series B, **14**: 141–163.

Posthuma, L., II Suter, G. W., Traas, T. P. Eds. (2002a). *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers.

Posthuma, L., Traas, T. P., II Suter, G. W. (2002b). General Introduction to Species Sensitivity Distributions. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 119–132.

R Development Core Team (2006). `R`: A Language and Environment for Statistical Computing. `R` Foundation for Statistical Computing, Vienna, Austria. Available at: `http://www.R-project.org` [Accessed: 1st May 2007].

Raimondo, S., Montague, B. J. and Barron, M. G. (2007). Determinants of Variability in Acute to Chronic Toxicity Ratios in Aquatic Invertebrates and Fish. *Environmental Toxicology and Chemistry*, **26**: 2019–2023.

Raimondo, S., Vivian, D. N., Delos, C. and Barron, M. G. (2008). Protectiveness of Species Sensitivity Distribution Hazard Concentrations for Acute Toxicity Used in Endangered Species Risk Assessment. *Environmental Toxicology and Chemistry*, **27**: 2599–2607.

Rand, G. M. Ed. (1995). *Fundamentals of Aquatic Toxicology: Effects, Environmental Fate and Risk Assessment.* **2nd ed.** Philadelphia: Taylor and Francis.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis.* **2nd ed.** Belmont (CA): Duxbury Press.

RIVM (National Institute of Public Health and the Environment). (1989). Environmental Risk Limits. RIVM Report: 601640001. Bilthoven, The Netherlands.

Roex, E. W. M., Van Gestel, C. A. M., Van Wezel, A. P. and Van Straalen, N. M. (2000). Ratios Between Acute Aquatic Toxicity and Effects on Population Growth Rates in Relation to Toxicant Mode of Action. *Environmental Toxicology and Chemistry*, **19**: 685–693.

Rubach, M. N., Baird, D. J. and Van den Brink, P. J. (In press). A New Method for Ranking Mode Specific Sensitivity of Freshwater Arthropods to Insecticides and its Relationship to Biological Traits. *Environmental Toxicology and Chemistry.* DOI: 10.1002/etc.55.

Schervish, M. J. (1995). *Theory of Statistics.* New York: Springer-Verlag.

Schroer, A. F. W., Belgers, J. D. M., Brock, T. C. M., Matser, A. M., Maund, S. J. and Van den Brink, P. J. (2004). Comparison of Laboratory Single Species and Field Population-Level Effects of the Pyrethroid Insecticide Lamda-Cyhalothrin on Freshwater Invertebrates. *Archives of Environmental Contamination and Toxicology*, **46**: 324–335.

Shao, Q. (2000). Estimation for Hazardous Concentrations Based on NOEC Toxicity Data: An Alternative Approach. *Environmetrics*, **11**: 583–595.

Solomon, K. R. and Takacs, P. (2002). Probabilistic Risk Assessment Using Species Sensitivity Distributions. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 285–313.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B*, **64**: 1–34.

Staples, C. A., Woodburn, K. B., Klecka, G. M., Mihaich, E. M., Hall, A. T., Ortego, L., Caspers, N. and Hentges, S. G. (2008). Comparison of Four Species Sensitivity Distribution Methods to Calculate Predicted No Effect Concentrations for Bisphenol A. *Human and Ecological Risk Assessment*, **14**: 455–478.

Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, **69**: 730–737.

Stephan, C. E., Mount, D. I., Hansen, D. J., Gentile, J. H., Chapman, G. A. and Brungs, W. A. (1985). Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses. Report No. PB 85-227049. U.S. Environmental Protection Agency, Office of Research and Development, Duluth, MN.

Stephan, C. E. (2002). Use of Species Sensitivity Distributions of Water Quality Criteria for Aquatic Life by the U.S. Environmental Protection Agency. In:

Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 211–220.

Stuart, A., Ord, J. K. and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics. Volume 2A: Classical Inference and the Linear Model.* **6th ed.** London: Arnold Publishers.

II Suter, G. W. (2002). North American History of Species Sensitivity Distributions. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 11–17.

Trass, T. P., Van De Meent, D., Posthuma, L., Hamers, T., Kater, B. J., De Zwart, D. and Aldenberg, T. (2002). The Potentially Affected Fraction as a Measure of Ecological Risk. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 315–344.

UN (United Nations) (2009). United Nations Flash Environmental Assessment Tool (FEAT). Version 1.0. Available at: `http://ochaonline.un.org/OchaLinkClick.aspx?link=ocha&docId=1135169` [Accessed: 1st June 2009].

US EPA (U.S. Environmental Protection Agency). (1998). Guidelines for Ecological Risk Assessment; Notice. *Federal Register*, **63**: 26846–26924.

US EPA (U.S. Environmental Protection Agency). (2002). Guidelines Establishing Test Procedures for the Analysis of Pollutants; Whole Effluent Toxicity Test Methods; Final Rule. *Federal Register*, **67**: 69951–69972.

US EPA (U.S. Environmental Protection Agency). (2004). Overview of the Ecological Risk Assessment Process in the Office of Pesticide Programs. Office of Prevention, Pesticides and Toxic Substances, Washington, DC.

US EPA (U.S. Environmental Protection Agency). (2007). ECOTOX User Guide: ECOTOXicology Database System. Version 4.0. Available at: `http:/www.epa.gov/ecotox/` [Accessed: 1st May 2007].

Van den Brink, P. J., Blake, N., Brock, T. C. M. and Maltby, L. (2006). Predictive Value of Species Sensitivity Distributions for Effects of Herbicides in Freshwater Ecosystems. *Human and Ecological Risk Assessment*, **12**: 645–674.

Van Straalen, N. M., and Denneman, C. A. J. (1989). Ecotoxicological Evaluation of Soil Quality Criteria. *Ecotoxicology and Environmental Safety*, **18**: 241–25

Van Straalen, N. M. (2002). Theory of Ecological Risk Assessment Based on Species Sensitivity Distributions. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 37–48.

Van Straalen, N. M. and Van Leeuwen, C. J. (2002). European History of Species Sensitivity Distributions. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 19–34.

Varian, H. R. (1975). A Bayesian Approach to Real Estate Assessment. In: Fienberg, E. and Zellner, A. (1975). *Studies in Bayesian Econometrics and Statistics in Honour of Leonard J. Savage.* Amsterdam: North-Holland Publishing Company, 195–208.

Verdonck, F. A. M., Aldenberg, T., Jaworska, J., and Vanrolleghem, P. A. (2003). Limitations of Current Risk Characterization Methods in Probabilistic Environmental Risk Assessment. *Environmental Toxicology and Chemistry*, **22**: 2209–2213.

Versteeg, D. J., Belanger, S. E. and Carr, G. J. (1999). Understanding Single-Species and Model Ecosystem Sensitivity: Data-Based Comparison. *Environmental Toxicology and Chemistry*, **18**: 1329–1346.

Van Vlaardingen, P. L. A., Traas, T. P., Wintersen, A. M. and Aldenberg, T. (2004). E_TX 2.0: A Program to Calculate Hazardous Concentrations and Fraction Affected, Based on Normally Distributed Toxicity Data. RIVM Report: 601501028/2004. Bilthoven, The Netherlands.

Wagner, C. and Løkke, H. (1991). Estimation of Ecotoxicology Protection Levels from NOEC Toxicity Data. *Water Research*, **25**: 1237–1242.

Warren-Hicks, W. J., Parkhurst, B. J. and Butcher, J. B. (2002). Methodology for Aquatic Ecological Risk Assessment. In: Posthuma, L., II Suter, G. W. and Traas, T. P. Eds. *Species Sensitivity Distributions in Ecotoxicology.* Boca Raton: Lewis Publishers, 345–382.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* **2nd ed.** New York: Springer.

Wheeler, J. R., Grist, E. P. M., Leung, K. M. Y, Morritt, D. and Crane, M. (2002). Species Sensitivity Distributions: Data and Model Choice. *Marine Pollution Bulletin*, **45**: 192–202.

Zeeman, M. G. (1995). Ecotoxicity Testing and Estimation Methods Developed Under Section 5 of The Toxic Substances Control Act (TSCA). In: Rand, G. M. Ed. *Fundamentals of Aquatic Toxicology: Effects, Environmental Fate and Risk Assessment.* **2nd ed.** Philadelphia: Taylor and Francis, 703–716.

Zellner, A. (1986). Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association*, **81**: 446–451.

Zieliński, R. (2005). Estimating Quantiles With Linex Loss Function. Applications to VaR Estimation. *Applicationes Mathematicae*, **32**: 367–373.

Appendix A

Posterior Distributions & [EFSA] Estimators

A.1 Posterior Distributions

Here we give details of the posterior distributions derived for the unknown SSD parameters θ in the species exchangeable context; all relevant notation was provided in Section 2.5. Within the Bayesian paradigm, the data generating mechanism for toxicity data $\mathbf{Y}$ based on substance $\mathcal{S}$ can be written identically for each behavioural model as $y_j \,|\, \mu, \sigma^2 \sim N(\mu, \sigma^2)$ for $j \in J_{\mathcal{S}}$.

For behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$, the likelihood function for the data conditioned on $\theta = (\mu, \sigma^2)$ is immediately defined from the above distributional assumption. In the case of $\mathcal{M}_3$, the data generating mechanism is augmented such that the toxicity data for the N additional substances $\mathcal{G}$ is used under the assumption that the log-SSD variance parameter σ^2 is homogeneous between these substances and $\mathcal{S}$. Hence, one additionally has $y_{ij} \,|\, \mu_i, \sigma^2 \sim N(\mu_i, \sigma^2)$ for $j \in J_i$ and $i \in \mathcal{G}$, and we denote $\theta = (\mu, \sigma^2, \mu_i : i \in \mathcal{G})$. For $\mathcal{M}_4$, the likelihood function is also immediately defined, under the assumption that σ is fixed and known; we denote the unknown SSD parameter as $\theta \equiv \mu$.

In all behavioural models, it is assumed that the non-informative prior distribution for μ (extendable to $(\mu_i : i \in \mathcal{G})$ also) is $p(\mu) \propto 1$ for $\mu \in \mathbb{R}$. For $\mathcal{M}_1$, the prior distribution for σ^2 is given as $p(\sigma^2) \propto \sigma^{-2}$ for $\sigma^2 \in \mathbb{R}^+$. These two distributions

are independent Jeffreys' priors, which together constitute a practical Jeffreys prior for θ; see Section 2.6.3 and Kass and Wasserman (1996) for further details. The behavioural model for $\mathcal{M}_2$ implies that $\sigma^2 \,|\, \alpha, \beta$ has an inverse-gamma distribution for $\sigma^2 \in \mathbb{R}^+$ with shape and scale parameters $\alpha > 0$ and $\beta > 0$, which are estimated separately from an additional toxicity database. As for $\mathcal{M}_1$, the independent Jeffreys prior distribution for $\sigma^2 > 0$ (treating μ separately) is used for $\mathcal{M}_3$. For $\mathcal{M}_4$, the prior distribution is singularly defined by the distribution of μ, *a priori*.

For a fully specified likelihood function $\ell(\theta \,|\, \mathbf{Y})$ (in the case of $\mathcal{M}_3$ one would augment $\mathbf{Y}$ as $\{\mathbf{Y}, \mathbf{Y}_G\}$ instead) and prior distribution function $p(\theta)$, one can update the prior distribution using Bayes' Theorem, to admit the posterior $p(\theta \,|\, \mathbf{Y})$:

$$p(\theta \,|\, \mathbf{Y}) = \frac{\ell(\theta \,|\, \mathbf{Y}) p(\theta)}{\int_\Theta \ell(\theta \,|\, \mathbf{Y}) p(\theta)\, d\theta}.$$

Under the behavioural model of $\mathcal{M}_3$, it is necessary for the nuisance parameters $(\mu_1, \ldots, \mu_N)$ to be integrated out since they have no bearing on the assessment pertaining to $\mathcal{S}$; this leaves the marginalised posterior distribution function.

It is straightforward to deduce that for all behavioural models, *a posteriori*, $p(\mu, \sigma^2 \,|\, \mathbf{Y}) = p(\mu \,|\, \sigma^2; \mathbf{Y}) p(\sigma \,|\, \mathbf{Y})$, such that $\mu \,|\, \sigma^2; \mathbf{Y} \sim N(\bar{y}, \sigma^2/n)$ for $\mu \in \mathbb{R}$ and $\sigma^2 \,|\, \mathbf{Y}$ is distributed according to the behavioural model as

$\mathcal{M}_1$**:** $\sigma^2 \,|\, \mathbf{Y} \sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{n-1}{2} s^2\right)$.

$\mathcal{M}_2$**:** $\sigma^2 \,|\, \mathbf{Y} \sim \mathcal{IG}\left(\frac{2\alpha+n-1}{2}, \frac{2\alpha+n-1}{2} s^2_{\text{adj}}\right)$.

$\mathcal{M}_3$**:** $\sigma^2 \,|\, \mathbf{Y}, \mathbf{Y}_G \sim \mathcal{IG}\left(\frac{\varsigma}{2} + n - 1, \frac{\varsigma+n-1}{2} s^{*2}_p\right)$.

for $\sigma^2 \in \mathbb{R}^+$; where s^2_{adj} is defined by Equation 2.14; s^{*2}_p is defined by Equation 2.17; and $\varsigma = \sum_{i \in \mathcal{G}} (n_i - 1)$.

It should be noted that while the prior distributions for each behavioural model are improper, i.e. they do not integrate to unity, the posterior distributions are in fact proper and well defined on the appropriate domain. This property allows us to circumvent many, but not all, of the problems encountered by undefined prior distributions.

A.2 [EFSA] Estimators

Here we give details of the derivation of the standard [EFSA] estimators from a Bayesian perspective, under the species exchangeable context. Recall from Appendix A.1 that the posterior distributions of (μ, σ^2) for $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are of the form

$$\begin{aligned} \mu \,|\, \sigma^2; \mathbf{Y} &\sim N(\bar{y}, \sigma^2/n) \quad \mu \in \mathbb{R}; \\ \sigma^2 \,|\, \mathbf{Y} &\sim \mathcal{IG}\,(\lambda_0, \lambda_1) \quad \sigma^2 \in \mathbb{R}^+ \end{aligned}$$

for some λ_0 and λ_1 which are functions of the data and prior distribution parameters.

Furthermore, recall from Section 2.6.4 that the MFE $= \mathbb{E}^{\theta \,|\, \mathbf{Y}} \Phi\left(\frac{\delta_p(\mathbf{Y}) - \mu}{\sigma}\right)$ where $\delta_p(\mathbf{Y})$ is a possible decision rule (cf. Equation 2.10). Hence, the MFE is equal to

$$\begin{aligned} &\int_0^\infty \mathbb{P}[\mu + \sigma Z < \delta_p \,|\, \sigma^2; \mathbf{Y}] f\left(\sigma^2 \,|\, \mathbf{Y}\right) d\sigma^2 \\ &= \int_0^\infty \mathbb{P}\left[Z' < \frac{\delta_p - \bar{y}}{\sigma\sqrt{1 + 1/n}} \,\middle|\, \sigma^2; \mathbf{Y} \right] f\left(\sigma^2 \,|\, \mathbf{Y}\right) d\sigma^2, \end{aligned}$$

where Z and Z' are both independently standard normal random variables. By definition of the normal and inverse-gamma distribution, we can explicitly write the MFE as

$$\frac{1}{\sqrt{2\pi}} \frac{\lambda_1^{\lambda_0}}{\Gamma(\lambda_0)} \int_0^\infty \int_{-\infty}^{\tilde{u}(\delta_p \,|\, \mathbf{Y})} \sigma^{-2(\lambda_0 + 3/2)} \exp\left\{ -\sigma^{-2} \left(\lambda_1 + \tfrac{u^2}{2} \right) \right\} du d\sigma^2,$$

where $\tilde{u}(\delta_p \,|\, \mathbf{Y}) = \left\{ u \in \mathbb{R} : \delta_p \geq \bar{y} + \sqrt{1 + \frac{1}{n}} u \right\}$. Next we perform integration with respect to σ^2, and subsequently changing variables to $t = u\sqrt{\lambda_0/\lambda_1}$ leads one to obtain

$$\frac{1}{\sqrt{2\pi\lambda_0}} \frac{\Gamma\left(\lambda_0 + \frac{1}{2}\right)}{\Gamma(\lambda_0)} \int_{-\infty}^{\sqrt{\frac{\lambda_0}{\lambda_1}} \tilde{u}(\delta_p \,|\, \mathbf{Y})} \left(1 + \frac{t^2}{2\lambda_0} \right)^{-(2\lambda_0 + 1)/2} dt.$$

The integrand is observed to be the density function of a Student-t distribution with $2\lambda_0$ degrees of freedom, implying that the MFE is $F_{t_{2\lambda_0}}\left(\tilde{u}(\delta_p \,|\, \mathbf{Y}) \sqrt{\lambda_0 \lambda_1^{-1}}\right)$ where $F_{t_{2\lambda_0}}$ is the corresponding CDF.

If one controls the MFE to be $p\%$, then an [EFSA] decision rule is

$$\delta_p(\mathbf{Y})_{[\text{EFSA}]} = \bar{y} - \sqrt{1 + \frac{1}{n}} \sqrt{\frac{\lambda_1}{\lambda_0}} t_{2\lambda_0, p},$$

where $t_{2\lambda_0,p}$ is the $(100-p)$-th percentile of a Student-t distribution with $2\lambda_0$ degrees of freedom. Substituting λ_0 and λ_1 with the shape and scale parameters for the posterior distribution of σ^2 conditional on the relevant behavioural model concludes the derivation.

Appendix B

Hyper-parameter Estimation

In this appendix we give details of the different methods for estimating the hyper-parameters of the various behavioural models, and models for non-exchangeability.

B.1 Notation

We denote $\mathcal{G}$ as the total collection of substances available in the toxicity database. $\mathcal{G}_2$ is the group of substances relevant to estimating the log-SSD variance heterogeneity parameters α and β, which are specified through the hierarchical model as the shape and scale parameters of an inverse-gamma distribution representing the population with which σ^2 is *a priori* sampled from under the behavioural model $\mathcal{M}_2$.

In a species exchangeable context, for each substance $i \in \mathcal{G}$ and species $j \in J_i$, log-tolerance values y_{ij} are assumed to be conditionally and independently normally distributed with mean μ_i and variance σ_i^2, where J_i is the collection of species of which have been tested with substance i, such that $|J_i| = n_i$. In addition, we denote $\bar{y}_i$ and s_i as the usual sample mean and standard deviation of the log-toxicity data for substance i. In the interest of clarity, we extend this notation by alternatively writing the precision $\tau_i = 1/\sigma_i^2$ where appropriate, and $\mu_{\mathcal{G}}$ and $\tau_{\mathcal{G}}$ as shorthand for the vectors of the μ_i and τ_i for $i \in \mathcal{G}$ respectively.

When a model for the non-exchangeability of a single special species is appropriate, we use a subset of $\mathcal{G}$, denoted $\mathcal{G}_1$, which contains all substances relevant to

the estimation of the non-exchangeability parameters: (k', ϕ') for the EFSA model (denoted $\mathcal{D}_1$); and (k, ϕ) for the model used in this research (denoted $\mathcal{D}_2$) which leads to tractable decision rules. The model under $\mathcal{D}_1$ for all substances $i \in \mathcal{G}_1$ is

$$\begin{aligned} y_{ij} \,|\, \mu_i, \sigma_i^2 &\sim N(\mu_i, \sigma_i^2) \text{ for } j \in J_i^*; \\ y_i^\dagger \,|\, \mu_i, \sigma_i^2, k', \phi' &\sim N\left(\mu_i - k'\sigma_i, [\phi'\sigma_i]^2\right), \end{aligned}$$

where J_i^* is the collection of ordinary species of which tolerance values have been recorded for substance i, such that $|J_i^*| = n_i^* \equiv n_i - 1$. Also, $y_i^\dagger$ is the log-tolerance value for the special species assessed with substance $i \in \mathcal{G}_1$. The model under $\mathcal{D}_2$ for all substances $i \in \mathcal{G}_1$ is

$$\begin{aligned} y_{ij} \,|\, \mu_i, \sigma_i^2 &\sim N(\mu_i, \sigma_i^2) \text{ for } j \in J_i^*; \\ y_i^\dagger \,|\, \mu_i, \sigma_i^2, k, \phi &\sim N\left(\mu_i - k, [\phi\sigma_i]^2\right). \end{aligned}$$

A restriction is made on $\mathcal{G}_1$ such that $n_i^* \geq 2 \; \forall i \in \mathcal{G}_1$. In addition, we take the assumption that $\mathcal{G}_1 \subseteq \mathcal{G}_2$; discussion on these assumptions can be found in Section 5.5. We will denote v_1 and v_2 as the number of substances in $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. For clarity, under $\mathcal{D}_1$ and $\mathcal{M}_2$ we denote the heterogeneity parameters as (α', β'). Finally, we define $\bar{y}_i^*$ and s_i^* as the usual sample mean and standard deviation of the log-tolerance values for all species in J_i^*.

B.2 Estimation of α & β: EFSA Method

Here we give details of a frequentist methodology proposed by EFSA (2005) for estimating hyper-parameters α and β whilst working under the context of species exchangeability and behavioural model $\mathcal{M}_2$. Begin by noting that the full likelihood function of the unknown parameters for substances in $\mathcal{G}_2$ is:

$$\ell(\mu_{\mathcal{G}_2}, \tau_{\mathcal{G}_2}) = \prod_{i \in \mathcal{G}_2} \prod_{j \in J_i} \sqrt{\frac{\tau_i}{2\pi}} \exp\left\{-\frac{\tau_i}{2}(y_{ij} - \mu_i)^2\right\}. \tag{B.1}$$

From within the frequentist paradigm, we augment the model by invoking that each τ_i is conditionally independently sampled from a gamma distribution with shape α and rate β. The *marginalised likelihood* function is obtained by multiplying Equation B.1 by $\prod_{i\in\mathcal{G}_2} p(\tau_i \,|\, \alpha, \beta)$ and integrating with respect to the 'nuisance parameters' $\tau_{\mathcal{G}_2}$. In the integral, τ_i only features as

$$\tau_i^{\alpha+(n_i/2)-1} \exp\left\{-\frac{\tau_i}{2}\left[\sum_{j\in J_i}(y_{ij}-\mu_i)^2 + 2\beta\right]\right\},$$

which is proportional to a gamma distribution with shape parameter $\alpha + (n_i/2)$ and rate parameter $\beta + \frac{1}{2}\sum_{j\in J_i}(y_{ij}-\mu_i)^2$. Hence,

$$\ell(\mu_{\mathcal{G}_2}, \alpha, \beta) = \prod_{i\in\mathcal{G}_2} (2\pi)^{-\frac{n_i}{2}} \frac{\beta^\alpha \Gamma(\alpha + n_i/2)}{\Gamma(\alpha)[\beta + \frac{1}{2}\sum_{j\in J_i}(y_{ij}-\mu_i)^2]^{(\alpha+n_i/2)}}.$$

Maximisation with respect to μ_i is achieved at $\hat{\mu}_i = \bar{y}_i$, independently of α and β, thus the profile marginal log-likelihood for α and β is

$$L(\alpha,\beta) = \sum_{i\in\mathcal{G}_2}\Big\{-\frac{n_i}{2}\log(2\pi) + \alpha\log\beta + \log\Gamma(\alpha+n_i/2) \\ - \log\Gamma(\alpha) - (\alpha+n_i/2)\log\left[\beta + \tfrac{1}{2}(n_i-1)s_i^2\right]\Big\}. \quad \text{(B.2)}$$

Equation B.2 is easily maximised using suitable software, such as the `optim()` function in `R` (2006).

B.2.1 Example: RIVM Fish Database

For the RIVM fish toxicity database we analyse in this research (see Section 4.1), we determined that $\hat{\vartheta} \equiv (\hat{\alpha}, \hat{\beta}) = (1.05, 0.088)$. Also, using standard likelihood theory (Rice, 1995) the covariance matrix is estimated by $\left[-L''(\hat{\vartheta})\right]^{-1}$, where $L''(\cdot)$ is the Hessian matrix of the log-likelihood function given by Equation B.2. For the RIVM fish database, we retrieve

$$\left[-L''(\hat{\vartheta})\right]^{-1} = \begin{pmatrix} 0.012288 & 0.001615 \\ 0.001615 & 0.000265 \end{pmatrix}.$$

A $100(1-x)\%$ confidence region for ϑ is specified by the region

$$\left\{\vartheta : (\vartheta - \hat{\vartheta})^T L''(\hat{\vartheta})(\vartheta - \hat{\vartheta}) \leq \chi^2_{2,x}\right\},$$

where $\chi^2_{2,x}$ is the $100(1-x)$-th percentile of a Chi-square distribution with 2 degrees of freedom.

B.3 Estimation of k' & ϕ': EFSA Method

Here we give details of a frequentist methodology proposed by EFSA (2005) for estimating hyper-parameters k' and ϕ' whilst working under the context of non-exchangeable species model $\mathcal{D}_1$, and behavioural model $\mathcal{M}_1$. The hyper-parameter estimates proposed were intended for application with a deterministic procedure.

For all substances $i \in \mathcal{G}_1$, define $t_i = (y_i^\dagger - \bar{y}_i^*)/s_i^*$. Then, from the non-exchangeability model $\mathcal{D}_1$, one deduces that

$$\begin{aligned} \bar{y}_i^* \,|\, \mu_i, \sigma_i &\sim N(\mu_i, \sigma_i^2/n_i^*); \\ \sqrt{n_i^* - 1}\frac{s_i^*}{\sigma_i} \,\Big|\, \sigma_i, \phi', &\sim \chi_{n_i^*-1}, \end{aligned}$$

where χ_π denotes a Chi distribution with π degrees of freedom. Therefore, the numerator in t_i is normally distributed with mean $-k'\sigma_i$ and standard deviation $\sigma_i\sqrt{\phi'^2 + 1/n_i^*}$. Hence $t_i/\sqrt{\phi'^2 + 1/n_i^*}$ has a non-central t-distribution with $\pi = n_i^* - 1$ degrees of freedom and non-centrality parameter $\eta = -k'/\sqrt{\phi'^2 + 1/n_i^*}$.

Since each statistic t_i is completely independent of μ_i and σ_i, we can determine the likelihood function for k' and ϕ', for all substances i such that the special species is assessed, to be

$$\prod_{i \in \mathcal{G}_1} \frac{1}{\sqrt{\phi'^2 + \frac{1}{n_i^*}}} f_{T_{\pi,\eta}}\left(t_i \Big/ \sqrt{\phi'^2 + 1/n_i^*}\right),$$

where $f_{T_{\pi,\eta}}$ denotes the PDF of the non-central t-distribution with π degrees of freedom and non-centrality parameter η. Maximisation of this function is simple using suitable software, such as that discussed previously, and leads to estimates of k' and ϕ'.

Note that one substance had to be removed from the RIVM database in order to perform this maximisation. This was because $s^* = 0$ which leads to an undefined value of t.

B.4 $\mathcal{MAP}$ Estimators of k & ϕ

Here we give details of a method for estimating hyper-parameters k and ϕ whilst working under the context of non-exchangeable species model $\mathcal{D}_2$, and behavioural model $\mathcal{M}_1$. Begin by noting that the full likelihood function for the unknown parameters for substances in $\mathcal{G}_1$ is:

$$\begin{aligned}
&\ell(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1}) \\
&\quad \propto \prod_{i \in \mathcal{G}_1} \phi^{-1} \tau_i^{n_i/2} \exp\left\{ -\frac{\tau_i}{2} \left[\phi^{-2}(y_i^\dagger - \mu + k)^2 + \sum_{j \in J_i^*} (y_{ij} - \mu_i)^2 \right] \right\} \\
&\quad = \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \tau_i^{n_i/2} \exp\left\{ -\frac{\tau_i}{2} \left[(\phi^{-2} + n_i^*)(\hat{\mu}_i - \mu_i)^2 + (n_i - 1)\hat{\sigma}_i^2 \right] \right\}, \qquad \text{(B.3)}
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mu}_i &= \frac{\phi^{-2}(y_i^\dagger + k) + n_i^* \bar{y}_i^*}{\phi^{-2} + n_i^*}; \qquad \text{(B.4)} \\
\hat{\sigma}_i^2 &= \frac{1}{n_i - 1} \left[\phi^{-2}(y_i^\dagger + k - \hat{\mu}_i)^2 + n_i^*(\hat{\mu}_i - \bar{y}_i^*)^2 + (n_i^* - 1) s_i^{*2} \right]. \qquad \text{(B.5)}
\end{aligned}$$

Multiplying the likelihood function by the joint prior distribution for k, ϕ, $\mu_{\mathcal{G}_1}$ and $\tau_{\mathcal{G}_2}$, which was defined in Section 5.5, yields the un-normalised posterior distribution. Note that because we use the precision τ as opposed to the variance σ^2 in the interest of clarity, it is necessary to transform the prior distribution for each τ_i; which is determined to be $p(\tau_i) \propto \tau_i^{-1}$ for $\tau_i > 0$, independently for each $i \in \mathcal{G}_1$. The un-normalised posterior distribution is defined as

$$\begin{aligned}
&p(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1} \mid \mathbf{Y}) \\
&\quad \propto \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \tau_i^{n_i/2 - 1} \exp\left\{ -\frac{\tau_i}{2} \left[(\phi^{-2} + n_i^*)(\hat{\mu}_i - \mu_i)^2 + (n_i - 1)\hat{\sigma}_i^2 \right] \right\}
\end{aligned}$$

for $\mu_i \in \mathbb{R}$, $\tau_i \in \mathbb{R}^+$, $k \in \mathbb{R}$, $\phi \in \mathbb{R}^+$. Integrating the un-normalised posterior distribution with respect to the 'nuisance parameters' $\mu_{\mathcal{G}_1}$ and $\tau_{\mathcal{G}_1}$, yields the un-normalised marginalised posterior distribution for the hyper-parameters. Hence,

$$p(k, \phi \,|\, \mathbf{Y}) \propto \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \frac{\Gamma(\hat{\alpha}_i)}{\hat{\beta}_i^{\hat{\alpha}_i}} \frac{1}{\sqrt{\phi^{-2} + n_i^*}},$$

where $\hat{\alpha}_i = \frac{1}{2}(n_i - 1)$ and $\hat{\beta}_i = \hat{\alpha}_i \hat{\sigma}_i^2$. Maximising this function with respect to its arguments determines the joint $\mathcal{MAP}$ estimator.

B.5 $\mathcal{MAP}$ Estimators of k' & ϕ'

Here we give details of a method for estimating hyper-parameters k' and ϕ' whilst working under the context of non-exchangeable species model $\mathcal{D}_1$, and behavioural model $\mathcal{M}_1$.

Essentially the likelihood function for this model is the same as in the final line of Equation B.3, except now $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are implicit functions of the different non-exchangeability hyper-parameters and τ_i, as we must replace k by $k'/\sqrt{\tau_i}$ and ϕ by ϕ' in Equations B.4 and B.5; we therefore denote these two respective equations as $\hat{\mu}_i(\tau_i)$ and $\hat{\sigma}_i^2(\tau_i)$ respectively.

The posterior distribution for k', ϕ', $\mu_{\mathcal{G}_1}$, and $\tau_{\mathcal{G}_1}$ maintains the same form with the changes made for $\hat{\mu}_i$ and $\hat{\sigma}_i^2$, and ϕ replaced by ϕ'. Integration of the full un-normalised posterior with respect to μ_i is a tractable calculation; however integration with respect to each τ_i must be approximated numerically which may be done straightforwardly by numerical quadrature to high accuracy. Hence, the un-normalised posterior distribution for k' and ϕ' is

$$p(k', \phi' \,|\, \mathbf{Y}) \propto \phi^{-v_1} \prod_{i \in \mathcal{G}_1} \frac{1}{\sqrt{\phi'^{-2} + n_i^*}} \int_0^\infty \tau_i^{\hat{\alpha}_i - 1} \exp\left\{ -\frac{\tau_i}{2}(n_i - 1)\hat{\sigma}_i^2(\tau_i) \right\} d\tau_i.$$

This posterior distribution is maximised in a similar fashion as before with respect to its arguments to determine the joint $\mathcal{MAP}$ estimator.

B.6 $\mathcal{MAP}$ Estimators of k, ϕ, α & β

Here we give details of a method for estimating hyper-parameters k, ϕ, α and β whilst working under the context of non-exchangeable species model $\mathcal{D}_2$, and behavioural model $\mathcal{M}_2$. Our derivation follows on from Appendix B.4. For $\mathcal{M}_2$, we use the additional $v_2 - v_1$ substances in $\mathcal{G}_2 \backslash \mathcal{G}_1$. The likelihood function defined by Equation B.3 is augmented such that

$$\begin{aligned} &\ell(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1}, \mu_{\mathcal{G}_2 \backslash \mathcal{G}_1}, \tau_{\mathcal{G}_2 \backslash \mathcal{G}_1}) \\ &\quad = \ \ell(k, \phi, \mu_{\mathcal{G}_1}, \tau_{\mathcal{G}_1}) \prod_{i \in \mathcal{G}_2 \backslash \mathcal{G}_1} \tau_i^{n_i/2} \exp\left\{ -\frac{\tau_i}{2} \left[n_i (\bar{y}_i - \mu_i)^2 + (n_i - 1) s_i^2 \right] \right\}. \end{aligned} \quad \text{(B.6)}$$

We next multiply by the joint posterior distribution, defined in Section 5.5, recalling that $\mathcal{G}_1 \subseteq \mathcal{G}_2$ so that the prior distribution for each τ_i is a gamma distribution with shape parameter α and rate parameter β. As for $\mathcal{M}_1$, we integrate out the nuisance parameters in order to obtain the un-normalised posterior distribution for the remaining hyper-parameters, which, via an extension of the earlier calculations, is determined to be

$$p(\alpha, \beta, k, \phi \,|\, \mathbf{Y}) \propto \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \right]^{v_2} \phi^{-v_1} \left(\prod_{i \in \mathcal{G}_1} \frac{1}{\sqrt{\phi^{-2} + n_i^*}} \right) \left(\prod_{i \in \mathcal{G}_2} \frac{\Gamma(\tilde{\alpha}_i)}{\tilde{\beta}_i^{\tilde{\alpha}_i}} \right), \quad \text{(B.7)}$$

where $\tilde{\alpha}_i = \alpha + \hat{\alpha}_i$ and $\tilde{\beta}_i = \beta + \hat{\beta}_i$ for $i \in \mathcal{G}_2 \subseteq \mathcal{G}_1$. As was the case earlier, maximisation with respect to its arguments determines the joint $\mathcal{MAP}$ estimator.

B.7 $\mathcal{MAP}$ Estimators of k', ϕ', α' & β'

Here we give details of a method for estimating hyper-parameters k', ϕ', α' and β' whilst working under the context of non-exchangeable species model $\mathcal{D}_2$, and behavioural model $\mathcal{M}_1$.

We follow a similar strategy to that of Appendix B.5 whereby we modify the estimation procedure provided in Appendix B.6. Begin by noticing that the likelihood function for the model here is essentially the same as defined by Equation B.6, except the part pertaining to $\mathcal{M}_1$ is as described in Appendix B.5, i.e. we must

replace k by $k'/\sqrt{\tau_i}$ and ϕ by ϕ' in Equations B.4 and B.5.

We multiply by the identical posterior distribution as in Appendix B.6 with k and ϕ replaced by k' and ϕ' respectively. As per before, the integration with respect to $\mu_{\mathcal{G}_2}$ is tractable, however it is required that one numerically approximate the integrals for each τ_i. Hence, the un-normalised posterior distribution for the hyper-parameters is

$$p(k', \phi', \alpha', \beta' \,|\, \mathbf{Y}) \propto \left[\frac{\beta'^{\alpha'}}{\Gamma(\alpha')}\right]^{v_2} \phi^{-v_1} \left(\prod_{i \in \mathcal{G}_2 \setminus \mathcal{G}_1} \int_0^\infty \tau_i^{\hat{\alpha}_i - 1} \exp\left\{-\tau_i \hat{\beta}_i\right\} \, d\tau_i\right) \left(\prod_{i \in \mathcal{G}_1} \int_0^\infty \tau_i^{\tilde{\alpha}_i - 1} \exp\left\{-\frac{\tau_i}{2}\left[2\beta' + (n_i - 1)\hat{\sigma}_i^2(\tau_i)\right]\right\} \, d\tau_i\right).$$

This posterior distribution is maximised in a similar fashion as before with respect to its arguments to determine the joint $\mathcal{MAP}$ estimator.

Appendix C

Deriving Bayes Rules

In this appendix we give details about the derivation of certain Bayes rules (and other estimators where appropriate) which were discussed in Chapter 3. Our operational procedure is to determine the decision rule $\delta_p^*(\mathbf{Y})$ which minimises the posterior expected loss; this is what we defined to be the Bayes rule (cf. Equation 3.1). The relevant posterior distributions for the different behavioural models are discussed in Appendix A for species exchangeable contexts; the corresponding versions for species non-exchangeable models are described throughout Chapter 5, and can be used, if required, to easily augment the decision rules listed here.

C.1 [SEL]: $\mathcal{M}_1$ & $\mathcal{M}_2$

The posterior expected loss is straightforwardly given by $\mathbb{E}^{\theta\,|\,\mathbf{Y}}[\psi_p(\theta) - \delta_p]^2$. It is then simple to deduce (for example consult Berger, 1985, p. 161) that this quantity is minimised at $\delta_p^*(\mathbf{Y}) = \mathbb{E}^{\theta\,|\,\mathbf{Y}}\psi_p(\theta)$. Using standard properties of conditional expectation yields the Bayes rule as

$$\delta_p^*(\mathbf{Y})_{[\mathrm{SEL}]} = \bar{y} - \kappa_p(n,\alpha)\hat{\sigma}, \tag{C.1}$$

where the assessment shift-factor is defined by

$$\kappa_p(n,\alpha) \equiv K_p\hat{\sigma}^{-1}\mathbb{E}^{\sigma^2\,|\,\mathbf{Y}}\sigma = K_p\sqrt{\frac{2\alpha+n-1}{2}}\frac{\Gamma\left(\frac{2\alpha+n-2}{2}\right)}{\Gamma\left(\frac{2\alpha+n-1}{2}\right)}$$

and $\hat{\sigma} = s_{\text{adj}}$. By setting $\alpha = \beta = 0$, Equation C.1 yields the Bayes rule for $\mathcal{M}_1$ (since $s_{\text{adj}} = s$); otherwise setting (α, β) to the values estimated from the additional toxicity database (see for example Appendices B.2) yields the Bayes rule for $\mathcal{M}_2$.

C.2 [AJ] Estimator: $\mathcal{M}_2$

In this section we give the derivation of the [AJ] estimator via similar arguments to those described in Section 2.6.3; i.e. defining it to be the $100(1-\gamma)$-th percentile of the posterior distribution of $\psi_p(\theta)$. Under the behavioural model $\mathcal{M}_2$, the posterior distribution of $\psi_p(\theta) \,|\, \mathbf{Y}$ is a scaled non-central t-distribution with density function

$$f_{\psi_p}(t \,|\, \alpha, \beta; \mathbf{Y}) = \frac{1}{s_{\text{adj}}/\sqrt{n}} f_{T_{2\alpha+n-1},\eta}\left(\frac{\bar{y} - t}{s_{\text{adj}}/\sqrt{n}}\right), \tag{C.2}$$

where $f_{T_{2\alpha+n-1},\eta}$ is the PDF of a non-central t-distribution with $2\alpha + n - 1$ degrees of freedom and non-centrality parameter $\eta = K_p\sqrt{n}$.

By definition, the $100(1-\gamma)$-th percentile of the distribution given by Equation C.2 is the value $\delta_p^{(\gamma)}(\mathbf{Y} \,|\, \alpha, \beta)$ satisfying

$$\int_{-\infty}^{\delta_p^{(\gamma)}(\mathbf{Y} \,|\, \alpha, \beta)} f_{\psi_p}(t \,|\, \alpha, \beta; \mathbf{Y}) dt = 1 - \gamma.$$

Therefore, the [AJ] estimator is

$$\delta_p^{(\gamma)}(\mathbf{Y} \,|\, \alpha, \beta)_{[\text{AJ}]} = \bar{y} - \frac{1}{\sqrt{n}} F^{-1}_{T_{2\alpha+n-1},\eta}(\gamma) s_{\text{adj}}.$$

C.3 [GAL]: $\mathcal{M}_1$ & $\mathcal{M}_2$

Here we outline the derivation of the [GAL] Bayes rule. For notational convenience we will drop the θ dependence parameterisation and denote $\psi_p \equiv \psi_p(\theta)$.

The posterior expected loss can be written as $\mathbb{E}^{\psi_p \,|\, \mathbf{Y}} L(\psi_p, \delta_p)$, and a Bayes rule is defined as the decision rule which minimises this quantity. Thus substituting $L(\psi_p, \delta_p)$ by the GAL function (as defined by Equation 3.6), and differentiating with respect to the decision rule δ_p, implies that a local stationary point is the

solution to

$$\int_{-\infty}^{\delta_p} C_2 f_{\psi_p}(t \,|\, \mathbf{Y}) d\psi_p - \int_{\delta_p}^{\infty} C_1 f_{\psi_p}(t \,|\, \mathbf{Y}) d\psi_p = 0,$$

or equivalently,

$$C_1 \mathbb{P}[\psi_p \geq \delta_p] = C_2 \mathbb{P}[\psi_p \leq \delta_p]$$

This equation is satisfied if and only if

$$\mathbb{P}[\psi_p \leq \delta_p] = \frac{C_1}{C_1 + C_2}.$$

Hence, this stationary point occurs at the $100C_1/(C_1 + C_2)$-th percentile of the posterior distribution of $\psi_p(\theta)$. It is straightforward to confirm this point is in fact a unique minimum, and thus the [GAL] Bayes rule.

The posterior distributions for θ under behavioural models $\mathcal{M}_1$ and $\mathcal{M}_2$ are provided in Appendix A.1. Determination of the posterior distribution of $\psi_p(\theta)$ is then obtainable by routine distribution theory (e.g. see Appendix C.2). In each situation, the [GAL] Bayes rule can be interpreted as an [AJ] decision rule if one sets $1 - \gamma = C_1/(C_1 + C_2)$, or equivalently, $\gamma = C_2/(C_1 + C_2)$. Note, however, that Aldenberg and Jaworska (2000) did consider the behavioural model $\mathcal{M}_2$.

C.4 [LINEX]: $\mathcal{M}_1$–$\mathcal{M}_4$

Here we derive the Bayes rule for scaled LINEX loss function for all behavioural models in order to maximise clarity. As a starting point, we derive the Bayes rule under $\mathcal{M}_2$, and describe how this is extended for other behavioural models. The posterior expected loss is simply given by

$$\mathbb{E}^{\theta \,|\, \mathbf{Y}} \left[\exp\left\{ \lambda \frac{(\delta_p - \psi_p(\theta))}{\sigma} \right\} - \lambda \frac{(\delta_p - \psi_p(\theta))}{\sigma} - 1 \right].$$

The first term in this expectation cannot be tractably determined. However, a closed form integral expression can be obtained as

$$\exp\left\{\lambda K_p + \frac{\lambda^2}{2n}\right\} \frac{\left(\frac{2\alpha+n-1}{2}\right)^{\frac{2\alpha+n-1}{2}}}{\Gamma\left(\frac{2\alpha+n-1}{2}\right)} \times \int_0^\infty t^{\frac{2\alpha+n-3}{2}} \exp\left\{-\left(\frac{2\alpha+n-1}{2}\right)t + \lambda t^{1/2}\frac{(\delta_p - \bar{y})}{s_{\text{adj}}}\right\} dt,$$

where α and β are estimated accordingly from a suitable toxicity database. The latter two terms in the posterior expected loss are equal to

$$-\lambda\left[(\delta_p - \bar{y})\frac{\Gamma\left(\frac{2\alpha+n}{2}\right)}{\Gamma\left(\frac{2\alpha+n-1}{2}\right)}\left(\frac{2\alpha+n-1}{2}\right)^{-1/2} s_{\text{adj}}^{-1}\right] - 1.$$

We next differentiate the posterior expected loss with respect to δ_p, and set this equal to zero in order to determine a local turning point. This leads to the Bayes rule being the solution to

$$\int_0^\infty t^{(2\alpha+n-2)/2} \exp\left\{\lambda\sqrt{t}\frac{(\delta_p - \bar{y})}{s_{\text{adj}}} - \left(\frac{2\alpha+n-1}{2}\right)t\right\} dt = \Gamma\left(\frac{2\alpha+n}{2}\right)\left[\frac{2\alpha+n-1}{2}\right]^{-\frac{2\alpha+n}{2}} \exp\left\{-\lambda\left[K_p + \frac{\lambda}{2n}\right]\right\} \quad \text{(C.3)}$$

for δ_p. The right-hand side of Equation C.3 is independent of the data $\mathbf{Y}$, which implies that the left-hand side must also be independent; hence, $(\delta_p - \bar{y})/s_{\text{adj}}$ must be a constant. By denoting this constant as $-\kappa_p^*$, one can therefore conclude that the Bayes rule is defined as

$$\delta_p^*(\mathbf{Y} \mid \alpha, \beta, \lambda)_{[\text{LINEX}]} = \bar{y} - \kappa_p^*(n, \alpha, \lambda)s_{\text{adj}},$$

where $\kappa_p^*(n, \alpha, \lambda)$ (the assessment shift-factor) is the solution to Equation C.3 for $\kappa_p^* \equiv (\delta_p - \bar{y})/s_{\text{adj}}$, as required. Although not done here, it is straightforward to confirm that $\kappa_p^*(n, \alpha, \lambda)$ is unique for fixed parameters.

Having defined the scaled [LINEX] Bayes rule for $\mathcal{M}_2$, it is now relatively easy to justify the Bayes rule for the other behavioural models considered in this thesis. By

setting $\alpha = \beta = 0$, one obtains the scaled [LINEX] Bayes rule for $\mathcal{M}_1$, as has been demonstrated previously by consequence of the posterior distributions of θ being members of the normal inverse-gamma family. For $\mathcal{M}_3$, the method is exactly the same except that the posterior distribution of $\sigma^2 \,|\, \mathbf{Y}, \mathbf{Y}_G$ is different, as described earlier. It is straightforward to deduce that replacing 2α by ς and s_{adj} by s_p^* will lead to the counterpart scaled [LINEX] Bayes rule. For $\mathcal{M}_4$ we assume σ is known, and so the Bayes rule reduces to being the solution to

$$\delta_p^*(\mathbf{Y} \,|\, \sigma, \lambda)_{[\text{LINEX}]} = -\frac{\sigma}{\lambda} \ln \left(\mathbb{E}^{\mu \,|\, \sigma^2; \mathbf{Y}} \left[\exp \left\{ -\frac{\lambda}{\sigma} \psi_p(\theta) \right\} \right] \right).$$

Standard properties of moment generating functions (for example, consult Rice 1995) for normal random variables identifies the Bayes rule to be

$$\delta_p^*(\mathbf{Y} \,|\, \lambda)_{[\text{LINEX}]} = \bar{y} - \left(K_p + \frac{\lambda}{2n} \right) \sigma.$$

C.5 Scaled [SEL]: $\mathcal{M}_1$

Here we derive the Bayes rules under the behavioural model $\mathcal{M}_1$ for the SEL loss function such that loss is placed on the discrepancy $(\psi_p(\theta) - \delta_p(\mathbf{Y}))/\sigma$. The posterior expected loss is given by

$$\mathbb{E}^{\theta \,|\, \mathbf{Y}} L(\psi_p(\theta), \delta_p; \sigma) = \mathbb{E}^{\theta \,|\, \mathbf{Y}} \left[\frac{\psi_p(\theta) - \delta_p}{\sigma} \right]^2.$$

Next we differentiate this with respect to δ_p to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\delta_p} \mathbb{E}^{\theta \,|\, \mathbf{Y}} L(\psi_p(\theta), \delta_p; \sigma) = 2\delta_p \mathbb{E}^{\theta \,|\, \mathbf{Y}} \left[\sigma^{-2} \right] - 2\mathbb{E}^{\theta \,|\, \mathbf{Y}} \left[\psi_p(\theta) \sigma^{-2} \right].$$

It is straightforward to show that the turning point obtained by equating this to zero is a minimum and hence a Bayes rule. This can be written as

$$\delta_p^\star(\mathbf{Y})_{[\text{SEL}]} = \mathbb{E}^{\theta \,|\, \mathbf{Y}}[\mu] - K_p \frac{\mathbb{E}^{\theta \,|\, \mathbf{Y}}[\sigma^{-1}]}{\mathbb{E}^{\theta \,|\, \mathbf{Y}}[\sigma^{-2}]}.$$

By determining the necessary moments of the normal inverse-gamma posterior distribution, the scaled [SEL] Bayes rule for $\mathcal{M}_1$ is

$$\delta_p^\star(\mathbf{Y})_{[\text{SEL}]} = \bar{y} - K_p\sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}s.$$

C.6 Scaled [GAL]: $\mathcal{M}_1$

Here we derive the Bayes rules under the behavioural model $\mathcal{M}_1$ for the SEL loss function such that loss is placed on the discrepancy $(\psi_p(\theta) - \delta_p(\mathbf{Y}))/\sigma$. The loss function for scaled [SEL] was given by Equation 3.15. The posterior expected loss can be written as

$$\mathbb{E}^{\theta\,|\,\mathbf{Y}}L(\psi_p(\theta),\delta_p;\sigma) = \\ C_1\int_{\psi_p(\theta)\geq\delta_p}\left(\frac{\psi_p(\theta)-\delta_p}{\sigma}\right)p(\theta\,|\,\mathbf{Y})d\theta + C_2\int_{\psi_p(\theta)<\delta_p}\left(\frac{\psi_p(\theta)-\delta_p}{\sigma}\right)p(\theta\,|\,\mathbf{Y})d\theta,$$

where integration is with respect to $\theta = (\mu, \sigma^2)$. By recalling that the posterior distribution can be written as $p(\theta\,|\,\mathbf{Y}) = p(\mu\,|\,\sigma^2;\mathbf{Y})p(\sigma^2\,|\,\mathbf{Y})$, we differentiate the posterior expected loss with respect to δ_p we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\delta_p}\mathbb{E}^{\theta\,|\,\mathbf{Y}}L(\psi_p(\theta),\delta_p;\sigma) &= -C_1\int_0^\infty\frac{1}{\sigma}p(\sigma^2\,|\,\mathbf{Y})\int_{\delta_p+K_p\sigma}^\infty p(\mu\,|\,\sigma^2;\mathbf{Y})d\mu d\sigma^2 \\ &\quad +C_2\int_0^\infty\frac{1}{\sigma}p(\sigma^2\,|\,\mathbf{Y})\int_{-\infty}^{\delta_p+K_p\sigma}p(\mu\,|\,\sigma^2;\mathbf{Y})d\mu d\sigma^2.\end{aligned}$$

A turning point is obtained by equating this to zero, which is straightforwardly shown to be a minimum. This occurs for δ_p which satisfies

$$\frac{C_1}{C_1+C_2}\int_0^\infty\frac{1}{\sigma}p(\sigma^2\,|\,\mathbf{Y})d\sigma^2 = \frac{C_1}{C_1+C_2}\int_0^\infty\frac{1}{\sigma}p(\sigma^2\,|\,\mathbf{Y})d\sigma^2\int_{-\infty}^{\delta_p+K_p\sigma}p(\mu\,|\,\sigma^2;\mathbf{Y})d\mu\sigma^2.$$

The left hand-side term is simply determined as a moment of the inverse-gamma distribution. Denote this moment as

$$M = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\left(\sqrt{\frac{n-1}{2}}s\right)^{-1}.$$

The outer integrand of the right hand-side is equal to M times the probability density function of an inverse-gamma distribution function with shape and scale parameters $n/2$ and $(n-1)s^2/2$ respectively. Hence, the right-hand side integral is equal to $M \times \mathbb{P}[\psi(\theta) \leq \delta_p \,|\, \mathbf{Y}]$ with a re-parameterised distribution of θ (normal inverse-gamma). Using the result from Appendix C.2, the Bayes rule yielded is

$$\delta_p^\star(\mathbf{Y} \,|\, C_1, C_2)_{[\text{GAL}]} = \bar{y} - \frac{\sqrt{n-1}}{n} F^{-1}_{T_{n, K_p\sqrt{n}}}\left(\frac{C_2}{C_1 + C_2}\right) s$$

where $F_{T_{n,K_p\sqrt{n}}}(\cdot)$ is the cumulative distribution function of a non-central T distribution with n degrees of freedom and non-centrality parameter $K_p\sqrt{n}$.

Appendix D

Details of Bayes Factors

Here we give details of how to calculate the Bayes factors discussed in Section 5.7.2. To derive the Bayes factor in favour of $\mathcal{D}_1$ against $\mathcal{D}_2$ based upon the evidence for a single substance $\mathcal{S}$, we fix the hyper-parameters as the $\mathcal{MAP}$ estimates, for reasons discussed in Section 5.7.2. A description of the models for $\mathcal{D}_1$ and $\mathcal{D}_2$ is provided in Section 5.7. We begin by deriving the Bayes factor for $\mathcal{M}_1$. The marginal probability of the data for model $\mathcal{D}_2$, $\mathbb{P}[\mathbf{Y} \,|\, \mathcal{D}_2]$, is given by

$$\int_0^\infty \int_{-\infty}^\infty \ell(\mu, \tau \,|\, k, \phi; \mathbf{Y}) p(\mu, \tau)\, d\mu\, d\tau = c\phi^{-1} \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-n/2} \tau^{n/2-1} \times \exp\left\{ -\frac{\tau}{2} \left[(\phi^{-2} + n^*)(\hat{\mu} - \mu)^2 + (n-1)\hat{\sigma}^2 \right] \right\} d\mu\, d\tau,$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are defined by Equations 5.2 and 5.3 respectively; and c is the undefined normalising constant of the improper non-informative prior distribution. It is simple to deduce that the marginal probability fully integrates out to

$$\frac{c(2\pi)^{-(n-1)/2}}{\phi\sqrt{\phi^{-2} + n^*}} \frac{\Gamma(\hat{\alpha})}{\hat{\beta}^{\hat{\alpha}}},$$

where $\hat{\alpha} = \frac{1}{2}(n-1)$ and $\hat{\beta} = \hat{\alpha}\hat{\sigma}^2$.

The corresponding marginal probability of the data for model $\mathcal{D}_1$ is derived in a similar manner, except now $\hat{\mu}$ and $\hat{\sigma}^2$ are implicit functions of the different non-exchangeability hyper-parameters and τ; this was explained in Appendix B.5 in the context of determining the $\mathcal{MAP}$ estimators. Hence, it is a straightforward

extension to obtain the marginal probability for $\mathcal{D}_1$ as

$$\frac{c(2\pi)^{-(n-1)/2}}{\phi'\sqrt{\phi'^{-2}+n^*}}\int_0^\infty \tau_i^{\hat{\alpha}-1}\exp\left\{-\frac{\tau}{2}(n-1)\hat{\sigma}^2(\tau)\right\}\,d\tau,$$

where $\hat{\mu}(\tau)$ and $\hat{\sigma}^2(\tau)$ are defined in Section B.5, with k replaced by $k'/\sqrt{\tau}$, and ϕ replaced by ϕ' respectively.

Subject to the argument made earlier that the ratio of the undefined normalising constants cancel by division, the Bayes factor B_{12} is defined as

$$\frac{\phi\sqrt{\phi^{-2}+n^*}}{\phi'\sqrt{\phi'^{-2}+n^*}}\frac{\hat{\beta}^{\hat{\alpha}}}{\Gamma(\hat{\alpha})}\int_0^\infty \tau^{\hat{\alpha}-1}\exp\left\{-\frac{\tau}{2}(n-1)\hat{\sigma}^2(\tau)\right\}\,d\tau.$$

The integral is the same as that in Appendix B.5, which is done straightforwardly using numerical quadrature to high accuracy.

The Bayes factor B_{12} for $\mathcal{M}_2$ is a straightforward extension of the derivation for $\mathcal{M}_1$. First, the marginal probability for both $\mathcal{D}_1$ and $\mathcal{D}_2$ is adapted by altering the prior distribution $p(\mu,\tau)$ conditional on the fixed estimates of the appropriate hyper-parameters (α',β') for $\mathcal{D}_1$ and (α,β) for $\mathcal{D}_2$. The Bayes rule is calculated in a similar manner by taking the ratio of these two marginal probabilities, yielding

$$B_{12}=\frac{\beta'^{\alpha'}}{\beta^{\alpha}}\frac{\Gamma(\alpha)}{\Gamma(\alpha')}\frac{\phi\sqrt{\phi^{-2}+n^*}}{\phi'\sqrt{\phi'^{-2}+n^*}}\frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})}\int_0^\infty \tau^{\tilde{\alpha}'-1}\exp\left\{-\tfrac{1}{2}\tau[2\beta'+(n-1)\hat{\sigma}^2(\tau)]\right\}\,d\tau,$$

where $\tilde{\alpha}=\alpha+\hat{\alpha}$; $\tilde{\beta}=\beta+\hat{\beta}$; and $\tilde{\alpha}'=\alpha'+\hat{\alpha}$.

As per the Bayes rule for $\mathcal{M}_1$, the integral is straightforwardly evaluated by numerical quadrature to high accuracy.

Appendix E

Analysis of $V(u; K^*)$

Here we analytically determine $V(u; K^*)$, including $V'(u; K^*)$ and $V''(u; K^*)$, for some of the different shaped SSDs described in Section 6.4. Note that for this appendix, we use the term 'SSD' to refer to distribution over log-concentration. For each SSD we also define the standardised SSD $f_Z(z)$ such that the population mean and variance is 0 and 1 respectively. In addition, we exemplify these functions by plotting $V(u; K^*)$, $V'(u; K^*)$ and $-V''(u; K^*)$ with $K^* = 1$ in: Figure E.1a (normal), Figure E.1b (logistic), Figure E.2 (SN, with $\alpha = \pm 3$), and Figure E.3a (Student-t, $\nu = 3$). In light of the discussion of the Student-t SSD during Section E.3b, we also examine $K^* = 4$.

In Section 6.7.1 we defined $V(u; K^*) = 1 - F_Z\left(F_Z^{-1}(1-u) - K^*\right)$. The first and second derivatives of $V(u; K^*)$ with respect to $u \in (0, 1)$ are given by Equations 6.9 and 6.10 respectively.

E.1 Normal Distribution

$$\begin{aligned} f_Z(z) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}z^2\right\}; \\ V'(u; K^*) &= \exp\left\{-\tfrac{1}{2}K^*\left[K^* - 2\Phi^{-1}(1-u)\right]\right\}; \\ V''(y; K^*) &= -K^*\sqrt{2\pi}\exp\left\{\tfrac{1}{2}\left[\Phi^{-1}(1-u)\right]^2 + K^*\Phi^{-1}(1-u) - \tfrac{1}{2}K^{*2}\right\}, \end{aligned}$$

where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the cumulative distribution and quantile functions re-

spectively.

E.2 Logistic Distribution

For notational convenience, let $q(u) = F_Z^{-1}(1-u) \equiv -s \ln\left(\frac{u}{1-u}\right)$ and $s = \sqrt{3}/\pi$.

$$\begin{aligned} f_Z(z) &= \frac{e^{-z/s}}{s\left(1+e^{-z/s}\right)^2}; \\ V'(u;K^*) &= e^{K^*/s}\left(\frac{1+e^{-q(u)/s}}{1+e^{-(q(u)-K^*)/s}}\right)^2; \\ V''(u;K^*) &= -e^{q(u)/s}\frac{\left(1+e^{-q(u)/s}\right)^4}{\left(1+e^{-(q(u)-K^*)/s}\right)^2}\left\{\frac{\tanh\left(\frac{K^*}{2s}\right)\operatorname{sech}^2\left(\frac{-q(u)}{2s}\right)}{1+\tanh\left(\frac{K^*}{2s}\right)\tanh\left(\frac{-q(u)}{2s}\right)}\right\}. \end{aligned}$$

E.3 Skew-Normal Distribution

For notational convenience, let $q(u) = F_Z^{-1}(1-u)$. Additionally, we also define the following functions:

$$\begin{aligned} \zeta &= \frac{2}{\pi}\left(\frac{\alpha}{\alpha^2+1}\right); \\ \varphi_\zeta(x;\alpha,K^*) &= \phi\left(\alpha\left[\sqrt{1-\zeta}(x-K^*)+\sqrt{\zeta}\right]\right); \\ \Phi_\zeta(x;\alpha,K^*) &= \Phi\left(\alpha\left[\sqrt{1-\zeta}(x-K^*)+\sqrt{\zeta}\right]\right), \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are were defined in Appendix E.1.

$$\begin{aligned} f_Z(z) &= 2\sqrt{1-\zeta}\varphi_\zeta\left(z;0,0\right)\Phi_\zeta\left(z;\alpha,0\right); \\ V'(u;K^*) &= \frac{\varphi_\zeta\left(q(u);0,K^*\right)\Phi_\zeta\left(q(u);\alpha,0\right)}{\varphi_\zeta\left(q(u);0,0\right)\Phi_\zeta\left(q(u);\alpha,K^*\right)}; \\ V''(u;K^*) &= \frac{1}{2}\frac{\varphi_\zeta\left(q(u);0,K^*\right)\Phi_\zeta\left(q(u);\alpha,0\right)}{\varphi_\zeta\left(q(u);0,0\right)\Phi_\zeta\left(q(u);\alpha,K^*\right)} \\ &\quad\times\left\{\alpha\left[\frac{\varphi_\zeta\left(q(u);\alpha,0\right)}{\Phi_\zeta\left(q(u);\alpha,0\right)}-\frac{\varphi_\zeta\left(q(u);\alpha,K^*\right)}{\Phi_\zeta\left(q(u);\alpha,K^*\right)}\right]-\sqrt{1-\zeta}K^*\right\}. \end{aligned}$$

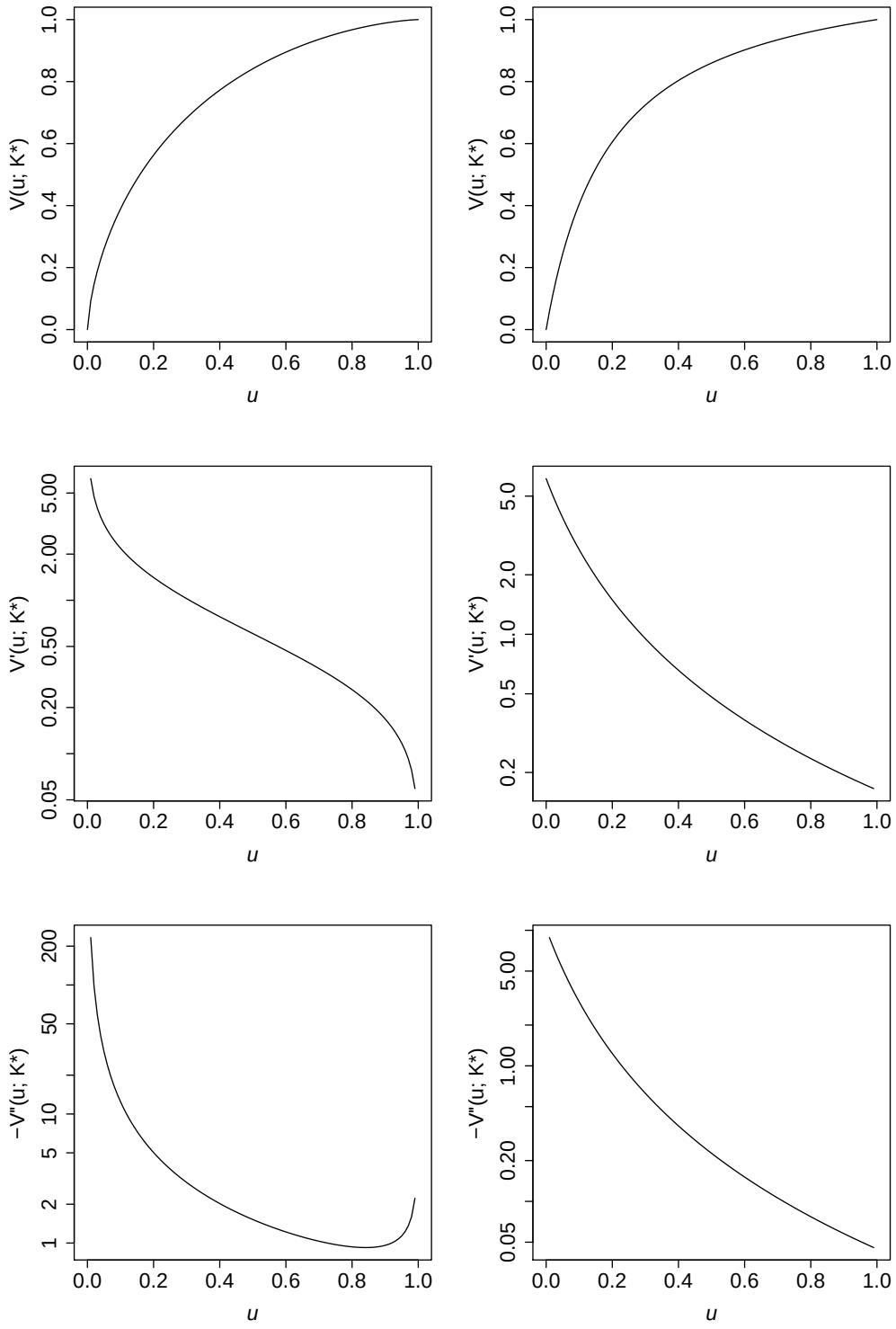
E.4 Student's t-Distribution

For notational convenience, let $q(u) = F_Z^{-1}(1-u)$ and

$$C_\nu = \frac{1}{\sqrt{(\nu-2)\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}.$$

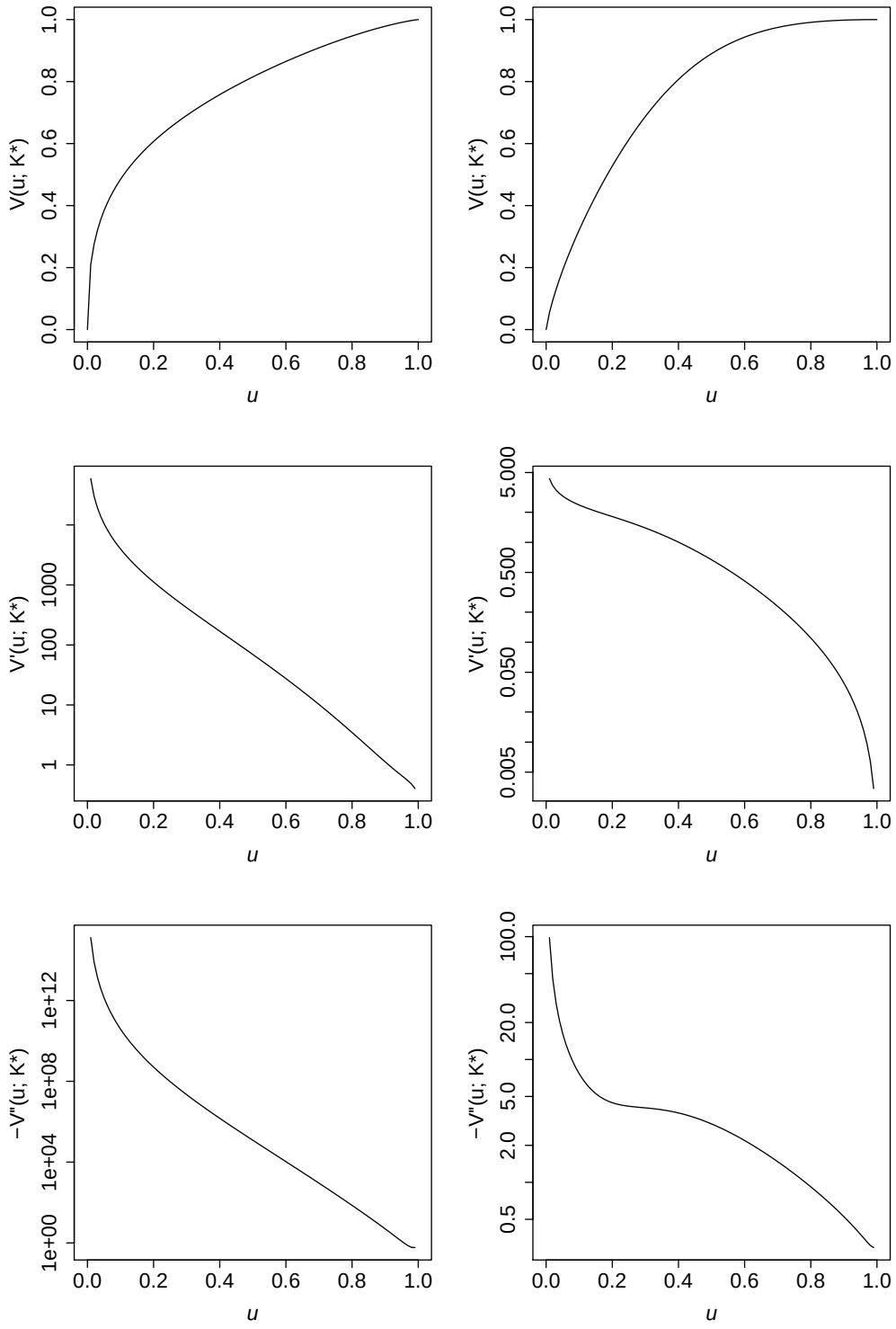
$$\begin{aligned} f_Z(u) &= C_\nu \left(1 + \frac{z^2}{\nu-2}\right)^{\left(\frac{\nu+1}{2}\right)}; \\ V'(u;K^*) &= \left(\frac{\nu-2+(q(u)-K^*)^2}{\nu-2+q(u)^2}\right)^{-\left(\frac{\nu+1}{2}\right)}; \end{aligned}$$

$$\begin{aligned} V''(u;K^*) = \frac{1}{C_\nu}\left(\frac{\nu+1}{\nu-2}\right)\left(1+\frac{q(u)^2}{\nu-2}\right)^{\nu+1}\left(1+\frac{(q(u)-K^*)^2}{\nu-2}\right)^{-\frac{\nu+1}{2}} \\ \left\{(q(u)-K^*)\left(1+\frac{(q(u)-K^*)^2}{\nu-2}\right)^{-1} - q(u)\left(1+\frac{q(u)^2}{\nu-2}\right)^{-1}\right\}. \end{aligned}$$



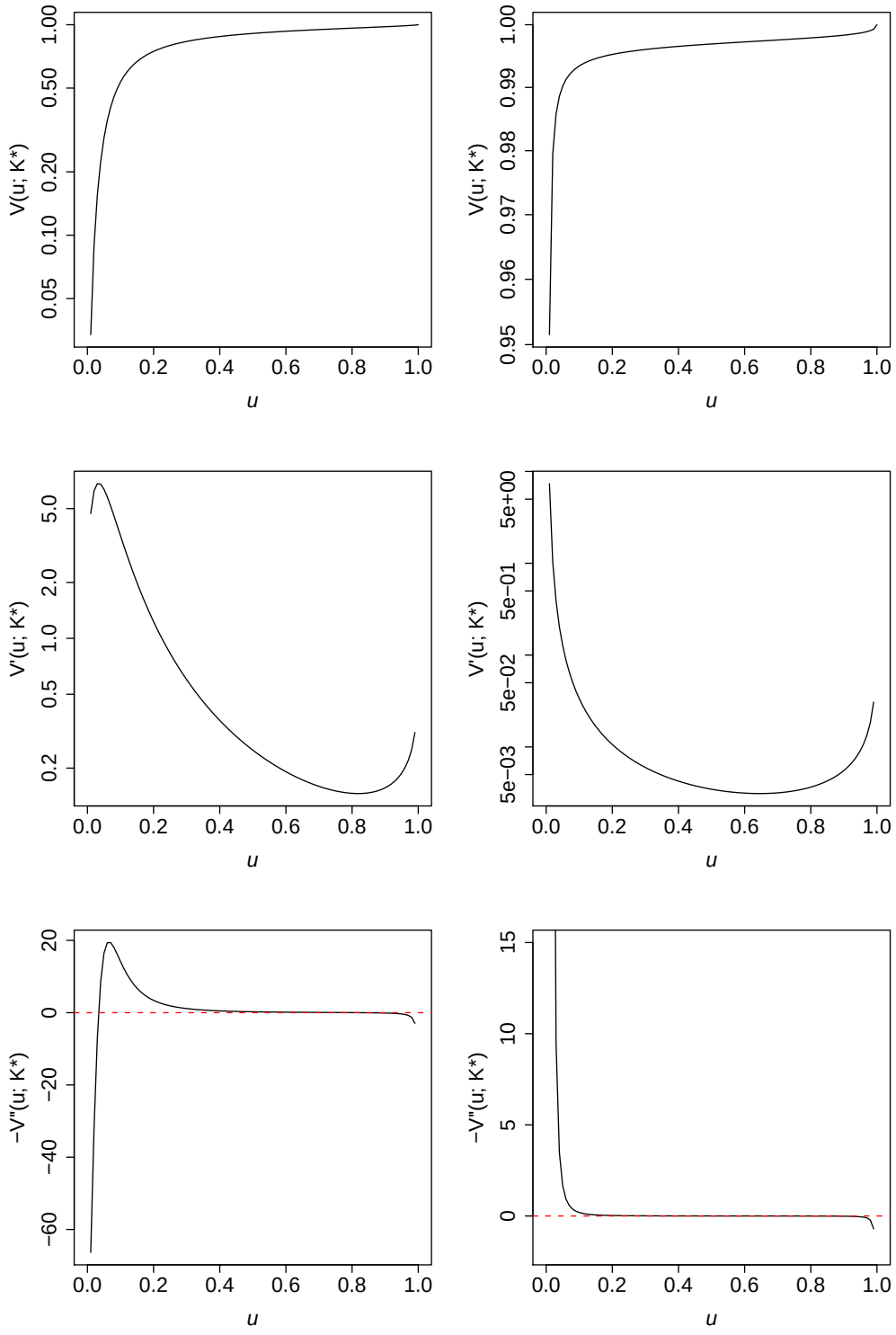
(a) Normal (b) Logistic

Figure E.1: Example analysis of $V(u; K^*)$ for the normal and logistic SSD. Top: $V(u; K^* = 1)$; middle: $V'(u; K^* = 1)$; top: $-V''(u; K^* = 1)$.



(a) SN ($\alpha = -3$) (b) SN ($\alpha = +3$)

Figure E.2: Example analysis of $V(u; K^*)$ for the standardised SN SSD with $\alpha = \pm 3$. Top: $V(u; K^* = 1)$; middle: $V'(u; K^* = 1)$; top: $-V''(u; K^* = 1)$.



(a) Student-t ($K^* = 1$)

(b) Student-t ($K^* = 4$)

Figure E.3: Example analysis of $V(u; K^*)$ for the standardised Student-t SSD with $\nu = 3$. Top: $V(u; K^*)$; middle: $V'(u; K^*)$; top: $-V''(u; K^*)$. Red-dashed lines correspond to $V''(u; K^*) = 0$.

Appendix F

Numerical Evaluation of $\mathbf{MFE}(n, i, K^*, k', \phi')$

For species non-exchangeable model $\mathcal{D}_1$ (Equations 5.7) with non-exchangeability parameters (k', ϕ') fixed, the MFE evaluated at the generalised decision rule

$$T(n, i, K) = Y_{(i:n)} - K,$$

under a log-normal SSD with location μ and standard deviation σ, was described in Section 6.8 as being equal to

$$\begin{aligned}\mathrm{MFE}(n, i, K^*, k', \phi') = 1 - \int_{-\infty}^{\infty} \Bigg\{ & F_{\beta(i,n-i)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{n \neq i\}} \\ & + \left[F_{\beta(i-1,n-i+1)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{i \neq 1\}} + \mathbf{1}_{\{i=1\}}\right. \\ & \left. - F_{\beta(i,n-i)}\left(\Phi(z + K^*)\right) \mathbf{1}_{\{n \neq i\}}\right] \Phi\left(\frac{z + K^* + k'}{\phi'}\right) \Bigg\} \phi(z) dz,\end{aligned}$$

where $K^* = K/\sigma$ is the standardised assessment shift and $\mathbf{1}_{\mathcal{A}}$ is the indicator function yielding value one if event $\mathcal{A}$ is true, zero otherwise.

For sufficiently large K^*, the integral is close to one for the majority of decision rules we are concerned with (by noting that large PNECs are unacceptable). In fact, by linearity, the integral is separable into three integrals each yielding a value of approximately one (not including the indicator functions). The precision in calculating each integral and subsequently summing is severely affected by some

choices of (n, i) and large K^*.

To improve accuracy of the integration[1], we rewrite the integrand such that floating point arithmetic is done to higher precision. As an example, consider one of the terms in the integral (above) which can be written as $F_{\beta(a,b)}\left(\Phi(z+c)\right)\Phi(dz+e)\phi(z)$ for some a and b which are positive integers, and c, d and e which are positive real numbers. This can be re-expressed as

$$\left[1 - F_{\beta(b,a)}\left(Q(z+c)\right) - Q(dz+e) + F_{\beta(b,a)}\left(Q(z+c)\right)Q(dz+e)\right]\phi(z),$$

where $Q(z) = 1-\Phi(z)$. Integration of this expression would consequently yield $1-\epsilon$, where ϵ is very small for certain choices of a, b, c and e. Consequently, numerical integration can be performed to a higher precision. Applying this principle to all terms in the integrand yields

$$\begin{aligned}
&\mathrm{MFE}(n, i, K^*, k', \phi') = \\
&\quad \int_{-\infty}^{\infty} F_{\beta(n-i+1,i-1)}\left(Q(z+K^*)\right)\mathbf{1}_{\{i\neq 1\}}\phi(z)dz \\
&- \int_{-\infty}^{\infty} F_{\beta(n-i+1,i-1)}\left(Q(z+K^*)\right)Q\left(\frac{z+K^*+k'}{\phi'}\right)\mathbf{1}_{\{i\neq 1\}}\phi(z)dz \\
&+ \int_{-\infty}^{\infty} F_{\beta(n-i,i)}\left(Q(z+K^*)\right)Q\left(\frac{z+K^*+k'}{\phi'}\right)\mathbf{1}_{\{n\neq i\}}\phi(z)dz \\
&+ \int_{-\infty}^{\infty} Q\left(\frac{z+K^*+k'}{\phi'}\right)\mathbf{1}_{\{n=i\}}\phi(z)dz.
\end{aligned}$$

The final integral (neglecting the indicator function) is straightforwardly determined to be

$$\int_{-\infty}^{\infty} Q\left(\frac{z+K^*+k'}{\phi'}\right)\phi(z)dz = \Phi\left(\frac{-(K^*+k')}{\sqrt{1+\phi^2}}\right).$$

Numerical evaluation of the penultimate integral was observed to be sensitive for large n when $i = 1$. By properties of the Beta CDF, the integrand for this term

[1]Numerical integration was performed using the `integrate` function in `R` version 2.9.2 (R, 2006).

can be re-written as

$$\int_{-\infty}^{\infty} \left[Q\left(z + K^*\right)\right]^{n-1} Q\left(\frac{z + K^* + k'}{\phi'}\right) \phi(z) dz,$$

which can be evaluated to a reasonably high accuracy. Performing numerical quadrature on each integral separately leads to suitably stable and accurate evaluation of the MFE for each assessment procedure.

Acronyms

ACR Acute to Chronic Ratio (p. 13)

AF Assessment Factor (p. 15)

AIC Akaike Information Criterion (p. 102)

AJ Aldenberg and Jaworska (p. 28)

AL Aldenberg and Luttik (p. 28)

AS Assessment Shift (p. 25)

ATS Adjusted Toxicity Statistic (p. 25)

BIC Bayesian Information Criterion (p. 101)

BF Bayes Factor (p. 141)

BNM Bimodal Normal Mixture (p. 176)

CDF Cumulative Distribution Function (p. 19)

ERA Ecotoxicological Risk Assessment (p. 8)

EC$_x$ Effect Concentration to x% of the species population (p. 12)

EFSA European Food Safety Authority (p. 4)

EP Exponential Power (p. 175)

EU European Union (p. 1)

FEAT Flash Environmental Assessment Tool (p. 45)

FOSD First Order Stochastic Dominance (p. 193)

GAL Generalised Absolute Loss (p. 51)

HC$_p$ Hazardous Concentration to p% of community (p. 20)

LA Luttik and Aldenberg (p. 13)

LC$_x$ Lethal Concentration to x% of the species population (p. 12)

LINEX LINear EXponential (p. 55)

M Method-of-Moments (p. 26)

MCMC Markov Chain Monte Carlo (p. 132)

MFE Mean Fraction Exceeded (p. 31)

NOEC No Observed Effect Concentration (p. 13)

PEC Predicted Environmental Concentration (p. 9)

PNEC Predicted No Effect Concentration (p. 9)

PAF Potentially Affected Fraction (p. 19)

PDF Probability Distribution Function (p. 167)

RCR Risk Characterisation Ratio (p. 8)

RIVM The Dutch National Institute for Public Health and the Environment (p. 23)

SEL Squared Error Loss (p. 48)

SN Skew-Normal (p. 174)

SOSD Second Order Stochastic Dominance (p. 193)

SSD Species Sensitivity Distribution (p. 17)

US EPA United States of America Environmental Protection Agency (p. 37)